

Gi-Chul Yang
Sio-long Ao
Len Gelman
Editors

IAENG Transactions on Engineering Technologies

Special Volume of the World Congress
on Engineering 2012

Lecture Notes in Electrical Engineering

Volume 229

For further volumes:
<http://www.springer.com/series/7818>

Gi-Chul Yang · Sio-long Ao
Len Gelman
Editors

IAENG Transactions on Engineering Technologies

Special Volume of the World Congress
on Engineering 2012

 Springer

Editors

Gi-Chul Yang
Department of Multimedia Engineering
Mokpo National University
Chonnam
Republic of South Korea

Len Gelman
School of Engineering, Applied
Mathematics and Computing
Cranfield University
Cranfield, Bedfordshire
UK

Sio-long Ao
IAENG Secretariat
Hong Kong
Hong Kong SAR

ISSN 1876-1100

ISBN 978-94-007-6189-6

DOI 10.1007/978-94-007-6190-2

Springer Dordrecht Heidelberg New York London

ISSN 1876-1119 (electronic)

ISBN 978-94-007-6190-2 (eBook)

Library of Congress Control Number: 2012956302

© Springer Science+Business Media Dordrecht 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference on “Advances in Engineering Technologies and Physical Science” was held in London, U.K., 4–6 July, 2012, under the World Congress on Engineering 2012 (WCE 2012). The WCE 2012 is organized by the International Association of Engineers (IAENG); the Congress details are available at: <http://www.iaeng.org/WCE2012>. IAENG is a non-profit international association for engineers and computer scientists, which was founded originally in 1968. The World Congress on Engineering serves as good platforms for the engineering community to meet with each other and to exchange ideas. The conferences have also struck a balance between theoretical and application development. The conference committees have been formed with over 300 committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The congress is truly international meeting with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than 1,100 manuscript submissions for the WCE 2012. All submitted papers have gone through the peer review process, and the overall acceptance rate is 54.61 %.

This volume contains 58 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include Applied and Engineering Mathematics, Computational Statistics, Mechanical Engineering, Bioengineering, Internet Engineering, Wireless Networks, Knowledge Engineering, Computational Intelligence, High Performance Computing, Manufacturing Engineering, and industrial applications. The book offers the state of art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent reference work for researchers and graduate students working on engineering technologies and physical science and applications.

Gi-Chul Yang
Sio-long Ao
Len Gelman

Contents

Inventory Control Under Parametric Uncertainty of Underlying Models	1
Nicholas A. Nechval, Konstantin N. Nechval and Maris Purgailis	
Periodic Solution and Strange Attractor in Impulsive Hopfield Networks with Time-Varying Delays	17
Yanxia Cheng, Yan Yan and Zhanji Gui	
Solving Stiff Ordinary Differential Equations Using Extended Block Backward Differentiation Formulae	31
Siti Ainor Mohd Yatim, Zarina Bibi Ibrahim, Khairil Iskandar Othman and Mohamed Suleiman	
On Fast Algorithms for Triangular and Dense Matrix Inversion.	45
Ryma Mahfoudhi and Zaher Mahjoub	
Adding Relation Between Two Levels of a Linking Pin Organization Structure Maximizing Communication Efficiency of Information	57
Kiyoshi Sawada	
Bayesian Inference for the Parameters of Two-Parameter Exponential Lifetime Models Based on Type-I and Type-II Censoring	67
Husam Awni Bayoud	
Analysing Metric Data Structures Thinking of an Efficient GPU Implementation.	81
Roberto Uribe-Paredes, Enrique Arias, Diego Cazorla and José Luis Sánchez	

Exploratory Analysis of Ergonomics Importance at Workplace and Safety Culture Amongst Occupational Safety and Health Practitioners	93
Md Sirat Rozlina, Mohamed Shaharoun Awaluddin, Syed Hassan Syed Abdul Hamid and Zakuan Norhayati	
Least Squares Data Fitting Subject to Decreasing Marginal Returns	105
Ioannis C. Demetriou	
The Further Development of Stem Taper and Volume Models Defined by Stochastic Differential Equations.	121
Petras Rupšys	
Computing Compressible Two-Component Flow Systems Using Diffuse Interface Method	135
A. Ballil, S. A. Jolgam, A. F. Nowakowski and F. C. G. A. Nicolleau	
Turbulent Boundary Layer Gas–Solid Flow Based on Two-Fluid Model	147
Hassan Basirat Tabrizi	
Molten Carbonate Fuel Cell as a Reducer of CO₂ Emissions from Gas Turbine Power Plants	159
Jaroslaw Milewski, Rafal Bernat and Janusz Lewandowski	
Computational Contact Modelling of Hip Resurfacing Devices	171
Murat Ali and Ken Mao	
Transport Phenomena in Engineering Problems: CFD-Based Computational Modeling	187
Maksim Mezhericher	
Investigating the Effects on the Low Speed Response of a Pressure Charged IC Engine Through the Application of a Twin-Entry Turbine Housing	201
Alex Kusztelan, Denis Marchant, Yufeng Yao and Yawei Wang	
A Data Mining Approach to Recognize Objects in Satellite Images to Predict Natural Resources	215
Muhammad Shahbaz, Aziz Guergachi, Aneela Noreen and Muhammad Shaheen	

Handling the Data Growth with Privacy Preservation in Collaborative Filtering 231
 Xiwei Wang and Jun Zhang

Machine Learning-Based Missing Value Imputation Method for Clinical Datasets 245
 M. Mostafizur Rahman and D. N. Davis

Opto-Electronic Hybrid Integrated Platform for High-Speed Telecom/Datacom Applications: Microwave Design and Optimization 259
 Wei Han, Marc Rensing, Xin Wang, Peter O’Brien and Frank H. Peters

Direct Torque Control of In-Wheel BLDC Motor Used in Electric Vehicle 273
 Alireza Tashakori Abkenar and Mehran Motamed Ektesabi

Different Detection Schemes Using Enhanced Double Weight Code for OCDMA Systems. 287
 Feras N. Hasoon, Mohammed H. Al-Mansoori and Sahbudin Shaari

Multigate RADFET Dosimeter for Radioactive Environment Monitoring Applications 301
 Fayçal Djéffal and Mohamed Meguellati

Multi-Objective-Based Approach to Optimize the Analog Electrical Behavior of GSDG MOSFET: Application to Nanoscale Circuit Design 315
 Toufik Bendib and Fayçal Djéffal

Performance Analysis of Series Hybrid Active Power Filter 327
 M. A. Mulla, R. Chudamani and A. Chowdhury

Protecting the End User Device in 4G Heterogeneous Networks 339
 Hani Alquhayz, Ali Al-Bayatti and Amelia Platt

Calibration Procedures for Indoor Location Using Fingerprinting 349
 Pedro Mestre, Luis Reigoto, Luis Coutinho, Aldina Correia, Joao Matias and Carlos Serodio

Handling the Congestion Control Problem of TCP/AQM Wireless Networks with PID Controllers 365
 Teresa Alvarez and Diego Martínez

On the Initial Network Topology Factor in Mobile Ad-Hoc Network	381
Ronit Nossenson and Adi Schwartz	
Internet Key Exchange Protocol Using ECC-Based Public Key Certificate	391
Sangram Ray and G. P. Biswas	
Intrusion Alert Correlation Framework: An Innovative Approach . . .	405
Huwaida Tagelsir Elshoush and Izzeldin Mohamed Osman	
An Interactive Shadow Removing Tool: A Granular Computing Approach	421
Abhijeet Vijay Nandedkar	
Resolution Enhancement for Digital Off-Axis Hologram Reconstruction	431
Nazeer Muhammad and Dai-Gyoung Kim	
A Novel Two-Scan Connected-Component Labeling Algorithm	445
Lifeng He, Yuyan Chao, Yun Yang, Sihui Li, Xiao Zhao and Kenji Suzuki	
Approaches to Bayesian Network Model Construction	461
Ifeyinwa E. Achumba, Djamel Azzi, Ifeanyi Ezebili and Sebastian Bersch	
Fertilization Operator for Multi-Modal Dynamic Optimization	475
Khalid Jebari, Abdelaziz Bouroumi and Aziz Ettouhami	
A Hardware Design for Binary Image Recognition	491
Saul Martinez-Diaz	
In-Situ Vibrational Spectroscopies, BTEM Analysis and DFT Calculations	501
Feng Gao, Chuanzhao Li, Effendi Widjaja and Marc Garland	
Convergence Speed of Generalized Longest-Edge-Based Refinement	511
José P. Suárez, Tania Moreno, Pilar Abad and Ángel Plaza	
Labeling the Nodes in the Intrinsic Order Graph with Their Weights	523
Luis González	

Solving VoIP QoS and Scalability Issues in Backbone Networks 537
 Martin Hruby, Michal Olsovsky and Margareta Kotocova

**Determining the Importance of Design Features
 on Usable Educational Websites.** 551
 Layla Hasan

A Flexible Dynamic Data Structure for Scientific Computing 565
 Josef Weinbub, Karl Rupp and Siegfried Selberherr

**Application of Curriculum Design Maturity Model at Private
 Institution of Higher Learning in Malaysia: A Case Study** 579
 Chee Ling Thong, Yusmadi Yah Jusoh, Rusli Abdullah
 and Nor Hayati Alwi

Reducing Job Failure Due to Churn in Dynamics Grids 591
 K. Abdelkader and J. Broeckhove

**Parallel Algorithm for Multiplying Integer Polynomials
 and Integers** 605
 Andrzej Chmielowiec

A Model to Improve Reliability in Cloud Computing 617
 P. Srivaramangai and Rengaramanujam Srinivasan

Natural Gas Price Forecasting: A Novel Approach. 627
 Prerna Mishra

**Structured Data Mining for Micro Loan Performance Prediction:
 The Case of Indonesian Rural Bank.** 641
 Novita Ikasari and Fedja Hadzic

Financial Forecasting Using the Kolmogorov–Feller Equation. 655
 Jonathan Blackledge, Marc Lamphiere, Kieran Murphy
 and Shaun Overton

**Surface Quality Improvement in CNC End Milling of Aluminum
 Alloy Using Nanolubrication System** 669
 Mohd Sayuti Ab Karim, Ahmed Aly Diah Mohammed Sarhan
 and Mohd Hamdi Abd Shukor

**A Price-Based Decision Policy to Mitigate the Tragedy
 of the Commons and Anti-Commons** 685
 M. Sumbwanyambe and A. L. Nel

Modeling Emergency Department Using a Hybrid Simulation Approach. 701
Norazura Ahmad, Noraida Abdul Ghani, Anton Abdulbasah Kamil and Razman Mat Tahar

The Challenge of Adopting Minimal Quantities of Lubrication for End Milling Aluminium 713
Brian Boswell and Mohammad Nazrul Islam

Fine-Tuning Negotiation Time in Multi-Agent Manufacturing Systems 725
W. L. Yeung

Adhesive Bonding of Attachments in Automotive Final Assembly 739
Loucas Papadakis, Vassos Vassiliou, Michalis Menicou, Manuel Schiel and Klaus Dilger

Uncertainty Components in Performance Measures 753
Sérgio Dinis Teixeira de Sousa, Eusébio Manuel Pinto Nunes and Isabel da Silva Lopes

Decision Making of Industrialized Building System: A Supply Chain Perspective on the Influence of Behavioral Economic Factors 767
Sharifah Akmam Syed Zakaria, Graham Brewer and Thayaparan Gajendran

Inventory Control Under Parametric Uncertainty of Underlying Models

Nicholas A. Nechval, Konstantin N. Nechval and Maris Purgailis

Abstract A large number of problems in inventory control, production planning and scheduling, location, transportation, finance, and engineering design require that decisions be made in the presence of uncertainty of underlying models. In the present paper we consider the case, where it is known that the underlying distribution belongs to a parametric family of distributions. The problem of determining an optimal decision rule in the absence of complete information about the underlying distribution, i.e., when we specify only the functional form of the distribution and leave some or all of its parameters unspecified, is seen to be a standard problem of statistical estimation. Unfortunately, the classical theory of statistical estimation has little to offer in general type of situation of loss function. In the paper, for improvement or optimization of statistical decisions under parametric uncertainty, a new technique of invariant embedding of sample statistics in a performance index is proposed. This technique represents a simple and computationally attractive statistical method based on the constructive use of the invariance principle in mathematical statistics. Unlike the Bayesian approach, an invariant embedding technique is independent of the choice of priors. It allows one to eliminate unknown parameters from the problem and to find the best invariant decision rules, which have smaller risk than any of the well-known decision rules. A numerical example is given.

N. A. Nechval (✉)

Department of Statistics, EVF Research Institute, University of Latvia,
Raina Blvd 19, Riga 1050, Latvia
e-mail: nechval@junik.lv

K. N. Nechval

Department of Applied Mathematics, Transport and Telecommunication Institute,
Lomonosov Street 1, Riga1019, Latvia
e-mail: konstan@tsi.lv

M. Purgailis

Department of Cybernetics, University of Latvia, Raina Blvd 19, Riga 1050, Latvia
e-mail: marispur@lanet.lv

Keywords Demand · Distribution · Inventory · Model · Optimization · Risk · Uncertainty

1 Introduction

Most of the inventory management literature assumes that demand distributions are specified explicitly. However, in many practical situations, the true demand distributions are not known, and the only information available may be a time-series of historic demand data. When the demand distribution is unknown, one may either use a parametric approach (where it is assumed that the demand distribution belongs to a parametric family of distributions) or a non-parametric approach (where no assumption regarding the parametric form of the unknown demand distribution is made).

Under the parametric approach, one may choose to estimate the unknown parameters or choose a prior distribution for the unknown parameters and apply the Bayesian approach to incorporating the demand data available. Scarf [1] and Karlin [2] consider a Bayesian framework for the unknown demand distribution. Specifically, assuming that the demand distribution belongs to the family of exponential distributions, the demand process is characterized by the prior distribution on the unknown parameter. Further extension of this approach is presented in [3]. Application of the Bayesian approach to the censored demand case is given in [4, 5]. Parameter estimation is first considered in [6] and recent developments are reported in [7, 8]. Liyanage and Shanthikumar [9] propose the concept of operational statistics and apply it to a single period newsvendor inventory control problem.

Within the non-parametric approach, either the empirical distribution or the bootstrapping method (e.g. see [10]) can be applied with the available demand data to obtain an inventory control policy.

Conceptually, it is useful to distinguish between “new-sample” inventory control, “within-sample” inventory control, and “new-within-sample” inventory control.

For the new-sample inventory control process, the data from a past sample of customer demand are used to make a statistical decision on a future time period for the same inventory control process.

For the within-sample inventory control process, the problem is to make a statistical decision on a future time period for the same inventory control process based on early data from that sample of customer demand.

For the new-within-sample inventory control process, the problem is to make a statistical decision on a future time period for the inventory control process based on early data from that sample of customer demand as well as on a past data sample of customer demand from the same process.

In this paper, we consider the case of the within-sample inventory control process, where it is known that the underlying distribution function of the customer demand belongs to a parametric family of distribution functions. However, unlike in the Bayesian approach, we do not assume any prior knowledge on the parameter values.

2 Cumulative Customer Demand

The primary purpose of this paper is to introduce the idea of cumulative customer demand in inventory control problems to deal with the order statistics from the underlying distribution. It allows one to use the available statistical information as completely as possible in order to improve statistical decisions for inventory control problems under parametric uncertainty.

Assumptions. The customer demand at the i th period represents a random variable $Y_i, i \in \{1, \dots, m\}$. For the cumulative customer demand, X , it is assumed that the random variables

$$X_1 = Y_1, \dots, X_k = \sum_{i=1}^k Y_i, \dots, X_l = \sum_{i=1}^l Y_i, \dots, X_m = \sum_{i=1}^m Y_i \quad (1)$$

represent the order statistics ($X_1 \leq \dots \leq X_m$) from the exponential distribution with the probability density function

$$f_\sigma(x) = \frac{1}{\sigma} \exp\left(-\frac{x}{\sigma}\right), \quad x \geq 0, \quad \sigma > 0, \quad (2)$$

and the probability distribution function

$$F_\sigma(x) = 1 - \exp\left(-\frac{x}{\sigma}\right). \quad (3)$$

Theorem 1 Let $X_1 \leq \dots \leq X_k$ be the first k ordered observations (order statistics) in a sample of size m from a continuous distribution with some probability density function $f_\theta(x)$ and distribution function $F_\theta(x)$, where θ is a parameter (in general, vector). Then the conditional probability density function of the l th order statistics $X_l (1 \leq k < l \leq m)$ given $X_k = x_k$ is

$$\begin{aligned} g_\theta(x_l|x_k) &= \frac{(m-k)!}{(l-k-1)!(m-l)!} \left[\frac{F_\theta(x_l) - F_\theta(x_k)}{1 - F_\theta(x_k)} \right]^{l-k-1} \\ &\quad \times \left[1 - \frac{F_\theta(x_l) - F_\theta(x_k)}{1 - F_\theta(x_k)} \right]^{m-l} \frac{f_\theta(x_l)}{1 - F_\theta(x_k)} \\ &= \frac{(m-k)!}{(l-k-1)!(m-l)!} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} (-1)^j \left[\frac{1 - F_\theta(x_l)}{1 - F_\theta(x_k)} \right]^{m-l+j} \\ &\quad \times \frac{f_\theta(x_l)}{1 - F_\theta(x_k)} \end{aligned}$$

$$\begin{aligned}
&= \frac{(m-k)!}{(l-k-1)!(m-l)!} \sum_{j=0}^{m-l} \binom{m-l}{j} (-1)^j \left[\frac{F_\theta(x_l) - F_\theta(x_k)}{1 - F_\theta(x_k)} \right]^{l-k-1+j} \\
&\quad \times \frac{f_\theta(x_l)}{1 - F_\theta(x_k)} \tag{4}
\end{aligned}$$

Proof From the marginal density function of X_k and the joint density function of X_k and X_l , we have the conditional density function of X_l , given that $X_k = x_k$, as

$$g_\theta(x_l|x_k) = g_\theta(x_l, x_k)/g_\theta(x_k). \tag{5}$$

This ends the proof.

Corollary 1.1 The conditional probability distribution function of X_l given $X_k = x_k$ is

$$\begin{aligned}
&P_\theta \{X_l \leq x_l | X_k = x_k\} \\
&= 1 - \frac{(m-k)!}{(l-k-1)!(m-l)!} \times \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} \frac{(-1)^j}{m-l+1+j} \\
&\quad \times \left[\frac{1 - F_\theta(x_l)}{1 - F_\theta(x_k)} \right]^{m-l+1+j} \\
&= \frac{(m-k)!}{(l-k-1)!(m-l)!} \sum_{j=0}^{m-l} \binom{m-l}{j} \frac{(-1)^j}{l-k+j} \left[\frac{F_\theta(x_l) - F_\theta(x_k)}{1 - F_\theta(x_k)} \right]^{l-k+j}. \tag{6}
\end{aligned}$$

Corollary 1.2 Let $X_1 \leq \dots \leq X_k$ be the first k ordered observations (order statistics) in a sample of size m from the exponential distribution (2). Then the conditional probability density function of the l th order statistics X_l ($1 \leq k < l \leq m$) given $X_k = x_k$ is

$$\begin{aligned}
g_\sigma(x_l|x_k) &= \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} (-1)^j \\
&\quad \times \frac{1}{\sigma} \exp\left(-\frac{(m-l+1+j)(x_l-x_k)}{\sigma}\right) \\
&= \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{m-l} \binom{m-l}{j} (-1)^j \\
&\quad \times \frac{1}{\sigma} \left[1 - \exp\left(-\frac{x_l-x_k}{\sigma}\right) \right]^{l-k-1+j} \exp\left(\frac{x_l-x_k}{\sigma}\right), \tag{7}
\end{aligned}$$

and the conditional probability distribution function of the l th order statistics X_l given $X_k = x_k$ is

$$\begin{aligned}
 P_\sigma \{X_l \leq x_l | X_k = x_k\} &= 1 - \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} \\
 &\quad \times \frac{(-1)^j}{m-l+1+j} \exp\left(-\frac{(m-l+1+j)(x_l-x_k)}{\sigma}\right) \\
 &= \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{m-l} \binom{m-l}{j} \\
 &\quad \times \frac{(-1)^j}{l-k+j} \left[1 - \exp\left(-\frac{x_l-x_k}{\sigma}\right)\right]^{l-k+j}. \quad (8)
 \end{aligned}$$

Corollary 1.3 If $l = k + 1$,

$$\begin{aligned}
 g_\sigma(x_{k+1}|x_k) &= (m-k) \frac{1}{\sigma} \exp\left(-\frac{(m-k)(x_{k+1}-x_k)}{\sigma}\right) \\
 &= (m-k) \sum_{j=0}^{m-k-1} \binom{m-k-1}{j} (-1)^j \\
 &\quad \times \frac{1}{\sigma} \left[1 - \exp\left(-\frac{x_{k+1}-x_k}{\sigma}\right)\right]^j \exp\left(\frac{x_{k+1}-x_k}{\sigma}\right), \quad (9)
 \end{aligned}$$

and

$$\begin{aligned}
 P_\sigma \{X_{k+1} \leq x_{k+1} | X_k = x_k\} &= 1 - \exp\left(-\frac{(m-k)(x_{k+1}-x_k)}{\sigma}\right) \\
 &= (m-k) \sum_{j=0}^{m-k-1} \binom{m-k-1}{j} \frac{(-1)^j}{1+j} \\
 &\quad \times \left[1 - \exp\left(-\frac{x_{k+1}-x_k}{\sigma}\right)\right]^{1+j}, \quad 1 \leq k \leq m-1. \quad (10)
 \end{aligned}$$

Corollary 1.4 If $l = k + 1$ and $Y_{k+1} = X_{k+1} - X_k$, then the probability density function of Y_{k+1} , $k \in \{1, \dots, m-1\}$, is given by

$$g_\sigma(y_{k+1}) = \frac{m-k}{\sigma} \exp\left(-\frac{(m-k)y_{k+1}}{\sigma}\right), \quad y_{k+1} \geq 0, \quad (11)$$

and the probability distribution function of Y_{k+1} is given by

$$G_{\sigma} \{y_{k+1}\} = 1 - \exp\left(-\frac{(m-k)y_{k+1}}{\sigma}\right). \quad (12)$$

Theorem 2 Let $X_1 \leq \dots \leq X_k$ be the first k ordered observations (order statistics) in a sample of size m from the exponential distribution (2), where the parameter σ is unknown. Then the predictive probability density function of the l th order statistics X_l ($1 \leq k < l \leq m$) is given by

$$\begin{aligned} g_{s_k}(x_l|x_k) &= \frac{k}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} (-1)^j \\ &\times \left[1 + (m-l+1+j) \frac{x_l - x_k}{s_k}\right]^{-(k+1)} \frac{1}{s_k}, \quad x_l \geq x_k, \quad s_k > 0, \end{aligned} \quad (13)$$

where

$$S_k = \sum_{i=1}^k X_i + (m-k)X_k \quad (14)$$

is the sufficient statistic for σ , and the predictive probability distribution function of the l th order statistics X_l is given by

$$\begin{aligned} P_{s_k} \{X_l \leq x_l | X_k = x_k\} &= 1 - \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} \\ &\times \frac{(-1)^j}{m-l+1+j} \left[1 + (m-l+1+j) \frac{x_l - x_k}{s_k}\right]^{-k}. \end{aligned} \quad (15)$$

Proof Using the technique of invariant embedding [11, 22], we reduce (7) to

$$\begin{aligned} g_{\sigma}(x_l|x_k) &= \frac{1}{\mathbf{B}(l-k, (m-l+1))} \sum_{j=0}^{l-k-1} \binom{l-k-1}{j} (-1)^j \\ &\times v \exp\left(-\frac{(m-l+1+j)(x_l - x_k)}{s_k} v\right) \frac{1}{s_k} = g_{s_k}(x_l|x_k, v), \end{aligned} \quad (16)$$

where

$$V = S_k/\sigma \quad (17)$$

is the pivotal quantity, the probability density function of which is given by

$$f(v) = \frac{1}{\Gamma(k)} v^{k-1} \exp(-v), \quad v \geq 0. \quad (18)$$

Then

$$g_{s_k}(x_l|x_k) = E\{g_{s_k}(x_l|x_k, v)\} = \int_0^{\infty} g_{s_k}(x_l|x_k, v) f(v) dv. \quad (19)$$

This ends the proof.

Corollary 2.1 If $l = k + 1$,

$$g_{s_k}(x_{k+1}|x_k) = k(m - k) \left[1 + (m - k) \frac{x_{k+1} - x_k}{s_k} \right]^{-(k+1)} \frac{1}{s_k}, \quad (20)$$

and

$$P_{s_k} \{X_{k+1} \leq x_{k+1} | X_k = x_k\} = 1 - \left[1 + (m - k) \frac{x_{k+1} - x_k}{s_k} \right]^{-k}, \quad (21)$$

Corollary 2.2 If $l = k + 1$ and $Y_{k+1} = X_{k+1} - X_k$, then the predictive probability density function of Y_{k+1} , $k \in \{1, \dots, m - 1\}$, is given by

$$g_{s_k}(y_{k+1}) = k(m - k) \left[1 + (m - k) \frac{y_{k+1}}{s_k} \right]^{-(k+1)} \frac{1}{s_k}, \quad y_{k+1} \geq 0, \quad (22)$$

and the predictive probability distribution function of Y_{k+1} is given by

$$G_{s_k}(y_{k+1}) = 1 - \left[1 + (m - k) \frac{y_{k+1}}{s_k} \right]^{-k}. \quad (23)$$

3 Inventory Control Models

This section deals with inventory items that are in stock during a single time period. At the end of the period, leftover units, if any, are disposed of, as in fashion items. Two models are considered. The difference between the two models is whether or not a setup cost is incurred for placing an order. The symbols used in the development of the models include:

c = setup cost per order,

c_1 = holding cost per held unit during the period,

c_2 = penalty cost per shortage unit during the period,

$g_{\sigma}(y_{k+1})$ = probability density function of customer demand, Y_{k+1} , during the $(k + 1)$ th period,

σ = scale parameter,

u = order quantity,

q = inventory on hand before an order is placed.

No-Setup Model (Newsvendor Model). This model is known in the literature as the *newsvendor* model (the original classical name is the *newsboy* model). It deals with stocking and selling newspapers and periodicals. The assumptions of the model are:

1. Demand occurs instantaneously at the start of the period immediately after the order is received.
2. No setup cost is incurred.

The model determines the optimal value of u that minimizes the sum of the expected holding and shortage costs. Given optimal $u (= u^*)$, the inventory policy calls for ordering $u^* - q$ if $q < u^*$; otherwise, no order is placed.

If $Y_{k+1} \leq u$, the quantity $u - Y_{k+1}$ is held during the $(k + 1)$ th period. Otherwise, a shortage amount $Y_{k+1} - u$ will result if $Y_{k+1} > u$. Thus, the cost per the $(k + 1)$ th period is

$$C(u) = \begin{cases} c_1 \frac{u - Y_{k+1}}{\sigma} & \text{if } Y_{k+1} \leq u, \\ c_2 \frac{Y_{k+1} - u}{\sigma} & \text{if } Y_{k+1} > u. \end{cases} \quad (24)$$

The expected cost for the $(k + 1)$ th period, $E_\sigma\{C(u)\}$, is expressed as

$$E_\sigma\{C(u)\} = \frac{1}{\sigma} \left(c_1 \int_0^u (u - y_{k+1}) g_\sigma(y_{k+1}) dy_{k+1} + c_2 \int_u^\infty (y_{k+1} - u) g_\sigma(y_{k+1}) dy_{k+1} \right). \quad (25)$$

The function $E_\sigma\{C(u)\}$ can be shown to be convex in u , thus having a unique minimum. Taking the first derivative of $E_\sigma\{C(u)\}$ with respect to u and equating it to zero, we get

$$\frac{1}{\sigma} \left(c_1 \int_0^u g_\sigma(y_{k+1}) dy_{k+1} - c_2 \int_u^\infty g_\sigma(y_{k+1}) dy_{k+1} \right) = 0 \quad (26)$$

or

$$c_1 P_\sigma\{Y_{k+1} \leq u\} - c_2 (1 - P_\sigma\{Y_{k+1} \leq u\}) = 0 \quad (27)$$

or

$$P_\sigma\{Y_{k+1} \leq u\} = \frac{c_2}{c_1 + c_2}. \quad (28)$$

It follows from (11), (12), (25), and (28) that

$$u^* = \frac{\sigma}{m - k} \ln \left(1 + \frac{c_2}{c_1} \right) \quad (29)$$

and

$$\begin{aligned} E_{\sigma}\{C(u^*)\} &= \frac{1}{\sigma} \left(c_2 E_{\sigma}\{Y_{k+1}\} - (c_1 + c_2) \int_0^{u^*} y_{k+1} g_{\sigma}(y_{k+1}) dy_{k+1} \right) \\ &= \frac{c_1}{m-k} \ln \left(1 + \frac{c_2}{c_1} \right). \end{aligned} \quad (30)$$

Parametric uncertainty. Consider the case when the parameter σ is unknown. To find the best invariant decision rule u^{BI} , we use the invariant embedding technique [11–22] to transform (24) to the form, which is depended only on the pivotal quantities V , V_1 , and the ancillary factor η .

Transformation of $C(u)$ based on the pivotal quantities V , V_1 is given by

$$C^{(1)}(\eta) = \begin{cases} c_1(\eta V - V_1) & \text{if } V_1 \leq \eta V, \\ c_2(V_1 - \eta V) & \text{if } V_1 > \eta V, \end{cases} \quad (31)$$

where

$$\eta = \frac{u}{S_k}, \quad (32)$$

$$V_1 = \frac{Y_{k+1}}{\sigma} \sim g(v_1) = (m-k) \exp[-(m-k)v_1], \quad v_1 \geq 0. \quad (33)$$

Then $E\{C^{(1)}(\eta)\}$ is expressed as

$$E\{C^{(1)}(\eta)\} = \int_0^{\infty} \left(c_1 \int_0^{\eta v} (\eta v - v_1) g(v_1) dv_1 + c_2 \int_{\eta v}^{\infty} (v_1 - \eta v) g(v_1) dv_1 \right) f(v) dv. \quad (34)$$

The function $E\{C^{(1)}(\eta)\}$ can be shown to be convex in η , thus having a unique minimum. Taking the first derivative of $E\{C^{(1)}(\eta)\}$ with respect to η and equating it to zero, we get

$$\int_0^{\infty} v \left(c_1 \int_0^{\eta v} g(v_1) dv_1 - c_2 \int_{\eta v}^{\infty} g(v_1) dv_1 \right) f(v) dv = 0 \quad (35)$$

or

$$\frac{\int_0^{\infty} v P(V_1 \leq \eta v) f(v) dv}{\int_0^{\infty} v f(v) dv} = \frac{c_2}{c_1 + c_2}. \quad (36)$$

It follows from (32), (34), and (36) that the optimum value of η is given by

$$\eta^* = \frac{1}{m-k} \left[\left(1 + \frac{c_2}{c_1} \right)^{1/(k+1)} - 1 \right], \quad (37)$$

the best invariant decision rule is

$$u^{\text{BI}} = \eta^* S_k = \frac{S_k}{m-k} \left[\left(1 + \frac{c_2}{c_1} \right)^{1/(k+1)} - 1 \right], \quad (38)$$

and the expected cost, if we use u^{BI} , is given by

$$E_{\sigma}\{C(u^{\text{BI}})\} = \frac{c_1(k+1)}{m-k} \left[\left(1 + \frac{c_2}{c_1} \right)^{1/(k+1)} - 1 \right] = \frac{c_1(k+1)u^{\text{BI}}}{S_k} = E\{C^{(1)}(\eta^*)\}. \quad (39)$$

It will be noted that, on the other hand, the invariant embedding technique [11–22] allows one to transform equation (25) as follows:

$$\begin{aligned} E_{\sigma}\{C(u)\} &= \frac{1}{\sigma} \left(c_1 \int_0^u (u - y_{k+1}) g_{\sigma}(y_{k+1}) dy_{k+1} + c_2 \int_u^{\infty} (y_{k+1} - u) g_{\sigma}(y_{k+1}) dy_{k+1} \right) \\ &= \frac{1}{s_k} \left(c_1 \int_0^u (u - y_{k+1}) v^2 (m-k) \exp\left(-\frac{v(m-k)y_{k+1}}{s_k}\right) \frac{1}{s_k} dy_{k+1} \right. \\ &\quad \left. + c_2 \int_u^{\infty} (y_{k+1} - u) v^2 (m-k) \exp\left(-\frac{v(m-k)y_{k+1}}{s_k}\right) \frac{1}{s_k} dy_{k+1} \right). \end{aligned} \quad (40)$$

Then it follows from (40) that

$$E\{E_{\sigma}\{C(u)\}\} = \int_0^{\infty} E_{\sigma}\{C(u)\} f(v) dv = E_{s_k}\{C^{(1)}(u)\}, \quad (41)$$

where

$$\begin{aligned} E_{s_k}\{C^{(1)}(u)\} &= \frac{k}{s_k} \left(c_1 \int_0^u (u - y_{k+1}) g_{s_k}^{\bullet}(y_{k+1}) dy_{k+1} \right. \\ &\quad \left. + c_2 \int_u^{\infty} (y_{k+1} - u) g_{s_k}^{\bullet}(y_{k+1}) dy_{k+1} \right) \end{aligned} \quad (42)$$

represents the expected predictive cost for the $(k + 1)$ th period. It follows from (42) that the cost per the $(k + 1)$ th period is reduced to

$$C^{(2)}(u) = \begin{cases} c_1 \frac{u - Y_{k+1}}{s_k/k} & \text{if } Y_{k+1} \leq u, \\ c_2 \frac{Y_{k+1} - u}{s_k/k} & \text{if } Y_{k+1} > u, \end{cases} \quad (43)$$

and the predictive probability density function of Y_{k+1} (compatible with (25)) is given by

$$g_{s_k}^\bullet(y_{k+1}) = (k + 1)(m - k) \left[1 + (m - k) \frac{y_{k+1}}{s_k} \right]^{-(k+2)} \frac{1}{s_k}, \quad y_{k+1} \geq 0. \quad (44)$$

Minimizing the expected predictive cost for the $(k + 1)$ th period,

$$E_{s_k}\{C^{(2)}(u)\} = \frac{k}{s_k} \left(c_1 \int_0^u (u - y_{k+1}) g_{s_k}^\bullet(y_{k+1}) dy_{k+1} + c_2 \int_u^\infty (y_{k+1} - u) g_{s_k}^\bullet(y_{k+1}) dy_{k+1} \right), \quad (45)$$

with respect to u , we obtain u^{BI} immediately, and

$$E_{s_k}\{C^{(2)}(u^{\text{BI}})\} = \frac{c_1(k + 1)}{m - k} \left[\left(1 + \frac{c_2}{c_1} \right)^{1/(k+1)} - 1 \right]. \quad (46)$$

It should be remarked that the cost per the $(k + 1)$ th period, $C^{(2)}(u)$, can also be transformed to

$$C^{(3)}(\eta) = \begin{cases} c_1 k \left(\frac{u}{s_k} - \frac{Y_{k+1}}{s_k} \right) & \text{if } \frac{Y_{k+1}}{s_k} \leq \frac{u}{s_k} \\ c_2 k \left(\frac{Y_{k+1}}{s_k} - \frac{u}{s_k} \right) & \text{if } \frac{Y_{k+1}}{s_k} > \frac{u}{s_k} \end{cases} = \begin{cases} c_1 k(\eta - W) & \text{if } W \leq \eta \\ c_2 k(W - \eta) & \text{if } W > \eta, \end{cases} \quad (47)$$

where the probability density function of the ancillary statistic $W = Y_{k+1}/S_k$ (compatible with (25)) is given by

$$g^\circ(w) = (k + 1)(m - k) [1 + (m - k)w]^{-(k+2)}, \quad w \geq 0. \quad (48)$$

Then the best invariant decision rule $u^{\text{BI}} = \eta^* S_k$, where η^* minimizes

$$E\{C^{(3)}(\eta)\} = k \left(c_1 \int_0^\eta (\eta - w) g^\circ(w) dw + c_2 \int_\eta^\infty (w - \eta) g^\circ(w) dw \right). \quad (49)$$

Comparison of statistical decision rules. For comparison, consider the maximum likelihood decision rule that may be obtained from (29),

$$u^{\text{ML}} = \frac{\widehat{\sigma}}{m - k} \ln \left(1 + \frac{c_2}{c_1} \right) = \eta_j^{\text{ML}} S_k, \quad (50)$$

where $\widehat{\sigma} = S_k/k$ is the maximum likelihood estimator of σ ,

$$\eta^{\text{ML}} = \frac{1}{m - k} \ln \left(1 + \frac{c_2}{c_1} \right)^{1/k}. \quad (51)$$

Since u^{BI} and u^{ML} belong to the same class,

$$\mathbf{C} = \{u : u = \eta S_k\}, \quad (52)$$

it follows from the above that u^{ML} is inadmissible in relation to u^{BI} .

Numerical example. If, say, $k = 1$ and $c_2/c_1 = 100$, we have that

$$\text{Rel. eff.}\{u^{\text{ML}}, u^{\text{BI}}, \sigma\} = E_\sigma\{C(u^{\text{BI}})\}/E_\sigma\{C(u^{\text{ML}})\} = 0.838. \quad (53)$$

Thus, in this case, the use of u^{BI} leads to a reduction in the expected cost of about 16.2% as compared with u^{ML} . The absolute expected cost will be proportional to σ and may be considerable.

Setup Model (s-S Policy). The present model differs from the one in (24) in that a setup cost c is incurred. Using the same notation, the total expected cost per the $(k + 1)$ th period is

$$\begin{aligned} E_\sigma\{\bar{C}(u)\} &= c + E_\sigma\{C(u)\} \\ &= c + \frac{1}{\sigma} \left(c_1 \int_0^u (u - y_{k+1}) g_\sigma(y_{k+1}) dy_{k+1} + c_2 \int_u^\infty (y_{k+1} - u) g_\sigma(y_{k+1}) dy_{k+1} \right). \end{aligned} \quad (54)$$

As shown above, the optimum value u^* must satisfy (28). Because c is constant, the minimum value of $E_\sigma\{\bar{C}(u)\}$ must also occur at u^* . In this case, $S = u^*$, and the value of $s(<S)$ is determined from the equation

$$E_{\sigma}\{C(s)\} = E_{\sigma}\{\bar{C}(S)\} = c + E_{\sigma}\{C(S)\}, \quad s < S. \quad (55)$$

This equation yields another value $s_1 (> S)$, which is discarded.

Assume that q is the amount on hand before an order is placed. How much should be ordered? This question is answered under three conditions: (1) $q < s$; (2) $s \leq q \leq S$; (3) $q > S$.

Case 1 ($q < s$). Because q is already on hand, its equivalent cost is given by $E_{\sigma}\{C(q)\}$. If any additional amount $u - q$ ($u > q$) is ordered, the corresponding cost given u is $E_{\sigma}\{\bar{C}(u)\}$, which includes the setup cost c , and we have

$$\min_{u>q} E_{\sigma}\{\bar{C}(u)\} = E_{\sigma}\{\bar{C}(S)\} < E_{\sigma}\{C(q)\}. \quad (56)$$

Thus, the optimal inventory policy in this case is to order $S - q$ units.

Case 2 ($s \leq q \leq S$). In this case, we have

$$E_{\sigma}\{C(q)\} \leq \min_{u>q} E_{\sigma}\{\bar{C}(u)\} = E_{\sigma}\{\bar{C}(S)\}. \quad (57)$$

Thus, it is not advantageous to order in this case and $u^* = q$.

Case 3 ($q > S$). In this case, we have for $u > q$,

$$E_{\sigma}\{C(q)\} < E_{\sigma}\{\bar{C}(u)\}. \quad (58)$$

This condition indicates that, as in Case 2, it is not advantageous to place an order—that is, $u^* = q$.

The optimal inventory policy, frequently referred to as the $s - S$ policy, is summarized as

$$\begin{aligned} &\text{if } q < s, \text{ order } S - q, \\ &\text{if } q \geq s, \text{ do not order.} \end{aligned} \quad (59)$$

The optimality of the $s - S$ policy is guaranteed because the associated cost function is convex.

Parametric uncertainty. In the case when the parameter σ is unknown, the total expected predictive cost for the $(k + 1)$ th period,

$$\begin{aligned} &E_{s_k}\{\bar{C}^{(1)}(u)\} = c + E_{s_k}\{C^{(1)}(u)\} \\ &= c + \frac{k}{s_k} \left(c_1 \int_0^u (u - y_{k+1}) g_{s_k}^{\bullet}(y_{k+1}) dy_{k+1} + c_2 \int_u^{\infty} (y_{k+1} - u) g_{s_k}^{\bullet}(y_{k+1}) dy_{k+1} \right), \end{aligned} \quad (60)$$

is considered in the same manner as above.

4 Conclusion and Future Work

In this paper, we develop a new frequentist approach to improve predictive statistical decisions for inventory control problems under parametric uncertainty of the underlying distributions for the cumulative customer demand. Frequentist probability interpretations of the methods considered are clear. Bayesian methods are not considered here. We note, however, that, although subjective Bayesian prediction has a clear personal probability interpretation, it is not generally clear how this should be applied to non-personal prediction or decisions. Objective Bayesian methods, on the other hand, do not have clear probability interpretations in finite samples. For constructing the improved statistical decisions, a new technique of invariant embedding of sample statistics in a performance index is proposed. This technique represents a simple and computationally attractive statistical method based on the constructive use of the invariance principle in mathematical statistics.

The methodology described here can be extended in several different directions to handle various problems that arise in practice. We have illustrated the proposed methodology for location-scale distributions (such as the exponential distribution). Application to other distributions could follow directly.

Acknowledgments This research was supported in part by Grant No. 06.1936, Grant No. 07.2036, Grant No. 09.1014, and Grant No. 09.1544 from the Latvian Council of Science and the National Institute of Mathematics and Informatics of Latvia.

References

1. Scarf H (1959) Bayes solutions of statistical inventory problem. *Ann Math Stat* 30:490–508
2. Karlin S (1960) Dynamic inventory policy with varying stochastic demands. *Manage Sci* 6: 231–258
3. Azoury KS (1985) Bayes solution to dynamic inventory models under unknown demand distribution. *Manage Sci* 31:1150–1160
4. Ding X, Puterman ML, Bisi A (2002) The censored newsvendor and the optimal acquisition of information. *Oper Res* 50:517–527
5. Lariviere MA, Porteus EL (1999) Stalking information: Bayesian inventory management with unobserved lost sales. *Manage Sci* 45:346–363
6. Conrad SA (1976) Sales data and the estimation of demand. *Oper Res Quart* 27:123–127
7. Agrawal N, Smith SA (1996) Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Res Logist* 43:839–861
8. Nahmias S (1994) Demand estimation in lost sales inventory systems. *Naval Res Logist* 41: 739–757
9. Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Oper Res Lett* 33:341–348
10. Bookbinder JH, Lordahl AE (1989) Estimation of inventory reorder level using the bootstrap statistical procedure. *IIE Trans* 21:302–312
11. Nechval NA, Nechval KN, Vasermanis EK (2003) Effective state estimation of stochastic systems. *Kybernetes (An International Journal of Systems & Cybernetics)* 32:666–678
12. Nechval NA, Berzins G, Purgailis M, Nechval KN (2008) Improved estimation of state of stochastic systems via invariant embedding technique. *WSEAS Trans Math* 7:141–159

13. Nechval NA, Nechval KN, Purgailis M (2011) Prediction of future values of random quantities based on previously observed data. *Eng Lett* 9:346–359
14. Nechval NA, Purgailis M, Nechval KN, Strelchonok VF (2012) Optimal predictive inferences for future order statistics via a specific loss function. *IAENG Int J Appl Math* 42:40–51
15. Nechval NA, Purgailis M, Cikste K, Berzins G, Nechval KN (2010) Optimization of statistical decisions via an invariant embedding technique. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2010, WCE 2010, London, 30 June–2 July 2010*, pp 1776–1782
16. Nechval NA, Purgailis M, Cikste K, Nechval KN (2010) Planning inspections of fatigued aircraft structures via damage tolerance approach. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2010, WCE 2010, London, 30 June–2 July 2010*, pp 2470–2475
17. Nechval NA, Purgailis M, Nechval KN, Rozevskis U (2011) Optimization of prediction intervals for order statistics based on censored data. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2011, WCE 2011, London, 6–8 July 2011*, pp 63–69
18. Nechval NA, Nechval KN, Purgailis M (2011) Statistical inferences for future outcomes with applications to maintenance and reliability. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2011, WCE 2011, London, 6–8 July 2011*, pp 865–871
19. Nechval NA, Purgailis M, Nechval KN, Bruna I (2012) Optimal inventory control under parametric uncertainty via cumulative customer demand. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, 4–6 July 2012*, pp 6–11
20. Nechval NA, Purgailis M, Nechval KN, Bruna I (2012) Optimal prediction intervals for future order statistics from extreme value distributions. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, 4–6 July 2012*, pp 1340–1345
21. Nechval NA, Nechval KN, Purgailis M (2011) Inspection policies in service of fatigued aircraft structures. In: Ao S-I, Gelman L (eds) *Electrical engineering and applied computing*, vol 90. Springer, Berlin, pp 459–472
22. Nechval NA, Nechval KN, Purgailis M (2013) Weibull prediction limits for a future number of failures under parametric uncertainty. In: Ao S-I, Gelman L (eds) *Electrical engineering and intelligent systems*, vol 130, LNEE. Springer, Berlin, pp 273–284

Periodic Solution and Strange Attractor in Impulsive Hopfield Networks with Time-Varying Delays

Yanxia Cheng, Yan Yan and Zhanji Gui

Abstract By constructing suitable Lyapunov functions, we study the existence, uniqueness and global exponential stability of periodic solution for impulsive Hopfield neural networks with time-varying delays. Our condition extends and generalizes a known condition for the global exponential periodicity of continuous Hopfield neural networks with time-varying delays. Further the numerical simulation shows that our system can occur many forms of complexities including gui strange attractor and periodic solution.

Keywords Hopfield neural network · Lyapunov functions · Pulse · Time-varying delay · Periodic solution · Strange attractor

1 Introduction

In recent years, stability of different classes of neural networks with time delay, such as Hopfield neural networks, cellular neural networks, bidirectional associative neural networks, Lotka-Volterra neural networks, has been extensively studied and various stability conditions have been obtained for these models of neural

Y. Cheng (✉)

The School of Science, Beijing Forestry University, Beijing 100083, People's Republic of China
e-mail: qcsj6463@163.com

Y. Yan

The School of Mathematics and Statistics, Hainan Normal University, Haikou,
Hainan 571158, People's Republic of China
e-mail: oishi19840923@163.com

Z. Gui

Department of Software Engineering, Hainan College of Software Technology, Qionghai,
Hainan 571400, People's Republic of China
e-mail: zhanjigui@sohu.com

networks. A citation will look like this, [1, 3, 6]. Here are some more citations [5, 10, 13, 16, 17].

Stability and convergence properties are generally regarded as important effects of delays. Both in biological and man-made neural systems, integration and communication delays are ubiquitous, and often become sources of instability. The delays in electronic neural networks are usually time varying, and sometimes vary violently with time due to the finite switching speed of amplifiers and faults in the electrical circuit. They slow down the transmission rate and tend to introduce some degree of instability in circuits. Therefore, fast response must be required in practical electronic neural-network designs. The technique to achieve fast response troubles many circuit designers. So, it is important to investigate the delay independent stability and decay estimates of the states of analog neural networks.

However, in implementation of networks, time delays are inevitably encountered because of the finite switching speed of amplifiers, see [2, 4, 7, 11, 12]. On the other hand, impulsive effect likewise exists in a wide variety of evolutionary processes in which states are changed abruptly at certain moments of time, involving such fields as medicine and biology, economics, mechanics, electronics and telecommunications, etc. Many interesting results on impulsive effect have been gained. Here are some more citations [2, 4, 7–9, 11, 12, 15, 18]. As artificial electronic systems, neural networks such as Hopfield neural networks, bidirectional neural networks and recurrent neural networks often are subject to impulsive perturbations which can affect dynamical behaviors of the systems just as time delays. Therefore, it is necessary to consider both impulsive effect and delay effect on the stability of neural networks.

In this chapter, we consider the following impulsive Hopfield neural networks with time-varying delays:

$$\begin{cases} \dot{x}_i(t) = -a_i x_i(t) + \sum_{j=1}^n a_{ij} g_j(x_j(t)) \\ \quad + \sum_{j=1}^n b_{ij} g_j(x_j(t - \tau_{ij}(t))) + I_i(t), & t \neq t_k, \\ \Delta x_i(t_k) = \gamma_{ik} x_i(t_k), & i = 1, 2, \dots, n, \quad k = 1, 2, \dots, \end{cases} \quad (1)$$

where n is the number of neurons in the network, $x_i(t)$ is the state of the i th neuron at time t , a_{ij} is the rate at which the i th neuron resets the state when isolated from the system, b_{ij} is the connection strength from the j th neuron to the i th neuron. $g(x) = (g_1(x_1), g_2(x_2), \dots, g_n(x_n))^T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the output of the i th neuron at time t , $I(t) = (I_1(t), I_2(t), \dots, I_n(t))^T \in \mathbb{R}^n$ is the ω -periodic external input to the i th neuron.

Throughout this chapter, we assume that

(**H**₁) For $j \in \{1, \dots, n\}$, $g_j(u)$ ($j = 1, 2, \dots, n$) is globally Lipschitz-continuous with the Lipschitz constant $L_j > 0$. That is,

$$|g_j(u_1) - g_j(u_2)| \leq L_j |u_1 - u_2|,$$

for all $u_1, u_2 \in \mathbb{R} = (-\infty, \infty)$.

(H₂) There exists a positive integer p such that, $t_{k+p} = t_k + \omega$, $\gamma_{i(k+p)} = \gamma_{ik}$, $k > 0$, $k = 1, 2, \dots$.

(H₃) $\tau_{ij}(t)$ ($i, j = 1, 2, \dots, n$) are continuously differentiable ω -periodic functions defined on \mathbb{R}^+ , $\tau = \sup_{0 \leq t \leq \omega} \tau_{ij}(t)$ and $\inf_{t \in \mathbb{R}^+} \{1 - \dot{\tau}_i(t)\} > 0$.

In order to describe the initial condition accompanying Eq. (1), we introduce the following notations.

Definition 1 A function $\phi : [-\tau, 0] \rightarrow \mathbb{R}$ is said to be a C^* -function if the following two conditions are satisfied:

- (a) ϕ is piecewise continuous with first kind discontinuity at the points t_k . Moreover, ϕ is left-continuous at each discontinuity point.
- (b) For all $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, p\}$, $\phi_i(t_k + 0) = \phi_i(t_k) + \gamma_{ik}\phi_i(t_k)$.

Let C^* denote the set of all the C^* -functions. Obviously, $(C^*, \mathbb{R}, +, \cdot)$ forms a vector space on \mathbb{R} . Now consider $(C^*, \mathbb{R}, +, \cdot)$ endowed with the norm defined by

$$\|\phi\|_\infty = \sup_{-\tau \leq \theta \leq 0} \|\phi(\theta)\| = \sup_{-\tau \leq \theta \leq 0} \max_{1 \leq i \leq n} |\phi_i(\theta)|.$$

Definition 2 A function $x : [-\tau, \infty] \rightarrow \mathbb{R}$ is said to be the special solution of Eq. (1) with initial condition $\phi \in C^*$ if the following two conditions are satisfied:

- (c) x is piecewise continuous with first kind discontinuity at the points t_k , $k \in \{1, \dots, p\}$.
- (d) x satisfies Eq. (1) for $t \geq 0$, and $x(\theta) = \phi(\theta)$ for $\theta \in [-\tau, 0]$.

Henceforth, we let $x(t, \phi)$ denote the special solution of Eq. (1) with initial condition $\phi \in C^*$

Definition 3 Equation (1) is said to be globally exponentially periodic if it possesses a periodic solution $x(t, \phi^*)$, and $x(t, \phi^*)$ is globally exponentially stable. That is, there exist positive constants ε and M such that every solution of Eq. (1) satisfies

$$\|x(t, \phi) - x(t, \phi^*)\|_\infty \leq M \|\phi - \phi^*\| e^{-\varepsilon t}, \text{ for all } t \geq 0.$$

2 Main Result

Now we define $\psi(t) = t - \tau_i(t)$, then $\psi^{-1}(t)$ has inverse function ν . Set

$$\delta_i = \max \left\{ \frac{1}{1 - \dot{\tau}_i(\psi_i^{-1}(t))} : t \in \mathbb{R} \right\}, i = 1, 2, \dots, n.$$

Theorem 4 Equation (1) is globally exponentially periodic if the following two conditions are satisfied:

- (H₄) $|1 + \gamma_{ik}| \leq 1$, for all $i \in \{1, \dots, n\}$, and $k \in \{1, \dots, p\}$,
 (H₅) There exist positive numbers $\alpha_1, \alpha_2, \dots, \alpha_n$, such that

$$\alpha_i a_i > L_i \sum_{j=1}^n \alpha_j (|b_{ji}| + \delta_j |c_{ji}|), i = 1, 2, \dots, n.$$

In order to prove Theorem 4, we need the following Lemma.

Lemma 5 Let $x(t, \phi), x(t, \varphi)$ be a pair of solutions of Eq. (1). If the two conditions given in Theorem 4 are satisfied, then there is a positive number ε such that,

$$\|x(t, \phi) - x(t, \varphi)\|_\infty \leq M(\varepsilon) \|\phi - \varphi\|_\infty e^{-\varepsilon t}, \text{ for all } t \geq 0.$$

where

$$M(\varepsilon) = \frac{1}{\min_{1 \leq j \leq n} \alpha_j} \sum_{i=1}^n \alpha_i \left[1 + \frac{1}{\varepsilon} L_j \delta_j |c_{ij}| (e^{\varepsilon \tau} - 1) \right].$$

Proof Let $x(t, \phi) = (x_1(t, \phi), x_2(t, \phi), \dots, x_n(t, \phi))^T$ and $x(t, \varphi) = (x_1(t, \varphi), x_2(t, \varphi), \dots, x_n(t, \varphi))^T$ be an arbitrary pair of solutions of Eq. (1). Let

$$\begin{aligned} \Delta x_i(t, \phi, \varphi) &= x_i(t, \phi) - x_i(t, \varphi), \\ \Delta g_j(x_j(t, \phi, \varphi)) &= g_j(x_j(t, \phi)) - g_j(x_j(t, \varphi)), \end{aligned}$$

$$\begin{aligned} V(t) &= \sum_{i=1}^n \alpha_i \left\{ |\Delta x_i(t, \phi, \varphi)| e^{-\varepsilon t} \right. \\ &\quad \left. + \sum_{j=1}^n \int_{t-\tau_j(t)}^t \frac{L_j |c_{ij}| |\Delta x_j(s, \phi, \varphi)|}{1 - \tau_j(\psi_j^{-1}(s))} e^{\varepsilon(s+\tau_j(\psi_j^{-1}(s)))} ds \right\}. \end{aligned} \quad (2)$$

We proceed by considering two possibilities.

Case 1. $t \neq t_k$ for all $k \in \{1, \dots, p\}$. From the second condition in Theorem 4, there is a small positive number ε such that

$$\alpha_i (a_i - \varepsilon) > L_i \sum_{j=1}^n \alpha_j (|b_{ji}| + \delta_j |c_{ji}| e^{\varepsilon \tau}), \quad (3)$$

where $i = 1, \dots, n$. Calculating the derivatives of $V(t)$ along the solutions of Eq. (1), we get

$$\begin{aligned}
 D^+V(t) = & \sum_{i=1}^n \alpha_i \left\{ e^{-\varepsilon t} D^+|\Delta x_i(t, \phi, \varphi)| \right. \\
 & + \sum_{j=1}^n \left[\frac{L_j |c_{ij}|}{1 - \tau_j(\psi_j^{-1}(t))} |\Delta x_j(t, \phi, \varphi)| \cdot e^{\varepsilon(t+\tau_j(\psi_j^{-1}(t)))} \right. \\
 & \left. \left. - L_j |c_{ij}| |\Delta x_j(t - \tau_j(t), \phi, \varphi)| e^{\varepsilon t} \right] \right. \\
 & \left. + \varepsilon e^{-\varepsilon t} |\Delta x_i(t, \phi, \varphi)| \right\}. \tag{4}
 \end{aligned}$$

Note that for $i = 1, \dots, n$,

$$\begin{aligned}
 \dot{x}_i(t, \phi) - \dot{x}_i(t, \varphi) = & -a_i \Delta x_i(t, \phi, \varphi) \\
 & + \sum_{j=1}^n b_{ij} \Delta g_j(x_j(t, \phi, \varphi)) \\
 & + \sum_{j=1}^n c_{ij} \Delta g_j(x_j(t - \tau_j(t), \phi, \varphi)),
 \end{aligned}$$

which plus (H_1) yields

$$\begin{aligned}
 D^+|x(t, \phi) - x(t, \varphi)| \leq & -a_i |\Delta x_i(t, \phi, \varphi)| + \sum_{j=1}^n L_j |b_{ij}| |\Delta x_j(t, \phi, \varphi)| \\
 & + \sum_{j=1}^n L_j |c_{ij}| |\Delta x_j(t - \tau_j(t), \phi, \varphi)|. \tag{5}
 \end{aligned}$$

Substituting Eq. (5) into Eq. (4), we obtain

$$\begin{aligned}
 D^+V(t) = & \sum_{i=1}^n \alpha_i \left[-a_i e^{-\varepsilon t} |\Delta x_i(t, \phi, \varphi)| \right. \\
 & + e^{-\varepsilon t} \sum_{j=1}^n L_j |b_{ij}| |\Delta x_j(t, \phi, \varphi)| \\
 & + e^{-\varepsilon t} \sum_{j=1}^n L_j |c_{ij}| |\Delta x_j(t - \tau_j(t), \phi, \varphi)| \\
 & + \varepsilon e^{-\varepsilon t} |\Delta x_i(t, \phi, \varphi)| \\
 & \left. + \sum_{j=1}^n \frac{L_j |c_{ij}|}{1 - \tau_j(\psi_j^{-1}(t))} |\Delta x_j(t, \phi, \varphi)| e^{\varepsilon(t+\tau_j(\psi_j^{-1}(t)))} \right]
 \end{aligned}$$

$$\begin{aligned}
& - \sum_{j=1}^n L_j |c_{ij}| |\Delta x_j(t - \tau_j(t), \phi, \varphi)| e^{\varepsilon t} \Big] \\
& \leq e^{-\varepsilon t} \sum_{i=1}^n \alpha_i \left[(\varepsilon - a_i) |\Delta x_i(t, \phi, \varphi)| \right. \\
& \quad + \sum_{j=1}^n L_j |b_{ij}| |\Delta x_j(t, \phi, \varphi)| \\
& \quad \left. + \sum_{j=1}^n L_j \delta_j |c_{ij}| e^{\varepsilon \tau} |\Delta x_j(t, \phi, \varphi)| \right] \\
& = e^{-\varepsilon t} \sum_{i=1}^n \left[\alpha_i (\varepsilon - a_i) \right. \\
& \quad \left. + L_i \sum_{j=1}^n (|b_{ji}| + \delta_i |c_{ji}| e^{\varepsilon \tau}) \right] \cdot |\Delta x_i(t, \phi, \varphi)| \leq 0. \tag{6}
\end{aligned}$$

Case 2. $t = t_k$, for some $k \in \{1, 2, \dots, p\}$. Then

$$\begin{aligned}
V(t+0) &= \sum_{i=1}^n \alpha_i \left[|\Delta x_i(t+0, \phi, \varphi)| e^{-\varepsilon t} \right. \\
& \quad \left. + \sum_{j=1}^n \int_{t-\tau_j(t)}^t \frac{L_j |c_{ij}| |\Delta x_j(s, \phi, \varphi)|}{1 - \dot{\tau}_j(\psi_j^{-1}(s))} e^{\varepsilon(s+\tau_j(\psi_j^{-1}(s)))} ds \right].
\end{aligned}$$

According to Eq. (2) and (H₄), we obtain

$$\begin{aligned}
V(t+0) - V(t) &= e^{-\varepsilon t} \sum_{i=1}^n \alpha_i \left(|\Delta x_i(t+0, \phi, \varphi)| - |\Delta x_i(t, \phi, \varphi)| \right) \\
&= -e^{-\varepsilon t} \sum_{i=1}^n \alpha_i (1 - |1 + \gamma_{ik}|) |\Delta x_i(t, \phi, \varphi)| \leq 0.
\end{aligned}$$

Namely, $V(t+0) \leq V(t)$.

Combining the above discussions, we obtain $V(t) \leq V(0)$ for all $t \geq 0$. This plus the inspections that

$$\begin{aligned}
V(t) &\geq e^{\varepsilon t} \sum_{i=1}^n \alpha_i |\Delta x_i(t, \phi, \varphi)| \\
&\geq \min_{1 \leq j \leq n} \alpha_j e^{\varepsilon t} \sum_{i=1}^n |\Delta x_i(t, \phi, \varphi)|
\end{aligned}$$

$$\geq \min_{1 \leq j \leq n} \alpha_j e^{\varepsilon t} \|x_i(t, \phi) - x_i(t, \varphi)\|_\infty, \tag{7}$$

and

$$\begin{aligned} V(0) &= \sum_{i=1}^n \alpha_i \left[|x_i(0, \phi) - x_i(0, \varphi)| \right. \\ &\quad \left. + \sum_{j=1}^n \int_{-\tau_j(0)}^0 \frac{L_j |c_{ij}|}{1 - \tau_j(\psi_j^{-1}(s))} \cdot \Delta x_j(s, \phi, \varphi) e^{\varepsilon(s+\tau_j(\psi_j^{-1}(s)))} ds \right] \\ &\leq \sum_{i=1}^n \alpha_i \left[|\phi_i(0) - \varphi_i(0)| \right. \\ &\quad \left. + \sum_{j=1}^n \int_{-\tau}^0 L_j \delta_j |c_{ij}| e^{\varepsilon \tau} |\phi_j(s) - \varphi_j(s)| e^{\varepsilon s} ds \right] \\ &\leq \sum_{i=1}^n \alpha_i \left[1 + \frac{L_j \delta_j}{\varepsilon} |c_{ij}| (e^{\varepsilon \tau} - 1) \right] \|\phi - \varphi\|_\infty. \end{aligned} \tag{8}$$

This implies that the conclusion of the Lemma hold by using Eqs. (6)–(8). \square

Proof of Theorem 4. First, we prove that Eq. (1) possesses an ω -periodic solution. For each solution $x(t, \phi)$ of Eq. (1) and each $t \geq 0$, we can define a function $x_t(\phi)$ in this fashion:

$$x_t(\phi)(\theta) = x(t + \theta, \phi), \text{ for } \theta \in [-\tau, 0].$$

On this basis, we can define a mapping $P : C^* \rightarrow C^*$ by

$$P\phi = x_\omega(\phi).$$

Let $x(t, \phi), x(t, \varphi)$ be an arbitrary pair of solutions of Eq. (1). Let ε be a positive number satisfying Eq. (3). Let $m \geq \frac{1}{\varepsilon \omega} \ln(2M(\varepsilon)) + 1$ be a positive integer. It follows from Lemma 5 that

$$\begin{aligned} &\|P^m \phi - P^m \varphi\| \\ &= \sup_{-\tau \leq \theta \leq 0} \|x(m\omega + \theta, \phi) - x(m\omega + \theta, \varphi)\|_\infty \\ &\leq M(\varepsilon) \sup_{-\tau \leq \theta \leq 0} e^{-\varepsilon(m\omega + \theta)} \|\phi - \varphi\|_\infty \\ &\leq M(\varepsilon) e^{-\varepsilon(m-1)\omega} \|\phi - \varphi\|_\infty \leq \frac{1}{2} \|\phi - \varphi\|_\infty, \end{aligned}$$

which shows that P^m is a contraction mapping on the Banach space C^* . According to the contraction mapping principle, P^m possesses a unique fixed point $\phi^* \in C^*$. Note that

$$P^m(P\phi^*) = P(P^m\phi^*) = P\phi^*.$$

which indicates that $P\phi^* \in C^*$ is also a fixed point of P^m . It follows from the uniqueness of fixed point of P^m that $P\phi^* = \phi^*$, viz. $x_\omega(\phi^*) = \phi^*$. Let $x(t, \phi^*)$ be the solution of Eq. (1) with initial condition ϕ^* , then

$$x_{t+\omega}(\phi^*)(\theta) = x_t(x_\omega(\phi^*)) = x_t(\phi^*) \quad \text{for } t \geq 0.$$

which implies

$$x(t + \omega, \phi^*) = x_{t+\omega}(\phi^*)(0) = x_t(x_\omega(0)) = x(t, \phi^*).$$

Thus, $x(t, \phi^*)$ is ω -periodic of Eq. (1).

On the other hand, it follows from Lemma 5 that every solution $x(t, \phi)$ of Eq. (1) satisfies

$$\|x(t, \phi) - x(t, \phi^*)\|_\infty \leq M(\varepsilon)\|\phi - \phi^*\|_\infty e^{-\varepsilon t},$$

for all $t \geq 0$. This shows that $x(t, \phi)$ is globally exponentially periodic. \square

3 An Illustrative Example

Consider the impulsive Hopfield neural network with time-varying delays:

$$\begin{aligned} \begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} &= \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \\ &+ \begin{pmatrix} 0.6 & 0.3 \\ 0.3 & -0.5 \end{pmatrix} \begin{pmatrix} \sin \frac{1}{\sqrt{2}}x_1(t) \\ \sin \frac{1}{2\sqrt{2}}x_2(t) \end{pmatrix} \\ &+ \begin{pmatrix} 0.8 & -0.5 \\ -0.6 & 0.6 \end{pmatrix} \begin{pmatrix} \sin \frac{1}{\sqrt{2}}x_1(t - \tau_1(t)) \\ \sin \frac{1}{2\sqrt{2}}x_2(t - \tau_2(t)) \end{pmatrix} \\ &+ \begin{pmatrix} 1 - \cos 2\pi t \\ 1 + \sin 2\pi t \end{pmatrix}, \\ \Delta x_1(t_k) &= \gamma_{1k}x_1(t_k), \\ \Delta x_2(t_k) &= \gamma_{2k}x_2(t_k). \end{aligned} \tag{9}$$

Obviously, the right hand side of Eq. (9) is 1-periodic (i.e. $\omega = 1$). Now we investigate the influence of the delay and the period T of impulsive effect on the Eq. (9). If $\tau(t) = \frac{1}{5}\pi$, $T = 1$, $\gamma_{1k} = \gamma_{2k} = 0.1$, then $p = 1$ in (H₂). According to Theorem 4, impulsive Hopfield neural networks Eq. (9) has a unique 1-periodic solution which is globally asymptotically stable (see Figs. 1, 2, 3, 4). In order to clearly observe the

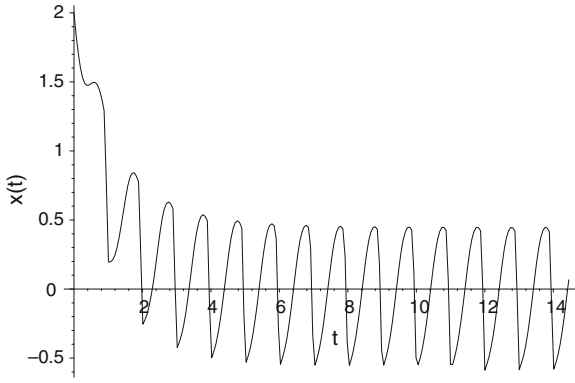


Fig. 1 Time-series of the $x_1(t)$ of Eq. (9) for $t \in [0, 16]$

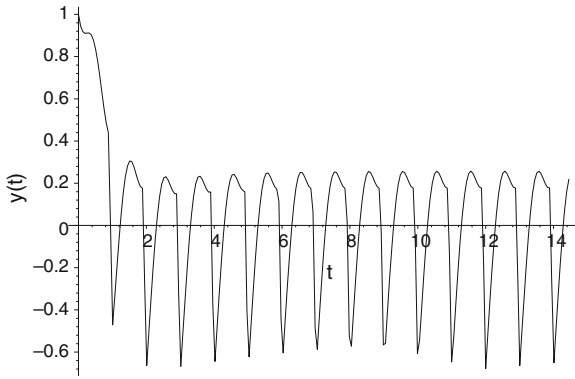


Fig. 2 Time-series of the $x_2(t)$ of Eq. (9) for $t \in [0, 16]$

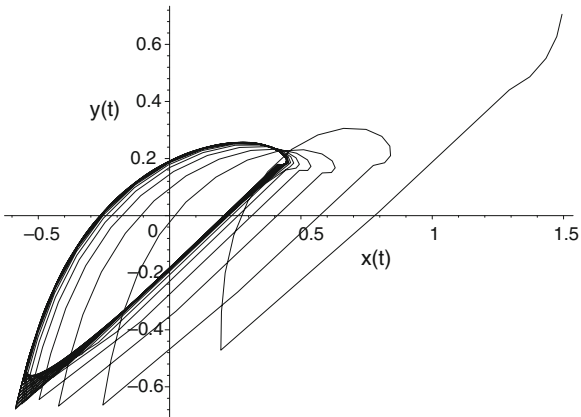


Fig. 3 Phase portrait of 1-periodic solutions of Eq. (9) for $t \in [0, 42]$

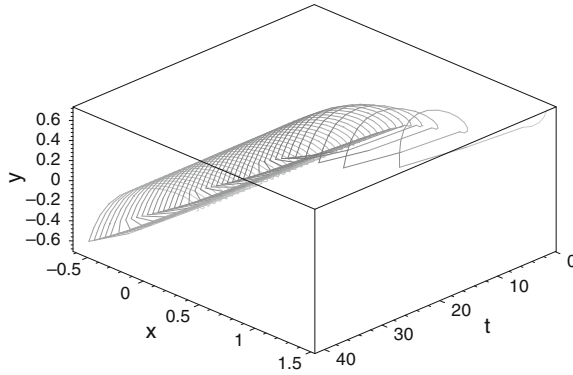


Fig. 4 Space figure of 1-periodic solutions of Eq. (9) by adding a time coordinate axes t

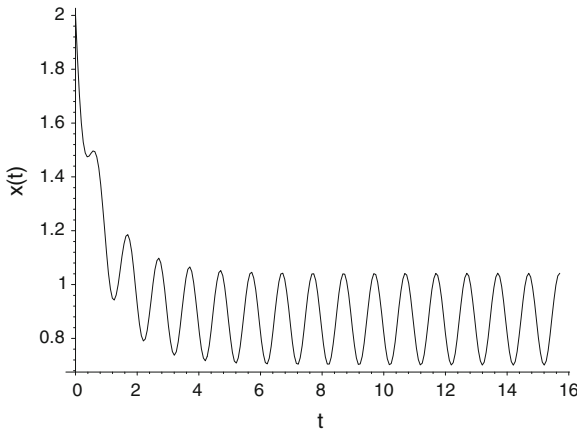


Fig. 5 Time-series of the $x_1(t)$ of Eq. (9) for $t \in [0, 16]$ with $\tau(t) = \frac{1}{5}\pi$

change trend of the solutions, we add a time coordinate axes to the Fig. 4 and change 2-D plan (Fig. 3) into 3-D space (Fig. 4).

If the effect of impulse is ignored, i.e. $\gamma_{1k} = 0, \gamma_{2k} = 0$, then Eq. (9) becomes periodic system. Obviously, the right hand side of Eq. (9) is 1-periodic. Numeric results show that Eq. (9) has a 1-periodic solution Fig. 5. Figures 6, 7, 8 show the dynamic behavior of the Eq. (9) with $\tau(t) = \frac{1}{5}\pi$.

Furthermore, If $\tau(t) = \frac{1}{5}\pi$ rises to $\tau(t) = \pi$ gradually, then periodic oscillation of Eq. (9) will be destroyed. Numeric results show that Eq. (9) still has a global attractor which may be gui chaotic strange attractor (see Figs. 9, 10, 11, 12). Every solutions of Eq. (9) will finally tend to the chaotic strange attractor.

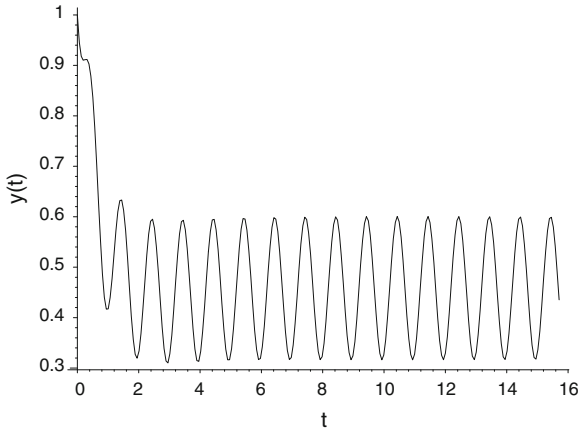


Fig. 6 Time-series of the $x_2(t)$ of Eq. (9) for $t \in [0, 16]$ with $\tau(t) = \frac{1}{5}\pi$

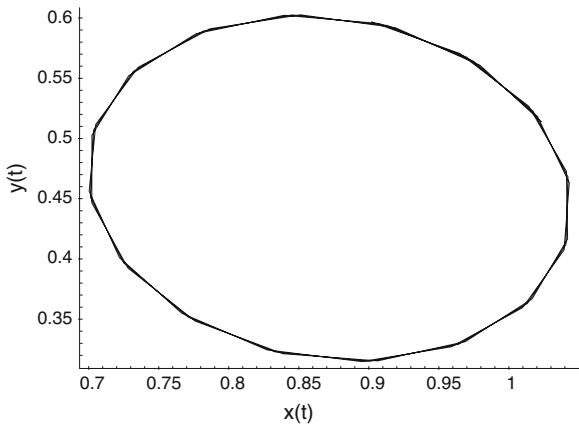


Fig. 7 Phase portrait of 1-periodic solutions of Eq.(9) with $\tau(t) = \frac{1}{5}\pi$

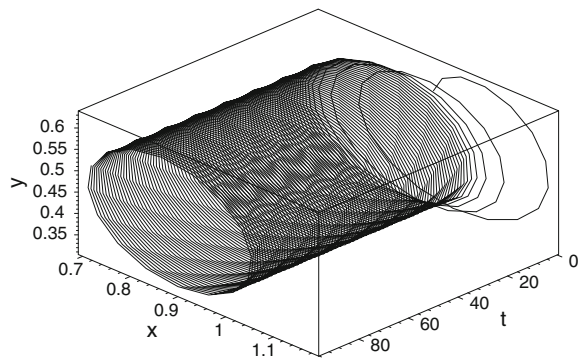


Fig. 8 Space figure of 1-periodic solutions of Eq. (9) by adding a coordinate axes t

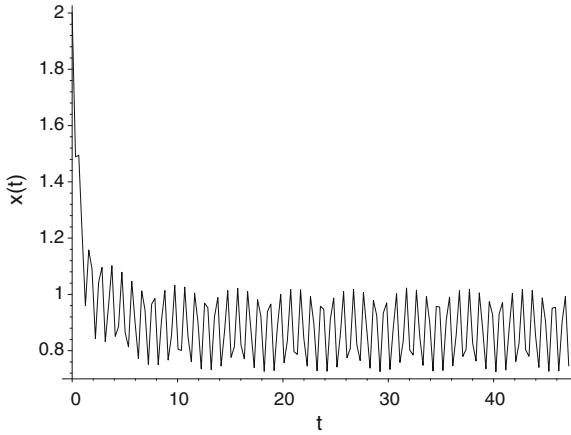


Fig. 9 Time-series of the $x_1(t)$ of Eq.(9) for $t \in [0, 48]$ with $\tau(t) = \pi$

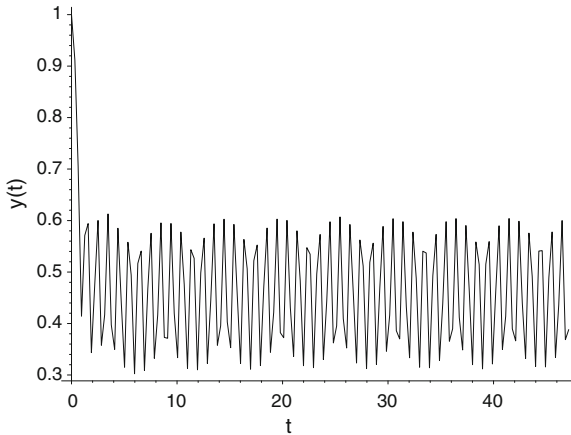


Fig. 10 Time-series of the $x_2(t)$ of Eq.(9) for $t \in [0, 48]$ with $\tau(t) = \pi$

4 Conclusion

We have established a sufficient condition for the existence and global exponential stability of a unique periodic solution in a class of HNNs with time-varying delays and periodic impulses, which assumes neither the differentiability nor the monotonicity of the activation functions.

Our condition extends and generalizes a known condition for the global exponential periodicity of pure continuous Hopfield neural networks with time-varying delays. Further the numerical simulation shows that our system can occur many forms of complexities including chaotic strange attractor and periodic solution.

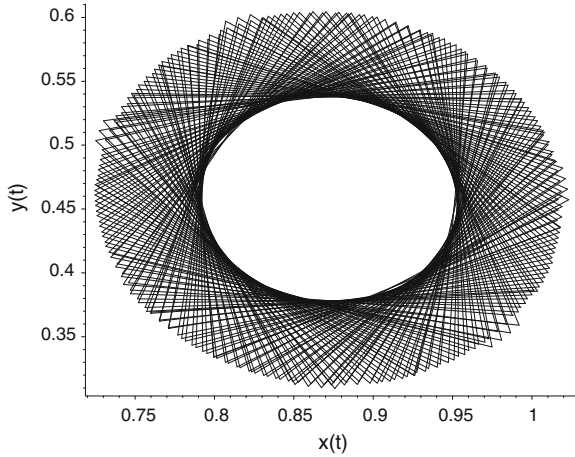


Fig. 11 Phase portrait of chaotic strange attractor of Eq.(9) with $\tau(t) = \pi$

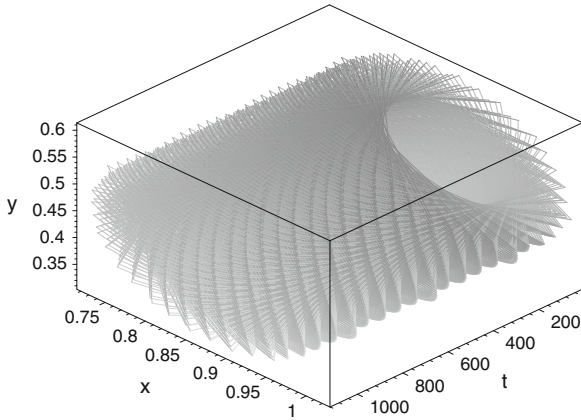


Fig. 12 Space figure of attractor of Eq. (9) by adding a coordinate axes t

In recent years, numerous results have been reported on the stability of discrete as well as continuous neural networks. It is worthwhile to introduce various impulsive neural networks and then establish the corresponding stability results that include some known results for pure discrete or continuous neural networks as special cases.

References

1. Abou-El-Ela AMA, Sadekand AI, Mahmoud AM (2012) Existence and uniqueness of a periodic solution for third-order delay differential equation with two deviating arguments. IAENG Int J Appl Math, 42:1 IJAM_42_1_02

2. Akca H, Alassar R, Covachev V, Covacheva Z, Al-Zahrani E (2004) Continuous-time additive Hopfield-type neural networks with impulses. *J Math Anal Appl* 290(2):436–451
3. Arika S, Tavsanoğlu V (2005) Global asymptotic stability analysis of bidirectional associative memory neural networks with constant time delays. *Neurocomputing* 68:161–176
4. Bainov DD, Simenov PS (1989) *Systems with impulse effect: stability theory and applications*. Ellis Horwood, Chichester
5. Cao J, Jiang Q (2004) An analysis of periodic solutions of bi-directional associative memory networks with time-varying delays. *Phys Lett A* 330(3–4):203–213
6. Cheng Y, Yan Y, Gui Z (2012) Existence and stability of periodic solution in impulsive Hopfield neural networks with time-varying delays. *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering WCE 2012, 4–6 July 2012 U.K, London*, pp 18–23
7. Gopalsamy K, Zhang BG (1989) On delay differential equation with impulses. *J Math Anal Appl* 139(1):110–C122
8. Gui Z, Ge W (2006) Periodic solution and chaotic strange attractor for shunting inhibitory cellular neural networks with impulses. *Chaos* 16(3):1–10
9. Gui Z, Yang XS (2006) Stability and existence of periodic solutions of periodic cellular neural networks. *Comput Math Appl* 52(12):1657–1670
10. Gui Z, Ge W (2007) Impulsive effect of continuous-time neural networks under pure structural variations. *Int J Bifurcat Chaos* 17(6):2127–2139
11. Gui Z, Yang XS, Ge W (2007) Periodic solution for nonautonomous bidirectional associative memory neural networks with impulses. *Neurocomputing* 70(13–15):2517–2527
12. Lakshmikantham V, Bainov DD, Simeonov P-S (1989) *Theory of impulsive differential equations*. World Scientific, Singapore
13. Li S, Liao X, Li C (2005) Hopf bifurcation of a two-neuron network with different discrete time delays. *Int J Bifurcat Chaos* 15(5):1589–1601
14. Li Y (2005) Global exponential stability of BAM neural networks with delays and impulses. *Chaos, Solitons Fractals* 24(1):279–285
15. Li Y, Xing W, Lu L (2006) Existence and global exponential stability of periodic solution of a class of neural networks with impulses. *Chaos, Solitons Fractals* 27(2):437–445
16. Song Y, Peng Y, Wei J (2008) Bifurcations for a predator-prey system with two delays. *J Math Anal Appl* 337(1):466–479
17. Xu S, Lam J, Zou Y (2005) Delay-dependent approach to stabilization of time-delay chaotic systems via standard and delayed feedback controllers. *Int J Bifurcat Chaos* 15(4):1455–1465
18. Yang X, Liao X, Evans DJ, Tang Y (2005) Existence and stability of periodic solution in impulsive Hopfield neural networks with finite distributed delays. *Phys Lett A* 343(1–3):108–116

Solving Stiff Ordinary Differential Equations Using Extended Block Backward Differentiation Formulae

Siti Ainor Mohd Yatim, Zarina Bibi Ibrahim, Khairil Iskandar Othman and Mohamed Suleiman

Abstract A comprehensive research on the existing Block Backward Differentiation Formulae (BBDF) was done. Based on the suitability in solving stiff ordinary differential equations (ODEs), BBDF of order 3 up 5 is collected using simplified strategy in controlling the step size and order of the method. Thus, Extended Block Backward Differentiation Formulae (EBBDF) is derived with the intention of optimizing the performance in terms of precision and computation time. The accuracy of the method are investigated using linear and non linear stiff initial value problems and its performance is compared with MATLAB's suite of ODEs solvers namely ode15s and ode23s.

Keywords BDF methods · Block methods · Initial value problem · Numerical analysis · Ordinary differential equations · Stiff ODEs

S. A. M. Yatim (✉)
School of Distance Education, Universiti Sains Malaysia USM,
11800 Penang, Malaysia
e-mail: ainormy@yahoo.com

Z. B. Ibrahim · M. Suleiman
Department of Mathematics, Universiti Putra Malaysia UPM,
43400 Serdang, Selangor, Malaysia
e-mail: zarinabb@science.upm.edu.my

M. Suleiman
e-mail: mohameds@science.upm.edu.my

K. I. Othman
Department of Mathematics, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
e-mail: khairil@tmsk.uitm.edu.my

1 Introduction

Applied problems arise from chemical, biological and physical phenomenon, particularly in many field of science and engineering [1–4], have inspired numbers of researches to develop effective and very accurate methods to solve stiff initial value problems (IVP) [5]. Many renowned method for solving stiff problems are based on backward differentiation formula (BDF) which also known as Gear’s method and was first introduced in [6]. The method was then developed to improve the results in terms of accuracy and computation time. The evidence is in the advancements of many other codes to meet the same objective of finding the most accurate approximation for IVPs. These codes include EPISODE, VODE, LSODE, MEBDF etc. [5–7].

Many in recent chapters have tried to describe and compare some of the best codes by taking into accounts their accuracy, rate of convergence, and computation time [8]. Consequently, some excellent codes which are both efficient and reliable for solving these particular classes of problems are made available. With the same objective, this chapter consider the numerical solution of the first order initial value problem,

$$y' = f(x, y) \quad (1)$$

with given initial values $y(a) = y_0$ in the given interval $x \in [a, b]$.

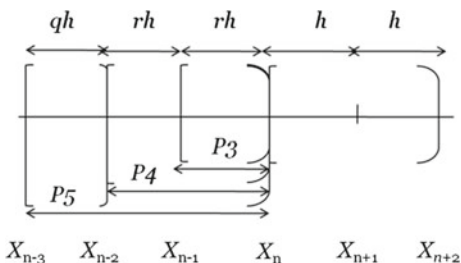
From the study on BBDF in [9], the competency of computing concurrent solution values at different points were revealed. The idea in [10, 11] was then extended by increasing the order of the method up to order 5. In this chapter, we extend and review the study of Extended Block Backward Differentiation Formulae (EBBDF) in [12]. Additionally, this chapter investigates the consistency and convergence of the method.

2 Derivation of Extended Block Backward Differentiation Formulae Method

2.1 Construction of EBBDF

Two values of y_{n+1} and y_{n+2} were computed simultaneously in block by using earlier blocks with each block containing a maximum of two points (Fig. 1). The orders of the method ($P3$, $P4$ and $P5$) are distinguished by the number of backvalues contained in total blocks. The ratio distance between current (x_n) and previous step (x_{n-1}) is represented as r and q in Fig. 1. In this chapter, the step size is given selection to decrease to half of the previous steps, or increase up to a factor of 1.9. For simplicity, q is assigned as 1, 2 and 10/19 for the case of constant, halving and increasing the step size respectively. The zero stability is achieved for each of these cases and explained in the next section.

Fig. 1 EBBDF method of order ($P3 - P5$)



We find approximating polynomials $P_k(x)$, by means of a k -degree polynomial interpolating the values of y at given points are $(x_{n-3}, y_{n-3}), (x_{n-2}, y_{n-2}), (x_{n-1}, y_{n-1}) \dots, (x_{n+2}, y_{n+2})$.

$$P_k = \sum_{j=0}^k y(x_{n+1-j}) \cdot L_{k,j}(x) \tag{2}$$

where

$$L_{k,j}(x) = \prod_{\substack{i=0 \\ i \neq j}}^k \frac{(x - x_{n+1-i})}{(x_{n+1-j} - x_{n+1-i})} \text{ for each } j = 0, 1, \dots, k.$$

The interpolating polynomial of the function $y(x)$ using Lagrange polynomial in (2) gives the following corrector for the first point y_{n+1}^P , and second point y_{n+2}^P . The resulting Lagrange polynomial for each order was given as follows:

For EBBDF of order $P3$ ($P = 3$)

$$\begin{aligned} P(x) &= P(x_{n+1} + sh) \\ &= \frac{(r + 1 + s)(s + 1)(s)}{2r + 4} y_{n+2} + \frac{(r + 1 + s)(s + 1)(s - 1)}{-1 - r} y_{n+1} \\ &\quad + \frac{(r + 1 + s)(s - 1)(s)}{2r} y_n + \frac{(1 + s)(s - 1)(s)}{-r(-1 - r)(-r - 2)} y_{n-1} \end{aligned} \tag{3}$$

For EBBDF of order $P4$ ($P = 4$)

$$\begin{aligned} P(x) &= P(x_{n+1} + sh) \\ &= \frac{(2r + 1 + s)(r + 1 + s)(1 + s)(s)}{2(2r + 2)(r + 2)} y_{n+2} \\ &\quad + \frac{(2r + 1 + s)(r + 1 + s)(1 + s)(s - 1)}{-(2r + 1)(r + 1)} y_{n+1} \end{aligned}$$

$$\begin{aligned}
& + \frac{(2r+1+s)(r+1+s)(s)(s-1)}{4r^2} y_n \\
& + \frac{(2r+1+s)(1+s)(s)(s-1)}{-r^2(-r-1)(-r-2)} y_{n-1} \\
& + \frac{(r+1+s)(1+s)(s)(s-1)}{2r^2(-2r-1)(-2r-2)} y_{n-2}
\end{aligned} \tag{4}$$

For EBBDF of order $P5$ ($P = 5$)

$$\begin{aligned}
P(x) &= P(x_{n+1} + sh) \\
&= \frac{(q+2r+1+s)(2r+1+s)(r+1+s)(1+s)s}{2(q+2r+2)(2r+2)(r+2)} y_{n+2} \\
&+ \frac{(q+2r+1+s)(2r+1+s)(r+1+s)(1+s)(s-1)}{-(q+2r+1)(2r+1)(r+1)} y_{n+1} \\
&+ \frac{(q+2r+1+s)(2r+1+s)(r+1+s)s(s-1)}{4(q+2r)r^2} y_n \\
&+ \frac{(q+2r+1+s)(2r+1+s)(1+s)s(s-1)}{-r^2(q+r)(-r-1)(-r-2)} y_{n-1} \\
&+ \frac{(q+2r+1+s)(r+1+s)(1+s)s(s-1)}{2qr^2(-2r-1)(-2r-2)} y_{n-2} \\
&+ \frac{(2r+1+s)(r+1+s)(1+s)s(s-1)}{-q(-q-r)(-q-2r)(-q-2r-1)(-q-2r-2)} y_{n-3}
\end{aligned} \tag{5}$$

By substituting $s = 0$ and $s = 1$ gives the corrector for the first and second point respectively. Therefore by letting $r = 1, q = 1, r = 2, q = 2$ and $r = 1, q = 10/19$ we produced the following equations for the first and second point of EBBDF.

EBBDF of order $P3$ ($p = 3$)

When $r = 1, q = 1$

$$\begin{aligned}
y_{n+1} &= 2hf_{n+1} - \frac{2}{3}y_{n+2} + 2y_n - \frac{1}{3}y_{n-1} \\
y_{n+2} &= \frac{6}{11}hf_{n+2} + \frac{18}{11}y_{n+1} - \frac{9}{11}y_n + \frac{2}{11}y_{n-1}
\end{aligned}$$

When $r = 2, q = 2$

$$\begin{aligned}
y_{n+1} &= 3hf_{n+1} + \frac{9}{8}y_{n+2} + \frac{9}{4}y_n - \frac{1}{8}y_{n-1} \\
y_{n+2} &= \frac{4}{7}hf_{n+2} + \frac{32}{21}y_{n+1} - \frac{4}{7}y_n + \frac{1}{21}y_{n-1}
\end{aligned}$$

When $r = 1, q = 10/19$

$$y_{n+1} = \frac{29}{19}hf_{n+1} - \frac{841}{1824}y_{n+2} + \frac{841}{380}y_n - \frac{361}{480}y_{n-1}$$

$$y_{n+2} = \frac{48}{91}hf_{n+2} + \frac{4608}{2639}y_{n+1} - \frac{576}{455}y_n + \frac{6859}{13195}y_{n-1}$$

EBBDF of order $P4$ ($P = 4$)

When $r = 1, q = 1$

$$y_{n+1} = \frac{6}{5}hf_{n+1} - \frac{3}{10}y_{n+2} + \frac{9}{5}y_n - \frac{3}{5}y_{n-1} + \frac{1}{10}y_{n-2}$$

$$y_{n+2} = \frac{12}{25}hf_{n+2} + \frac{48}{25}y_{n+1} - \frac{36}{25}y_n + \frac{16}{25}y_{n-1} - \frac{3}{25}y_{n-2}$$

When $r = 2, q = 2$

$$y_{n+1} = \frac{15}{8}hf_{n+1} - \frac{75}{128}y_{n+2} + \frac{225}{128}y_n - \frac{25}{128}y_{n-1} + \frac{3}{128}y_{n-2}$$

$$y_{n+2} = \frac{12}{23}hf_{n+2} + \frac{192}{115}y_{n+1} - \frac{18}{23}y_n + \frac{3}{23}y_{n-1} - \frac{2}{115}y_{n-2}$$

When $r = 1, q = 10/19$

$$y_{n+1} = \frac{1131}{1292}hf_{n+1} - \frac{14703}{82688}y_{n+2} + \frac{1279161}{516800}y_n$$

$$- \frac{183027}{108800}y_{n-1} + \frac{10469}{27200}y_{n-2}$$

$$y_{n+2} = \frac{1392}{3095}hf_{n+2} + \frac{89088}{40235}y_{n+1} - \frac{242208}{77375}y_n$$

$$+ \frac{198911}{77375}y_{n-1} - \frac{658464}{1005875}y_{n-2}$$

EBBDF of order $P5$ ($P = 5$)

When $r = 1, q = 1$

$$y_{n+1} = \frac{12}{13}hf_{n+1} - \frac{12}{65}y_{n+2} + \frac{24}{13}y_n - \frac{12}{13}y_{n-1}$$

$$+ \frac{4}{13}y_{n-2} - \frac{3}{65}y_{n-3}$$

$$y_{n+2} = \frac{60}{137}hf_{n+2} + \frac{300}{137}y_{n+1} - \frac{300}{137}y_n + \frac{200}{137}y_{n-1}$$

$$- \frac{75}{137}y_{n-2} + \frac{12}{137}y_{n-3}$$

When $r = 2, q = 2$

$$\begin{aligned}
 y_{n+1} &= \frac{105}{71}hf_{n+1} - \frac{3675}{9088}y_{n+2} + \frac{3675}{2272}y_n - \frac{1225}{4544}y_{n-1} \\
 &\quad + \frac{147}{2272}y_{n-2} - \frac{75}{9088}y_{n-3} \\
 y_{n+2} &= \frac{24}{49}hf_{n+2} + \frac{3072}{1715}y_{n+1} - \frac{48}{49}y_n + \frac{12}{49}y_{n-1} \\
 &\quad - \frac{16}{245}y_{n-2} + \frac{3}{343}y_{n-3}
 \end{aligned}$$

When $r = 1, q = 10/19$

$$\begin{aligned}
 y_{n+1} &= \frac{402}{449}hf_{n+1} - \frac{13467}{77228}y_{n+2} + \frac{13467}{7184}y_n - \frac{13467}{13021}y_{n-1} \\
 &\quad + \frac{4489}{8980}y_{n-2} - \frac{7428297}{44792240}y_{n-3} \\
 y_{n+2} &= \frac{516}{1189}hf_{n+2} + \frac{177504}{79663}y_{n+1} - \frac{5547}{2378}y_n + \frac{59168}{34481}y_{n-1} \\
 &\quad - \frac{5547}{5945}y_{n-2} + \frac{7428297}{23102270}y_{n-3}
 \end{aligned}$$

As similar to analysis for order of Linear Multistep Method (LMM) given in [13], we use the following definition to determine the order of EBBDF method.

Definition 1 The LMM [13] and the associated difference operator L defined by

$$L[z(x); h] = \sum_{k=0}^j [\alpha_k z(x + kh) - h\beta_k z'(x + kh)] \tag{6}$$

are said to be of order p if $c_0 = c_1 = \dots = c_p = 0, C_{p+1} \neq 0$. The general form for the constant C_q is defined as

$$C_q = \sum_{k=0}^j \left[k^q \alpha_k - \frac{1}{(q-1)!} k^{q-1} \beta_k \right], \quad q = 2, 3, \dots, p+1 \tag{7}$$

Consequently, BBDF method can be represented in standard form by an equation $\sum_{j=0}^k A_j y_{n+j} = h \sum_{j=0}^k B_j f_{n+j}$ where A_j and B_j are r by r matrices with elements $a_{l,m}$ and $b_{l,m}$ for $l, m = 1, 2, \dots, r$. Since EBBDF for variable order (P) is a block method, we extend the Definition 1 in the form of

$$L[z(x); h] = \sum_{k=0}^j [A_k z(x + kh) - hB_k z'(x + kh)] \tag{8}$$

And the general form for the constant C_q is defined as

$$C_q = \sum_{k=0}^j \left[k^q A_k - \frac{1}{(q-1)!} k^{q-1} B_k \right] \quad q = 2, 3, \dots, p+1 \quad (9)$$

A_k is equal to the coefficients of y_k where $k = n = (p-2), \dots, n+1, n+2$ and $P = 3, 4, 5$.

Throughout this section, we illustrate the effect of Newton-type scheme which in general form of

$$y_{n+1,n+2}^{(i+1)} - y_{n+1,n+2}^{(i)} = - \left[(i-A) - hB \frac{\partial F}{\partial y} y_{n+1,n+2}^{(i)} \right]^{-1} \left[(I-A) y_{n+1,n+2}^i - hBF \left(y_{n+1,n+2}^{(i)} \right) - \xi \right] \quad (10)$$

The general form of EBBDF method is

$$\begin{cases} y_{n+1} = \alpha_1 h f_{n+1} + \theta_1 y_{n+2} + \psi_1 \\ y_{n+2} = \alpha_1 h f_{n+2} + \theta_1 y_{n+1} + \psi_2 \end{cases} \quad (11)$$

With ψ_1 and ψ_2 are the back values. By setting,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, y_{n+1,n+2} = \begin{bmatrix} y_{n+1} \\ y_{n+2} \end{bmatrix}, B = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}, F_{n+1,n+2} = \begin{bmatrix} f_{n+1} \\ f_{n+2} \end{bmatrix}, \text{ and}$$

$$\xi_{n+1,n+2} = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix}$$

Equation (11) in matrix-vector form is equivalent to

$$(I-A)y_{n+1,n+2} = hBF_{n+1,n+2} + \xi_{n+1,n+2} \quad (12)$$

Equation (12) is simplified as

$$\hat{f}_{n+1,n+2} = (I-A)y_{n+1,n+2} - hBF_{n+1,n+2} - \xi_{n+1,n+2} = 0 \quad (13)$$

Newton iteration is performed to the system $\hat{f}_{n+1,n+2} = 0$, by taking the analogous form of (10) where $J_{n+1,n+2} = \left(\frac{\partial F}{\partial Y} \right) \left(Y_{n+1,n+2}^{(i)} \right)$, is the Jacobian matrix of F with respect to Y . Equation (10) is separated to three different matrices denoted as

$$E_{1,2}^{(i+1)} = y_{n+1,n+2}^{(i+1)} - y_{n+1,n+2}^{(i)} \quad (14)$$

$$\hat{A} = (I-A) - hB \frac{\partial F}{\partial Y} \left(y_{n+1,n+2}^{(i)} \right) \quad (15)$$

$$\hat{B} = (I - A)y_{n+1,n+2}^{(i)} - hBF \left(y_{n+1,n+2}^{(i)} \right) - \xi_{n+1,n+2} \quad (16)$$

Two-stage Newton iteration works to find the approximating solution to (1) with two simplified strategies based on evaluating the Jacobian ($J_{n+1,n+2}$) and LU factorization of \hat{A} [13].

2.2 Consistency and Convergence of EBBDF Methods

We first consider the general form of a block linear multistep method (LMM):

$$\sum_{j=0}^k A_j y_{n+j} = h \sum_{j=0}^k B_j F_{n+j} \quad (17)$$

where A_j and B_j are r by r matrices with elements α_{im}, β_{im} for $i.m = 0, 1, \dots, r$.

Equation (17) is applied for EBBDF method

$$\sum_{j=0}^k \alpha_{ij} y_{n+2-j} = h \sum_{j=0}^k \beta_j f_{n+2-j} \quad (18)$$

The expression (18) is expanded to give the following system of equation.

$$\begin{aligned} & \begin{bmatrix} \alpha_{1,0} & \alpha_{1,1} \\ \alpha_{2,0} & \alpha_{2,1} \end{bmatrix} \begin{bmatrix} y_{n+2} \\ y_{n+1} \end{bmatrix} + \dots + \begin{bmatrix} \alpha_{1,k-1} & \alpha_{1,k} \\ \alpha_{2,k-1} & \alpha_{2,k} \end{bmatrix} \begin{bmatrix} y_{n+3-k} \\ y_{n+2-k} \end{bmatrix} \\ & = h \left[\begin{bmatrix} \beta_{1,4} & \beta_{1,5} \\ \beta_{2,4} & \beta_{2,5} \end{bmatrix} \begin{bmatrix} f_{n+2} \\ f_{n+1} \end{bmatrix} + \dots + \begin{bmatrix} \beta_{1,k-1} & \beta_{1,k} \\ \beta_{2,k-1} & \beta_{2,k} \end{bmatrix} \begin{bmatrix} f_{n+3-k} \\ f_{n+2-k} \end{bmatrix} \right] \end{aligned} \quad (19)$$

With

$$A_0 = \begin{bmatrix} \alpha_{1,0} \\ \alpha_{2,0} \end{bmatrix}, \dots, A_k = \begin{bmatrix} \alpha_{1,k} \\ \alpha_{2,k} \end{bmatrix}, \text{ and } B_1 = \begin{bmatrix} \beta_{1,1} \\ \beta_{2,1} \end{bmatrix}, \dots, B_k = \begin{bmatrix} \beta_{1,k} \\ \beta_{2,k} \end{bmatrix}$$

Adopting the order procedure used in the single case for the block method, we associate with a linear difference operator $L[y(x_n), h]$, given as

$$L[y(x_n), h] = \sum_{j=0}^k [\alpha_j y(x_n + jh) - h\beta_j f(x_n + jh, y(x_n + jh))] \quad (20)$$

The use of appropriate Taylor expansions about a suitable x , reduces (20) to the form

$$L[y(x_n), h] = c_0 y(x_n) + c_1 h y^{(x_n)} + \dots + c_r h^r y^r(x_n) \quad (21)$$

where

$$c_0 = \sum_{j=0}^k A_j, c_1 = \sum_{j=0}^k j A_j - \sum_{j=0}^k B_j \text{ and,}$$

$$c_r = \frac{1}{r!} \sum_{j=0}^k j^r A_j - \frac{1}{(r-1)!} \sum_{j=0}^k j^{r-1} B_j$$

The EBBDF method is said to be of order p if $c_0 = c_1 = \dots = c_p = 0, c_{p+1} \neq 0$ and the local truncation error is

$$t_{n+k} = c_{p+1} h^{p+1} y^{p+1}(x_n) + O(h^{p+2}) \quad (22)$$

2.3 Stability Regions of the EBBDF Method

Definition 2 A method is said to be absolute stable in a region R for a given $h\lambda$ if for that $h\lambda$, all the roots r_s of the stability polynomial $\pi(r, h\lambda) = \rho(r) - h\lambda\sigma(r) = 0$, satisfy $|r_s| < 1, s = 1, 2, \dots, k$.

Definition 3 The LMM is said to be zero-stable if no root of the first characteristic polynomial $\rho(r)$ has modulus greater than one, and if every root with unit modulus is simple.

By applying test equation $y' = \lambda y$ to (1.1) we obtain,

$$\sum_{j=0}^k \alpha_{ij} y_{n+2-j} = h \sum_{j=0}^k \beta_j \lambda_{n+2-j} \quad (23)$$

The Eq. (23) is equivalent to

$$\sum_{j=0}^r A_j Y_j = 0 \quad (24)$$

where $A_j = [A_0, A_1, \dots, A_r], Y_j = [Y_0, Y_1, \dots, Y_r]$ and,

$$A_j = \begin{bmatrix} \alpha_{1,2j} & \alpha_{1,(2j+1)} - \beta_{1,(2j+1)} h\lambda \\ \alpha_{2,(2j)} - \beta_{2,(2j)} h\lambda & \alpha_{2,(2j+1)} \end{bmatrix}, Y_j = \begin{bmatrix} y_{n+2-2j} \\ y_{n+1-2j} \end{bmatrix}$$

The stability polynomial, $R(t, \hat{h})$ associated with the method of (18) is given by

$$\det \left(\sum_{j=0}^r A_j t^j \right) \quad (25)$$

while the absolute stability region of this method in the $h\lambda$ plane is determined by solving $\det \left(\sum_{j=0}^r A_j t^j \right) = 0$. The stability region was given by the set of points

determined by the boundary $t = e^{i\theta}$, $0 \leq \theta \leq 2\pi$. The stability region is obtained by finding the region for which $|t| < 1$.

2.4 Order and Step size Selection

The importance of choosing the step size is to achieve reduction in computation time and number of iterations. Meanwhile changing the order of the method is designed for finding the best approximation. Strategies proposed in [14] are applied in this study for choosing the step size and order. The strategy is to estimate the maximum step size for the following step. Methods of order $P - 1$, P , $P + 1$ are selected depending on the occurrence of every successful step. Consequently, the new step size h_{new} is obtained from which order produces the maximum step size.

The user initially will have to provide an error tolerance limit, TOL on any given step and obtain the local truncation error (LTE) for each iteration. The LTE is obtained from

$$LTE_k = y_{n+2}^{(P+1)} - y_{n+2}^{(P)}, P = 3, 4, 5$$

where $y_{n+2}^{(P+1)}$ is the $(P + 1)$ th order method and $y_{n+2}^{(P)}$ is the k th order method. By finding the LTEs, the maximum step size is defined as

$$h_{P-1} = h_{old} \times \left(\frac{TOL}{LTE_{P-1}} \right)^{\frac{1}{P}}, h_P = h_{old} \times \left(\frac{TOL}{LTE_P} \right)^{\frac{1}{P+1}}, h_{P+1} = h_{old} \times \left(\frac{TOL}{LTE_{P+1}} \right)^{\frac{1}{P+2}}$$

where h_{old} is the stepsize from previous block and h_{max} is obtained from the maximum stepsize given in above equations.

The successful step is dependent on the condition $LTE < TOL$. If this condition fails, the values of y_{n+1}, y_{n+2} are rejected, and the current step is reiterated with step size selection ($q = 2$). On the contrary, the step size increment for each successful step is defined as $h_{new} = c \times h_{max}$ and if $h_{new} > 1.9 \times h_{old}$ then $h_{new} = 1.9 \times h_{old}$. Where c is the safety factor, p is the order of the method while and is the step size from previous and current block respectively. In this chapter, c is set to be 0.8 so as to make sure the rejected step is being reduced.

3 Numerical Results

We carry out numerical experiments to compare the performance of EBBDF method with stiff ODE solvers in **MATLAB** mentioned earlier. This chapter considers the comparison of four different factors namely number of steps taken, average error, maximum error and computation time. These test problems are performed under different conditions of error tolerances—(a) 10^{-2} , (b) 10^{-4} and, (c) 10^{-6}

Table 1 Numerical results for problem 1

	VSVO	ODE15s	ODE23s
TOL	1.000e-002		
Total steps	29	41	37
Maximum error	6.4925e-003	0.0084	0.0045
Average error	8.8396e-004	9.1188e-004	0.0011
Time	0.0104	0.1094	0.0781
TOL	1.000e-004		
Total steps	74	94	182
Maximum error	1.9531e-006	1.6634e-004	2.5500e-004
Average error	4.6275e-007	1.9159e-005	4.7081e-005
Time	0.0107	0.1250	0.1563
TOL	1.000e-006		
Total steps	279	163	1194
Maximum error	9.8112e-007	3.0953e-006	1.0911e-005
Average error	4.8070e-007	1.3149e-006	1.1596e-006
Time	0.0125	0.2813	0.6250

The test problems and solution are listed below

Problem 1

$$y' = -1000y + 3000 - 2000e^{(-x)}y(0) = 0 \quad 0 \leq x \leq 20$$

With solution: $y(x) = 3 - 0.998e^{(-1000x)} - 2.002e^{(-x)}$

Problem 2

Table 2 Numerical results for problem 2

	VSVO	ODE15s	ODE23s
TOL	1.000e-002		
Total steps	35	45	137
Maximum error	3.0045e-004	0.0146	0.0030
Average error	4.6584e-005	0.0069	0.0014
Time	0.0111	0.0156	0.2031
TOL	1.000e-004		
Total steps	84	99	1211
Maximum error	1.1002e-005	2.7591e-004	6.3075e-005
Average error	2.5775e-006	7.6548e-005	2.4915e-005
Time	0.0123	0.0469	0.3281
TOL	1.000e-006		
Total steps	380	186	8289
Maximum error	8.9627e-008	6.1936e-006	1.5667e-006
Average error	2.4244e-008	2.1681e-006	5.3155e-007
Time	0.0199	0.0781	2.3750

$$y_1' = -2y_1 + y_2 + 2 \sin(x) \quad y_1(0) = 2 \quad 0 \leq x \leq 10$$

$$y_2' = 998y_1 - 999y_2 + 999(\cos(x) - \sin(x)) \quad y_2(0) = 3$$

With solution:

$$y_1(x) = 2e^{(-x)} + \sin(x)$$

$$y_2(x) = 2e^{(-x)} + \cos(x)$$

From Tables 1 and 2, among the three methods tested, our method, EBBDF method requires the shortest execution time, smallest maximum error and average error for each given tolerance level. Moreover, EBBDF outperformed ode15s and ode23s in term of total steps when the tolerance level is less than 10^{-4} .

4 Conclusion and Future Work

Extended Block Backward Differentiation Formulae (EBBDF) was derived in this chapter. From the numerical experiments, the comparisons between EBBDF and solvers in **MATLAB**, show improvements in term of computation time, average error as well as maximum error. Therefore, we conclude that EBBDF can serve as an alternative solver for solving stiff ODEs. Subsequently, the method can be extended to solve higher order ODEs in future. One can also apply parallel computing to improve the computation time.

References

1. Sack-Davis R (1980) Fixed leading coefficient implementation of SD-formulas for stiff ODEs. *ACM Trans Math Softw* 6(4):540–562
2. Yatim SAM, Ibrahim ZB, Othman KI, Suleiman MB (2011) Quantitative comparison of numerical method for solving stiff ordinary differential equations. *Math Prob Eng* 2011, ID 193961
3. Mahmood AS, Casasus L, Al-Hayani W (2005) The decomposition method for stiff systems of ordinary differential equations. *Appl Math Comput* 167(2):964–975
4. Ibanez J, Hernandez V, Arias E, Ruiz PA (2009) Solving initial value problems for ordinary differential equations by two approaches: BDF and piecewise-linearized methods. *Comput Phys Commun* 180(5):712–723
5. Enright WH, Hull TE, Lindberg B (1975) Comparing numerical methods for stiff systems of O.D.Es. *BIT* 15(1):10–48
6. Byrne GD, Hindmarsh AC, Jackson KR, Brown HG (1977) A comparison of two ode codes: gear and episode. *Comput Chem Eng* 1(2):133–147
7. Cash JR, Considine S (1992) An MEBDF code for stiff initial value problems. *ACM Trans Math Softw* 18(2):142–155
8. Abelman S, Patidar KC (2008) Comparison of some recent numerical methods for initial-value problems for stiff ordinary differential equations. *Comput Math Appl* 55(4):733–744

9. Ibrahim ZB, Suleiman MB, Othman KI (2008) Fixed coefficients block backward differentiation formulas for the numerical solution of stiff ordinary differential equations. *Eur J Sci Res* 21(3):508–520
10. Ibrahim ZB, Othman KI, Suleiman MB (2007) Variable step size block backward differentiation formula for solving stiff odes. In: *Proceedings of the world congress on engineering 2007, WCE 2007*, London, UK, pp 785–789, 2–4 July 2007
11. Yatim SAM, Ibrahim ZB, Othman KI, Suleiman MB (2010) Fifth order variable step block backward differentiation formulae for solving stiff ODEs. *Proc World Acad Sci Eng Tech* 62:998–1000
12. Yatim SAM, Ibrahim ZB, Othman KI, Suleiman MB (2012) Numerical solution of extended block backward differentiation formulae for solving stiff ODEs, *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012*, London, UK, pp 109–113, 4–6 July 2012
13. Lambert JD (1973) *Computational methods ordinary differential equation*. Wiley, New York
14. Hall G, Watt JM (1976) *Modern numerical methods for ordinary differential equations*. Clarendon Press, Oxford

On Fast Algorithms for Triangular and Dense Matrix Inversion

Ryma Mahfoudhi and Zaher Mahjoub

Abstract We first propose in this paper a recursive algorithm for triangular matrix inversion (TMI) based on the ‘Divide and Conquer’ (D&C) paradigm. Different versions of an original sequential algorithm are presented. A theoretical performance study permits to establish an accurate comparison between the designed algorithms. Our implementation is designed to be used in place of `dtrtri`, the level 3 BLAS TMI. Afterwards, we generalize our approach for dense matrix inversion (DMI) based on LU factorization (LUF). This latter is used in Mathematical software libraries such as LAPACK `xGETRI` and MATLAB `inv`. $A = LU$ being the input dense matrix, `xGETRI` consists, once the factors L and U are known, in inverting U then solving the triangular matrix system $XL = U^{-1}$ (i.e. $L^T X^T = (U^{-1})^T$, thus $X = A^{-1}$). Two other alternatives may be derived here (L and U being known) : (i) first invert L , then solve the matrix system $UX = L^{-1}$ for X ; (ii) invert both L and U , then compute the product $X = U^{-1}L^{-1}$. Each of these three procedures involves at least one triangular matrix inversion (TMI). Our DMI implementation aims to be used in place of the level 3 BLAS TMI-DMI. Efficient results could be obtained through an experimental study achieved on a set of large sized randomly generated matrices.

Keywords Dense matrix inversion · Divide and conquer · Level 3 BLAS · LU factorization · Recursive algorithm · Triangular matrix inversion

R. Mahfoudhi (✉) · Z. Mahjoub
Faculty of Sciences of Tunis, University of Tunis El Manar,
University Campus - 2092 Manar II, Tunis, Tunisia
e-mail: rimahayet@yahoo.fr

Z. Mahjoub
e-mail: zaher.mahjoub@fst.rnu.tn

1 Introduction

Triangular matrix inversion (TMI) is a basic kernel used in many scientific applications. Given its cubic complexity in terms of the matrix size, say n , several works addressed the design of practical efficient algorithms for solving this problem. Apart the standard TMI algorithm consisting in solving n linear triangular systems of size $n, n - 1, \dots, 1$ [1], a recursive algorithm, of same complexity, has been proposed by Heller in 1973 [2–4]. It uses the ‘Divide and Conquer’ (D&C) paradigm and consists in successive decompositions of the original matrix. Our objective here is two-fold i.e. (i) design an efficient algorithm for TMI that outperforms the BLAS routines and (ii) use our TMI kernel for dense matrix inversion (DMI) through LU factorization, thus deriving an efficient DMI kernel.

The remainder of the paper is organized as follows. In Sect. 2, we present the D&C paradigm. We then detail in Sect. 3 a theoretical study on diverse versions of Heller’s TMI algorithm. Section 4 is devoted to the generalization of the former designed algorithm for DMI. An experimental study validating our theoretical contribution is presented in Sect. 5.

2 Divide and Conquer Paradigm

There are many paradigms in algorithm design. Backtracking, Dynamic programming, and the Greedy method to name a few. One compelling type of algorithms is called Divide and Conquer (D&C). Algorithms of this type split the original problem to be solved into (equal sized) sub-problems. Once the sub-solutions are determined, they are combined to form the solution of the original problem. When the sub-problems are of the same type as the original problem, the same recursive process can be carried out until the sub-problem size is sufficiently small. This special type of D&C is referred to as D&C recursion. The recursive nature of many D&C algorithms makes it easy to express their time complexity as recurrences. Consider a D&C algorithm working on an input size n . It divides its input into a (called arity) sub-problems of size n/b . Combining and conquering are assumed to take $f(n)$ time. The base-case corresponds to $n = 1$ and is solved in constant time. The time complexity of this class of algorithms can be expressed as follows:

$$\begin{aligned} T(n) &= O(1) && \text{if } n = 1 \\ &= aT(n/b) + f(n) && \text{otherwise.} \end{aligned}$$

Let $f(n) = O(n^\delta)$ ($\delta \geq 0$), the master theorem for recurrences can in some instances be used to give a tight asymptotic bound for the complexity [1]:

- $a < b^\delta \Rightarrow T(n) = O(n^\delta)$
- $a = b^\delta \Rightarrow T(n) = O(n^\delta \log_b n)$

- $a > b^\delta \Rightarrow T(n) = O(n^{\log_b a})$

3 Recursive TMI Algorithms

We first recall that the well known standard algorithm (SA) for inverting a triangular matrix (either upper or lower), say A of size n , consists in solving n triangular systems. The complexity of (SA) is as follows [1]:

$$SA(n) = n^3/3 + n^2/2 + n/6 \tag{1}$$

3.1 Heller's Recursive Algorithm (HRA)

Using the D&C paradigm, Heller proposed in 1973 a recursive algorithm [2, 3] for TMI. The main idea he used consists in decomposing matrix A as well as its inverse B (both of size n) into 3 submatrices of size $n/2$ (see Fig. 1, A being assumed lower triangular). The procedure is recursively repeated until reaching submatrices of size 1. We hence deduce:

$$B_1 = A_1^{-1}, B_3 = A_3^{-1}, B_2 = -B_3 A_2 B_1 \tag{2}$$

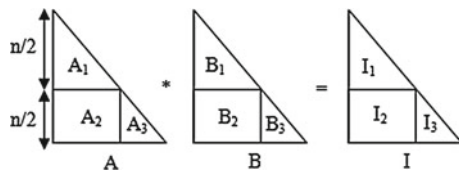
Therefore, inverting matrix A of size n consists in inverting 2 submatrices of size $n/2$ followed by two matrix products (triangular by dense) of size $n/2$. In [3] Nasri proposed a slightly modified version of the above algorithm. Indeed, since $B_2 = -B_3 A_2$ and $B_1 = -A_3^{-1} A_2 A_1^{-1}$, let $Q = A_3^{-1} A_2$. From (2), we deduce:

$$A_3 Q = A_2, B_2 A_1 = -Q \tag{3}$$

Hence, instead of two matrix products needed to compute matrix B_2 , we have to solve 2 matrix systems of size $n/2$ i.e. $A_3 Q = A_2$ and $(A_1)^T (B_2)^T = -Q^T$. We precise that both versions are of $n^3/3 + O(n^2)$ complexity [3].

Now, for sake of simplicity, we assume that $n = 2^q (q \geq 1)$. Let RA- k be the Recursive Algorithm designed by recursively applying the decomposition k times i.e. until reaching a threshold size $n/2^k (1 \leq k \leq q)$. The complexity of RA- k is as

Fig. 1 Matrix decomposition in Heller's algorithm



follows [3]:

$$RA - k(n) = n^3/3 + n^2/2^{k+1} + n/6 \tag{4}$$

3.2 Recursive Algorithm Using Matrix Multiplication (RAMM)

As previously seen, to invert a triangular matrix via block decomposition, one requires two recursive calls and two triangular matrix multiplications (TRMM) [5]. Thus, the complexity recurrence formula is:

$$RAMM(n) = 2RAMM(n/2) + 2TRMM(n/2) + O(n^2)$$

The idea consists in using the fast algorithm for TRMM presented below.

ALGORITHM 1	
RAMM	
Begin	
<i>If</i> ($n = 1$) <i>then</i>	
$B_1 = I/A_1$	
$B_3 = I/A_3$	
$B_2 = -B_3 * A_2 * B_1$	
<i>Else</i> /* splitting matrices into three blocks of sizes $n/2$	
$B_1 = RAMM(A_1)$	
$B_3 = RAMM(A_3)$	
$C = TRMM(-B_3, A_2)$	
$B_2 = TRMM(C, B_1)$	
<i>Endif</i>	
End	

- **TRMM algorithm**

To perform the multiplication of a triangular (resp. dense) by a dense (resp. triangular) via block decomposition in halves, we require four recursive calls and two dense matrix-matrix multiplications (MM) Fig. 2.

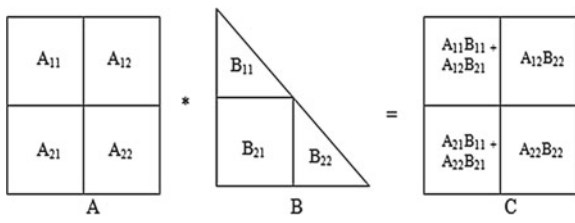


Fig. 2 Matrix decomposition in TRMM algorithm

ALGORITHM 2**TRMM**

Begin
If ($n=1$) **then**
 $A_{11} * B_{11} = C_{11}$
 $A_{11} * B_{12} = C_{12}$
 $A_{21} * B_{11} + A_{22} * B_{21} = C_{21}$
 $A_{21} * B_{12} + A_{22} * B_{22} = C_{22}$
Else /* *splitting matrices into four blocks of sizes $n/2$*
 $C_{11} = TRMM(A_{11}, B_{11})$
 $C_{12} = TRMM(A_{11}, B_{12})$
 $C_{21} = MM(A_{21}, B_{11}) + TRMM(A_{22}, B_{21})$
 $C_{22} = MM(A_{21}, B_{12}) + TRMM(A_{22}, B_{22})$
Endif
End

The complexity recurrence formula is thus :

$$TRMM(n) = 4TRMM(n/2) + 2MM(n/2) + O(n^2).$$

To optimize this algorithm, we will use a fast algorithm for dense MM i.e. Strassen algorithm.

- **MM algorithm**

In [6, 7], the author reported on the development of an efficient and portable implementation of Strassen MM algorithm. Notice that the optimal number of recursive levels depends on both the matrix size and the target architecture and must be determined experimentally.

3.3 Recursive Algorithm Using Triangular Systems Solving (RATSS)

In this version, we replace the two matrix products by two triangular systems solving of size $n/2$ (see Sect. 3.1). The algorithm is as follows:

ALGORITHM 3

RATSS

```

Begin
  If ( $n=1$ ) then
     $B_1 = 1/A_1$ ,
     $B_3 = 1/A_3$ 
     $Q = A_2/A_3$ 
     $B_2 = -Q/A_1$ 
  Else /* splitting matrices into four blocks of sizes  $n/2$ 
     $B_1 = RAMM(A_1)$ 
     $B_3 = RAMM(A_3)$ 
     $Q = TSS(A_3, A_2)$ 
     $B_2 = TSS(A_1^T, -Q^T)$ 
  Endif
End
  
```

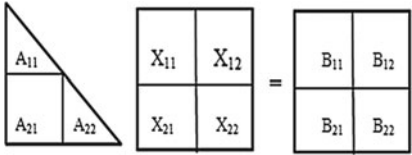
• **TSS algorithm**

We now discuss the implementation of solvers for triangular systems with matrix right hand side (or equivalently left hand side). This kernel is commonly named **trsm** in the BLAS convention. In the following, we will consider, without loss of generality, the resolution of a lower triangular system with matrix right hand side ($AX = B$). Our implementation is based on a block recursive algorithm in order to reduce the computations to matrix multiplications [8, 9].

ALGORITHM 4

TSS

```

Begin
  If ( $n=1$ ) then
     $X = B/A$ 
  Else /* splitting matrices into four blocks of sizes  $n/2$ 
    

$$\begin{bmatrix} A_{11} & & \\ A_{21} & A_{22} & \end{bmatrix}
 \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}
 =
 \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

     $X_{11} = TSS(A_{11}, B_{11})$ 
     $X_{12} = TSS(A_{11}, B_{12})$ 
     $X_{21} = TSS(A_{22}, B_{21} - MM(A_{21}, X_{11}))$ 
     $X_{22} = TSS(A_{22}, B_{22} - MM(A_{21}, X_{12}))$ 
  Endif
End
  
```

3.4 Algorithms Complexity

As well known, the complexity of the Strassen's Algorithm is $MM(n) = O(n^{\log_2 7})$

Besides, the cost RAMM(n) satisfies the following recurrence formula:

$$RAMM(n) = 2RAMM(n/2) + 2TRMM(n/2) + O(n^2).$$

Since

$$\begin{aligned} TRMM(n) &= 4TRMM(n/2) + 2MM(n/2) + O(n^2) \\ &= 4TRMM(n/2) + O(n^{\log_2 7}) + O(n^2) \\ &= n^2 + O(n^{\log_2 7}) + O(n^2) = O(n^{\log_2 7}) \end{aligned}$$

We therefore get :

$$\begin{aligned} RAMM(n) &= 2RAMM(n/2) + 2TRMM(n/2) + O(n^2) \\ &= n \log(n) + O(n^{\log_2 7}) + O(n^2) = O(n^{\log_2 7}) \end{aligned}$$

Following a similar way, we prove that $TRMM(n) = O(n^{\log_2 7})$

4 Dense Matrix Inversion

4.1 LU Factorization

As previously mentioned, three alternative methods may be used to perform a DMI through LU factorization (LUF). The first one requires two triangular matrix inversions (TMI) and one triangular matrix multiplication (TMM) i.e. an upper one by a lower one. The two others both require one triangular matrix inversion (TMI) and a triangular matrix system solving (TSS) with matrix right hand side or equivalently left hand side (Algorithm 4). Our aim is to optimize both LUF, TMI as well as TMM kernels [10].

4.2 Recursive LU Factorisation

To reduce the complexity of LU factorization, blocked algorithms were proposed in 1974 [11]. For a given matrix A of size n, the L and U factors verifying $A=LU$ may be computed as follows:

ALGORITHM 5

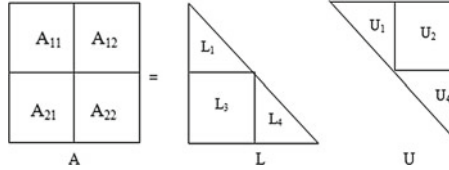
LUF

Begin

If ($n=1$) **Then**

$L=A$; $U=A$

Else /* split matrices into four blocks of sizes $n/2$



$$(L_1, [U_1, U_2]) = LUF([A_{11} \ A_{12}])$$

$$L_3 = A_{21} U_1^{-1}$$

$$H = A_{22} - L_3 U_2$$

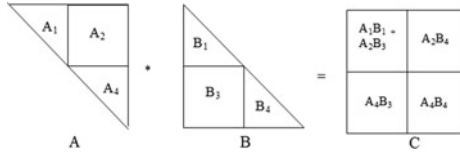
$$(L_4, U_4) = LUF(H)$$

Endif

End

4.3 Triangular Matrix Multiplication (TMM)

Block wise multiplication of an upper triangular matrix by a lower one, can be depicted as follows:



Thus, to compute the dense matrix $C = AB$ of size n , we need:

- Two triangular matrix multiplication (an upper one by a lower one) of size $n/2$
- Two multiplications of a triangular matrix by a dense one (TRMM) of size $n/2$.
- Two dense matrix multiplication (MM) of size $n/2$.

ALGORITHM 6

TMM

```

Begin
  If (n=1) Then
    C = A*B
  Else /* split matrices into four blocks of sizes n/2
    C1 = TMM(A1,B1)+MM(A2,B3)
    C2 = TRMM(B4, A2)
    C3 = TRMM(A4,B3)
    C4 = TMM(A4,B4)
  Endif
End
    
```

Clearly, if any matrix-matrix multiplication algorithm with $O(n^{\log_2 7})$ complexity is used, then the algorithms previously presented both have the same $O(n^{\log_2 7})$ complexity instead of $O(n^3)$ for the standard algorithms.

5 Experimental Study

5.1 TMI Algorithm

This section presents experiments of our implementation of the different versions of triangular matrix inversion described above. We determinate the optimal number of recursive levels for each one (as already precised, the optimal number of recursive levels depends on the matrix size and the target architecture and must be determined experimentally). The experiments (as well as the following on DMI) use BLAS library in the last level and were achieved on a 3 GHz, 4Go RAM PC. We used the g++ compiler under Ubuntu 11.01.

We recall that **dtrtri** refers to the BLAS triangular matrix inversion routine with double precision floating points. We named our routines RAMM, RATSS, see fig. 3.

Fig. 3 Time ratio
dtrtri/RATSS

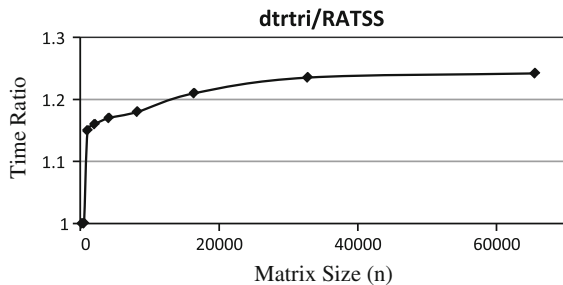


Table 1 Timing of triangular matrix inversion (seconds)

Matrix size	dtrtri	RAMM	RATSS	Time ratio dtrtri/RATSS
256	0.01	0.02	0.01	1
512	0.02	0.03	0.02	1
1024	0.23	0.25	0.2	1.15
2048	2.03	2.08	1.71	1.16
4096	15.54	15.58	13.27	1.17
8192	121.64	127.77	102.9	1.18
16384	978.17	981.35	810.68	1.21
32768	7902.14	7927.85	6396.87	1.23
65536	64026.02	64296.97	51548.52	1.24

We notice that for increasing matrix sizes, RATSS becomes even more efficient than **dtrtri** (improvement factor between 15 and 24%). On the other hand, **dtrtri** is better than RAMM, see table 1.

5.2 DMI Algorithm

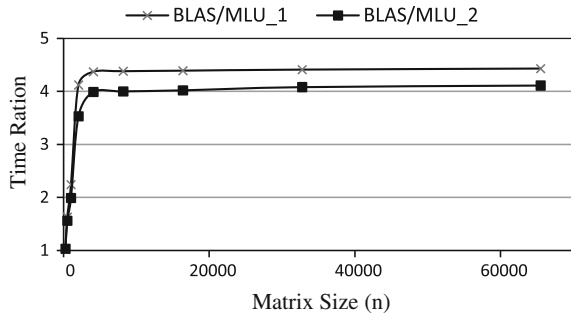
Table 2 provides a comparison between LU factorization-based algorithms i.e. MILU_1 (one TMI and one triangular matrix system solving), MILU_2 (two TMIs and one TMM), and the BLAS routine where the routine **dgetri** was used in combination with the factorization routine **dgetrf** to obtain the matrix inverse (see Fig. 4).

We remark that the time ratio increases with the matrix size i.e. MILU_1 and MILU_2 become more and more efficient than BLAS (the speed-up i.e. time ratio reaches 4.4 and more).

Table 2 Timing of dense matrix inversion (seconds)

Matrix size	BLAS	MILU_1	MILU_2	Time ratio $\frac{\text{BLAS}}{\text{MILU}_1}$	Time ratio $\frac{\text{BLAS}}{\text{MILU}_2}$
256	0.06	0.06	0.06	1.02	1.03
512	0.12	0.07	0.08	1.63	1.56
1024	1.46	0.65	0.73	2.24	1.99
2048	12.00	2.91	3.40	4.12	3.53
4096	96.01	21.97	24.06	4.37	3.99
8192	764.35	174.51	191.09	4.38	4.00
16384	5922.38	1349.06	1473.23	4.39	4.02
32768	50276.71	11400.61	12322.72	4.41	4.08
65536	401295.45	90585.88	97638.80	4.43	4.11

Fig. 4 Time ratio: BLAS/MLU_1 and BLAS/MLU_2



6 Conclusion and Future Work

In this paper we targeted and reached the goal of outperforming the efficiency of the well-known BLAS library for triangular and dense matrix inversion. It has to be noticed that our (recursive) algorithms essentially benefit from both (recursive) Strassen matrix multiplication algorithm, recursive solvers for triangular systems and the use of BLAS routines in the last recursion level. This performance was achieved thanks to (i) efficient reduction to matrix multiplication where we optimized the number of recursive decomposition levels and (ii) reusing numerical computing libraries as much as possible.

These results we obtained lead us to precise some attracting perspectives we intend to study in the future. We may particularly cite the following points.

- Achieve an experimental study on matrix of larger sizes.
- Study the numerical stability of these algorithms
- Generalize our approach to other linear algebra kernels

References

1. Quarteroni A, Sacco R, Saleri F (2007) *Méthodes numériques. Algorithmes, analyse et applications*, Springer, Milan
2. Heller D (1978) A survey of parallel algorithms in numerical linear algebra. *SIAM Rev* 20:740–777
3. Nasri W, Mahjoub Z (2002) Design and implementation of a general parallel divide and conquer algorithm for triangular matrix inversion. *Int J Parallel Distrib Syst Netw* 5(1):35–42
4. Aho AV, Hopcroft JE, Ullman JD (1975) *The design and analysis of computer algorithms*. Addison-Wesley, Reading
5. Mahfoudhi R (2012) A fast triangular matrix inversion. *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012*, pp 100–102
6. Steven H, Elaine M, Jeremy R, Anna T, Thomas T (1996) Implementation of Strassen’s algorithm for matrix multiplication. In: *Supercomputing '96 proceedings ACM/IEEE conference on supercomputing (CDROM)*

7. Strassen V (1969) Gaussian elimination is not optimal. *Numer Math* 13:354–356
8. Andersen BS, Gustavson F, Karaivanov A, Wasniewski J, Yalamov PY (2000) LAWRA—Linear algebra with recursive algorithms. *Lecture notes in computer science*, vol 1823/2000, pp 629–632
9. Dumas JG, Pernet C, Roch JL (2006) Adaptive triangular system solving. In: *Proceedings of the challenges in symbolic computation software*
10. Mahfoudhi R, Mahjoub Z (2012) A fast recursive blocked algorithm for dense matrix inversion. In: *Proceedings of the 12th international conference on computational and mathematical methods in science and engineering, cmmse 2012, La Manga, Spain*
11. Aho AV, Hopcroft JE, Ullman JD (1974) *The design and analysis of computer algorithms*. Addison-Wesley, Reading

Adding Relation Between Two Levels of a Linking Pin Organization Structure Maximizing Communication Efficiency of Information

Kiyoshi Sawada

Abstract This paper proposes a model of adding relation to a linking pin organization structure where every pair of siblings in a complete binary tree of height H is adjacent such that the communication of information in the organization becomes the most efficient. For a model of adding an edge between a node with a depth M and its descendant with a depth N , we formulated the total shortening distance which is the sum of shortening lengths of shortest paths between every pair of all nodes and obtained an optimal depth N^* which maximizes the total shortening distance for each value of M .

Keywords Communication efficiency · Complete binary tree · Linking pin · Organization structure · Shortest path · Total distance

1 Introduction

A linking pin organization structure is a structure in which relations between members of the same section are added to a pyramid organization structure and is called System 4 in Likert's organization classification [1]. In the linking pin organization structure there exist relations between each superior and his direct subordinates and those between members which have the same immediate superior.

The linking pin organization structure can be expressed as a structure where every pair of siblings which are nodes which have the same parent in a rooted tree is adjacent, if we let nodes and edges in the structure correspond to members and relations between members in the organization respectively [2, 3]. Then the height of the linking pin organization structure expresses the number of levels in the organization,

K. Sawada (✉)
Department of Policy Studies, University of Marketing and Distribution Sciences,
Kobe 651-2188, Japan
e-mail: Kiyoshi_Sawada@red.umds.ac.jp

and the number of children of each node expresses the number of subordinates of each member. Moreover, the path between a pair of nodes in the structure is equivalent to the route of communication of information between a pair of members in the organization, and adding edges to the structure is equivalent to forming additional relations other than those between each superior and his direct subordinates and between members which have the same direct subordinate [4].

The purpose of our study is to obtain an optimal set of additional relations to the linking pin organization such that the communication of information between every member in the organization becomes the most efficient. This means that we obtain a set of additional edges to the structure minimizing the sum of lengths of shortest paths between every pair of all nodes.

We have obtained an optimal depth for each of the following two models of adding relations in the same level to a complete K -ary linking pin structure of height H where every pair of siblings in a complete K -ary tree of height H is adjacent: (i) a model of adding an edge between two nodes with the same depth and (ii) a model of adding edges between every pair of nodes with the same depth [5]. A complete K -ary tree is a rooted tree in which all leaves have the same depth and all internal nodes have K ($K = 2, 3, \dots$) children [6]. Furthermore, we have proposed a model of adding relation between the top and a member in a complete K -ary linking pin structure of height H [7]. When an edge between the root and a node with a depth N is added, an optimal depth N^* is obtained by minimizing the sum of lengths of shortest paths between every pair of all nodes.

This paper proposes a model of adding an edge between a node with a depth M ($M = 0, 1, \dots, H - 2$) and its descendant with a depth N ($N = M + 2, M + 3, \dots, H$) in a complete binary (that is $K = 2$) linking pin structure of height H ($H = 2, 3, \dots$) [8]. This model corresponds to the formation of an additional relation between a superior and his indirect subordinate. Figure 1 shows an example of a complete binary linking pin structure of $H = 5$.

If $l_{i,j}$ ($= l_{j,i}$) denotes the distance, which is the number of edges in the shortest path from a node v_i to a node v_j ($i, j = 1, 2, \dots, 2^{H+1} - 1$) in the complete binary linking pin structure of height H , then $\sum_{i < j} l_{i,j}$ is the total distance. Furthermore, if $l'_{i,j}$ denotes the distance from v_i to v_j after adding an edge in this model, $l_{i,j} - l'_{i,j}$ is called the shortening distance between v_i and v_j , and $\sum_{i < j} (l_{i,j} - l'_{i,j})$ is called the *total shortening distance*. Minimizing the total distance is equivalent to maximizing the total shortening distance. When an edge between a node with a depth M and its descendant with a depth N is added to the complete binary linking pin structure of height H , an optimal depth N^* is obtained by maximizing the total shortening distance for each value of M .

In Sect. 2 we formulate the total shortening distance of the above model. In Sect. 3 we show an optimal depth N^* which maximizes the total shortening distance for each value of M and in Sect. 4 we illustrate an optimal depth N^* with numerical examples.

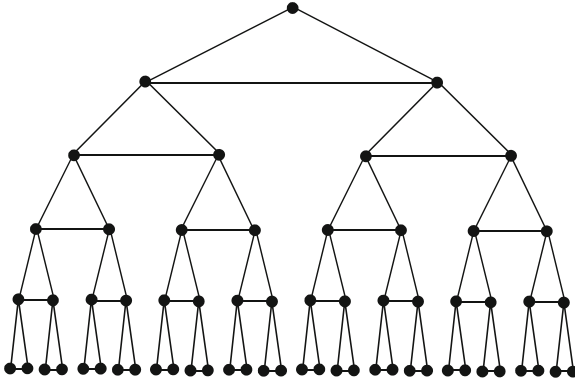


Fig. 1 An example of a complete binary linking pin structure of $H = 5$

2 Formulation of Total Shortening Distance

This section formulates the total shortening distance when an edge between a node with a depth M ($M = 0, 1, \dots, H - 2$) and its descendant with a depth N ($N = M + 2, M + 3, \dots, H$) is added to a complete binary linking pin structure of height H ($H = 2, 3, \dots$).

Let v_M denote the node with a depth M and let v_N denote the node with a depth N which gets adjacent to v_M . The set of descendants of v_N is denoted by V_1 . (Note that every node is a descendant of itself [6].) The set of descendants of v_M and ancestors of parent of v_N is denoted by V_2 . (Note that every node is an ancestor of itself [6].) Let V_3 denote the set obtained by removing V_1 and V_2 from the set of descendants of v_M . Let V_4 denote the set obtained by removing descendants of v_M from the set of all nodes of the complete binary linking pin structure.

The sum of shortening distances between every pair of nodes in V_1 and nodes in V_2 is given by

$$A_H(M, N) = W(H - N) \sum_{i=1}^{\lfloor \frac{N-M}{2} \rfloor} (N - M - 2i + 1), \tag{1}$$

where $W(h)$ denotes the number of nodes of a complete binary tree of height h ($h = 0, 1, 2, \dots$), and $\lfloor x \rfloor$ denotes the maximum integer which is equal to or less than x . The sum of shortening distances between every pair of nodes in V_2 is given by

$$B(M, N) = \sum_{i=1}^{\lfloor \frac{N-M}{2} \rfloor - 1} \sum_{j=1}^{\lfloor \frac{N-M}{2} \rfloor - i} (N - M - 2i - 2j + 1), \tag{2}$$

where we define $\sum_{i=1}^0 \cdot = 0$. The sum of shortening distances between every pair of nodes in V_1 and nodes in V_3 is given by

$$C_H(M, N) = W(H - N) \sum_{i=1}^{\lfloor \frac{N-M-1}{2} \rfloor} W(H - M - i)(N - M - 2i). \quad (3)$$

The sum of shortening distances between every pair of nodes in V_2 and nodes in V_3 is given by

$$\begin{aligned} D_H(M, N) = & \sum_{i=1}^{\lfloor \frac{N-M-1}{2} \rfloor - 1} W(H - M - i) \sum_{j=1}^{\lfloor \frac{N-M-1}{2} \rfloor - i} (N - M - 2i - 2j) \\ & + \sum_{i=1}^{\lfloor \frac{N-M-1}{2} \rfloor} W(H - N + i - 1) \sum_{j=1}^{\lfloor \frac{N-M-1}{2} \rfloor - i + 1} (N - M - 2i - 2j + 2), \end{aligned} \quad (4)$$

where we define $\sum_{i=1}^{-1} \cdot = 0$. The sum of shortening distances between every pair of nodes in V_3 is given by

$$\begin{aligned} E_H(M, N) = & \sum_{i=1}^{\lfloor \frac{N-M}{2} \rfloor - 1} W(H - N + i - 1) \\ & \times \sum_{j=1}^{\lfloor \frac{N-M}{2} \rfloor - i} W(H - M - j)(N - M - 2i - 2j + 1). \end{aligned} \quad (5)$$

The sum of shortening distances between every pair of nodes in V_1 and nodes in V_4 is given by

$$F_H(M, N) = (W(H) - W(H - M)) W(H - N)(N - M - 1). \quad (6)$$

The sum of shortening distances between every pair of nodes in V_2 and nodes in V_4 is given by

$$G_H(M, N) = (W(H) - W(H - M)) \sum_{i=1}^{\lfloor \frac{N-M}{2} \rfloor - 1} (N - M - 2i - 1). \quad (7)$$

The sum of shortening distances between every pair of nodes in V_3 and nodes in V_4 is given by

$$J_H(M, N) = (W(H) - W(H - M)) \sum_{i=1}^{\lfloor \frac{N-M-1}{2} \rfloor} W(H - N + i - 1)(N - M - 2i). \quad (8)$$

From the above equations, the total shortening distance $S_H(M, N)$ is given by

$$S_H(M, N) = A_H(M, N) + B(M, N) + C_H(M, N) + D_H(M, N) + E_H(M, N) + F_H(M, N) + G_H(M, N) + J_H(M, N). \quad (9)$$

3 An Optimal Depth N^* for Each Value of M

This section obtains an optimal depth N^* which maximizes the total shortening distance $S_H(M, N)$ for each value of M .

Let us classify $S_H(M, N)$ into two cases of $N = M + 2L$ where $L = 1, 2, \dots, \lfloor (H - M)/2 \rfloor$ and $N = M + 2L + 1$ where $L = 1, 2, \dots, \lfloor (H - M - 1)/2 \rfloor$. Since the number of nodes of a complete binary tree of height h is

$$W(h) = 2^{h+1} - 1, \quad (10)$$

$S_H(M, M + 2L)$ and $S_H(M, M + 2L + 1)$ become

$$\begin{aligned} & S_H(M, M + 2L) \\ &= \left(2^{H-M-2L+1} - 1\right) \sum_{i=1}^L (2L - 2i + 1) + \sum_{i=1}^{L-1} \sum_{j=1}^{L-i} (2L - 2i - 2j + 1) \\ &+ \left(2^{H-M-2L+1} - 1\right) \sum_{i=1}^{L-1} \left(2^{H-M-i+1} - 1\right) (2L - 2i) \\ &+ \sum_{i=1}^{L-2} \left(2^{H-M-i+1} - 1\right) \sum_{j=1}^{L-i-1} (2L - 2i - 2j) \\ &+ \sum_{i=1}^{L-1} \left(2^{H-M-2L+i} - 1\right) \sum_{j=1}^{L-i} (2L - 2i - 2j + 2) \\ &+ \sum_{i=1}^{L-1} \left(2^{H-M-2L+i} - 1\right) \sum_{j=1}^{L-i} \left(2^{H-M-j+1} - 1\right) (2L - 2i - 2j + 1) \\ &+ \left(2^{H+1} - 2^{H-M+1}\right) \left(2^{H-M-2L+1} - 1\right) (2L - 1) \end{aligned}$$

$$\begin{aligned}
& + \left(2^{H+1} - 2^{H-M+1}\right) \sum_{i=1}^{L-1} (2L - 2i - 1) \\
& + \left(2^{H+1} - 2^{H-M+1}\right) \sum_{i=1}^{L-1} \left(2^{H-M-2L+i} - 1\right) (2L - 2i) \\
= & 2^{2H-2M-3L+3} - 2^{2H-2M-L+2} - 3 \cdot 2^{2H-M-2L+2} + 2^{2H-M-L+3} - 2^{H-M-2L+1} \\
& - 5 \cdot 2^{H-M-L+1} + 2^{H-M+3} - (3L - 2)2^{H+1} - L, \tag{11}
\end{aligned}$$

and

$$\begin{aligned}
& S_H(M, M + 2L + 1) \\
= & \left(2^{H-M-2L} - 1\right) \sum_{i=1}^L (2L - 2i + 2) + \sum_{i=1}^{L-1} \sum_{j=1}^{L-i} (2L - 2i - 2j + 2) \\
& + \left(2^{H-M-2L} - 1\right) \sum_{i=1}^L \left(2^{H-M-i+1} - 1\right) (2L - 2i + 1) \\
& + \sum_{i=1}^{L-1} \left(2^{H-M-i+1} - 1\right) \sum_{j=1}^{L-i} (2L - 2i - 2j + 1) \\
& + \sum_{i=1}^L \left(2^{H-M-2L+i-1} - 1\right) \sum_{j=1}^{L-i+1} (2L - 2i - 2j + 3) \\
& + \sum_{i=1}^{L-1} \left(2^{H-M-2L+i-1} - 1\right) \sum_{j=1}^{L-i} \left(2^{H-M-j+1} - 1\right) (2L - 2i - 2j + 2) \\
& + \left(2^{H+1} - 2^{H-M+1}\right) \left(2^{H-M-2L} - 1\right) 2L \\
& + \left(2^{H+1} - 2^{H-M+1}\right) \sum_{i=1}^{L-1} (2L - 2i) \\
& + \left(2^{H+1} - 2^{H-M+1}\right) \sum_{i=1}^L \left(2^{H-M-2L+i-1} - 1\right) (2L - 2i + 1) \\
= & \frac{5}{3} \cdot 2^{2H-2M-3L+1} - \frac{5}{3} \cdot 2^{2H-2M-L+1} - 3 \cdot 2^{2H-M-2L+1} + 3 \cdot 2^{2H-M-L+1} \\
& - 2^{H-M-2L+1} - 2^{H-M-L+3} + 5 \cdot 2^{H-M+1} - 3L \cdot 2^{H+1} - 2L, \tag{12}
\end{aligned}$$

respectively.

Lemma 1

- (i) If $L = 1$, then $S_H(M, M + 2L) < S_H(M, M + 2L + 1)$.
- (ii) If $L \geq 2$, then $S_H(M, M + 2L) > S_H(M, M + 2L + 1)$.

Proof

(i) If $L = 1$, then

$$\begin{aligned} & S_H(M, M + 2L) - S_H(M, M + 2L + 1) \\ &= 2^{2H-M-1} \left(\frac{1}{2^{M+1}} + \frac{1}{2^{H-M-3}} - 1 \right) - 3 \cdot 2^{H-M} + 1 \\ &< 0. \end{aligned} \tag{13}$$

(ii) If $L \geq 2$, then

$$\begin{aligned} & S_H(M, M + 2L) - S_H(M, M + 2L + 1) \\ &= \frac{1}{3} \cdot 2^{2H-M-L+1} \left(3 - \frac{1}{2^M} - \frac{9}{2^L} + \frac{7}{2^{M+2L}} \right) \\ &\quad + 2^{H+2} \left(1 - \frac{1}{2^M} + \frac{1}{2^{M+1}} - \frac{1}{2^{M+L+1}} \right) + L \\ &> 0, \end{aligned} \tag{14}$$

where $L = 2, 3, \dots, \lfloor (H - M - 1)/2 \rfloor$.

Q.E.D.

Lemma 2 If $L \geq 2$, then $L^* = 2$ maximizes $S_H(M, M + 2L)$.

Proof If $L \geq 2$, then $L^* = 2$ maximizes $S_H(M, M + 2L)$ since

$$\begin{aligned} & S_H(M, M + 2L) - S_H(M, M + 2L + 2) \\ &= 2^{2H-M-L+1} \left(2 - \frac{1}{2^M} - \frac{9}{2^{L+1}} + \frac{7}{2^{M+2L+1}} \right) \\ &\quad + 2^H \left(6 - \frac{1}{2^{M+2L-1}} - \frac{1}{2^{M+L-2}} - \frac{1}{2^{M+L}} + \frac{1}{2^{M+2L+1}} \right) + 1 \\ &> 0, \end{aligned} \tag{15}$$

where $L = 2, 3, \dots, \lfloor (H - M)/2 \rfloor - 1$.

Q.E.D.

Lemma 3

- (i) If $M = 0$ and $H = 4$, then $S_H(M, M + 3) > S_H(M, M + 4)$.
- (ii) If $M = 0$ and $H \geq 5$, then $S_H(M, M + 3) < S_H(M, M + 4)$.
- (iii) If $M \geq 1$, then $S_H(M, M + 3) > S_H(M, M + 4)$.

Proof

(i) If $M = 0$ and $H = 4$, then

$$S_H(M, M + 3) - S_H(M, M + 4) = 2 > 0. \tag{16}$$

(ii) If $M = 0$ and $H \geq 5$, then

$$S_H(M, M + 3) - S_H(M, M + 4) = 2^{2H-3} \left(\frac{17}{2^H} - 1 \right) < 0. \quad (17)$$

(iii) If $M \geq 1$, then

$$\begin{aligned} & S_H(M, M + 3) - S_H(M, M + 4) \\ &= 2^{2H-M-2} \left(1 - \frac{1}{2^{M-2}} + \frac{1}{2^{M-1}} + \frac{1}{2^{M+1}} \right) + 2^{H-M-3} + 2^{H+1} \\ &> 0. \end{aligned} \quad (18)$$

Q.E.D.

Theorem 4 *Let N^* maximize $S_H(M, N)$ for each value of M , then we have the following:*

- (i) *If $M = H - 2$, then $N^* = M + 2$.*
- (ii) *If $M = H - 3$, then $N^* = M + 3$.*
- (iii) *If $M \leq H - 4$, then we have the following:*
 - (a) *If $M = 0$ and $H = 4$, then $N^* = M + 3$.*
 - (b) *If $M = 0$ and $H \geq 5$, then $N^* = M + 4$.*
 - (c) *If $M \geq 1$, then $N^* = M + 3$.*

Proof

- (i) If $M = H - 2$, then $N^* = M + 2$ trivially.
- (ii) If $M = H - 3$, then $N^* = M + 3$ since $S_H(M, M + 2) < S_H(M, M + 3)$ from (i) of Lemma 1.
- (iii) If $M \leq H - 4$, then $N^* = M + 3$ for $N \leq M + 3$ from (i) of Lemma 1 and $N^* = M + 4$ for $N \geq M + 4$ from (ii) of Lemma 1 and Lemma 2.
 - (a) If $M = 0$ and $H = 4$, then $N^* = M + 3$ since $S_H(M, M + 3) > S_H(M, M + 4)$ from (i) of Lemma 3.
 - (b) If $M = 0$ and $H \geq 5$, then $N^* = M + 4$ since $S_H(M, M + 3) < S_H(M, M + 4)$ from (ii) of Lemma 3.
 - (c) If $M \geq 1$, then $N^* = M + 3$ since $S_H(M, M + 3) > S_H(M, M + 4)$ from (iii) of Lemma 3.

Q.E.D.

Table 1 Optimal depth N^* for each value of M and the total shortening distance $S_H(M, N^*)$

M	$H = 3$		$H = 4$		$H = 5$		$H = 6$	
	N^*	$S_H(M, N^*)$	N^*	$S_H(M, N^*)$	N^*	$S_H(M, N^*)$	N^*	$S_H(M, N^*)$
0	3	10	3	54	4	298	4	1366
1	3	9	4	58	4	342	4	1582
2	–	–	4	25	5	154	5	918
3	–	–	–	–	5	57	6	346
4	–	–	–	–	–	–	6	121

4 Numerical Examples

Table 1 shows numerical examples of the optimal depth N^* for each value of M and the total shortening distance $S_H(M, N^*)$ in the case of $H = 3, 4, 5, 6$ and $M = 0, 1, 2, 3, 4$.

Table 1 reveals the optimal pair of depths $(M, N)^*$ which maximizes $S_H(M, N)$. If $H = 3$, then $(M, N)^* = (0, 3)$ and if $H = 4, 5, 6$, then $(M, N)^* = (1, 4)$.

5 Conclusions

This study considered the addition of relation to a linking pin organization structure such that the communication of information between every member in the organization becomes the most efficient. For a model of adding an edge between a node with a depth M and its descendant with a depth N to a complete binary linking pin structure of height H where every pair of siblings in a complete binary tree of height H is adjacent, we obtained an optimal depth N^* which maximizes the total shortening distance for each value of M . Theorem 4 reveals that the most efficient manner of adding relation between a superior and his indirect subordinate is to add the relation to a subordinate of the second, the third or the fourth level below the superior depending on the level of superior and the number of levels in the organization structure. Furthermore, we illustrate an optimal depth N^* for each value of M and the total shortening distance $S_H(M, N^*)$ with numerical examples.

References

1. Likert R, Likert JG (1976) New ways of managing conflict. McGraw-Hill, New York
2. Robbins SP (2003) Essentials of organizational behavior. 7th edn. Prentice Hall, Upper Saddle River
3. Takahara Y, Mesarovic M (2003) Organization structure: cybernetic systems foundation. Kluwer Academic/Plenum Publishers, New York
4. Sawada K, Wilson R (2006) Models of adding relations to an organization structure of a complete K-ary tree. Eur J Oper Res 174:1491–1500

5. Sawada K (2008) Adding relations in the same level of a linking pin type organization structure. *IAENG Int J Appl Math* 38:20–25
6. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) *Introduction to algorithms*. 2nd edn. MIT Press, Cambridge
7. Sawada K, Kawakatsu H, Mitsuishi T (2012) A model of adding relation between the top and a member of a linking pin organization structure with K subordinates. *Lecture notes in engineering and computer science: Proceedings of the international multiConference of engineers and computer scientists 2012, IMECS 2012, Hong Kong, 14–16 March 2012*, pp 1598–1601
8. Sawada K (2012) A model of adding relation between two levels of a linking pin organization structure. *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012*, pp 76–79

Bayesian Inference for the Parameters of Two-Parameter Exponential Lifetime Models Based on Type-I and Type-II Censoring

Husam Awni Bayoud

Abstract The parameters of the two-parameter exponential distribution are estimated in this chapter from the Bayesian viewpoint based on complete, Type-I and Type-II censored samples. Bayes point estimates and credible intervals of the unknown parameters are proposed under the assumption of suitable priors on the unknown parameters and under the assumption of the squared error loss function. Illustrative example is provided to motivate the proposed Bayes point estimates and the credible intervals. Various Monte Carlo simulations are also performed to compare the performances of the classical and Bayes estimates.

Keywords Bayes estimate · Censored samples · Credible interval · Maximum likelihood estimate · Mean squared error · Squared error loss function

1 Introduction

Let X_1, X_2, \dots, X_n be a random sample of size n from a two-parameter exponential distribution with a scale parameter θ and a location parameter λ , denoted by $E(\theta, \lambda)$, where θ and λ are independent. If the lifetime of a component is assumed to follow an exponential life model with parameters θ and λ then the parameter λ represents the component's guarantee lifetime, and the parameter $1/\theta$ represents the component's mean lifetime.

The probability density function (p.d.f) of X at x is given by:

$$f(x|\theta, \lambda) = \theta e^{-\theta(x-\lambda)}; \quad 0 \leq \lambda \leq x \quad \text{and} \quad \theta > 0 \quad (1)$$

H. A. Bayoud (✉)
Fahad Bin Sultan University, Tabuk 71454, Saudi Arabia
e-mail: hbayoud@fbsu.edu.sa; husam.awni@yahoo.com

This distribution plays an important role in survival and reliability analysis; see for example Balakrishnan and Basu [1].

In life testing experiments, it often happens that the experiment is censored in the sense that the experimenter may not be in a position to observe the life times of all items put on test because of time limitations and other restrictions on the data collection. The two most common censoring schemes are Type-I and Type-II censoring schemes. In Type-I censoring scheme, the experiment continues up to a preselected fixed time T but the number of failures is random, whereas in Type-II censoring scheme, the experimental time is random but the number of failures is fixed, k .

The estimation of the parameters of two-parameter exponential distribution based on Types I and II censored samples has been considered by several authors in the literature from the Bayesian point of view. El-Sayyed [2] has derived Bayes estimate and unbiased estimate for θ^{-1} . Singh and Prasad [3, 4] have considered the problem of estimating the scale parameter θ^{-1} from the Bayesian viewpoint when the scale parameter λ is known. Sarhan [5] has studied several empirical Bayes estimates for one parameter exponential distribution. Singh and Kumar [6, 7] proposed Bayes estimates for the scale parameter under multiply Type-II censoring scheme. Singh and Kumar [8] proposed Bayes point estimates for the scale parameter under Type-II censoring by using generalized non-informative prior and natural conjugate prior. Shi and Yan [9] proposed empirical Bayes estimate for the scale parameter under Type-I censored sample assuming known location parameter. Recently, Bayoud [10] has proposed Bayes estimates and credible intervals for the scale and location parameters based on Type-I censored sample under the assumption of squared error loss function.

It is noted that in many practical applications, the value of the parameter λ may not be known. Therefore, it is useful and important to consider the problem of estimating the parameter θ when λ is unknown.

This chapter aims to derive Bayes point estimates and credible intervals for scale and location parameters of a two-parameter exponential distribution in order to estimate the guarantee and the mean life time of that distribution. This will be performed based on complete, Type-I and Type-II censored samples. Bayes point estimates are proposed under the assumption of the squared error loss function. The scale parameter θ is assumed to follow exponential distribution with hyper parameter A , and the location parameter λ is assumed to follow uniform distribution from zero to B . Suggestions for choosing the hyper parameters A and B are provided.

The rest of this chapter is organized as follows: Sect. 2 describes the probability models that are needed in this work. Bayes point estimates for the scale and location parameters are proposed in Sect. 3 based on complete, Type-I and Type-II censored samples separately. Credible intervals are derived for the unknown parameters in Sect. 4. An illustrative example is provided in Sect. 5. Simulation studies are performed in Sect. 6. Finally, the main conclusions are included in Sect. 7.

2 Models

2.1 Complete Sample

Let $X_1, X_2, \dots, X_n \sim E(\theta, \lambda)$, with p.d.f given in (1). The likelihood function of the complete sample X_1, X_2, \dots, X_n given θ and λ is given by:

$$L(x_1, x_2, \dots, x_n | \theta, \lambda) = \theta^n e^{-\theta \sum_{i=1}^n (x_i - \lambda)} \tag{2}$$

Suitable priors on the unknown parameters are assumed in order to derive Bayes estimates and credible intervals.

The parameter θ is assumed to follow exponential distribution with p.d.f given by:

$$g(\theta) = Ae^{-A\theta}; \quad A > 0 \tag{3}$$

where the hyper parameter A is a preselected positive real number that is chosen to reflect our beliefs about the expected value of $1/\theta$, because the expected value of θ equals $1/A$.

The parameter λ is assumed to follow a uniform distribution with p.d.f given by:

$$p(\lambda) = \frac{1}{B}; \quad 0 \leq \lambda \leq B \tag{4}$$

where the hyper parameter B is a preselected positive real number that is chosen to reflect our beliefs about the lower bound of the x 's, which can be easily assumed to equal the minimum observed value, $x_{(1)}$.

The joint posterior p.d.f of θ and λ given $\{x_1, x_2, \dots, x_n\}$ is given by:

$$\begin{aligned} h_C(\theta, \lambda | x_1, x_2, \dots, x_n) &= \frac{L(x_1, x_2, \dots, x_n | \lambda, \theta) g(\theta) p(\lambda)}{\int_0^B \int_0^\infty L(x_1, x_2, \dots, x_n | \lambda, \theta) g(\theta) p(\lambda) d\theta d\lambda} \\ &= \frac{n\theta^n e^{-\theta[A + \sum_{i=1}^n (x_i - \lambda)]}}{C\Gamma(n)} \end{aligned} \tag{5}$$

where $C = \frac{1}{D^n} - \frac{1}{E^n}$ in which $D = A + \sum_{i=1}^n (x_i - B)$ and $E = A + \sum_{i=1}^n x_i$, $\theta > 0$ and $0 \leq \lambda \leq B$.

Therefore,

The marginal posterior p.d.f of θ given $\{x_1, x_2, \dots, x_n\}$ is given by:

$$\begin{aligned}
 h_{\theta,C}(\theta|x_1, x_2, \dots, x_n) &= \int_0^B h(\theta, \lambda|x_1, x_2, \dots, x_n) d\lambda \\
 &= \frac{\theta^{n-1}}{C\Gamma(n)} \left(e^{-D\theta} - e^{-E\theta} \right)
 \end{aligned}
 \tag{6}$$

where $\theta > 0$, D , E and C are defined in (5).

The marginal posterior p.d.f of λ given $\{x_1, x_2, \dots, x_n\}$ is given by:

$$\begin{aligned}
 h_{\lambda,C}(\lambda|x_1, x_2, \dots, x_n) &= \int_0^\infty h(\theta, \lambda|x_1, x_2, \dots, x_n) d\theta \\
 &= \frac{n^2}{C} \frac{1}{\left(A + \sum_{i=1}^n (x_i - \lambda) \right)^{n+1}}
 \end{aligned}
 \tag{7}$$

where $0 \leq \lambda \leq B \leq x_{(1)}$ and C is defined in (5).

2.2 Type-I Censored Sample

In Type-I censored scheme a random sample of n units is tested until a predetermined time T at which the test is terminated. Failure times of r units are observed, where r is a random variable. Thus the lifetime x_i is observed only if $x_i \leq T$; $i = 1, 2, \dots, n$.

$$\text{Let } \delta_i = \begin{cases} 0; & x_i > T \\ 1; & x_i \leq T \end{cases}$$

Therefore, $r = \sum_{i=1}^n \delta_i$ which is assumed to be greater than zero.

The likelihood function of the Type-I censored data is given by:

$$\begin{aligned}
 L_I(x_1, x_2, \dots, x_n|\theta, \lambda, T) &= \prod_{i=1}^n [f(x_i|\theta, \lambda)]^{\delta_i} [1 - F(T|\theta, \lambda)]^{1-\delta_i} \\
 &= \theta^r e^{-\theta \sum_{i=1}^n x_i \delta_i} e^{-\theta[T(n-k) - n\lambda]}
 \end{aligned}
 \tag{8}$$

The joint posterior p.d.f of θ and λ based on the Type -I censored sample is given by:

$$\begin{aligned}
 h_I(\theta, \lambda | x_1, x_2, \dots, x_n, T) &= \frac{L_I(x_1, x_2, \dots, x_n | \lambda, \theta) g(\theta) p(\lambda)}{\int_0^B \int_0^\infty L_I(x_1, x_2, \dots, x_n | \lambda, \theta) g(\theta) p(\lambda) d\theta d\lambda} \\
 &= \frac{ne^{-\theta \left[\sum_{i=1}^n x_i \delta_i + T(n-k) - n\lambda + A \right]}}{C_1 \Gamma(r)} \tag{9}
 \end{aligned}$$

where $\theta > 0, 0 \leq \lambda \leq B$ and $C_1 = \frac{1}{D_1} - \frac{1}{E_1} \neq 0$ in which $D_1 = \sum_{i=1}^n x_i \delta_i + A + T(n-r) - nB$ and $E_1 = \sum_{i=1}^n x_i \delta_i + A + T(n-r)$.

The marginal posterior p.d.f of θ given Type-I censored data is given by:

$$\begin{aligned}
 h_{\theta, I}(\theta | x_1, x_2, \dots, x_n, T) &= \int_0^B h_I(\theta, \lambda | x_1, x_2, \dots, x_n, T) d\lambda \\
 &= \frac{\theta^{r-1}}{C_1 \Gamma(r)} \left(e^{-D_1 \theta} - e^{-E_1 \theta} \right) \tag{10}
 \end{aligned}$$

where $\theta > 0, D_1, E_1$ and C_1 are defined in (9).

The marginal posterior p.d.f of λ given Type-I censored data is given by:

$$\begin{aligned}
 h_{\lambda, I}(\lambda | x_1, x_2, \dots, x_n, T) &= \int_0^\infty h_I(\theta, \lambda | x_1, x_2, \dots, x_n, T) d\theta \\
 &= \frac{nr}{C_1} \frac{1}{\left[\sum_{i=1}^n x_i \delta_i + T(n-r) - n\lambda + A \right]^{r+1}} \tag{11}
 \end{aligned}$$

where $0 \leq \lambda \leq B \leq x_{(1)}$ and C_1 is defined in (9).

2.3 Type-II Censored Sample

In Type-II censored scheme the number of failures k is determined at the beginning of the experiment, the time needed to observe those k failures equals $x_{(k)}$, the k th order statistic. The likelihood function of the Type-II censored data is given by:

$$\begin{aligned}
 L_{II} (x_{(1)}, x_{(2)}, \dots, x_{(k)}|\theta, \lambda) &= \frac{n!}{(n-k)!} \prod_{i=1}^k f(x_{(i)}|\theta, \lambda) [1 - F(x_{(k)})]^{n-k} \\
 &= \frac{n!}{(n-k)!} \theta^k e^{-\theta \left[\sum_{i=1}^k x_{(i)} + (n-k)x_{(k)} - n\lambda \right]} \tag{12}
 \end{aligned}$$

The joint posterior p.d.f of θ and λ based on the Type-II censored sample is given by:

$$\begin{aligned}
 h_{II} (\theta, \lambda|x_{(1)}, x_{(2)}, \dots, x_{(k)}) &= \frac{L_{II} (x_{(1)}, x_{(2)}, \dots, x_{(k)}|\lambda, \theta) g(\theta) p(\lambda)}{\int_0^B \int_0^\infty L_{II} (x_{(1)}, x_{(2)}, \dots, x_{(k)}|\lambda, \theta) g(\theta) p(\lambda) d\theta d\lambda} \\
 &= \frac{n\theta^k e^{-\theta \left[\sum_{i=1}^k x_{(i)} + (n-k)x_{(k)} - n\lambda + A \right]}}{C_2 \Gamma(k)} \tag{13}
 \end{aligned}$$

where $\theta > 0, 0 \leq \lambda \leq B$ and $C_2 = \frac{1}{D_2^k} - \frac{1}{E_2^k} \neq 0$ in which $D_2 = \sum_{i=1}^k x_{(i)} + A + (n-k)x_{(k)} - nB$ and $E_2 = \sum_{i=1}^k x_{(i)} + A + (n-k)x_{(k)}$

The marginal posterior p.d.f of θ given Type-II censored data is given by:

$$\begin{aligned}
 h_{\theta, II}(\theta|x_{(1)}, x_{(2)}, \dots, x_{(k)}) &= \int_0^B h_{II} (\theta, \lambda|x_{(1)}, x_{(2)}, \dots, x_{(k)}) d\lambda \\
 &= \frac{\theta^{k-1}}{C_2 \Gamma(k)} \left(e^{-D_2\theta} - e^{-E_2\theta} \right) \tag{14}
 \end{aligned}$$

where $\theta > 0, D_2, E_2$ and C_2 are defined in (13).

The marginal posterior p.d.f of λ given Type-II censored data is given by:

$$\begin{aligned}
 h_{\lambda, II}(\lambda|x_{(1)}, x_{(2)}, \dots, x_{(k)}) &= \int_0^\infty h_{II} (\theta, \lambda|x_{(1)}, x_{(2)}, \dots, x_{(k)}) d\theta \\
 &= \frac{nk}{C_2} \frac{1}{\left[\sum_{i=1}^k x_{(i)} + (n-k)x_{(k)} - n\lambda + A \right]^{k+1}} \tag{15}
 \end{aligned}$$

where $0 \leq \lambda \leq B \leq x_{(1)}$ and C_2 is defined in (13).

3 Classical and Bayes Point Estimates

In this section the maximum likelihood and Bayesian estimates (MLE and BE) are proposed for the unknown parameters based on the complete, Type-I and Type-II censored samples. The BE are derived under the assumption of the squared error loss function (SELF). However, the BE of a parameter equals the posterior mean of that parameter if the SELF is assumed.

3.1 Based on Complete Sample

In the case of complete sample, the BE of the unknown parameters θ and λ are respectively given by:

$$\begin{aligned} \hat{\theta}_C = E_{h_{\theta,C}}(\theta) &= \int_0^{\infty} \theta h_{\theta,C}(\theta|x_1, x_2, \dots, x_n) d\theta \\ &= \frac{n}{C} \left(\frac{1}{D^{n+1}} - \frac{1}{E^{n+1}} \right) \end{aligned} \tag{16}$$

$$\begin{aligned} \hat{\lambda}_C = E_{h_{\lambda,C}}(\lambda) &= \int_0^B \lambda h_{\lambda,C}(\lambda|x_1, x_2, \dots, x_n) d\lambda \\ &= \frac{1}{C} \left(\frac{B}{D^n} + \frac{1}{n(1-n)} \left(\frac{1}{D^{n-1}} - \frac{1}{E^{n-1}} \right) \right) \end{aligned} \tag{17}$$

The MLE of θ and λ based on the complete sample are respectively: $\hat{\theta}_{MLE,C} = \frac{n}{\sum_{i=1}^n (x_i - x_{(1)})}$ and $\hat{\lambda}_{MLE,C} = x_{(1)}$

3.2 Based on Type-I Censored Sample

In the case of Type-I censored sample, the BE of the unknown parameters θ and λ are respectively given by:

$$\begin{aligned}\hat{\theta}_I = E_{h_{\theta,I}}(\theta) &= \int_0^{\infty} \theta h_{\theta,I}(\theta|x_1, x_2, \dots, x_n) d\theta \\ &= \frac{r}{C_1} \left(\frac{1}{D_1^{r+1}} - \frac{1}{E_1^{r+1}} \right)\end{aligned}\quad (18)$$

$$\begin{aligned}\hat{\lambda}_I = E_{h_{\lambda,I}}(\lambda) &= \int_0^B \lambda h_{\lambda,I}(\lambda|x_1, x_2, \dots, x_n) d\lambda \\ &= \frac{1}{C_1} \left(\frac{B}{D_1^k} + \frac{1}{n(1-r)} \left(\frac{1}{D_1^{r-1}} - \frac{1}{E_1^{r-1}} \right) \right)\end{aligned}\quad (19)$$

The MLE of θ and λ based on the Type-I censored sample are respectively:

$$\hat{\theta}_{MLE,I} = \frac{r}{\sum_{i=1}^n x_i \delta_i + T(n-r) - nx_{(1)}} \quad \text{and} \quad \hat{\lambda}_{MLE,I} = x_{(1)}$$

3.3 Based on Type-II Censored Sample

In the case of Type-II censored sample, the BE of the unknown parameters θ and λ are respectively given by:

$$\begin{aligned}\hat{\theta}_{II} = E_{h_{\theta,II}}(\theta) &= \int_0^{\infty} \theta h_{\theta,II}(\theta|x_{(1)}, x_{(2)}, \dots, x_{(k)}) d\theta \\ &= \frac{k}{C_2} \left(\frac{1}{D_2^{k+1}} - \frac{1}{E_2^{k+1}} \right)\end{aligned}\quad (20)$$

$$\begin{aligned}\hat{\lambda}_{II} = E_{h_{\lambda,II}}(\lambda) &= \int_0^B \lambda h_{\lambda,II}(\lambda|x_{(1)}, x_{(2)}, \dots, x_{(k)}) d\lambda \\ &= \frac{1}{C_2} \left(\frac{B}{D_2^k} + \frac{1}{n(1-k)} \left(\frac{1}{D_2^{k-1}} - \frac{1}{E_2^{k-1}} \right) \right)\end{aligned}\quad (21)$$

The MLE of θ and λ based on the Type-II censored sample are respectively:

$$\hat{\theta}_{MLE,II} = \frac{k}{\sum_{i=1}^k x_{(i)} + x_{(k)}(n-k) - nx_{(1)}} \quad \text{and} \quad \hat{\lambda}_{MLE,II} = x_{(1)}$$

4 Credible Intervals

4.1 Based on Complete Sample

Based on the complete sample x_1, x_2, \dots, x_n and by using the posterior density function of θ that is defined in (6), the equal-tailed $(1 - \alpha)100\%$ credible interval for θ denoted by (θ_L, θ_U) can be obtained numerically by solving the following integral equations:

$$\int_0^{\theta_L} \frac{\theta^{n-1}}{C\Gamma(n)} (e^{-D\theta} - e^{-E\theta}) d\theta = \alpha/2 \quad \text{and} \quad \int_{\theta_U}^{\infty} \frac{\theta^{n-1}}{C\Gamma(n)} (e^{-D\theta} - e^{-E\theta}) d\theta = \alpha/2$$

Similarly, by using the posterior density function of λ that is defined in (7), the equal-tailed $(1 - \alpha)100\%$ credible interval for λ can be easily derived as:

$$\left(-\frac{1}{n} \left[\left(\frac{C\alpha}{2} + F^{-n} \right)^{-1/n} - F \right], -\frac{1}{n} \left[\left(\frac{2}{C\alpha} - (F - nB)^{-n} \right)^{-1/n} - F \right] \right)$$

in which $F = \sum_{i=1}^n x_i + A$.

4.2 Based on Type-I Censored Sample

Based on a Type-I censored sample and by using the posterior density function defined in (10), the equal-tailed $(1 - \alpha)100\%$ credible interval for θ denoted by $(\theta_{L,I}, \theta_{U,I})$ can be obtained numerically by solving the following integral equations:

$$\int_0^{\theta_{L,I}} \frac{\theta^{r-1}}{C_1\Gamma(r)} (e^{-D_1\theta} - e^{-E_1\theta}) d\theta = \alpha/2 \quad \text{and} \\ \int_{\theta_{U,I}}^{\infty} \frac{\theta^{r-1}}{C_1\Gamma(r)} (e^{-D_1\theta} - e^{-E_1\theta}) d\theta = \alpha/2$$

Similarly, by using the posterior density function of λ that is defined in (11), the equal-tailed $(1 - \alpha)100\%$ credible interval for λ can be easily derived as:

$$\left(-\frac{1}{n} \left[\left(\frac{C_1\alpha}{2} + F_1^{-r} \right)^{-1/r} - F_1 \right], -\frac{1}{n} \left[\left(\frac{2}{C_1\alpha} - [F_1 - nB]^{-r} \right)^{-1/r} - F_1 \right] \right)$$

in which $F_1 = A + \sum_{i=1}^n x_i \delta_i + T(n - r)$.

4.3 Based on Type-II Censored Sample

Based on a Type-II censored sample and by using the posterior density function defined in (14), the equal-tailed $(1 - \alpha)100\%$ credible interval for θ denoted by $(\theta_{L,II}, \theta_{U,II})$ can be obtained numerically by solving the following integral equations:

$$\int_0^{\theta_{L,II}} \frac{\theta^{k-1}}{C_2\Gamma(k)} \left(e^{-D_2\theta} - e^{-E_2\theta} \right) d\theta = \alpha/2 \quad \text{and}$$

$$\int_{\theta_{U,II}}^B \frac{\theta^{k-1}}{C_2\Gamma(k)} \left(e^{-D_2\theta} - e^{-E_2\theta} \right) d\theta = \alpha/2$$

Similarly, by using the posterior density function of λ that is defined in (15) the equal-tailed $(1 - \alpha)100\%$ credible interval for λ can be easily derived as:

$$\left(-\frac{1}{n} \left[\left(\frac{C_2\alpha}{2} + F_2^{-k} \right)^{-1/k} - F_2 \right], -\frac{1}{n} \left[\left(\frac{2}{C_2\alpha} - [F_2 - nB]^{-k} \right)^{-1/k} - F_2 \right] \right)$$

in which $F_2 = A + \sum_{i=1}^k x_{(i)} + (n - k)x_{(k)}$.

5 Numerical Example

Using Mathematica 5, if U has a Uniform(0,1) distribution, then x that satisfies $U = 1 - e^{-\theta(x-\lambda)}$ follows $E(\theta, \lambda)$. Let Data I = {9.25012, 9.67048, 9.98415, 8.35142, 8.26661, 11.1222, 8.79416, 8.16523, 11.3372, 8.68471, 10.478, 11.0089} be a random sample generated from $E(0.5, 8)$.

Table 1 summarizes the values of MLE, BE and credible interval for the scale and location parameters. Those estimates were computed based on the complete, Type-I and Type-II censored samples. The hyper parameters A and B were assumed to equal one over of the available sample's mean and the minimum observation respectively.

Table 1 MLE, BE and 95 % credible interval for θ and λ based on data I

	Complete sample ($n = 12$)			Type-I CS ($T = 11$)			Type-II CS ($k = 6$)		
	MLE	BE	95%CI	MLE	BE	95%CI	MLE	BE	95%CI
$\theta = 0.5$	0.70	0.70	(0.40, 1.06)	0.36	0.54	(0.28, 0.86)	0.66	0.66	(0.29, 1.15)
$\lambda = 8$	8.17	8.04	(7.76, 9.60)	8.17	7.99	(7.61, 9.56)	8.17	8.01	(7.67, 8.92)

It becomes apparent from Table 1 that the MLEs and BEs give almost the same results for estimation the scale parameter θ based on the complete and Type-II censored samples. It can be also seen from Table 1 that BE performs, in terms of the mean square error MSE, better than the MLE for estimation the scale parameter θ based on Type-I censored sample. On another hand, BE dominates, in terms of MSE, the MLE for estimation the location parameter λ based on complete, Type-I and Type-II censored samples. Moreover, the proposed credible interval gives reasonable results for estimation the parameters θ and λ in all cases.

6 Simulation Studies

In this section, the performance of the MLEs and the proposed BEs of λ and θ is investigated through various simulation studies based on complete, Type-I (with arbitrary $T = 5$) and Type-II (with arbitrary $r = 3$) censored schemes. Simulation studies are carried out on various exponential distributions with $(\theta, \lambda) = (0.5, 2), (3, 0.3), (1, 1)$ and $(2, 0)$. The hyper parameters A and B are assumed to equal one over the available sample’s mean and the minimum observed value, $x_{(1)}$ respectively. The main reason for doing this is to allow us to compare the proposed

Table 2 Expected MLE and BE along with their MSE based on complete sample

θ	λ	n	$\hat{\theta}_{MLE,C}$	$\hat{\theta}_{B,C}$	$\hat{\lambda}_{MLE,C}$	$\hat{\lambda}_{B,C}$
0.5	1	5	0.84 (0.41)	0.77 (0.22)	1.38 (0.30)	1.08 (0.15)
		30	0.53 (0.01)	0.53 (0.01)	1.07 (0.01)	1.00 (0.01)
	4	5	0.85 (0.46)	0.82 (0.36)	4.40 (00.32)	4.01 (0.20)
		30	0.54 (0.01)	0.54 (0.01)	4.07 (0.01)	4.00 (0.00)
3.0	1	5	4.93 (15.42)	2.56 (0.65)	1.07 (0.01)	0.97 (0.01)
		30	3.20 (0.45)	2.95 (0.29)	1.01 (0.00)	1.00 (0.00)
	4	5	4.88 (13.38)	3.76 (3.50)	4.07 (0.01)	3.99 (0.01)
		30	3.21 (0.41)	3.13 (0.35)	4.01 (0.00)	4.00 (0.00)
5.5	1	5	8.59 (33.72)	3.25 (5.37)	1.03 (0.00)	0.96 (0.00)
		30	5.87 (1.28)	5.01 (0.81)	1.01 (0.00)	1.00 (0.00)
	4	5	9.25 (51.13)	5.97 (5.00)	4.04 (0.00)	4.00 (0.00)
		30	5.81 (1.06)	5.55 (0.79)	4.01 (0.00)	4.00 (0.00)

Table 3 Expected MLE and BE along with their MSE based on Type-I censored sample

θ	λ	n	$\hat{\theta}_{MLE,I}$	$\hat{\theta}_{B,I}$	$\hat{\lambda}_{MLE,I}$	$\hat{\lambda}_{B,I}$
0.5	1	5	0.84 (0.42)	0.77 (0.23)	1.38 (0.30)	1.07 (0.15)
		30	0.54 (0.01)	0.53 (0.01)	1.07 (0.01)	1.00 (0.01)
	4	5	0.80 (0.37)	0.77 (0.30)	4.40 (00.32)	3.97 (0.18)
		30	0.52 (0.01)	0.52 (0.01)	4.07 (0.01)	4.00 (0.01)
3.0	1	5	4.93 (15.4)	2.56 (0.65)	1.07 (0.01)	0.97 (0.01)
		30	3.20 (0.45)	2.95 (0.29)	1.01 (0.00)	1.00 (0.00)
	4	5	4.83 (13.61)	3.73 (3.36)	4.07 (0.01)	3.99 (0.01)
		30	3.18 (0.39)	3.10 (0.33)	4.01 (0.00)	4.00 (0.00)
5.5	1	5	8.59 (33.72)	3.25 (5.37)	1.03 (0.00)	0.96 (0.00)
		30	5.87 (1.28)	5.01 (0.81)	1.01 (0.00)	1.00 (0.00)
	4	5	9.49 (69.27)	5.97 (5.79)	4.04 (0.00)	3.99 (0.00)
		30	5.83 (1.45)	5.56 (1.1)	4.01 (0.00)	4.00 (0.00)

Table 4 Expected MLE and BE along with their MSE based on Type-II censored sample

θ	λ	n	$\hat{\theta}_{MLE,II}$	$\hat{\theta}_{B,II}$	$\hat{\lambda}_{MLE,II}$	$\hat{\lambda}_{B,II}$
0.5	1	5	1.38 (3.09)	0.95 (0.49)	1.38 (0.30)	1.10 (0.16)
		30	1.54 (6.23)	0.86 (0.38)	1.07 (0.01)	0.99 (0.01)
	4	5	1.35 (4.19)	1.12 (1.37)	4.40 (00.32)	4.03 (0.22)
		30	1.47 (7.25)	1.16 (1.44)	4.07 (0.01)	4.00 (0.01)
3.0	1	5	9.80 (615.3)	2.09 (1.06)	1.07 (0.01)	0.93 (0/01)
		30	8.67 (244.9)	1.95 (1.32)	1.01 (0.00)	0.98 (0.00)
	4	5	9.76 (774.9)	4.31 (5.98)	4.07 (0.01)	3.98 (0.01)
		30	9.39 (1170)	4.10 (4.94)	4.01 (0.00)	4.00 (0.00)
5.5	1	5	13.5 (214.2)	2.40 (9.76)	1.03 (0.00)	0.92 (0.01)
		30	14.2 (375.1)	2.27 (10.5)	1.01 (0.00)	0.98 (0.00)
	4	5	14.4 (426.3)	5.63 (3.76)	4.04 (0.00)	3.98 (0.00)
		30	15.2 (499.6)	5.60 (3.40)	4.01 (0.00)	4.00 (0.00)

BEs with the MLEs directly. 1000 simulated datasets are generated from $E(\theta, \lambda)$ by using Mathematica 5. For the purpose of comparison, the average value of the MLE and the proposed BE along with the mean squared error (MSE) in parentheses are reported by assuming $n = 5$ and 30 based on complete, Type-I and Type-II censored samples in Tables 2, 3 and 4 respectively. Estimators with the smallest MSE values are preferred.

It becomes apparent from Tables 2, 3 and 4 that the proposed BEs behave better than the existing MLEs based on the complete, Type-I and Type-II censored samples as the MSE values of the proposed BEs are less than those of the MLEs. It has been shown in Table 4 that the MSE values of the MLEs of θ are so high whereas the MSE values of the proposed BEs are so small relatively to those of the MLEs, this motivates the using the proposed BE based on Type-II censored samples. It can be also observed that when n increases, the MSE of the proposed BEs and of the MLEs decreases, which is expected.

7 Conclusions

In this chapter, Bayes procedures for estimating the scale and location parameters, θ and λ , of a two parameter exponential distribution were developed based on complete, Type-I and Type-II censored samples. Prior probability distributions for the parameters θ and λ were assumed to be exponential and uniform distributions respectively. Bayes point estimates and credible intervals for θ and λ were proposed in the cases of complete, Type-I and Type-II censored samples under the squared error loss. It was shown from a random dataset that, the MLE and Bayes estimates gave excellent and almost equivalent results for estimation the parameter θ in the case of complete and Type-II censored samples. Furthermore, and based on a random dataset, excellent results were obtained from the proposed credible intervals for estimation the scale and the location parameters.

Bayes estimates are highly recommended to estimate the scale and location parameters of two-parameter exponential distribution based on Type-I and Type-II samples as simulation studies showed that the MSE values of the proposed Bayes estimates are much less than those of the existing MLEs .

References

1. Balakrishnan N, Basu AP (1995) The exponential distribution: theory, methods and applications. Gordon and Breach Publishers, Newark, NJ
2. El-Sayyed GM (1967) Estimation of the parameter of an exponential distribution. *J Roy Stat Soc: Ser B (Stat Methodol)* 29:525–532
3. Singh RS, Prasad B (1989) Uniformly strongly consistent prior distribution and empirical Bayes estimators with asymptotic optimality and rates in a non-exponential family. *Sankhya A*51:334–342
4. Prasad B, Singh RS (1990) Estimation of prior distribution and empirical Bayes estimation in a non-exponential family. *J Stat Plan Infer* 24:81–86
5. Sarhan AM (2003) Empirical Bayes estimates in exponential reliability model. *Appl Math Comput* 135:319–332
6. Singh U, Kumar A (2005) Shrinkage estimators for exponential scale parameter under multiply Type II censoring. *Aust J Stat* 34:39–49
7. Singh U, Kumar A (2005) Bayes estimator for one parameter exponential distribution under multiply-II censoring. *Indian J Math Math Sci* 1:23–33
8. Singh U, Kumar A (2007) Bayesian estimation of the exponential parameter under a multiply Type-II censoring scheme. *Austrian J Stat* 36(3):227–238
9. Shi Y, Yan W (2010) The EB estimation of scale-parameter for two-parameter exponential distribution under the Type-I censoring life test. *J Phys Sci* 14:25–30
10. Bayoud HA (2012) Bayesian analysis of Type-I censored sample from two-parameter exponential distribution. In: *Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE (2012), 4–6 July 2012 U.K , London*, pp 291–296

Analysing Metric Data Structures Thinking of an Efficient GPU Implementation

Roberto Uribe-Paredes, Enrique Arias, Diego Cazorla
and José Luis Sánchez

Abstract Similarity search is becoming a field of interest because it can be applied to different areas in science and engineering. In real applications, when large volumes of data are processing, query response time can be quite high. In this case, it is necessary to apply mechanisms to significantly reduce the average query response time. For that purpose, modern GPU/Multi-GPU systems offer a very impressive cost/performance ratio. In this paper, the authors make a comparative study of the most popular pivot selection methods in order to establish a set of attractive features from the point of view of future GPU implementations.

Keywords Clustering-based methods · Comparative study · Data structures · Metric spaces · Pivot-based methods · Range queries · Similarity search.

1 Introduction

In the last decade, the search of similar objects in a large collection of stored objects in a metric database has become a most interesting problem. This kind of search can be found in different applications such as voice and image recognition, data

R. Uribe-Paredes (✉)
Computer Engineering Department, University of Magallanes,
Avenida Bulnes, 01855 Punta Arenas, Chile
e-mail: roberto.uribeparedes@gmail.com

E. Arias · D. Cazorla · J. L. Sánchez
Computing Systems Department, University of Castilla-La Mancha,
Ave. Espaa s/n, 020071 Albacete, Spain
e-mail: enrique.arias@uclm.es

D. Cazorla
e-mail: diego.cazorla@uclm.es

J. L. Sánchez
e-mail: jose.sgarcia@uclm.es

mining, plagiarism detection and many others. A typical query for these applications is the *range search* which consists in obtaining all the objects that are at some given distance from the consulted object.

1.1 Similarity Search in Metric Spaces

Similarity is modeled in many interesting cases through metric spaces, and the search of similar objects through range search or nearest neighbors. A metric space (\mathbb{X}, d) is a set \mathbb{X} and a distance function $d : \mathbb{X}^2 \rightarrow \mathbb{R}$, such that $\forall x, y, z \in \mathbb{X}$ fulfills the properties of positiveness [$d(x, y) \geq 0$, and $d(x, y) = 0 \iff x = y$], symmetry [$d(x, y) = d(y, x)$] and triangle inequality [$d(x, y) + d(y, z) \geq d(x, z)$].

In a given metric space (\mathbb{X}, d) and a finite data set $\mathbb{Y} \subseteq \mathbb{X}$, a series of queries can be made. The basic query is the *range query* (x, r) , a query being $x \in \mathbb{X}$ and a range $r \in \mathbb{R}$. The range query around x with range r (or radius r) is the set of objects $y \in \mathbb{Y}$ such that $d(x, y) \leq r$. A second type of query that can be built using the range query is *k nearest neighbors (kNN)*, the query being $x \in \mathbb{X}$ and object k . k nearest neighbors to x are a subset \mathbb{A} of objects \mathbb{Y} , such that if $|\mathbb{A}| = k$ and an object $y \in \mathbb{A}$, there is no object $z \notin \mathbb{A}$ such that $d(z, x) \leq d(y, x)$.

Metric access methods, metric space indexes or *metric data structures* are different names for data structures built over a set of objects. The objective of these methods is to minimize the amount of distance evaluations made to solve the query. Searching methods for metric spaces are mainly based on dividing the space using the distance to one or more selected objects.

Metric space data structures can be grouped into two classes [1], *clustering-based* and *pivots-based* methods. The *clustering-based* structures divide the space into areas, where each area has a so-called centre. Some data is stored in each area, which allows easy discarding the whole area by just comparing the query with its centre. Algorithms based on clustering are better suited for high-dimensional metric spaces. Some clustering-based indexes are *BST* [2], *GHT* [3], *M-Tree* [4], *GNAT* [5], *EGNAT* [6] and many others.

There exist two criteria to define the areas in clustering-based structures: *hyperplanes* and *covering radius*. The former divides the space into *Voronoi* partitions and determines the hyperplane the query belongs to according to the corresponding centre. The covering radius criterion divides the space into spheres that can be intersected and one query can belong to one or more spheres.

In the *pivots-based* methods, a set of pivots is selected and the distances between the pivots and database elements are precalculated. When a query is made, the query distance to each pivot is calculated and the triangle inequality is used to discard the candidates. Its objective is to filter objects during a request through the use of a triangular inequality, without really measuring the distance between the object under request and the discarded object. Mathematically, these construction and searching processes can be expressed as follows:

- Let $\{p_1, p_2, \dots, p_k\}$ a set of pivots, $p_i \in \mathbb{X}$. For each element y of the database \mathbb{Y} the distance to the k pivots ($d(y, p_1), \dots, d(y, p_k)$) is stored. Given a query q and a range r , the distance ($d(q, p_1), \dots, d(q, p_k)$) to the k pivots is calculated.
- If for some pivot p_i the expression $|d(q, p_i) - d(y, p_i)| > r$ is holding, then for triangle inequality $d(q, y) > r$, and therefore it is unnecessary to explicitly evaluate $d(q, y)$. All the objects not discarded by this rule have to be directly compared to the query q .

Some pivots-based indexes are *LAESA* [7], *FQT* and its variants [8], *Spaghettis* and its variants [9], *FQA* [9], *SSS-Index* [10] and others.

Array-type structures implement these concepts directly. The difference among the array-type structures lies on extra structures used to reduce the computational cost to obtain the number of candidates keeping invariable the evaluation of distances.

Many indexes are trees and the children of each node define areas of space. Range queries traverse the tree, entering into all the children whose areas cannot be proved to be disjoint with the query region.

The increased size of databases and the emergence of new data types create the need to process a large volume of data. Then, new research topics appear such as efficient use of computational resources (storage and its hierarchy, processors, network, etc) that allows us to reduce the execution time and to save energy. In this sense, recent appearance of GPUs for general purpose computing platforms offers powerful parallel processing capabilities at a low price and energy cost. However, this kind of platforms has some constraints related to the memory hierarchy.

The present work analyses, by means of a set of experiments, the results obtained for several metric structures in order to obtain those attractive features [11] from the point of view of a future GPU-based implementation: selection of pivots and centres techniques, needed storage and simplicity of the data structure.

The paper is structured as follows. In Sect. 2 the metric structures considered in this paper are described. In Sect. 3 the features to be evaluated are presented. Section 4 outlines the experimental results and discussion. Finally, the conclusions and future work are commented in Sect. 5.

2 Metric Structures

The metric structures considered in this comparative study are:

Generic Metric Structure (GMS). This structure represents the most basic structure: it is an array-type structure based on pivots, which are obtained randomly. From this generic structure could be derived the rest of structures based on arrays and the choice of the pivots could be carried out according to *SSS-Index* or *MSD* methods. These pivot selection techniques will be introduced later.

Spaghettis [12]. It is an array-type structure based on pivots and does not assume any pivot selection method. However, each entry in the array, that represents distances

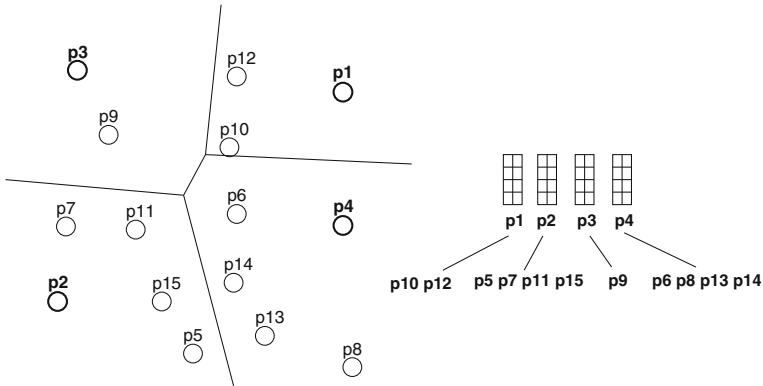


Fig. 1 Construction of *EGNAT* structure: data space and metric structure

between an element in the database and the pivots, is sorted with respect to this distance, obtaining a reduction on the execution time by means of a binary search. In this work, the array is sorted considering only the first pivot.

SSS-Index. *SSS-Index (Sparse Spatial Selection)* [10] is basically the generic structure varying the way in which the pivots are selected. The selection methods will be introduced later.

LAESA. Like *SSS-Index*, it is a structure similar to the generic one, but the selection of pivots is carried out by a method called *Maximum Sum of Distances (MSD)*.

EGNAT. *Evolutionary GNAT* [6] is a clustering tree-type structure derived from *GNAT* structure. This method pretends to exploit the secondary memory hierarchy (see Fig. 1). This structure is far from the array-type of the generic structure.

The choice of these metric structures is motivated because they are representative of this field of knowledge, and we have considered structures based on pivots and on clustering, array-type and tree-type.

With respect to the choice of pivot selection, we have considered the following:

Randomly. As the name suggests, this method consists in selecting randomly the set of pivots of the database.

Sparse Spatial Selection (SSS). *Sparse Spatial Selection* [10] is a method to select a dynamic set of pivots or centres distributed in the space. Let (\mathbb{X}, d) be a metric space, $\mathbb{U} \subset \mathbb{X}$ and M the largest distance between all pairs of objects, i.e. $M = \max\{d(x, y) / x, y \in \mathbb{U}\}$. Initially, the set of pivots contains the first element of the collection. After that, an element $x_i \in \mathbb{U}$, is selected as a pivot if and only if the distance between it and the rest of selected pivots is greater than or equal to $M * \alpha$, being α a constant whose optimum values are close to 0.4 [10] (see Fig. 2).

Maximum Sum of Distances (MSD). *MSD (Maximum Sum of Distances)* is used in *LAESA (Linear Approximating Search Algorithm)* [7, 13]. The underlying idea is

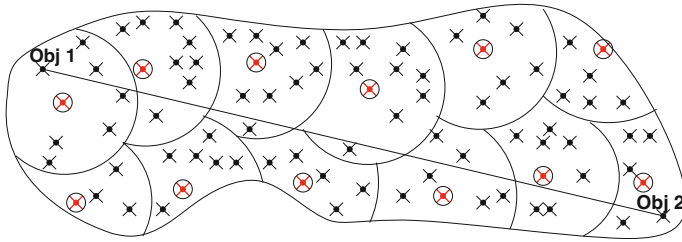


Fig. 2 Partition of the space using SSS methods

to select pivots considering that the distance between them is always the maximum. Starting with a base pivot arbitrarily selected, the distance between the objects and the selected pivot is calculated, and then the new base pivot to be selected is the one located to the maximum distance. The distances are added in a vector to calculate the next base pivot. This is an iterative process that ends when the required number of base pivots is obtained.

3 Metric Structures Features to be Evaluated

In the literature it is possible to find a wide range of metric structures for similarity searching [1, 14].

In this work a set of representative metric structures have been considered based on pivots, clustering, array-type or tree-type. We have considered this variety of structures in order to determine, experimentally, if the cost in the searching process compensates the complexity of the implementation, taking into account that the decision taken here will condition the future implementation on a GPU-based platform.

The relevant features considered in this work are:

Execution time. The execution time is a key factor in order to determine the best implementation. In the literature lot of papers are found talking about evaluation of distances [6, 10], but they do not consider execution time (floating point operations and I/O operations), memory accesses, etc.

Distance evaluations. In general, the reduction on evaluation of distances has been considered as the main goal of the new structures design, and evidently, it has a direct impact on the execution time. However, the high processing capacity of current computational platforms implies that distance evaluation is not always the operation with a higher computational cost. For instance, in GPU-based platforms, sorting operation affects to the execution time more than the evaluation of distances.

Storage requirements. A very interesting feature to evaluate is the memory needed to store a structure, even more if memory constraints are considered as is the case of GPU platforms. We have only addressed main memory, being secondary memory out

of the scope of this paper. The point is, “how much storage I am willing to sacrifice versus performance?”.

4 Experimental Evaluation

In this section, the case studies used as benchmarks and the testbed considered in this paper are described. Moreover, preliminary results are presented.

4.1 Case of Studies and Platform

We considered two datasets: a subset of the Spanish dictionary and a color histograms database, obtained from the Metric Spaces Library (see www.sisap.org). The Spanish dictionary we used is composed of 86,061 words. The edit distance was used. Given two words, this distance is defined as the minimum number of insertions, deletions or substitutions of characters needed to make one of the words equal to the other. The second space is a color histogram. It is a set of 112,682 color histograms (112-dimensional vectors) from an image database. Any quadratic form can be used as a distance, so we chose Euclidean distance as the simplest meaningful alternative.

The results presented in this section belong to a set of experiments with the following features:

- For all data structures considered in this paper, a set of tests were carried out using pivots from 1 to 1362 for word space and from 1 to 244 for color histograms (see Fig. 3). From all the results, only the best results have been plotting.
- For word space, each experiment has 8,606 queries over a *Spaghettis* with 77,455 objects. For vector space, we have used a dataset of 101,414 objects and 11,268 queries.

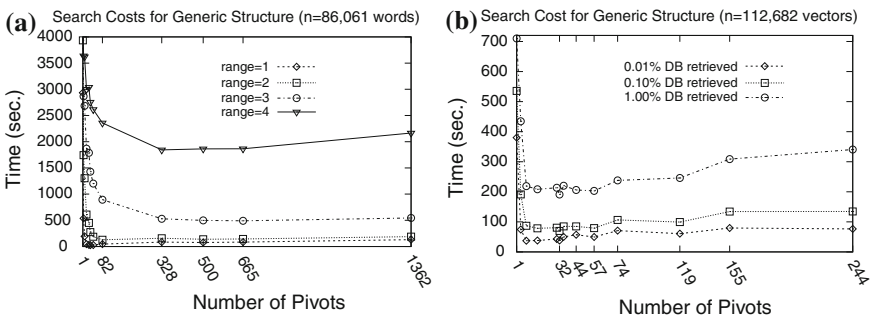


Fig. 3 Execution time for the implementation considering a generic metric structure (GMS). **a** General result for Spanish dictionary **b** General result for color histograms

- For each query, a range search between 1 and 4 was considered for the first space, and for vectors space we have chosen ranges which allow to retrieve 0.01, 0.1 and 1 % from the dataset.

We have chosen this experimental environment because is the most usual environment to evaluate this kind of algorithms. Also, these datasets are representative of discrete and continuous searching, respectively.

The hardware platform used is called Marte and belongs to the Albecete Research Institute of Informatics (I3A: <http://www.i3a.uclm.es>). Marte is a 2 Quadcore Xeon E5530 at 2.4GHz and 48GB of main memory, and Ubuntu 9.10 (64 bits) Linux Operating System. The compilation has been done using gcc 4.3.4 compiler.

4.2 Experimental Results and Discussion

Although results are usually shown considering the search ranges in the X axis, in this paper we have considered a different approach. In order to compare the behaviour of different pivot-based structures, in our opinion, it is more interesting to show the results against the number of pivots, typically 4, 8, 16 and 32, but also 1 and a number of pivots bigger than 32, especially when we need to compare with *SSS-Index*. This structure does not allow to choose the number of pivots (they are calculated depending on several parameters such as the value of α and the kind of search space) and usually uses a big number of pivots.

Figure 3 shows an overview of the behaviour of the generic structure based on pivots for both datasets.

Usually in metric structures, the performance of a structure increases with the number of pivots. Nevertheless, as can be seen in Fig. 3 for the generic structure, the performance increases till a point (that depends on the range considered, e.g. 32 pivots for range 1) from where the performance remains the same or decreases. This behaviour is common to all the structures as shown in Fig. 4. In this figure only the results close to the best one are shown.

Notice that when using the *SSS* structure we cannot select a priori an exact number of pivots. This is the reason why the minimum number of pivots is 44 (for word space) and 35 (for color histograms). For word space, as the distance is discrete, there are not values between 328 and 665, so the value 500 does not exist in *SSS-Index*. The value 500 neither is shown in *EGNAT* because the needed structure is bigger than the RAM memory, swapping is needed and consequently performance is poor.

Analysing the results in Fig. 4 we can conclude that for small ranges *Spaghettis* has the overall best performance considering both datasets. The reason is that the use of binary search allows a quick search of the first element in the database inside the range. *GMS* is very close to *Spaghettis* performance, and the other 3 structures have a bad behaviour in one of the two datasets, color histograms in *MSD* and *SSS*, and the Spanish dictionary in *EGNAT*. Nevertheless when we consider bigger search ranges, the advantage of *Spaghettis* is lost; in this case the price in time of the binary search is not worthy because less elements are discarded.

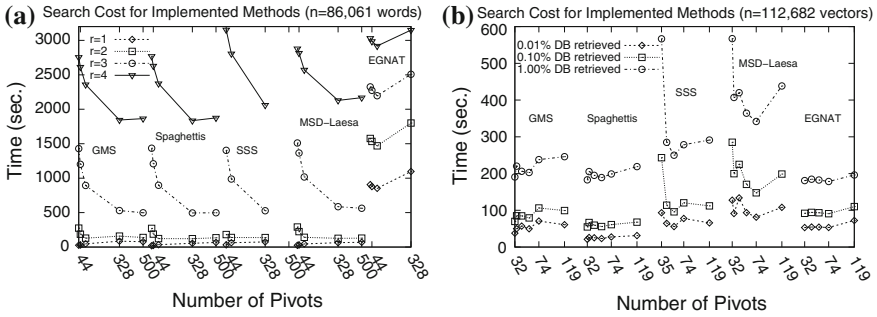


Fig. 4 Execution time for the implementation considering all structures. **a** General result for Spanish dictionary. **b** General result for color histograms

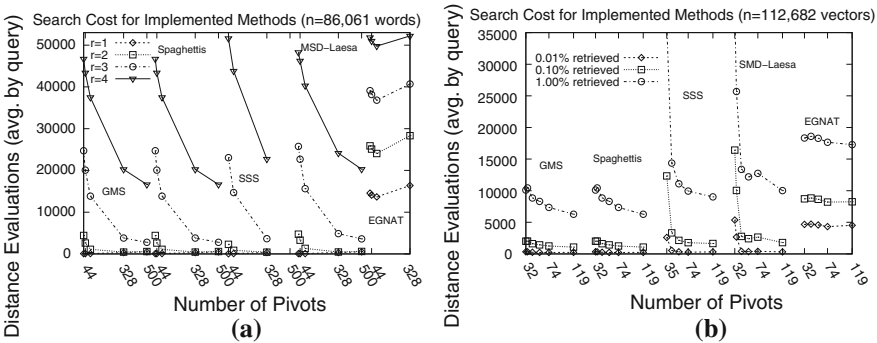


Fig. 5 Distance evaluations for the implementation considering all structures. **a** General result for Spanish dictionary. **b** General result for color histograms

Figure 5 shows the same scenario for distance evaluations. The number of evaluations decreases when the number of pivots increases. This means that, comparing Figs. 4 and 5, at some point is not worthy to increase the number of pivots because the time consumed in their management is bigger than the time consumed in the distance evaluations we save. We can also conclude that using more pivots is better for big ranges, and it has little influence for small ranges.

Tables 1 and 2 show, in detail, the execution time (in seconds) of the best cases depending on the range or on the data retrieved percentage, respectively. In these tables several modifications of the generic structure are considered. In these modifications the pivots were not selected randomly but following the pivots selection methods used by the other structures. Thus, first we get a subset of pivots from the database randomly or using *SSS* and then *MSD* is applied to get the number of pivots for the best performance case (32 or 44 depending on the range). Only modifications of the structure with a good performance are considered in the tables (e.g. “MSD x on *SSS* y ” cases are not included in color histograms because they have a poor performance).

The results obtained for the modified generic structures are good. For small ranges *Spaghettis* is still better, but when the range increases the new structures have better

Table 1 Execution time for the best methods on word space (column: range; row: data structure)

Index	1	2	3	4
Spaghettis 32	18.24	270.39	1434.78	2769.06
MSD 32 on GMS 665	25.32	255.37	1453.01	2783.26
MSD 32 on GMS 1362	25.78	251.16	1436.40	2802.90
MSD-Laesa 32	25.84	291.08	1510.68	2879.07
GMS 32	26.18	274.37	1428.74	2754.02
MSD 32 on SSS 665	27.61	249.92	1489.16	2953.54
MSD 44 on random 1362	27.79	168.91	1200.94	2647.71
MSD 32 on SSS 1362	27.86	269.72	1506.16	2910.85
SSS-Index 44 ($\alpha = 0.55$)	31.93	180.91	1404.33	3153.15

Table 2 Execution time for the best methods on color histograms (column: data retrieved percentage; row: data structure)

Index	0.01	0.1	1.0
Spaghettis 32	21.00	54.90	182.87
GMS 32	37.77	69.89	190.74
MSD 32 on GMS 119	39.28	71.85	190.26
MSD 32 on GMS 1014	46.55	96.10	246.91
EGNAT 32	53.01	91.86	180.89
SSS-Index 57 ($\alpha = 0.6$)	55.77	95.91	249.85
MSD-Laesa 35	91.33	199.75	406.99

performance. The advantage of using *MSD* over a big number of pivots randomly chosen is that it allows to choose the best pivots and the exact number of pivots desired and, consequently, it allows to determine the size of the structure which is an important factor to consider when we need to fit the structure in a virtual page or in GPU memory.

Looking forward to the GPU implementation, the size of the structure is a very important factor. A structure that does not fit into GPU memory will not have a good performance. Figure 6 shows that *EGNAT* structure is much bigger than pivot-based structures. As expected, the size of the structure in pivot-based structures is directly proportional to the number of pivots. In order to have a more detailed view, the bigger values were removed from the table (e.g. in color histograms *EGNAT* with 119 centres needs 2 GBytes, in word space *EGNAT* with 328 centres needs 6 GBytes).

Tree-type structures have a good performance when the radius increases, and they are very stable with respect to the number of pivots or centres. This means that we can get a good performance even selecting a small number of centres. The problem with this kind of structure is that when a new node in the tree is created, there is no guarantee that it will be completed, leading to a situation in which the size of the structure can grow a lot depending on how objects are distributed in subtrees. In the tree-based structure used in this paper we obtained that less than 20% of the nodes were completed.

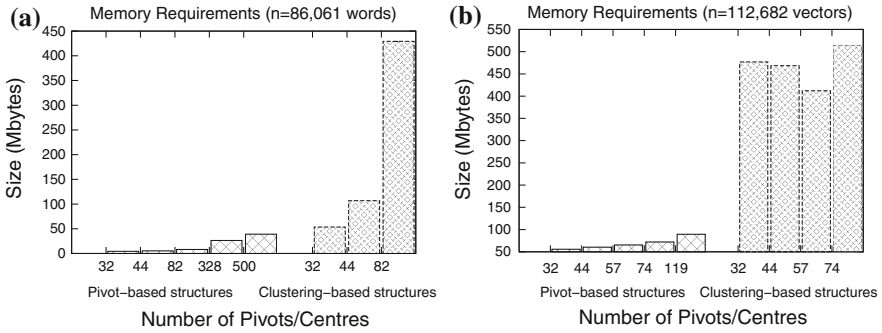


Fig. 6 Memory requirements for pivot-based structures (GMS, SSS, MSD, Spaghettis) and clustering-based structures (EGNAT). **a** General result for Spanish dictionary. **b** General result for color histograms

5 Conclusions and Future Works

In this work a comparative study of different metric structures has been carried out.

Different types of metric structures and pivot selection methods have been considered in order to make a good comparison. The comparison has been made according to three criteria: execution time, evaluation of distances and storage requirements.

According to the experimental results, it is not possible to select a metric structure as the best one, because it depends on the space distribution of the database. Three structures are candidates to be eligible as the best: *Spaghettis*, *Generic structure + MSD* and *EGNAT*. However from the point of view of a future GPU implementation the best one is *Generic + MSD* due to:

1. By using a generic structure it is not necessary to apply a binary search like *Spaghettis*. Binary search operation is very expensive in a GPU-based platform in comparison with the evaluation of distances.
2. Using a generic structure the storage requirements are lower than using a *EGNAT* structure.
3. Thanks to the combination of generic structure and *MSD* pivot selection, it is possible to reduce the number of pivots till satisfying the memory constraints inherent to the GPU-based platforms.

To sum up, using the generic structure we will take benefits in terms of execution time, storage and, in addition, the code is more simple.

As we said in the introduction, the work presented in this paper allows us to choose the best option from the point of view of a parallel implementation of the similarity search method based on metric structures on a GPU-based platform, representing that the future work.

Acknowledgments This work has been supported by the Ministerio de Ciencia e Innovación, project SATSIM (Ref: CGL2010-20787-C02-02), Spain and Research Center, University of Magallanes, Chile. Also, this work has been partially supported by CAPAP-H3 Network (TIN2010-12011-E).

References

1. Chávez E, Navarro G, Baeza-Yates R, Marroquín JL (2001) Searching in metric spaces. *ACM Comput Surv* 33(3):273–321
2. Kalantari I, McDonald G (1983) A data structure and an algorithm for the nearest point problem. *IEEE Trans Software Eng* 9(5):631–634
3. Uhlmann J (1991) Satisfying general proximity/similarity queries with metric trees. *Inf Process Lett* 40:175–179
4. Ciaccia P, Patella M, Zezula P (1997) M-tree : an efficient access method for similarity search in metric spaces. In: *Proceedings of 23rd international conference on VLDB*, 426–435
5. Brin S (1995) Near neighbor search in large metric spaces. In: *Proceedings of the 21st VLDB conference*, Morgan Kaufmann Publishers, 574–584, 1995
6. Navarro G, Uribe-Paredes R (2011) Fully dynamic metric access methods based on hyperplane partitioning. *Inf Syst* 36(4):734–747
7. Micó L, Oncina J, Vidal E (1994) A new version of the nearest-neighbor approximating and eliminating search (aesa) with linear preprocessing-time and memory requirements. *Pattern Recogn Lett* 15:9–17
8. Baeza-Yates R, Cunto W, Manber U, Wu S (1994) Proximity matching using fixed queries trees. In: *5th Combinatorial Pattern Matching (CPM'94)*. LNCS 807:198–212
9. Chávez E, Marroquín J, Navarro G (2001a) Fixed queries array: a fast and economical data structure for proximity searching. *Multimedia Tools Appl* 14(2):113–135
10. Pedreira O, Brisaboa NR (2007) Spatial selection of sparse pivots for similarity search in metric spaces. In: *Proceedings of the 33rd conference on current trends in theory and practice of computer science (SOFSEM (2007) LNCS, vol 4362*. Czech Republic, Springer, Harrachov, pp 434–445
11. Uribe-Paredes R, Cazorla D, Sánchez JL, Arias E (2012) A comparative study of different metric structures: thinking on gpu implementations. In: *Proceedings of the world congress on engineering (2012) WCE 2012*. Lecture notes in engineering and computer science, England, London, pp 312–317
12. Chávez E, Marroquín J, Baeza-Yates R (1999) Spaghettis: an array based algorithm for similarity queries in metri spaces. In: *Proceedings of 6th international symposium on String Processing and Information Retrieval (SPIRE'99)*, IEEE CS Press, pp 38–46
13. Micó L, Oncina J, Carrasco R (1996) A fast branch and bound nearest neighbor classifier in metric spaces. *Pattern Recogn Lett* 17:731–739
14. Hetland M (2009) The basic principles of metric indexing. In: Coello CA, Dehuri S, Ghosh, S (eds) *Swarm intelligence for multi-objective problems in data mining*, Studies in computational intelligence, vol 242. Springer Berlin, pp 199–232

Exploratory Analysis of Ergonomics Importance at Workplace and Safety Culture Amongst Occupational Safety and Health Practitioners

Md Sirat Rozlina, Mohamed Shaharoun Awaluddin,
Syed Hassan Syed Abdul Hamid and Zakuan Norhayati

Abstract This paper is a revised and extended version of a paper that was presented at WCE 2012. The article reports on a study to identify key components which can be used to relate ergonomics awareness and safety culture. These components can be used to facilitate the research which is aimed at determining the elements that influence the ergonomics awareness and the relationship with safety culture in an organization. A survey was done using a sample of 108 OSH practitioners in manufacturing companies in Malaysia. Exploratory Factor Analysis was used to determine the importance of ergonomics at their workplace and their beliefs on the importance of safety culture to be inculcated at their companies. 3 factors for ergonomics importance were identified: (i) Implication of & Need for improvement, (ii) Fitting the job to the workers and (iii) Basic ergonomics consideration. Safety culture questions were focused on the OSH practitioners perceptions on safety climate importance. Three constructs were designed: (i) commitment and leadership, (ii) motivation and (iii) safety management system practice. This finding is significant in order to study the influence of the perceptions of OSH practitioners on ergonomics importance at workplace to the safety culture.

M. S. Rozlina (✉) · M. S. Awaluddin
UTM Razak School of Engineering and Advanced Technology, Universiti Teknologi Malaysia,
International campus, Block H, Level 1, Jalan Semarak, 54100 Kuala Lumpur, Malaysia
e-mail: rozlina@mail.fkm.utm.my

M. S. Awaluddin
e-mail: awaludin@ic.utm.my

S. H. S. A. Hamid
Department of Occupational Safety and Health (DOSH), 2nd Level,
Block D3, Complex D, 62530 Putrajaya, Malaysia
e-mail: sabdulhamid@mohr.gov.my

Z. Norhayati
Faculty of Management and Human Resource, Universiti Teknologi Malaysia,
UTM, 81310 Skudai, Johor, Malaysia
e-mail: norhayatimz@utm.my

Keywords Commitment · Ergonomics awareness · Exploratory factor analysis (EFA) · Leadership · Motivation · Safety culture · Occupational safety and health practitioners

1 Introduction

Safety culture is defined as ‘a set of value, perceptions and attitudes and patterns of behavior [1–4]. Ergonomics is a scientific discipline concerning with the understanding of interactions among humans and other elements of a system and it will contribute to job satisfaction [5]. In safety, all situations must be in ergonomics compliance such as equipment, control panel and Personal Protective Equipment (PPE) (e.g. helmet, goggle, belt, shoes). Thus, ergonomics awareness is very important and it is prove to have substantial impact on the industry, organization, management, employees and overall well-being of the system [6]. Ergonomics awareness helps in ergonomics application and contributes significantly to human well-being and safety.

2 Conceptual Background

Even though ergonomics is listed under safety and health, they are actually in two different issues.

For example, safety hazard is easy to identify based on the moving machine, breaking of any ropes, height while ergonomics hazard is difficult to be identified. It is based on work methods such as repetitive movement, heavy lifting, awkward posture [7]. It needs high awareness to put ergonomics risk as priority [8]. The effect of safety is shown as acute effect while ergonomics is chronic effect—takes time to get the effect [7]. Many ergonomically related injuries can progress to long-term or may give permanent disabilities if not taken care of properly. The ergonomics risk may appear after retirement [8]. The effect of the safety hazard can be seen in forms of injury such as wounds, cuts, burns while ergonomics hazard can be seen in forms of Musculoskeletal Disorders (MSDs) and Carpal Tunnel Syndrome (CTS) [8]. The injuries is primary injuries based on Heinrich (1941) while ergonomics is secondary injuries. For example, overexertion often results in muscular or thermal exhaustion; but lifting too much weight can tear tendons from the bone, and introduce bone fragments into soft tissue [9]. Once damaged, tendons and ligaments heal slowly. This type of secondary injuries would not have happened without the original injury [9]. Medical certificate (MC) and compensation will be given by company when safety accident occur but for ergonomics there are trend of, absenteeism and MC that can show the occupational ergonomics risk [8]. For the regulations, safety has clear regulations while ergonomics is lack of specific standard, for example, for awkward position situation depends on the capability of the company’s Safety and health practitioner which refer to OSHA Sec 15 [10], general duties of employers: to ensure so far as is practicable, the safety, health and welfare at work of all his employees.

The study of differences between safety and ergonomics may emphasize why OSH practitioners do not aware the immediate importance of ergonomics. Hence, ergonomics awareness is important amongst OSH practitioners to ensure the implementation of ergonomics. The study on ergonomics awareness was determined through some literature review and discussions with experts. There are a variety of methods that have been used to assess ergonomics awareness. Unfortunately, no standard measurement available that can be used to build a construct. Some issues that OSH practitioners need to be aware are the extent of implications of ergonomics, the extent of suitability of jobs to the workers, equipment used [11], workspace and workplace design, assessment tools and administrative awareness [12].

The awareness of implication of ergonomics can be measured by the extent of their beliefs on ergonomics effect to the workers such as the implication of high force towards workers [11, 13], repetitive motions [11, 14], workspace [11], and long term exposure [12, 13]. The ergonomics awareness also can be measured by looking at the suitability of jobs regarding the type of tasks according to standards suitable for ergonomics [10, 11, 15, 16].

Equipment specification suitable to workers, hand jigs and fixtures [11], workspace and workplace design is important to be evaluated as this can determine the ergonomics awareness technically [12]. Based on improvement awareness, the OSH practitioners need to determine the improvement methods, work study using time measurement to decide their capability of doing work and continuous improvement [14, 16].

Safety culture is a critical factor in implementation of safety and health and it is believed to give a positive impact to the companies [1, 3, 17–21] such as increase productivity and profit by reducing rejects, cost and reducing stress. It can be achieved through comfortable work environment, designed tools, man-machine interface and work method.

The purpose of inculcating a safety culture is to develop a nature of safe work consistently and guided by a well-defined set of core values that protect and promote the health and well-being of the individual and the environment [4]. Safety culture require a development of individual safety beliefs, attitudes and behaviors [22] and it must be initiated by OSH practitioners at the workplace.

The strength of safety culture can be measured by looking at some of the attributes such as commitment, leadership, communication, employee involvement, motivation and safety practice.

Commitment of top management to OSH is vital in all undertakings to reduce injuries, diseases, fatalities in order to improve efficiency, productivity and business performance [3, 19–21, 23]. The commitment by top management including OSH practitioners could be shown through the development of policy, approval of budget, time allocation to attend meeting, observation and direct advice on OSH to workers.

Leadership is one of the key success of safety culture and it can be demonstrated together with management commitment [24]. The activities does not have a great difference with commitment. This involves the top management commitment in providing resources, motivation, priorities and accountabilities [25], developing communication skill, listening, mentoring, facilitating and observational skill [23].

All can be achieved through suitable training for employees and top management in order to enhance the safety awareness and ergonomics awareness [26–28].

Motivation can influence behavior towards safety [1, 20, 24, 29]. Rewards such as financial or recognition reward can be a motivation to the workers. Employee involvement can be an approach to motivation. It will develop feelings of self-worth, belonging and values by involving them in training, consultation about PPE and job rotation [3, 19, 21, 24, 30]. Employee empowerment allows them the freedom of power to suggest and implement good practice at work place [3, 23].

Safety practice is about how to manage safety in a systematic way [3, 10, 18]. It covers and emphasizes the formalization of safety policy, formulation of safety procedure, describing how safety problems are identified, investigated, assessed, controlled and implemented [23].

3 Problem Identification

The function of managing safety is usually assigned to a person in charge (Sect. 29 [10]) namely Safety and Health Officer (SHO). In Malaysia, such people may also be designated post such as Safety and Health and Environment Officer (SHE), Health, Safety, Environment and Security Officer (SHES), safety engineer and the like. For the research purpose, the terminology of Safety and Health Practitioner (OSH practitioner) will be used in a broader context. They are well-trained to manage the risk, and proactively intervening in unsafe situations [26].

The responsibility of OSH practitioners is very high as the employer give the authority to OSH practitioners to ensure the highest safety and health standards at the workplace and he/she constantly interacts the employer regarding acts and regulations (regulation 18, safety and health regulation under [10]). They are also representatives of the companies to initiate any activities and steps to be taken including to advise the employer in any matter related to safety and health.

However, the role of OSHA regarding ergonomics has been ill-defined. In Malaysia, there is no specific act, regulations or guideline available to explain ergonomics implementation in general, unlike safety issues [31]. Ergonomics is important at least in theory but its actual awareness among Malaysian OSH practitioners has not been investigated. Human factors or ergonomics is believed to play a fundamental role in organization health and safety performance and this indirectly is also associated with safety culture.

The objective of the study is to investigate the extent of ergonomics awareness and its influence in inculcating safety culture amongst OSH practitioners. Ergonomics awareness is measured by their attitude in determining ergonomics importance at workplace. Basic knowledge is not studied in this paper as it is well informed that their basic knowledge on ergonomics awareness is adequate among the OSH practitioners.

4 Methodology

Some researchers in psychology such as [32, 33] suggested that attitudes included three components: cognitive, affective and conative (behavioral). Chang and Liao [34] summarized the three components whereby the cognitive represents the beliefs or idea associated with a particular subject. The affective component is the individual's evaluation of the object and emotion associated with the object. The conative illustrates the action or intention toward action directed at the object. Shafel and Shafel [35] concluded that attitude also affects behavioural intentions, which represent 'a plan of action that is arrived at through conscious, deliberately processing'. Davidson et al. [36] found that 'intention was better predictors of behavior'. Chang and Liao [34] called it as behavior intention and used this methodology in their research to measure attitude of their case study object in the aviation field. For this paper, the authors developed the question on the basis of cognitive components, representing the beliefs of respondents. It is used to measure attitudes of OSH practitioners on the importance of ergonomics on some issues.

A seven-point likert scale was employed to the both questions of ergonomics importance at workplace and safety culture to respond to those items. (1 = not relevant, 2 = not important at all, to 7 = critical).

4.1 Procedure of Collecting Data

250 mails were delivered to OSH practitioners in manufacturing industries and 108 completed replied were received. This number of response is considered adequate as the trend is similar in other parts of the world, even in developed nations [37, 38].

4.2 Demographic Data

SHO were asked on their position, level of education, year of work experience gained in company or other companies, year of work experience as OSH practitioners and training obtained for past three years. Respondents in companies include those in electrical and electronic (27.8%), chemical or apart (15.7%), metal, machines and equipment (13%), rubber or plastic based (12%), automotive and accessories (7.4%), wooden product including furniture (4.6%), printing and publishing (2.8%), paper and paper based (0.9%), textile and leather (0.9%) and others (food manufacturing, medical products)(14.8%). Education levels were in the following categories: SPM (11.1%), Diploma (28.7%), Degree (47.1%) and Post Graduate degree (13%). Most of them were called Safety and Health Officer (SHO) (50.9%), Safety, Health and Environment Officer (SHE) (38.9%), Health and Safety, Environment and Security Officer (SHES) (4.6%), engineer (2.8%), and others (safety and health executive,

safety and environment affairs manager, and ergonomist) (2.8%). Based on their work experiences in company/ companies, most of them have 16–25 year experience (41.7%) and more than 25 years (21.3%). The others were 0–5 year (19.4%) and 6–15 years (17.6%).

4.3 Content Validity

In this study, all the measurement items were developed and constructed based on literature review and validated by relevant representatives from NIOSH, academicians, DOSH and companies. This is important to determine that the items represent the domain of the construct.

4.4 Exploratory Factor Analysis

EFA is used to identify how many latent variables underlie the complete set of items and reducing those items to a smaller, more manageable set of underlying factors [34]. The presence of meaningful patterns among 29 ergonomics beliefs on importance at the workplace items and simplified the importance contained in a small set of factors or dimensions. The EFA can be used when researchers have measurements on collection of variables and would like to have some idea about what construct might be used to explain the inter-correlation among these variables [39].

The questions of ergonomics importance at workplace were verified and modified from the work done as mentioned in conceptual background.

EFA was done on the 29 items of ergonomics. The Overall-Keiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis. After deleting items which has low factor loading and reliability, 20 items were identified to be appropriate for further analysis. The KMO for ergonomics importance was 0.92 (superb according to Field, 2000) with factor loading values ranging from 0.58 to 0.82. The Bartlett Test of sphericity reached statistical significance with $\chi^2(108) = 1644.21$, $p < 0.0001$ indicating that the correlation between the items were sufficiently large for Principle Component Analysis (PCA). The three factors solution explained a total of 57.35% of the variance, with factor 1 contributing 54.58%, factor 2 contributing 8.06% and factor 3 contributing 6.66%. The reliability analysis, measured by Cronbach alpha α values ranged from 0.80 to 0.93 and were considered as having high internal consistency for three-factor safety culture. Factor analysis, percent of variance and Cronbach alpha value can be seen in Table 1. These 20 items with three new factors namely as: (1) ‘implication of and need for improvement’ (10 items), (2) ‘fitting the jobs to workers’ (7 items) and (3) ‘basic ergonomics considerations’ (3 items).

Safety culture variables were derived and modified from previous work done by [3, 17, 19, 40] and some literature relating to the field of safety culture and safety management [1, 17, 21, 24, 27, 30]. Altogether 22 items were developed and analysed

Table 1 Factor analysis of ergonomics importance at workplace

Factor and items	Factor loading	% of variance	Cronbach alpha (α)
Implication of and need for improvement		54.58	0.926
1 ..high force against time (example: involving high force within 30 min continuously or more than 2 h within 8 work hours-lifting goods at warehouses)	0.81		
2 .. repetitive movement (example: involving repetitive job with 2 times in a minute at one time,—assembly work at workplace , using spanner repetitively as a tool in long duration)	0.78		
3 .. improvements based on ergonomics analysis (example: RULA, REBA, OCRA)	0.77		
4 .. effect of work on workers (example: duration, shift)	0.76		
5 .. work study considering for allowances in time measurement for a task (example: allowance for emergency, going to toilet, doing other works, administrative work)	0.72		
6 .. continuous improvements (example: always include ergonomics issue in safety meeting agenda)	0.67		
7 .. suitable number of workers for each production line (example: give ergonomics consideration in terms of workers psychological effect)	0.67		
8 .. the importance of work space provision (example: location of control switch and suitable workspace and workplace for the workers who has short hand or leg)	0.66		
9 .. the importance of improving long term exposure to unergonomics workplace design (example: protection from hard surface through suitable foot wear and anti fatigue mats)	0.57		
Fitting the job to workers		8.06	0.915
1 .. according to age, suitability or health condition	0.77		
2 ..improvement based on common sense.	0.78		
3 ..improvement based on standards (example: guideline by DOSH, ISO or ILO references)	0.72		
4 .. specification of equipment suitable to workers (adjustable equipment; size of chair, width of seat)	0.69		
5 .. checking the suitability of equipment for a given task (example: machine that use one hand or two hand operation)	0.66		
6 .. hand tools to handle work piece such as jigs and fixtures.	0.63		
7 .. the guidelines for ergonomically designed seating and furniture	0.59		
Basic ergonomics considerations		6.65	0.804
1 ..anthropometric data in purchasing equipment (example: purchasing chair for office work)	0.82		
2 .. anthropometric data in workspace design (example: work piece is arranged according to importance or the primer, secunder and tertier access zone, workers can move comfort in workspace)	0.81		
3 .. anthropometric data in layout design (example: seating work that involving sequences, need the work layout to be arranged in semi-circle towards worker)	0.60		

Total variance 65.636%, KMO = 0.919, Bartlett test $\chi^2 = 1644.205$, df = 210, significance level (p) = 0.0001

using SPSS. The KMO for safety culture was 0.92. The Bartlett Test of sphericity with $\chi^2(108) = 1447.59$, $p < 0.0001$. The three factors solution explained a total of 57.35% of the variance, with factor 1 contributing 57.352%, factor 2 contributing 8.24% and factor 3 contributing 6.11%. The reliability analysis, measured by cronbach alpha (α) values ranged from 0.92 to 0.94 and were considered as having internal consistency for three—factor safety culture. Factor analysis, percent of variance and Cronbach alpha value can be seen in Table 2. After EFA, the items become 17 items with the three new factors namely as: (1) Commitment and leadership (7 items), (2) Motivation (6 items), and (3) Safety Management System Practice (4 items).

5 Discussions

EFA is used to identify suitable factors or dimensions for the beliefs on ergonomics importance at the workplace towards safety culture. Based on the final results on empirical study, three crucial factors relating to awareness of ergonomics importance at workplace were identified: (1) Implication of and Need for Improvement, (2) Fitting The Job To Workers and (3) Basic Ergonomics Considerations.

5.1 Ergonomics Awareness Factor

Implication of and Need for Improvement is important as it needs employer to be aware on implications of not being aware of the ergonomics risk and mentioned briefly in regulation 18 (Duties of Safety and Health Officers Regulation 1997) [10] and Regulation 11 (Functions of Safety and Health Committee) under OSHA 1994 [10, 13] to inspect any machinery, plant, equipment, or any manual work that may cause injuries and to review the effectiveness of safety and health programs.

Fitting the job to the workers or other word is to 'fit the job to the man' and is the guiding philosophy of ergonomics because it is about human engineering and workspace design relating to the design tasks to suit the characteristics of workers. It is the underlying assumptions that can be specified around which the job can be designed for any jobs [11].

Basic ergonomics considerations are some issues of awareness that emphasized the importance of ergonomics related to the physical or namely as anthropometric data [11, 15, 41] such as consideration of equipment design suited to the workers while purchasing equipment mentioned in 15 (2)(b) OSHA 1994 [10], layout design and workspace design under regulation 20 and regulation 24 (Safety, Health and Welfare Regulation 1970) under FMA 1967 [31].

Table 2 Factor analysis and reliability analysis on safety culture

Factor and items	Factor loading	% of variance	Cronbach alpha (α)
Commitment and leadership		57.35	0.91
1 Developing teamwork spirit	0.79		
2 Top management approved the use of new technology for generating an ergonomics environment	0.77		
3 New employee is instilled with the importance of ergonomics in the workplace	0.71		
4 Give suitable rewards to workers who give suggestions on safety and health improvement	0.71		
5 Ensure employees are both involved and empowered	0.66		
6 Analysis and ergonomics improvements assisted by consultation	0.64		
7 Give enough knowledge (training) to the safety and health practitioner in the organization.	0.63		
Motivation		8.24	0.92
1 Meeting periodically held between managers and workers to take decisions affecting organization of work	0.86		
2 Employees view safety and health (including ergonomics) as the natural, normal and acceptable way of doing things	0.77		
3 Top management provide financial support for ergonomics issue	0.75		
4 Incentive offered to workers to suggest improvement in working conditions	0.71		
5 All organization level changed to ergonomics behavior	0.69		
6 Incentive offered to workers to put in practice and procedures of action	0.58		
Safety management system practice		6.11	0.87
1 Organization levels comment on each other on safety and health issue to identify corrective action	0.83		
2 Safety policy contains commitment to continuous improvement, attempting to improve objective already achieved	0.80		
3 Safety and health policy (including ergonomics) is co-ordinated with HR policies	0.75		
4 Standards of action or work procedures elaborated on basis of risk evaluation	0.69		

Total variance = 71.70, KMO = 0.915, *Bartlett Test* $\chi^2 = 1447.59$, df = 136, significance level (p) = 0.0001

5.2 Safety Culture Factor

For the safety culture construct, the elements identified in this chapter included: (1) commitment and leadership, (2) motivation and (3) safety management system practice.

Commitment and leadership covers employee involvement [3, 19] and commitment by top management [3, 19–21, 30], leadership [26–28], subsequently would give an impact to employee empowerment [21]. It is also covered attending OSH committee chair, supporting for the development and implementation of safety programs by physical and spiritual, approving financial and technology used [42] in order to get the employee to be involved and empowered in safety activities [29, 34, 42]. Leadership aspect includes the way top management control the safe operating procedure (SOP), show the safe way to do task, listen and communicate actively with members of team.

Motivation part is emphasized by job satisfaction [6, 28] by encouragement of practicing what they obtained in training [29]. Safety culture can be successful if top management appreciate the employees and give incentives for the safe behavior [18, 24, 29], which in turns the workers will feel free to discuss, openly, without barrier on safety programs, risk or any matter related to safety and health.

Safety management system is one of the factor that can develop safety culture [1, 17–20, 24, 27] which is measured by policy, procedures, financial budget, continuous improvement [17, 19, 28].

However, it is possible to confirm these construct model of Ergonomics Awareness to Safety Culture by Confirmatory Factor Analysis (CFA). AMOS will be used to confirm the Exploratory Factor Analysis (EFA).

6 Conclusions

The empirical study has identified three principal elements on ergonomics awareness that will have significant impact on safety culture measured by using three measurement variables. The finding of this study has contributed theoretically to a growing body of knowledge on ergonomics awareness amongst OSH practitioners in terms of their beliefs on ergonomics importance towards safety culture. This paper suggests the constructs of ergonomics awareness to safety culture and evaluated by EFA. To some extent, further work will be needed to confirm the theoretical model through Confirmatory Factor Analysis (CFA). This paper is a revised and extended version of a paper that was presented at WCE 2012 [31].

References

1. Ahasan R, Imbeau D (2003) Who belongs to ergonomics? *Work Study* 52(3):123–128
2. Pearson JC, Nelson PE, Titsworth S, Harter L (2000) *Introduction to human communication*, 8th edn. Mc Graw Hill Companies Inc., Boston
3. Muniz BF, Montes-Peon JM, Vasquez- Ordas CJ (2007) Safety management system: development and validation of a multidimensional scale. *J Loss Prev Process Ind* 20:52–68
4. DOSH, Occupational Safety and Health Master Plan 2015 (OSH-MP 15) (2010) Ministry of Human Resources, Malaysia. <http://www.dosh.gov.my/>

5. Vink P, Emada AS, Zink KJ (2008) Defining stakeholder involvement in participatory design processes. *Appl Ergonomics* 39:519–526
6. Gungor C (2009) A human factors and ergonomics awareness survey of professional personnel in the American furniture industry, Msc. Thesis, Mississippi State University, US
7. Punnet L, Cherniack M, Henning R, Morse T, Faghri P (2009) A conceptual framework for integrating workplace health promotion and occupational ergonomics programs. *Public Health Reports*, no. 124, pp 16–25
8. Cooper MD, Philips RA (2004) Exploratory analysis of the safety climate and safety behaviour relationship. *J Saf Res* 35:497–512
9. Sullivan JM, Ward LC, Tripp D, French JD, Adams H, Stanish DW (2005) Secondary prevention of work disability: community-based psychosocial intervention for musculoskeletal disorder. *J Occup Rehabil* 15(3):377–393
10. OSHA, Occupational Safety and Health Act 1994 (Act 514) & Regulations and Orders. International Law Book Services (ILBS), Kuala Lumpur, 2011
11. Bridger RS (2009) Introduction to ergonomics, 3rd edn. CRC Press, Boca Raton
12. Stranks J (2007) Human factors and behavioural safety. Butterworth-Heinemann, Burlington
13. Coluci MZ, Alexandre NM, Rosecrance J (2009) Reliability and validity of an ergonomics-related job factors questionnaire. *Int J Ind Ergonomics* 39:995–1001
14. Helander M (2006) A guide to human factors and ergonomics. CRC Press, Florida
15. Karwowski W (2006) Handbook on standards and guidelines in ergonomics and human factors. CRC Press, New York
16. Tayyari F, Smith JL (2003) Occupational ergonomics: principle and applications. Kluwer Academic Publishers, Massachusetts
17. Cooper M (2000) Towards a model of safety culture. *Saf Sci* 36:111–136
18. Gill GK, Shergill GS (2004) Perceptions of safety management and safety culture in the aviation industry in New Zealand. *J Air Transp Manag* 10:233–239
19. Choudhry RM, Fang D, Mohamed S (2007) The nature of safety culture: a survey of the state of the art. *Saf Sci* 45:993–1012
20. Guldenmund FW (2007) The use of questionnaire in safety culture research—an evaluation. *Saf Sci* 45:723–743
21. Hsu SH, Lee CC, Wu MC, Takano K (2008) A cross-cultural study of organizational factors on safety: Japanese vs Taiwanese oil refinery plants. *Accid Anal Prev* 40:24–34
22. Zohar D (2000) A group level mode of safety culture: testing the effect of group climate on microaccidents in manufacturing job. *Appl Psychol* 85:587–596
23. Taylor JB (2010) Safety culture: assessing and changing the behaviour of organisations. Psychological and behavioural of risk. Gower Publishing Limited, Great Britain
24. Hoivik D, Tharaldsen JE, Baste V, Moen BE (2009) What is most important for safety climate: the company belonging or the local working environment?—a study from Norwegian offshore industry. *Saf Sci* 47:1324–1331
25. Petersan D (1998) Safety policy, leadership and culture, in encyclopaedia of occupational health and safety, Geneva, International Labour Office (ILO), pp 59.2–59.4
26. IOSH (2004) Promoting a positive safety culture, institutional of occupational safety and health technical guidance, August 2004
27. Fitzgerald M (2005) Safety performance improvement through culture change. Part B. *Process Saf Environ Prot* 83(B4):324–330
28. Cabrera DD, Fernaud EH, Diaz RI (2007) An evaluation of a new instrument to measure organizational safety culture values and practice. *Accid Anal Prev* 39:1202–1211
29. Mearns K, Kirwan B, Kennedy RJ (2009) Developing a safety culture measurement Toolkit (SCMT) for European ANSPs, Eighth USA/ Europe Air Traffic Management Research and Development Seminar (ATM 2009), pp 1–9
30. Erensal YC, Albayrak YE (2008) Transferring appropriate manufacturing technologies for developing countries. *J Manuf Technol Manag* 19(2):158–171

31. Rozlina MS, Awaluddin MS, Norhayati Z, Hamid SHSA (2012) Perceptions of ergonomics importance at workplace and safety culture amongst safety and health (SH) practitioners in Malaysia. Lecture notes in engineering and computer science: proceedings of the world congress on engineering WCE 2012, 4–6 July 2012. U.K , London, pp 372–376
32. Eagly AH, Chaiken S (1993) *The Psychology of Attitudes*. Harcourt Brace Jovanovich, Fort Worth
33. Wolf RGAM (1993) *Instrument development in the affective domain: measuring attitudes and values in corporate and school settings: second Edition (2nd edn)*. Kluwer Academic Publishers, Boston
34. Chang YH, Liao MY (2009) The effect of aviation safety education on passenger cabin safety awareness. *Saf Sci* 47:1337–1345
35. Shaftel J, Shaftel TL (2005) The influence of effective teaching in accounting on student attitudes. *Behav Perform Account Educat* 20(3):231–246
36. Davidson AR, Yantis S, Norwood M, Montano D (1985) Amount of information about the attitude object and attitude behavior consistency. *J Pers Soc Psychol* 49:1184–1198
37. Whybark D (1997) GMRG survey research in operations management. *Int J Oper Prod Manag* 17(7):686–696
38. Ahmed S, Hassan MH, Taha Z (2004) State of implementation of TPM in SMIs: a survey study in Malaysia. *J Qual Maint Eng* 10(2):93–106
39. Lee ACAH (2009) *A first course in factor analysis, 2nd edn*. Lawrence Erlbaum Associates, New Jersey
40. Hahn SE, Murphy LR (2008) A short scale for measuring safety climate. *J Saf Sci* 46(7):1047–1066
41. Kroemer KHE, Grandjean E (1997) *Fitting the task to the human: a textbook of occupational ergonomics, 5th edn*. Taylor and Francis, London
42. Bentley T, David T (2010) Incorporating organisational safety culture within ergonomics practice. *Ergonomics* 53(10):1167–1174

Least Squares Data Fitting Subject to Decreasing Marginal Returns

Ioannis C. Demetriou

Abstract Let data of a univariate process be given. If the data are related by a sigmoid curve, but the sigmoid property has been lost due to the errors of the measuring process, then the least sum of squares change to the data that provides nonnegative third divided differences is proposed. The method is highly suitable for estimating points on a sigmoid curve of unspecified parametric form subject to increasing marginal returns or subject to diminishing marginal returns. It is a structured quadratic programming calculation, which is solved very efficiently by a special least squares algorithm that takes into account the form of the constraints. Some numerical results illustrate the method on a variety of data sets. Moreover, two applications of the method on real economic data demonstrate its modeling capability. The first one concerns renewable energy consumption data, which exhibit a sigmoid pattern. The second one concerns technological substitutions among the PDP computers to the VAX computers.

Keywords Approximation · Decreasing marginal returns · Divided difference · Least squares data fitting · Optimization · Renewable energy consumption · Sigmoid · Technological substitution

1 Introduction

Applications of sigmoid curves are common in science, technology, economics and medicine [1, 7, 12, 15, 21]. For, example, a biological growth follows a sigmoid curve or logistic curve, which best models growth and decline over time [14]. Since the adoption of technology and technology-based products is similar to biological growth, many growth curve models have been developed to forecast the penetration

I. C. Demetriou (✉)

Department of Economics, University of Athens, 8 Pessmazoglou street, Athens 10559, Greece
e-mail: demetri@econ.uoa.gr

of these products with the logistic curve and the Gompertz curve the most frequently referenced [17, 18]. Other examples with the sigmoid assumption come from economic substitution [19], from production and distribution cost data for analysis of operations of a firm [9], from decision making [11] and from image processing [10], for instance.

We consider the general problem where measurements from a sigmoid process are to provide estimation to an underlying sigmoid function $f(x)$, but the measurements include random errors. If it is known that the data can be modeled by growth curves or sigmoid curves or that they allow a certain sigmoid form that depends on a few parameters rather than having to estimate unknown function values, then the analysis is usually simplified by existing parametric methods [6, 13]. In this paper we propose a method for estimating points on a sigmoid curve *of unspecified parametric form, when the process is subject to increasing marginal returns or subject to diminishing marginal returns*. The method may be applied to a variety of situations, where the analyst takes the view that the “sigmoid” property of the (unknown) underlying function has been lost due to errors in the data.

Let $\{\phi_i : i = 1, 2, \dots, n\}$ be a sequence of measurements (data) of smooth function values $\{f(x_i) : i = 1, 2, \dots, n\}$, where the abscissae $\{x_i : i = 1, 2, \dots, n\}$ are in strictly ascending order, and let $\phi[x_{i-1}, x_i, x_{i+1}, x_{i+2}]$ designate the third divided difference relative to the four abscissae x_{i-1}, x_i, x_{i+1} and x_{i+2} :

$$\begin{aligned} \phi[x_{i-1}, x_i, x_{i+1}, x_{i+2}] &= \frac{\phi_{i-1}}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})(x_{i-1} - x_{i+2})} \\ &+ \frac{\phi_i}{(x_i - x_{i-1})(x_i - x_{i+1})(x_i - x_{i+2})} \\ &+ \frac{\phi_{i+1}}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)(x_{i+1} - x_{i+2})} \\ &+ \frac{\phi_{i+2}}{(x_{i+2} - x_{i-1})(x_{i+2} - x_i)(x_{i+2} - x_{i+1})}, \\ &i = 2, 3, \dots, n - 2. \end{aligned} \tag{1}$$

The sequence of the third differences

$$\{\phi[x_{i-1}, x_i, x_{i+1}, x_{i+2}], i = 2, 3, \dots, n - 2\}$$

is an appropriate description of the third derivative of $f(x)$ and if the data are error free, then the number of sign changes in (1) is no greater than the number of sign changes in the third derivative of $f(x)$. However, due to errors of measurement it is possible that the sequence $\{\phi[x_{i-1}, x_i, x_{i+1}, x_{i+2}], i = 2, 3, \dots, n - 2\}$ contains far more sign changes than the sequence

$$\{f[x_{i-1}, x_i, x_{i+1}, x_{i+2}], i = 2, 3, \dots, n - 2\}.$$

We assume that no sign changes occur in the third derivative of the underlying function. Thus, if the third divided differences of the data show sign irregularities, we take the view of [5] that some smoothing should be possible in order to recover the missing property. Specifically, we address the problem of making least changes to the data subject to nonnegative third divided differences. We define “least change” with respect to the L_2 norm, which means that we seek a vector \underline{y} that minimizes the sum of the squares

$$\Phi(\underline{y}) = \sum_{i=1}^n (y_i - \phi_i)^2 \quad (2)$$

subject to the constraints

$$y[x_{i-1}, x_i, x_{i+1}, x_{i+2}] \geq 0, \quad i = 2, 3, \dots, n-2, \quad (3)$$

where we regard the data $\{\phi_i : i = 1, 2, \dots, n\}$ and the best fit $\{y_i : i = 1, 2, \dots, n\}$ as components of the n -vectors $\underline{\phi}$ and \underline{y} , respectively. In order to simplify our notation, we denote the constraint normals with respect to \underline{y} by $\{\underline{a}_i : i = 2, 3, \dots, n-2\}$ and we set $y[x_{i-1}, x_i, x_{i+1}, x_{i+2}] = \underline{y}^T \underline{a}_i$, for $i = 2, 3, \dots, n-2$. It is important to note that the constraints on \underline{y} are linear and have linearly independent normals. Also, the second derivative matrix with respect to \underline{y} of the objective function (2) is twice the unit matrix. Thus, the problem of minimizing (2) subject to (3) is a strictly convex quadratic programming problem that has a unique solution. There exist several general algorithms (see, for example, [8]) and two special algorithms [2, 4] that may be applied to this problem after appropriate modifications.

Since the i th third divided difference can be expressed as the difference of two consecutive second divided differences divided by the difference between those arguments which are not in common

$$y[x_{i-1}, x_i, x_{i+1}, x_{i+2}] = \frac{1}{(x_{i+2} - x_{i-1})} (y[x_i, x_{i+1}, x_{i+2}] - y[x_{i-1}, x_i, x_{i+1}]) \quad (4)$$

the constraints (3) imply the inequalities

$$y[x_i, x_{i+1}, x_{i+2}] \geq y[x_{i-1}, x_i, x_{i+1}], \quad i = 2, 3, \dots, n-2. \quad (5)$$

The essential concept in restrictions (5) is process subject to non-decreasing marginal returns, marginal return being the term used for the change in return due to an increase in x . Criterion (3) or the equivalent criterion (5) provides a property that allows a sigmoid shape for the underlying function, as we explain next. Indeed, without loss of generality we assume that there is an index k inside the interval $[2, n-2]$ such that $\{y[x_i, x_{i+1}, x_{i+2}] \leq 0, i = 1, 2, \dots, k-2\}$ and $\{y[x_i, x_{i+1}, x_{i+2}] \geq 0, i = k-1, k, \dots, n-2\}$. It follows that there is a concave region of the fit on $[x_1, x_k]$ and a convex region on $[x_{k-1}, x_n]$. In the concave region, the fit exhibits non-increasing returns

$$y[x_i, x_{i+1}] \geq y[x_{i+1}, x_{i+2}], \quad i = 1, 2, \dots, k-2$$

and in the convex region exhibits non-decreasing returns

$$y[x_i, x_{i+1}] \leq y[x_{i+1}, x_{i+2}], \quad i = k - 1, k, \dots, n - 2,$$

where $y[x_i, x_{i+1}] = (y_{i+1} - y_i) / (x_{i+1} - x_i)$ is the first divided difference relative to x_i and x_{i+1} . It follows that our assumption on non-decreasing second divided differences seems suitable for applications to sigmoid data fitting. Further, it is interesting to note that if we replace the third differences (3) by the analogous second or first differences, we obtain the best convex fit [5], or the best monotonic fit [20] to the data, the latter problem especially having found numerous applications in various subjects during the last 60 years.

This paper is an extended version of [3] and it is organized as follows. In Sect. 2 we outline a quadratic programming method for the optimization calculation. In Sect. 3 we consider two examples on real economic data and reveal important properties of the process. The first is an application to the U.S.A. renewable energy consumption data during the period 1980-2010. The second is an application to technological substitutions among the PDP computers to the VAX computers between the years 1984 and 1991. The results are briefly analyzed and the modeling capability of the method is demonstrated. In Sect. 4 we present numerical results that demonstrate the accuracy of the calculation and the smoothing quality of the method. In Sect. 5 we present some concluding remarks and discuss on the possibility of future directions of this research.

The method may also be applied to the problem where inequalities (3) are replaced by the reversed ones, in which case we obtain a convex / concave fit. The latter problem may be treated computationally as the former one after an overall change of sign of $\underline{\phi}$.

2 An Outline of the Method of Calculation

It is straightforward to calculate the solution of the problem of Sect. 1 by standard quadratic programming methods. However, because each of the constraint functions $y[x_{i-1}, x_i, x_{i+1}, x_{i+2}]$, for $i = 2, 3, \dots, n - 2$, depends on only four adjacent components of \underline{y} and because of the tractability of the least squares objective function, we have developed a special version of the quadratic programming algorithm of [4] that is faster than general algorithms.

Our algorithm generates a finite sequence of subsets $\{A_k : k = 1, 2, \dots\}$ of the constraint indices $\{2, 3, \dots, n - 2\}$ with the property

$$\underline{y}^T \underline{a}_i = 0, \quad i \in A_k. \quad (6)$$

For each k , we denote by $\underline{y}^{(k)}$ the vector that minimizes (2) subject to the equations (6) and we call each constraint in (6) an active constraint. All the active constraints

constitute the active set. Since the constraint normals are linearly independent, unique Lagrange multipliers $\{\lambda_i^{(k)} : i \in A_k\}$ are defined by the first order optimality condition

$$2(\underline{y}^{(k)} - \underline{\phi}) = \sum_{i \in A_k} \lambda_i^{(k)} \underline{a}_i, \tag{7}$$

while, by strict complementarity, $\lambda_i^{(k)} = 0, i \notin A_k$. The method chooses A_k so that each $\lambda_i^{(k)}$ satisfies the conditions

$$\lambda_i^{(k)} \geq 0, i \notin A_k. \tag{8}$$

If A_k is not the final set of the mentioned sequence, A^* say, then the quadratic programming algorithm makes adjustments to A_k until the solution is reached. The Karush-Kuhn-Tucker conditions [8]:p.200 provide necessary and sufficient conditions for optimality. They state that \underline{y} is optimal if and only if the constraints (3) are satisfied and there exist nonnegative Lagrange multipliers $\lambda_i \geq 0, i \in A^*$ such that (7) holds, after we replace $\underline{y}^{(k)}$ by \underline{y} , $\lambda_i^{(k)}$ by λ_i and A_k by A^* .

The calculation begins from any vector $\underline{y}^{(1)}$ such that $\lambda_i^{(1)} \geq 0, for i \in A_1$, where a suitable choice for A_1 is provided by [4]. If the constraints (3) hold at $\underline{y} = \underline{y}^{(1)}$, then the calculation terminates because the Karush-Kuhn-Tucker conditions are satisfied. We assume that at the k th iteration, $A_k, \underline{y}^{(k)}$ and $\underline{\lambda}^{(k)}$ are available, but $\underline{y}^{(k)}$ violates some of the constraints (3). Then, the index of the most violated constraint, ℓ say, is added to A_k and new values of the Lagrange multipliers are calculated. Now, if there are negative multipliers indexed in A_k , then an index, κ say, that is always different from ℓ , is picked from A_k , the κ th constraint is dropped from the active set and A_k is set to $A_k \setminus \{\kappa\}$. The algorithm continues iteratively dropping constraints from the active set until it eventually recovers the inequalities (8). Then a new iteration starts, while current A_k is distinct from all its previous instances at this step and, in exact arithmetic, the value of (2) moves strictly upwards. Since there is only a finite number of sets A_k , the algorithm cannot cycle indefinitely between its steps. This approach is well suited to our problem, while a particular advantage is that only $O(n)$ computer operations are needed for updating the matrices associated with the calculation of $\underline{y}^{(k)}$ and $\underline{\lambda}^{(k)}$.

3 Applications

Substitution in economics is the process at which one product supplants another as it enters the market [19]:p. 273. In competitive strategy several important questions are raised on how to best defend against a substitute, or how to promote substitution. Although the rate of penetration of substitutes differs from product to product, the

path of substitution for successful substitutes looks like an S-curve, where demand is plotted against time.

In this section we present two applications of our method to real data from economic substitutions. First, the set of annual data of renewable energy consumption in quadrillion Btu (energy unit equal to about 1055 joules) in the U.S.A. for the period 1980-2010 (Release Date Report: March 2010 by the Energy Information Administration) is used to illustrate the modeling performance of our method in calculating the best fit. The data are presented in the first two columns of Table 1. For purposes of analysis we are not interested in the physical details of the process, but only in what they imply for the shape of the relationship over time. Since we have to estimate values for an unknown consumption function, initially we make an attempt to distinguish any trends by a primary analysis of the scaled first, second and third differences of the data. These differences are presented in columns 3, 4 and 5, respectively, of Table 1, rounded to four decimal places. The first differences show a slight convex trend, but the second and third differences appear to fluctuate irregularly around zero, as it is shown in Fig. 1. Furthermore, as the trend indicates, the data first seem to increase less than proportionately, then to decrease and then to increase more than proportionately. Therefore we take the view that the underlying consumption function follows the shape of a concave / convex curve.

The method of Sect. 2 was applied to these data (columns 1 and 2 in Table 1). Without any preliminary analysis the data were fed to the computer program and within 17 active set changes the solution was reached. The best fit is presented in the sixth column of Table 1 and the corresponding Lagrange multipliers are presented in the seventh column. Fig. 2 shows the data and the fit. Furthermore, the sequences of the scaled first, second and third divided differences of the best fit are presented in the last three columns of Table 1. We can immediately notice the non-decreasing property of the sequence of the second divided differences and the correspondence between the zero Lagrange multipliers and the non-zero third divided differences. The zero Lagrange multipliers show that all the constraints, but those corresponding to the years 1983, 1999 and 2000, are active. As points with zero third divided differences lie on a parabola, it follows that the calculated consumption curve consists of three overlapping parabolae. A sensitivity analysis would conclude that the best fit is strongly dependent upon the placement of all active constraints on [1984, 1998] and on [2001, 2007], because the associated Lagrange multipliers are away from zero.

The piecewise monotonicity of the first differences (column 8) and the sign change of the second differences (column 9) may lead one's search for estimating the inflection point of the consumption curve. Indeed, the first differences decrease monotonically until 2001 and increase monotonically subsequently, indicating a lower turning point of the marginal consumption curve in the interval [2000, 2001]. The essential feature for the consumption curve is that its secant-slope (cf. first differences), though decreasing until 2001 starts increasing afterwards. Moreover, it is positive up to 1992, then negative up to 2001 and positive afterwards. A rationalization of this is the idea that after 2001, the intensity of the use of renewable energy is increased annually, either because of increased energy demands or because other energy types are being replaced by renewable ones leading to larger and larger consumption increments

Table 1 Least squares fit by nonnegative third divided differences to U.S.A. renewable energy consumption data (in Btu) per year

Year x_i	Data			Best Fit					
	Consumption φ_i	1st differences	2nd differences	3rd differences	Best fit y_i	Lagrange multiplier	1st differences	2nd differences	3rd differences
1980	2.5853	0.1273	-0.0902	0.3029	2.5742	0.0667	0.1328	-0.0165	0.0000
1981	2.7126	0.0371	0.2127	-0.3679	2.7070	0.2336	0.1163	-0.0165	0.0000
1982	2.7496	0.2498	-0.1552	0.1305	2.8233	0.0587	0.0998	-0.0165	0.0000
1983	2.9994	0.0945	-0.0247	-0.1162	2.9231	0.0000	0.0833	-0.0165	0.0091
1984	3.0940	0.0699	-0.1409	0.1559	3.0064	0.5830	0.0667	-0.0075	0.0000
1985	3.1638	-0.0711	0.0150	0.1694	3.0731	2.3521	0.0593	-0.0075	0.0000
1986	3.0927	-0.0561	0.1844	0.0068	3.1323	5.0696	0.0518	-0.0075	0.0000
1987	3.0367	0.1283	0.1912	-0.8971	3.1841	7.8506	0.0443	-0.0075	0.0000
1988	3.1650	0.3196	-0.7058	1.1444	3.2284	10.3146	0.0368	-0.0075	0.0000
1989	3.4846	-0.3863	0.4385	-0.3471	3.2653	13.7771	0.0294	-0.0075	0.0000
1990	3.0983	0.0523	0.0914	-0.2543	3.2947	17.0604	0.0219	-0.0075	0.0000
1991	3.1506	0.1437	-0.1628	0.2715	3.3165	19.1687	0.0144	-0.0075	0.0000
1992	3.2943	-0.0191	0.1087	-0.1777	3.3310	19.8823	0.0069	-0.0075	0.0000
1993	3.2752	0.0896	-0.0690	0.1596	3.3379	18.8250	-0.0005	-0.0075	0.0000
1994	3.3648	0.0206	0.0906	-0.2650	3.3374	16.1612	-0.0080	-0.0075	0.0000
1995	3.3853	0.1111	-0.1744	0.0498	3.3294	12.2266	-0.0155	-0.0075	0.0000
1996	3.4965	-0.0633	-0.1246	0.3601	3.3139	8.1166	-0.0230	-0.0075	0.0000
1997	3.4331	-0.1879	0.2355	-0.2604	3.2909	4.6843	-0.0304	-0.0075	0.0000
1998	3.2453	0.0476	-0.0249	-0.3846	3.2605	1.8382	-0.0379	-0.0075	0.0000
1999	3.2929	0.0227	-0.4095	0.8993	3.2226	0.0000	-0.0454	-0.0075	0.0556
2000	3.3156	-0.3868	0.4898	-0.5334	3.1772	0.0000	-0.0528	0.0482	0.0012
2001	2.9289	0.1030	-0.0436	0.1679	3.1244	0.6650	-0.0047	0.0494	0.0000
2002	3.0319	0.0594	0.1243	-0.1878	3.1197	1.4678	0.0447	0.0494	0.0000

(continued)

Table 1 (continued)

Data			Best fit						
Year	Consumption φ_i	1st differences	2nd differences	3rd differences	Best fit y_i	Lagrange multiplier	1st differences	2nd differences	3rd differences
2003	3.0913	0.1837	-0.0635	0.1176	3.1644	1.9700	0.0941	0.0494	0.0000
2004	3.2750	0.1202	0.0541	-0.0572	3.2584	2.2709	0.1434	0.0494	0.0000
2005	3.3953	0.1743	-0.0031	0.4806	3.4019	2.3309	0.1928	0.0494	0.0000
2006	3.5696	0.1712	0.4774	-1.0293	3.5947	1.9994	0.2422	0.0494	0.0000
2007	3.7408	0.6487	-0.5518	0.7116	3.8369	0.7004	0.2916	0.0494	0.0000
2008	4.3895	0.0968	0.1598	-	4.1284	-	0.3409	0.0494	-
2009	4.4863	0.2566	-	-	4.4694	-	0.3903	-	-
2010	4.7430	-	-	-	4.8597	-	-	-	-

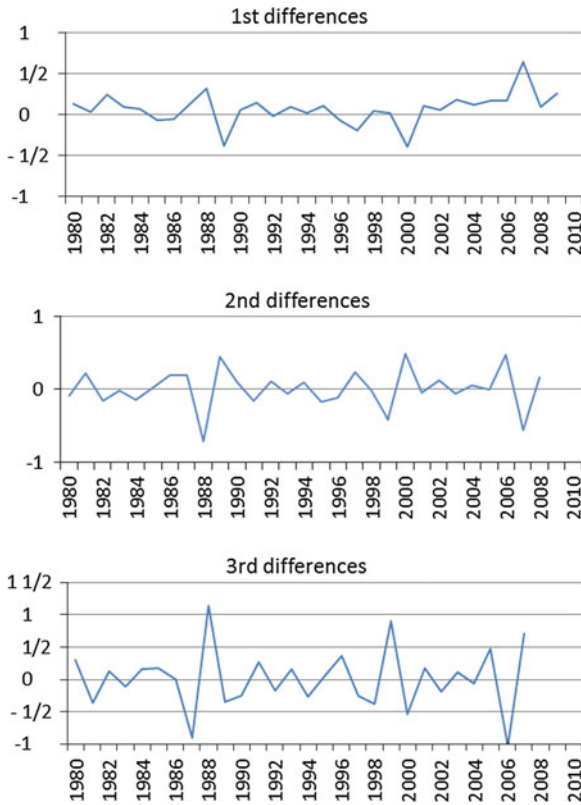
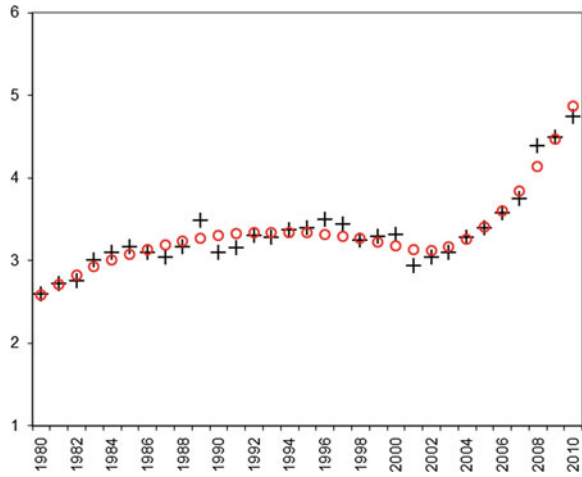


Fig. 1 First, second and third divided differences of the data given in Table 1 (columns 3, 4 and 5 respectively). The continuous line is only for illustration

of renewable energy. The size of the Lagrange multipliers shows that the strongest resistance of the energy market to the renewable energy entering is during [1991, 1994]. Moreover, since the first twenty second differences are negative and the last nine are positive the best fit consists of one concave section on the interval [1980, 2001] and one convex section on the interval [2000, 2010]. Hence the inflection point of the consumption curve lies in the interval [2000, 2001]. Furthermore, the analysis suggests that any estimation of the upper limit in the maximum possible value of energy market penetration rate, which is highly desirable in estimating substitution processes [18], is rather immature at this stage of the renewable energy consumption process.

The second application is a fit to data provided by Modis [16] and I am particularly grateful to Dr. Modis for providing me this data set. Modis analyzes technological substitutions among computer products of Digital Equipment CorporationTM in Europe and discusses on the limitations of Fisher and Pry’s model [7] on these data. The transitions are from the PDP computers to the VAX computers between the years

Fig. 2 Graphical representation of the data given in Table 1. The data of column 1 annotate the x-axis. The data of column 2 are denoted by (+) and the best fit of column 6 by (o)



1984 and 1991, and the data have been derived on 31 trimesters. The first column of Table 2 displays the trimesters and the second column presents the percentage of substitution of PDP by VAX products. The data were fed to our computer program and the best fit subject to non-positive third divided differences is presented in the sixth column of Table 2. All the other columns are explained in Table 1. Especially for the differences presented in columns 3, 4 and 5, we notice that the first and second differences show concave trends, while the third differences exhibit deviations along their range that need investigation. Further, Fig. 3 displays the data and the fit. The computer program terminated at the optimum within 33 active set changes. Without entering a theoretical justification of the results, we note that our method provides an informative description of the substitution process. Indeed, in view of the active constraints, it reveals the ranges of convexity and concavity as well as the rates of marginal change, and, where a Lagrange multiplier is large, the problem is particularly dependent upon the associated constraint.

There exist 26 active constraints, the non-active ones being the 11th and 12th associated with zero Lagrange multipliers. It follows that the calculated substitution curve consists of two overlapping parabolae, one on the interval [1, 13] and one on the interval [13, 31]. The first differences increase monotonically until the 13th trimester and decrease monotonically subsequently, indicating an upper turning point of the marginal substitution curve in the interval [12, 13]. Moreover, they are negative on the first four trimesters, positive up to the 23rd trimester and negative afterwards, which is indicative of the penetration rates on these ranges. Also, we see that the PDP to VAX transition took 16 trimesters to go from about 4 to 90 %. The sequence of the second differences is non-increasing, where the first twelve second differences are positive and the last sixteen are negative. It follows that the best fit consists of one convex section on the interval [1, 14] and one concave section on the interval [13, 31]. Hence, the inflection point of the substitution curve lies in the interval [12, 13].

Table 2 Least squares fit by non-positive third divided differences from the PDP computers to the VAX computers between the years 1984 and 1991

Data		Best fit							
Trimester	%Substitution	1st differences	2nd differences	3rd differences	Best fit y_i	Lagrange multiplier	1st differences	2nd differences	3rd differences
x_i	ϕ_i	differences	differences	differences	y_i	multiplier	differences	differences	differences
1	3.55	6.750	-9.910	13.670	12.690	54.840	-4.534	1.231	0.000
2	10.30	-3.160	3.760	-5.200	8.156	151.652	-3.303	1.231	0.000
3	7.14	0.600	-1.440	-1.450	4.852	276.710	-2.072	1.231	0.000
4	7.74	-0.840	-2.890	16.830	2.780	400.252	-0.841	1.231	0.000
5	6.90	-3.730	13.940	-33.810	1.938	492.508	0.390	1.231	0.000
6	3.17	10.210	-19.870	31.750	2.328	548.425	1.621	1.231	0.000
7	13.38	-9.660	11.880	-11.650	3.949	511.415	2.852	1.231	0.000
8	3.72	2.220	0.230	1.230	6.800	399.958	4.083	1.231	0.000
9	5.94	2.450	1.460	8.710	10.883	243.711	5.314	1.231	0.000
10	8.39	3.910	10.170	-17.460	16.196	89.510	6.545	1.231	0.000
11	12.30	14.080	-7.290	21.560	22.741	0.000	7.776	1.231	-1.014
12	26.38	6.790	14.270	-23.930	30.516	0.000	9.007	0.217	-1.079
13	33.17	21.060	-9.660	-0.280	39.523	127.627	9.224	-0.862	0.000
14	54.23	11.400	-9.940	24.340	48.747	349.983	8.362	-0.862	0.000
15	65.63	1.460	14.400	-33.100	57.109	615.940	7.500	-0.862	0.000
16	67.09	15.860	-18.700	32.140	64.609	910.613	6.638	-0.862	0.000
17	82.95	-2.840	13.440	-31.840	71.247	1163.785	5.776	-0.862	0.000
18	80.11	10.600	-18.400	28.330	77.024	1356.937	4.914	-0.862	0.000
19	90.71	-7.800	9.930	-5.850	81.938	1437.439	4.053	-0.862	0.000
20	82.91	2.130	4.080	-10.510	85.991	1423.772	3.191	-0.862	0.000
21	85.04	6.210	-6.430	3.930	89.181	1340.785	2.329	-0.862	0.000
22	91.25	-0.220	-2.500	6.200	91.510	1190.039	1.467	-0.862	0.000

(continued)

Table 2 (continued)

Data		Best fit						
Trimester	%Substitution	1st	2nd	3rd	Lagrange	1st	2nd	3rd
x_i	φ_i	differences	differences	differences	multiplier	differences	differences	differences
23	91.030	-2.720	3.700	-4.340	983.214	0.605	-0.862	0.000
24	88.310	0.980	-0.640	-2.480	751.943	-0.257	-0.862	0.000
25	89.290	0.340	-3.120	-2.030	520.435	-1.119	-0.862	0.000
26	89.630	-2.780	-5.150	21.250	304.149	-1.981	-0.862	0.000
27	86.850	-7.930	16.100	-26.080	123.336	-2.843	-0.862	0.000
28	78.920	8.170	-9.980	4.990	28.776	-3.704	-0.862	0.000
29	87.090	-1.810	-4.990	-	-	-4.566	-0.862	-
30	85.280	-6.800	-	-	-	-5.428	-	-
31	78.480	-	-	-	73.684	-	-	-

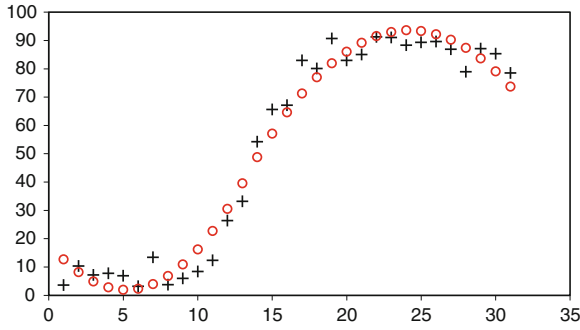


Fig. 3 Graphical representation of the data given in Table 2, as in Fig. 2.

The discussion suggests that our method is able to describe the substitution process everywhere except possibly at the beginning of the data, where substitution is still quite immature.

4 Numerical Results

In this section we present numerical results that show the smoothing performance of our method and the efficiency of the calculation. We have developed a FORTRAN 77 version of the algorithm outlined in Sect. 2 that calculates the solution of the problem of Sect. 1. A PC with an Intel 733 MHz processor was used with the Compaq Visual FORTRAN 6.1 compiler in double precision arithmetic. The data $\{\phi_i : i = 1, 2, \dots, n\}$ were random perturbations of values of the negative modified Gompertz function

$$f(x) = -66.7 - 173.0(0.01157)^{0.7607x/2-1}, 0 \leq x \leq 28, \tag{9}$$

which resembles an elongated ‘S’, consisting of a concave part for $0 \leq x \leq 12.932$ and a convex part subsequently that are joined by a straight line. Note that such data with errors are notoriously difficult to rectify. The data errors in our experiment were simulated by adding to $f(x_i)$ a number from the uniform distribution over $[-r, r]$, where $n = 100, 200, 500, 1000$ and for each n the x_i are equidistant values in $[0, 28]$. The choices $r = 0.5, 5, 10$ have provided a variety of data sets and substantial differences in the active set changes. We have measured the following parameters and present their values in Table 3:

- NACT*, the number of constraints at the optimal active set A^* ;
- The CPU time in seconds required to reach the solution;
- $\rho = \max_{i \in A^*} |\underline{y}^T \underline{a}_i| / \|\underline{a}_i\|_2$, the maximum absolute component of the normalized active constraint residuals at the solution. Since ρ is zero in exact arithmetic,

Table 3 Number of active constraints, CPU times, program accuracy and smoothing performance: data are provided by (9)

<i>n</i>	NACT	CPU	Digits of accuracy				Approximation quality indicators				
			ρ	σ	<i>MaxL</i>	<i>MinL</i>	<i>NORM2</i>	<i>MaxD</i>	<i>PRERR</i>	<i>INFL</i>	
<i>r</i> = 0.5											
100	91	0.05	13.2	12.4	-1.1	2.4	2.9	0.71	0.47	13.86	
200	190	0.43	13.2	12.2	-1.0	3.7	4.1	0.83	0.55	12.66	
500	491	3.30	13.2	8.9	-0.9	4.3	6.9	1.10	0.70	13.24	
1000	991	14.00	13.3	10.1	-0.8	5.3	9.7	1.10	0.72	11.86	
<i>r</i> = 5											
100	91	0.05	13.3	12.4	-1.6	2.1	25.0	5.70	3.60	11.60	
200	192	0.43	13.2	11.9	-1.5	1.8	38.0	5.30	3.40	11.68	
500	490	3.40	13.2	9.9	-1.3	3.6	62.0	5.90	3.70	13.24	
1000	991	12.00	13.2	9.8	-1.1	4.5	89.0	5.80	3.60	11.69	
<i>r</i> = 10											
100	93	0.05	13.3	12.3	-1.9	1.7	49.0	11.00	6.60	11.31	
200	193	0.38	13.2	11.6	-1.7	1.6	75.0	11.00	6.50	11.82	
500	492	3.40	13.1	10.5	-1.6	2.9	120.0	12.00	6.90	13.24	
1000	989	13.00	13.2	8.5	-1.3	3.3	180.0	11.00	6.60	11.46	

Table 3 presents $-\log \rho$ rather than ρ that shows the number of decimal places to which ρ is zero;

$\sigma = \max_{1 \leq i \leq n} |y_i - \phi_i - \frac{1}{2} \sum_{k \in A^*} \lambda_k a_{ik}|$, the maximum absolute component of the residuals of the Karush-Kuhn-Tucker conditions, where a_{ik} is the i th component of the k th constraint gradient. Since σ is zero in exact arithmetic, Table 3 presents $-\log \sigma$ rather than σ ;

MaxL, *MinL*, the number of decimal places to which the maximum and the minimum Lagrange multiplier are zero;

NORM2, the distance between the data and the best fit;

MaxD = $\max_{1 \leq i \leq n} |\phi_i - y_i|$, the maximum deviation;

PRERR = $100 \times \text{MaxD} / (\max_{1 \leq i \leq n} \phi_i - \min_{1 \leq i \leq n} \phi_i)$, the percent relative error to the scale of values taken on by the data;

INFL, the central abscissa of the first positive second divided difference of the best fit, providing an estimation of the inflection point of the best fit.

Parameters ρ and σ provide a measure of the accuracy of the computer program. Parameters *NORM2*, *MaxD* and *PRERR* are the actual smoothing quality indicators that the user has available at the end of the calculation. *MaxL* and *MinL* give a sense of the sensitivity of the problem upon the constraints. We can see in Table 3 that the number of constraints in the optimal active sets is quite close to $n - 3$, which is the number of constraints in (3). The CPU time ranged from 0.05 to 13.0 seconds as n ranged from 100 to 1000. The accuracy of the active constraints, i.e. ρ , has been as good as the machine accuracy. The accuracy of the Karush-Kuhn-Tucker residuals, i.e. σ , ranged from 9 to 12 decimal digits, which indicates an accurate and stable

calculation. The *MaxD* indicator provided a satisfactory bound to the magnitude of the error contaminated the function values. Most importantly, in all the experiments, the sign change in the sequence of the second divided differences effectively indicated the interval where the inflection point belongs, as we can see by comparing *INFL* with 12.932, namely the inflection point of (9).

5 Concluding Remarks

We have proposed a method that gives the best least squares fit to data contaminated by random errors subject to nonnegative third divided differences. The method is suitable when the data exhibit a sigmoid trend, where a concave region is followed by a convex one. The method is also suitable when it would be better to employ non-positive instead of nonnegative divided differences, in which case a convex region precedes a concave one. The fit consists of a certain number of overlapping parabolae, which not only provides flexibility in data fitting, but also helps managing further operations with the data fit like interpolation, extrapolation, differentiation and integration. Moreover, the interval of the inflection point of the fit is provided automatically by the calculation. This data fitting procedure may be used in many situations, when analyzing processes that are subject to diminishing marginal returns, as for example in modeling product lifecycles. Analogously, for increasing marginal returns, as for example in estimating cost and production functions. The accompanying FORTRAN program is suitable for calculations that involve several thousand data points and it would be most useful for real problem applications. In addition, there is nothing to prevent combining certain features of our method with the logistic curve or the Gompertz curve and other parametric sigmoid forms if there exists an opportunity for improved practical analyses.

Acknowledgments This work was partially supported by the University of Athens under Research Grant 11105.

References

1. Bengisu M, Nekhili R (2006) Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change* 73:835–844
2. Cullinan MP (1990) Data smoothing using non-negative divided differences and 12 approximation. *IMA J. of Numerical Analysis* 10:583–608
3. Demetriou IC (2012) "Applications of the discrete least squares 3-convex fit to sigmoid data", *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012 WCE 2012*, 4–6 July, 2012. U.K, London, pp 285–290
4. Demetriou IC, Lipitakis EA "Least squares data smoothing by nonnegative divided differences", Unpublished manuscript
5. Demetriou IC, Powell MJD (1991) The minimum sum of squares change to univariate data that gives convexity. *IMA J. of Numerical Analysis* 11:433–448

6. Dierckx P (1995) *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford
7. Fisher JC, Pry RH (1971) A simple substitution model of technological change. *Technological Forecasting and Social Change* 2:75–88
8. Fletcher R (1987) *Practical Methods of Optimization*. J. Wiley and Sons, Chichester, U.K.
9. Fubrycky WJ, Thuesen GJ, Verna D (1998) *Economic Decision Analysis*, 3rd edn. Prentice Hall, Upper Saddle River, New Jersey
10. Gonzalez RC, Wintz P (1987) *Digital Image Processing*, 2nd edn. Addison Wesley Publishing Company, Reading, Mass
11. Lindley DV (1985) *Making Decisions*, 2nd edn. J. Wiley and Sons, London, U.K.
12. McKenna CJ, Rees R (1996) *Economics: A Mathematical Introduction*. Oxford University Press, New York
13. Meade N, Islam T (1991) Forecasting with growth curves: an empirical comparison. *International Journal of Forecasting* 11:199–215
14. Medawar PB (1941) The Laws of biological growth. *Nature* 148:772–774
15. Modis T (1992) *Predictions - Society's Telltale Signature Reveals the Past and Forecasts the Future*. Simon and Schuster, New York
16. Modis T (1993) Technological substitutions in the computer industry. *Technological Forecasting and Social Change* 43:157–167
17. Morrison JS (1995) Life-cycle approach to new product forecasting. *The Journal of Business Forecasting Methods and Systems* 14:3–5
18. Morrison JS (1996) How to use diffusion models in new product forecasting. *The Journal of Business Forecasting Methods and Systems* 15:6–9
19. Porter ME (1985) *Competitive Advantage, Creating and Sustaining Superior Performance*. The Free Press, Collier Macmillan Publishers, London, U.K
20. Robertson T, Wright FT, Dykstra RL (1988) *Order Restricted Statistical Inference*. John Wiley and Sons, New York
21. West JB (1985) *Respiratory Physiology - the Essentials*, 3rd edn. Williams and Wilkins, Baltimore

The Further Development of Stem Taper and Volume Models Defined by Stochastic Differential Equations

Petras Rupšys

Abstract Stem taper process measured repeatedly among a series of individual trees is standardly analyzed by fixed and mixed regression models. This stem taper process can be adequately modeled by parametric stochastic differential equations (SDEs). We focus on the segmented stem taper model defined by the Gompertz, geometric Brownian motion and Ornstein-Uhlenbeck stochastic processes. This class of models enables the representation of randomness in the taper dynamics. The parameter estimators are evaluated by maximum likelihood procedure. The SDEs stem taper models were fitted to a data set of Scots pine trees collected across the entire Lithuanian territory. Comparison of the predicted stem taper and stem volume with those obtained using regression based models showed a predictive power to the SDEs models.

Keywords Diameter · Geometric Brownian motion · Gompertz process · Ornstein-Uhlenbeck process · Taper · Transition probability density · Stochastic differential equation · Volume

1 Introduction

Deterministic and stochastic differential equations are probably the most commonly mathematical representations for describing continuous time processes [1, 2]. Biological experiments often imply repeated measurements on a series of experimental units. Stem taper process is usually measured repeatedly among a collection of individual trees. Traditionally, the relationship between volume, height and diameter has been modeled based on simple linear and nonlinear regressions. The base

P. Rupšys (✉)
Department of Forest Management, Aleksandras Stulginskis University,
LT-53361 Kaunas, Lithuania
e-mail: petras.rupsys@asu.lt

assumption of these regression models is that the observed variations from the regression curve are constant at different values of a covariate would be realistic if the variations were due to measurement errors. Instead, it is unrealistic, as the variations are due to random changes on growth rates induced by random environmental perturbations. Stochastic differential equations (SDEs) models do not have such weakness [3]. We propose to model these variations using SDEs that are deduced from the standard deterministic growth function by adding random variations to the growth dynamics [3–12]. Due to the specific characteristics of diameter dynamics, we thus consider SDEs models with drift and diffusion terms that can depend linearly or nonlinearly on state variables.

There is a long history characterizing the stem profile (taper) of trees. Mathematically defining stem taper is necessary for the accurate prediction of stem volume. Taper equations do just this and are important to foresters and forest scientists because they provide a flexible alternative to conventional volume equations. These equations are widely used in forestry to estimate diameter at any given height along a tree bole and therefore to calculate total or merchantable stem volume. One crucial element in these models is the functional response that describes the relative diameter of tree stem consumed per relative height for given quantities of diameter at breast height D and total tree height H . The most commonly studied stem taper relations range from simple taper functions to more complex forms [13–20]. Taper curve data consist of repeated measurements of a continuous diameter growth process over height of individual trees. These longitudinal data have two characteristics that complicate their statistical analysis: (a) within-individual tree correlation that appears with data measured on the same tree and (b) independence but extremely high variability between the experimental taper curves of the different trees. Mixed models provide one of powerful tools to analysis of longitudinal data. These models incorporate the variability between individual trees by means of the expression of the model's parameters and in terms of both fixed and random effects. Each parameter in the model may be represented by a fixed effect that stands for the mean value of the parameter as well as a random effect that expresses the difference between the value of the parameter fitted for each specific tree and the mean value of the parameter—the fixed effect. Random effects are conceptually random variables. They are modeled as such in terms of describing their distribution. This helps to avoid the problem of overparameterisation. A large number of mixed-effect taper regression models have been completed, and the study is still one of the important issues in progress [16, 17, 19].

The increasing popularity of mixed-effects models lies in their ability to model total variation, splitting it into its within- and between-individual tree components. In this paper, we propose to model these variations using SDEs that are deduced from the standard deterministic growth function by adding random variations to the growth dynamics [6–12]. Although numerous sophisticated models exist for stem profile [18, 20], relatively few models have been produced using SDEs [3, 12].

The basis of the work is a deterministic segmented taper model, which uses different SDEs for various parts of the stem to overcome local bias. In this paper an effort has been made to present a class of SDEs stem taper models and to show that they are quite viable and reliable for predicting not just diameter outside the bark

for a given height but merchantable stem volume as well. Our main contribution is to expand stem taper and stem volume models by using SDEs and to show how an adequate model can be made. In this paper attention is restricted to homogeneous SDEs in the Gompertz, geometric Brownian motion and Ornstein-Uhlenbeck type.

2 Stem Taper Models

Consider a one-dimensional stochastic process $Y(x)$ evolving in M different experimental units (e.g. trees) randomly chosen from a theoretical population (tree species). We suppose that the dynamics of the relative diameter $Y^i = d/D^i$ versus the relative height $x^i = h/H^i$ ($x^i \in [0; 1]$) is expressed by the Itô stochastic differential equation [21], where d is the diameter outside the bark at any given height h , D^i is the diameter at breast height outside the bark of i th tree, H^i is the total tree height from ground to tip of i th tree. In this paper is used a class of the SDEs that are reducible to the Ornstein-Uhlenbeck process. The stochastic processes used in this work incorporated environmental stochasticity, which accounts for variability in the diameter growth rate that arises from external factors (such as soil structure, water quality and quantity, and levels of various soil nutrients) that equally affect all the trees in the stands.

The first utilized stochastic process of the relative diameter dynamics is defined in the following Gompertz form [6, 8]

$$dY^i(x^i) = [\alpha_G Y^i(x^i) - \beta_G Y^i(x^i) \ln(Y^i(x^i))]dx^i + \sigma_G Y^i(x^i)dW_G^i(x^i), \quad (1)$$

where $P(Y^i(x_0^i) = y_0^i) = 1, i = 1, \dots, M, Y^i(x^i)$ is the value of the diameter growth process at the relative height $x^i \geq x_0^i, \alpha_G, \beta_G,$ and σ_G are fixed effects parameters (identical for the entire population of trees), y_0^i is non-random initial relative diameter. The $W_G^i(x^i), i = 1, \dots, M$ are mutually independent standard Brownian motions. The second stochastic process of the relative diameter dynamics is defined in the following geometric Brownian motion form [23]

$$dY^i(x^i) = \alpha_{GB} Y^i(x^i)dx^i + \sigma_{GB} Y^i(x^i)dW_{GB}^i(x^i), \quad (2)$$

where $P(Y^i(x_0^i) = y_0^i) = 1, i = 1, \dots, M, \alpha_{GB},$ and σ_{GB} are fixed effects parameters (identical for the entire population of trees) and $W_{GB}^i(x^i)$ are mutually independent standard Brownian motions. The third stochastic process of the relative diameter dynamics is defined in the following Ornstein-Uhlenbeck form [22]

$$dY^i(x^i) = \left(\alpha_O - \frac{Y^i(x^i)}{\beta_O} \right) dx^i + \sigma_O dW_O^i(x^i) \quad (3)$$

where $P(Y^i(x_0^i) = y_0^i) = 1, i = 1, \dots, M, \alpha_O, \beta_O,$ and σ_O are fixed effects parameters (identical for the entire population of trees) and $W_O^i(x^i)$ are mutually independent standard Brownian motions.

In this paper is used a segmented stochastic taper process which consists of three different SDEs defined by (1)–(3). Max and Burkhardt [24] proposed a segmented polynomial regression model that uses two joining points 0.15, 0.75 to link three different stem sections. Following this idea the stem taper models (with two joining points: 0.15, 0.75 or $\frac{1.3}{H^i}, 0.75$) are defined in the two different forms

$$dY^i(x^i) = \begin{cases} [\alpha_G Y^i(x^i) - \beta_G Y^i(x^i) \ln(Y^i(x^i))]dx^i + \sigma_G Y^i(x^i)dW_G^i(x^i), & x^i \leq 0.15 \\ \alpha_{GB} Y^i(x^i)dx^i + \sigma_{GB} Y^i(x^i)dW_{GB}^i(x^i), & 0.15 < x^i \leq 0.75 \\ [\alpha_G Y^i(x^i) - \beta_G Y^i(x^i) \ln(Y^i(x^i))]dx^i + \sigma_G Y^i(x^i)dW_G^i(x^i), & x^i > 0.75 \end{cases} \quad (4)$$

$$dY^i(x^i) = \begin{cases} [\alpha_G Y^i(x^i) - \beta_G Y^i(x^i) \ln(Y^i(x^i))]dx^i + \sigma_G Y^i(x^i)dW_G^i(x^i), & x^i \leq \frac{1.3}{H^i} \\ \alpha_{GB} Y^i(x^i)dx^i + \sigma_{GB} Y^i(x^i)dW_{GB}^i(x^i), & \frac{1.3}{H^i} < x^i \leq 0.75 \\ [\alpha_G Y^i(x^i) - \beta_G Y^i(x^i) \ln(Y^i(x^i))]dx^i + \sigma_G Y^i(x^i)dW_G^i(x^i), & x^i > 0.75 \end{cases} \quad (5)$$

Using Eqs. (4), (5) and either fixing the stem butt and top or assuming that the stem butt and top were free, we define five stem taper models.

Model 1: Equation (4) and $P(Y^i(x_0^i) = \gamma) = 1, i = 1, \dots, M$ (the stem butt and top of the i th tree are free), γ is additional fixed-effects parameter (identical for the entire population of trees).

Model 2: Equation (4) and $P(Y^i(x_0^i) = y_0^i) = 1, i = 1, \dots, M$ (the stem butt of the i th tree is fixed and the top is free).

Model 3: Equation (4) and $P(Y^i(x_0^i) = y_0^i) = 1, P(Y^i(1) = 0) = 1, i = 1, \dots, M$ (the stem butt and top of the i th tree are fixed).

Model 4: Equation (5) and $P(Y^i(\frac{1.3}{H^i}) = 1) = 1, i = 1, \dots, M$ (the diameter at breast height of the i th tree is fixed and top is free).

Model 5: Equation (5) and $P(Y^i(\frac{1.3}{H^i}) = 1) = 1, P(Y^i(1) = 0) = 1, i = 1, \dots, M$ (the diameter at breast height and top of the i th tree are fixed).

Assume that tree i is measured at $n_i + 1$ discrete relative height points $(x_0, x_1, \dots, x_{n_i})$ $i = 1, \dots, M$. Let \underline{y}^i be the vector of relative diameters for tree i , $\underline{y}^i = (y_0^i, y_1^i, \dots, y_{n_i}^i)$, where $y^i(x_j^i) = y_j^i, \underline{y} = (\underline{y}^1, \underline{y}^2, \dots, \underline{y}^M)$ is the n -dimensional total relative diameter vector, $n = \sum_{i=1}^M (n_i + 1)$. Therefore, we need to estimate fixed-effects parameters $\gamma, \alpha_G, \beta_G, \sigma_G, \alpha_{GB}, \beta_{GB}, \sigma_{GB}, \alpha_O, \beta_O, \sigma_O$ using all the data in \underline{y} simultaneously.

Models 2 and 3 use one tree-specific prior relative diameter y_0^i (this known initial condition additional needs stem diameter measured at a stem height of 0 m). Models 4 and 5 use known relative diameter at breast height, 1. The transition probability

density function of the relative diameter stochastic processes $Y^i(x_j^i)$, $x_j^i \in [0; 1]$, $i = 1, \dots, M$, $j = 0, \dots, n_i$ defined by Eqs. (1)–(3) can be deduced in the following form: for the Gompertz stochastic process [8]

$$p_G(y^i, x | y_z^i) = \frac{1}{y^i \sqrt{2\pi} v_G(x, z)} \exp\left(-\frac{1}{2v_G(x, z)} \left(\ln y^i - \mu_G(x, z, y_z^i)\right)^2\right) \quad (6)$$

$$\mu_G(x, z, y_z^i) = \ln y_z^i e^{-\beta_G(x-z)} + \frac{1 - e^{-\beta_G(x-z)}}{\beta_G} \left(\alpha_G - \frac{\sigma_G^2}{2}\right) \quad (7)$$

$$v_G(x, z) = \frac{1 - e^{-2\beta_G(x-z)}}{2\beta_G} \sigma_G^2 \quad (8)$$

for the geometric Brownian motion [22]

$$p_{GB}(y^i, x | y_z^i) = \frac{1}{\sigma_{GB} y^i \sqrt{2\pi(x-z)}} \exp\left(-\frac{\left(\ln\left(\frac{y^i}{y_z^i}\right) - \left(\alpha_{GB} - \frac{1}{2}\sigma_{GB}^2\right)(x-z)\right)^2}{2\sigma_{GB}^2(x-z)}\right) \quad (9)$$

and for the Ornstein-Uhlenbeck stochastic process [23]

$$p_O(y^i, x | y_z^i) = \frac{1}{\sqrt{2\pi} v_O(x, z)} \exp\left(-\frac{(y^i - \mu_O(x, z, y_z^i))^2}{2v_O(x, z)}\right) \quad (10)$$

$$\mu_O(x, z, y_z^i) = y_z^i \exp\left(-\frac{x-z}{\beta_O}\right) + \alpha_O \beta_O \left(1 - \exp\left(-\frac{x-z}{\beta_O}\right)\right) \quad (11)$$

$$v_O(x, z) = \frac{\sigma_O^2 \beta_O}{2} \left(1 - e^{-\frac{2(x-z)}{\beta_O}}\right) \quad (12)$$

The conditional mean and variance functions $m(x^i | \cdot)$ and $v(x^i | \cdot)$ (x^i is the relative height of the i th tree) of the stochastic processes (1)–(3) are defined by

$$\begin{aligned} m_G(x^i | y_0^i, \alpha_G, \beta_G, \sigma_G) \\ = y_0^i e^{-\beta_G x^i} \exp\left(\frac{1 - e^{-\beta_G x^i}}{\beta_G} \left(\alpha_G - \frac{\sigma_G^2}{2}\right) + \frac{\sigma_G^2}{4\beta_G} (1 - e^{-2\beta_G x^i})\right) \end{aligned} \quad (13)$$

$$\begin{aligned} w_G(x^i | y_0^i, \alpha_G, \beta_G, \sigma_G) \\ = \exp\left(2\left(\ln y_0^i e^{-\beta_G x^i} + \frac{1 - e^{-\beta_G x^i}}{\beta_G} \left(\alpha_G - \frac{\sigma_G^2}{2}\right)\right) + \frac{\sigma_G^2}{2\beta_G} (1 - e^{-2\beta_G x^i})\right) \\ \times \left(\exp\left(\frac{\sigma_G^2}{2\beta_G} (1 - e^{-2\beta_G x^i})\right) - 1\right) \end{aligned} \quad (14)$$

for the Gompertz stochastic process [8],

$$m_{GB}(x^i | y_0^i, \alpha_{GB}) = y_0^i e^{\alpha_{GB} x^i} \tag{15}$$

$$w_{GB}(x^i | y_0^i, \alpha_{GB}, \sigma_{GB}) = (y_0^i)^2 e^{2\alpha_{GB} x^i} (e^{\sigma_{GB}^2 x^i} - 1) \tag{16}$$

for the geometric Brownian motion [22] and for the Ornstein-Uhlenbeck process the conditional mean and variance functions $m(x^i | \cdot)$ and $v(x^i | \cdot)$ are defined by [23]

$$m_O(x^i | y_0^i, \alpha_O, \beta_O) = y_0^i \exp\left(-\frac{x^i}{\beta_O}\right) + \alpha_O \beta_O \left(1 - \exp\left(-\frac{x^i}{\beta_O}\right)\right) \tag{17}$$

$$w_O(x^i | \beta_O, \sigma_O) = \frac{\sigma_O^2 \beta_O}{2} \left(1 - e^{-\frac{2x^i}{\beta_O}}\right) \tag{18}$$

Using the transition probability densities (6), (9) and (10) of SDEs (1)–(3), the transition probability density functions of the relative diameter stochastic process, for Models 1–5 take the form, respectively

$$p_1(y_j^i, x_j^i | y_{j-1}^i) = \begin{cases} p_G(y_j^i, x_j^i | y_{j-1}^i), & x_j^i \leq 0.15 \\ p_{GB}(y_j^i, x_j^i | y_{j-1}^i), & 0.15 < x_j^i \leq 0.75 \\ p_O(y_j^i, x_j^i | y_{j-1}^i), & x_j^i > 0.75 \end{cases}, i = 1, \dots, M, y_0^i = \gamma, j = 1, \dots, n_i \tag{19}$$

$$p_2(y_j^i, x_j^i | y_{j-1}^i) = \begin{cases} p_G(y_j^i, x_j^i | y_{j-1}^i), & x_j^i \leq 0.15 \\ p_{GB}(y_j^i, x_j^i | y_{j-1}^i), & 0.15 < x_j^i \leq 0.75 \\ p_O(y_j^i, x_j^i | y_{j-1}^i), & x_j^i > 0.75 \end{cases}, i = 1, \dots, M, j = 1, \dots, n_i \tag{20}$$

$$p_3(y_j^i, x_j^i | y_{j-1}^i, y_{j+1}^i) = \begin{cases} p_G(y_j^i, x_j^i | y_{j-1}^i), & x_j^i \leq 0.15 \\ p_{GB}(y_j^i, x_j^i | y_{j-1}^i), & 0.15 < x_j^i \leq 0.75 \\ p_O(y_j^i, 1 - x_j^i | y_{j+1}^i), & x_j^i > 0.75 \end{cases}, i = 1, \dots, M, j = 1, \dots, n_i - 1 \tag{21}$$

$$\begin{aligned}
 & p_4(y_j^i, x_j^i | y_{j-1}^i) \\
 &= \begin{cases} PG(y_j^i, \frac{1.3}{H^i} - x_j^i | y_{j-1}^i), x_j^i \leq \frac{1.3}{H^i} \\ PG_B(y_j^i, x_j^i | y_{j-1}^i), \frac{1.3}{H^i} < x_j^i \leq 0.75, i = 1, \dots, M, j = 1, \dots, n_i \\ PO(y_j^i, x_j^i | y_{j-1}^i), x_j^i > 0.75 \end{cases} \quad (22)
 \end{aligned}$$

$$\begin{aligned}
 & p_5(y_j^i, x_j^i | y_{j-1}^i, y_{j+1}^i) \\
 &= \begin{cases} PG(y_j^i, \frac{1.3}{H^i} - x_j^i | y_{j-1}^i), x_j^i \leq \frac{1.3}{H^i} \\ PG_B(y_j^i, x_j^i | y_{j-1}^i), \frac{1.3}{H^i} < x_j^i \leq 0.75, i = 1, \dots, M, j = 1, \dots, n_i - 1 \\ PO(y_j^i, 1 - x_j^i | y_{j+1}^i), x_j^i > 0.75 \end{cases} \quad (23)
 \end{aligned}$$

Using the conditional mean and variance functions (13)–(18) we define the trajectories of diameter’ and its variance’ for Models 1–5 in the following form, respectively

$$\begin{aligned}
 & d_1(h, D, H) \\
 &= \begin{cases} D \cdot m_G(\frac{h}{H} | \hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \frac{h}{H} \leq 0.15 \\ D \cdot m_{GB}(\frac{h}{H} - 0.15 | m_G(0.15 | \hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \hat{\alpha}_{GB}, \hat{\beta}_{GB}), 0.15 < \frac{h}{H} \leq 0.75 \\ D \cdot m_O(\frac{h}{H} - 0.75 | m_{GB}(0.75 - 0.15 | \hat{\alpha}_{GB}, \hat{\sigma}_{GB}), \hat{\alpha}_O, \hat{\beta}_O), \frac{h}{H} > 0.75 \end{cases} \quad (24)
 \end{aligned}$$

$$\begin{aligned}
 & w_1(h, D, H) \\
 &= \begin{cases} D^2 \cdot w_G(\frac{h}{H} | \hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \frac{h}{H} \leq 0.15 \\ D^2 \cdot (w_G(0.15 | \hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G) + w_{GB}(\frac{h}{H} - 0.15 | \hat{\beta}_{GB}, \hat{\sigma}_{GB})), 0.15 < \frac{h}{H} \leq 0.75 \\ D^2 \cdot (w_G(0.15 | \hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G) + w_{GB}(0.75 - 0.15 | \hat{\beta}_{GB}, \hat{\sigma}_{GB}) \\ + w_O(\frac{h}{H} - 0.75 | \hat{\beta}_O, \hat{\sigma}_O)), \frac{h}{H} \geq 0.75 \end{cases} \quad (25)
 \end{aligned}$$

$$\begin{aligned}
 & d_2(h, D, H, d_0) \\
 &= \begin{cases} D \cdot m_G(\frac{h}{H} | \frac{d_0}{D}, \hat{\alpha}, \hat{\beta}_G, \hat{\sigma}_G), \frac{h}{H} \leq 0.15 \\ D \cdot m_{GB}(\frac{h}{H} - 0.15 | m_G(0.15 | \frac{d_0}{D}, \hat{\alpha}, \hat{\beta}_G, \hat{\sigma}_G), \hat{\alpha}_{GB}, \hat{\beta}_{GB}), 0.15 < \frac{h}{H} \leq 0.75 \\ D \cdot m_O(\frac{h}{H} - 0.75 | m_{GB}(0.75 - 0.15 | \hat{\alpha}_{GB}, \hat{\sigma}_{GB}), \hat{\alpha}_O, \hat{\beta}_O), \frac{h}{H} > 0.75 \end{cases} \quad (26)
 \end{aligned}$$

$w_2(h, D, H, d_0)$

$$= \begin{cases} D^2 \cdot w_G\left(\frac{h}{H} \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \frac{h}{H} \leq 0.15 \\ D^2 \cdot (w_G(0.15 \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.) + w_{GB}\left(\frac{h}{H} - 0.15 \left| \hat{\beta}_{GB}, \hat{\sigma}_{GB} \right.\right)), 0.15 < \frac{h}{H} \leq 0.75 \\ D^2 \cdot (w_G(0.15 \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.) + w_{GB}(0.75 - 0.15 \left| \hat{\beta}_{GB}, \hat{\sigma}_{GB} \right.) \\ + w_O\left(\frac{h}{H} - 0.75 \left| \hat{\beta}_O, \hat{\sigma}_O \right.\right)), \frac{h}{H} \geq 0.75 \end{cases} \quad (27)$$

$d_3(h, D, H, d_0)$

$$= \begin{cases} D \cdot m_G\left(\frac{h}{H} \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \frac{h}{H} \leq 0.15 \\ D \cdot m_{GB}\left(\frac{h}{H} - 0.15 \left| m_G(0.15 \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \hat{\alpha}_{GB}, \hat{\beta}_{GB} \right), 0.15 < \frac{h}{H} \leq 0.75 \\ D \cdot m_O\left(1 - \frac{h}{H} \left| 0, \hat{\alpha}_O, \hat{\beta}_O \right.\right), \frac{h}{H} > 0.75 \end{cases} \quad (28)$$

$w_3(h, D, H, d_0)$

$$= \begin{cases} D^2 \cdot w_G\left(\frac{h}{H} \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \frac{h}{H} \leq 0.15 \\ D^2 \cdot (w_G(0.15 \left| \frac{d_0}{D}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.) + w_{GB}\left(\frac{h}{H} - 0.15 \left| \hat{\beta}_{GB}, \hat{\sigma}_{GB} \right.\right)), 0.15 < \frac{h}{H} \leq 0.75 \\ D^2 \cdot w_O\left(1 - \frac{h}{H} \left| \hat{\beta}_O, \hat{\sigma}_O \right.\right), \frac{h}{H} \geq 0.75 \end{cases} \quad (29)$$

$d_4(h, D, H)$

$$= \begin{cases} D \cdot m_G\left(\frac{1.3}{H} - \frac{h}{H} \left| 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \frac{h}{H} \leq \frac{1.3}{H} \\ D \cdot m_{GB}\left(\frac{h}{H} - \frac{1.3}{H} \left| m_G\left(0. \left| 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \hat{\alpha}_{GB}, \hat{\beta}_{GB} \right.\right), \frac{1.3}{H} < \frac{h}{H} \leq 0.75 \\ D \cdot m_O\left(\frac{h}{H} - 0.75 \left| m_{GB}(0.75 - 0.15 \left| \hat{\alpha}_{GB}, \hat{\sigma}_{GB} \right.\right), \hat{\alpha}_O, \hat{\beta}_O \right.\right), \frac{h}{H} > 0.75 \end{cases} \quad (30)$$

$w_4(h, D, H)$

$$= \begin{cases} D^2 \cdot w_G\left(\frac{1.3}{H} - \frac{h}{H} \left| 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G \right.\right), \frac{h}{H} \leq \frac{1.3}{H} \\ D^2 \cdot w_{GB}\left(\frac{h}{H} - \frac{1.3}{H} \left| 1., \hat{\beta}_{GB}, \hat{\sigma}_{GB} \right.\right), \frac{1.3}{H} < \frac{h}{H} \leq 0.75 \\ D^2 \cdot (w_{GB}(0.75 - \frac{1.3}{H} \left| 1., \hat{\beta}_{GB}, \hat{\sigma}_{GB} \right.) + w_O\left(\frac{h}{H} - 0.75 \left| \hat{\beta}_O, \hat{\sigma}_O \right.\right)), \frac{h}{H} \geq 0.75 \end{cases} \quad (31)$$

$$d_5(h, D, H) = \begin{cases} D \cdot m_G(\frac{1.3}{H} - \frac{h}{H} \mid 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \frac{h}{H} \leq \frac{1.3}{H} \\ D \cdot m_{GB}(\frac{h}{H} - \frac{1.3}{H} \mid m_G(0. \mid 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \hat{\alpha}_{GB}, \hat{\beta}_{GB}), \frac{1.3}{H} < \frac{h}{H} \leq 0.75 \\ D \cdot m_O(1 - \frac{h}{H} \mid 0, \hat{\alpha}_O, \hat{\beta}_O), \frac{h}{H} > 0.75 \end{cases} \quad (32)$$

$$w_5(h, D, H) = \begin{cases} D^2 \cdot w_G(\frac{1.3}{H} - \frac{h}{H} \mid 1., \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G), \frac{h}{H} \leq \frac{1.3}{H} \\ D^2 \cdot w_{GB}(\frac{h}{H} - \frac{1.3}{H} \mid 1., \hat{\beta}_{GB}, \hat{\sigma}_{GB}), \frac{1.3}{H} < \frac{h}{H} \leq 0.75 \\ D^2 \cdot w_O(1 - \frac{h}{H} \mid \hat{\beta}_O, \hat{\sigma}_O), \frac{h}{H} \geq 0.75 \end{cases} \quad (33)$$

where $\hat{\gamma}, \hat{\alpha}_G, \hat{\beta}_G, \hat{\sigma}_G, \hat{\alpha}_{GB}, \hat{\sigma}_{GB}, \hat{\alpha}_{OU}, \hat{\beta}_O, \hat{\sigma}_O$ are maximum likelihood estimators.

In this paper, we apply the theory of a one-stage maximum likelihood estimator for stem taper Models 1–5. As all models have closed form transition probability density functions (19)–(23), the log-likelihood function for Models 1, 2, 4, and Models 3, 5 are given, respectively

$$L_k(\gamma, \alpha_G, \beta_G, \sigma_G, \alpha_{GB}, \sigma_{GB}, \alpha_{OU}, \beta_O, \sigma_O) = \sum_{i=1}^M \sum_{j=1}^{n_i} \ln(p_k(y_j^i, x_j^i \mid y_{j-1}^i)), k = 1, 2, 4 \quad (34)$$

$$L_k(\alpha_G, \beta_G, \sigma_G, \alpha_{GB}, \sigma_{GB}, \alpha_{OU}, \beta_O, \sigma_O) = \sum_{i=1}^M \sum_{j=1}^{n_i-1} \ln(p_k(y_j^i, x_j^i \mid y_{j-1}^i, y_{j+1}^i)), k = 3, 5 \quad (35)$$

To assess the standard errors of the maximum likelihood estimators for stem taper Models 1–5, a study of the Fisher [25] information matrix was performed. The asymptotic variance of the maximum likelihood estimator is given by the inverse of the Fisher’ information matrix, which is the lowest possible achievable variance among the competing estimators. By defining $p_k(\theta^k) \equiv \ln(L_k(\theta^k))$, where $k = 1, 2, 3, 4, 5, \theta^1 = (\gamma, \alpha_G, \beta_G, \sigma_G, \alpha_{GB}, \beta_{GB}, \sigma_{GB}, \alpha_O, \beta_O, \sigma_O), \theta^k = (\alpha_G, \beta_G, \sigma_G, \alpha_{GB}, \beta_{GB}, \sigma_{GB}, \alpha_O, \beta_O, \sigma_O), k = 2, 3, 4, 5, L_k(\theta^k)$ is defined by Eqs. (34), (35), the vector $p_k(\theta^k)' \equiv \frac{\partial p_1(\theta^k)}{\partial \theta^k}$, and the matrix $p_k(\theta^k)'' \equiv \left[\frac{\partial^2 p_k(\theta^k)}{\partial \theta_i^k \partial \theta_j^k} \right]^T$, we get that $n^{1/2}(\hat{\theta}_n^k - \theta^k) \xrightarrow{d} N(0, [i(\theta^k)]^{-1})$, where the Fisher’ information matrix is

$$i(\theta^k) = E(p'(\theta^k)p'(\theta^k)^T) = -E(p''(\theta^k)) \quad (36)$$

The standard errors of the maximum likelihood estimators are defined by the diagonal elements of the matrix $[i(\theta^k)]^{-1}$, $k = 1, 2, 3, 4, 5$.

The performance statistics of the stem taper equations for the diameter and the volume included four statistical indices: mean absolute prediction bias (MAB) [18], precision (P) [18], the least squares-based Akaike' [26] information criterion (AIC) and a coefficient of determination (R^2). The AIC can generally be used for the identification of an optimum model in a class of competing models.

3 Results and Discussion

We focus on the modeling of Scots pine (*Pinus Sylvestris*) tree data set. Scots pine trees dominate Lithuanian forests, growing on *Arenosols* and *Podzols* forest sites and covering 725,500 ha. Stem measurements for 300 Scots pine trees were used for volume and stem profile models analysis. All section measurements include of 3,821 data points. Summary statistics for the diameter outside the bark at breast height (D), total height (H), volume (V) and age (A) of all the trees used for parameters estimate and models comparison are presented in Table 1.

To test the compatibility between taper and volume equations of all used stem taper models, the observed and predicted volume values from the sampled trees were calculated in the following form

$$V_i = \frac{\pi}{40000} \left(\sum_{k=1}^{n_i-2} \frac{(d_{ik}^2 + d_{ik+1}^2 + d_{ik} \cdot d_{ik+1}) \cdot L_{ik}}{3} + \frac{d_{in_i-1}^2 \cdot L_{in_i-1}}{3} \right) \quad (37)$$

Using the estimation data set, the parameters of SDEs stem taper Models 1–5 were estimated by the maximum likelihood procedure. Estimation results are presented in Table 2. All parameters of the Models 1–5 are highly significant ($p < 0.001$).

To test the reliability of all the tested stem taper models, the observed and predicted volume values for the sampled trees were calculated by Eq. (37). Table 3 lists the fit statistics for the new developed stem taper and volume models. The best values of the fit statistics were produced by the stem taper Models 2 and 3 with fixed tree butt.

Another way to evaluate and compare the stem taper and volume models is to examine the graphics of the residuals at different predicted diameters and volumes. The residuals are the differences between the measured and predicted diameters

Table 1 Summary statistics

Data	Tree number	Min	Max	Mean	St. Dev.
D (cm)	300	6.3	53.8	24.6	9.9
H (m)	300	5.6	34.5	20.6	5.4
V (m ³)	300	0.01	3.21	0.58	0.57
A (yr)	300	23	161	77.2	25.8

Table 2 Estimated parameters (standard errors in parentheses)

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
α_G	-2.5496 (0.0923)	-2.3414 (0.0396)	-2.3414 (0.0396)	2.5567 (0.1624)	2.5567 (0.1624)
β_G	22.5646 (1.3237)	19.7111 (0.6374)	19.7111 (0.6374)	-6.5159 (1.6853)	-6.5159 (1.6853)
σ_G	0.3546 (0.0105)	0.3389 (0.0083)	0.3389 (0.0083)	0.3299 (0.0144)	0.3299 (0.0144)
α_{GB}	-1.0255 (0.0213)	-1.0243 (0.0212)	-1.0243 (0.0212)	-1.0750 (0.0142)	-1.0750 (0.0142)
σ_{GB}	0.2703 (0.0045)	0.2688 (0.0044)	0.2688 (0.0044)	0.1954 (0.0031)	0.1954 (0.0031)
α_O	-1.6762 (0.0236)	-1.6676 (0.0241)	2.5288 (0.0464)	-1.6691 (0.0237)	2.5273 (0.0463)
β_O	17.2475 (7.0385)	11.4237 (3.7533)	0.3535 (0.0173)	15.9351 (6.2025)	0.3514 (0.0171)
σ_O	0.2140 (0.0050)	0.2147 (0.0051)	0.1916 (0.0046)	0.2148 (0.0051)	0.1917 (0.0046)
γ	1.2937 (0.0236)	-	-	-	-

Table 3 Fit statistics for all the tested stem taper and volume models^a

Model	MAB	P	AIC	R ²	Count
<i>Taper models</i>					
M. 1	1.2679	1.6769	35523	0.9769	3821
M. 2	1.1033	1.4757	34404	0.9825	3821
M. 3	0.9565	1.3803	33869	0.9848	3821
M. 4	1.2043	1.8096	36059	0.9730	3821
M. 5	1.0924	1.7657	35870	0.9743	3821
<i>Volume models</i>					
M. 1	0.0445	0.0735	153	0.9837	300
M. 2	0.0438	0.0664	58	0.9881	300
M. 3	0.0412	0.0654	77	0.9873	300
M. 4	0.0421	0.0696	124	0.9852	300
M. 5	0.0423	0.0698	125	0.9851	300

^a The best values of fit statistics for all the taper and volume models are in bold

and volumes. Graphical diagnostics of the residuals for the stem taper and volume predictions indicated that the residuals calculated using the SDEs stem taper Model 3 had more homogeneous variance than the other models.

Taper profiles for three randomly selected Scots pine trees (diameters outside the bark at breast heights of 6.3 cm, 17.0 cm, 40.7 cm, and total tree heights of 6.8 m, 21.1 m, 30.3 m) were constructed using SDEs stem taper Models 2, 3 and are plotted in Fig. 1. Figure 1 includes the stem taper curves and the standard deviation curves. It is clear that all of the taper curves followed the stem data very closely.

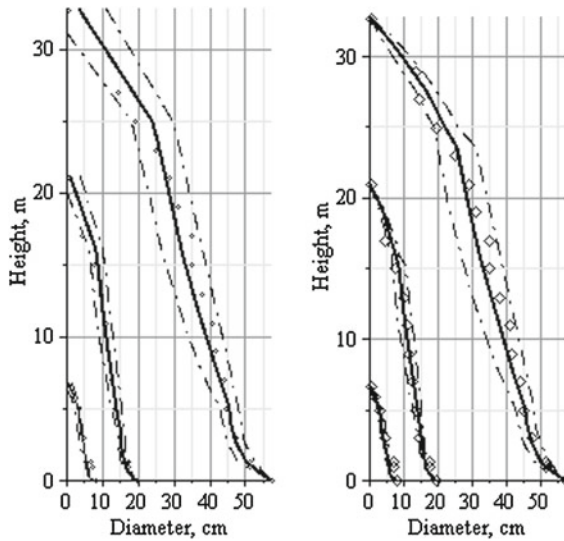


Fig. 1 Stem tapers and standard deviations for three randomly selected trees generated using the SDEs stem taper Models 2 (*left*) and 3 (*right*): *solid line*—taper curve; *dash dot line*—standard deviation of a tree diameter; *diamond*—observed data

4 Conclusion and Future Work

The new taper models were developed using SDEs. Comparison of the predicted stem taper and volume values calculated using SDEs Models 2 and 3 with the values obtained using the other models revealed a comparable predictive power of the stem taper Model 3 with fixed stem butt.

The SDEs approach allows us to incorporate new tree variables, mixed-effect parameters, and new forms of stochastic dynamics.

The variance functions developed here can be applied generate weights in every linear and nonlinear least squares regression stem taper model the weighted least squares form.

Finally, stochastic differential equation methodology may be of interest in diverse of areas of research that are far beyond the modelling of tree taper and volume.

References

1. Jana D, Chakraborty S, Bairagi N (2012) Stability, nonlinear oscillations and bifurcation in a delay-induced predator-prey system with harvesting. *Eng Lett* 20(3):238–246
2. Boughamoura W, Trabelsi F (2011) Variance reduction with control variate for pricing asian options in a geometric Levy model. *IAENG Int J Appl Math* 41(4):320–329

3. Bartkevičius E, Petrauskas E, Rupšys P, Russetti G (2012) Evaluation of stochastic differential equations approach for predicting individual tree taper and volume, Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering, WCE 2012, 4–6 July 2012, vol 1. London, pp 611–615
4. Suzuki T (1971) Forest transition as a stochastic process. Mit Forstl Bundesversuchsanstalt Wein 91:69–86
5. Tanaka K (1986) A stochastic model of diameter growth in an even-aged pure forest stand. J Jpn For Soc 68:226–236
6. Rupšys P, Petrauskas E, Mažeika J, Deltuvas R (2007) The gompertz type stochastic growth law and a tree diameter distribution. Baltic For 13:197–206
7. Rupšys P, Petrauskas E (2009) Forest harvesting problem in the light of the information measures. Trends Appl Sci Res 4:25–36
8. Rupšys P, Petrauskas E (2010) The bivariate Gompertz diffusion model for tree diameter and height distribution. For Sci 56:271–280
9. Rupšys P, Petrauskas E (2010) Quantifying tree diameter distributions with one-dimensional diffusion processes. J Biol Syst 18:205–221
10. Rupšys P, Bartkevičius E, Petrauskas E (2011) A univariate stochastic gompertz model for tree diameter modeling. Trends Appl Sci Res 6:134–153
11. Rupšys P, Petrauskas E (2012) Analysis of height curves by stochastic differential equations. Int J Biomath 5(5):1250045
12. Rupšys P, Petrauskas E, Bartkevičius E, Memgaudas R (2011) Re-examination of the taper models by stochastic differential equations. Recent advances in signal processing, computational geometry and systems theory pp 43–47
13. Kozak A, Munro DD, Smith JG (1969) Taper functions and their application in forest inventory. For Chron 45:278–283
14. Max TA, Burkhart HE (1976) Segmented polynomial regression applied to taper equations. For Sci 22:283–289
15. Kozak A (2004) My last words on taper equations. For Chron 80:507–515
16. Trincado J, Burkhart HE (2006) A generalized approach for modeling and localizing profile curves. For Sci 52:670–682
17. Yang Y, Huang S, Meng SX (2009) Development of a tree-specific stem profile model for White spruce: a nonlinear mixed model approach with a generalized covariance structure. Forestry 82:541–555
18. Rupšys P, Petrauskas E (2010) Development of q-exponential models for tree height, volume and stem profile. Int J Phys Sci 5:2369–2378
19. Westfall JA, Scott CT (2010) Taper models for commercial tree species in the Northeastern United States. For Sci 56:515–528
20. Petrauskas E, Rupšys P, Memgaudas R (2011) Q-exponential variable form of a stem taper and volume models for Scots pine *Pinus Sylvestris* in Lithuania. Baltic For 17(1):118–127
21. Itô K (1936) On stochastic processes. Jpn J Math 18:261–301
22. Oksendal BK (2002) Stochastic differential equations, an introduction with applications. Springer, New York, p 236
23. Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian motion. Physical Rev 36:823–841
24. Max TA, Burkhart HE (1976) Segmented polynomial regression applied to taper equations. For Sci 22:283–289
25. Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A-Math Phys Eng Sci 222:309–368
26. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Computing Compressible Two-Component Flow Systems Using Diffuse Interface Method

A. Ballil, S. A. Jolgam, A. F. Nowakowski and F. C. G. A. Nicolleau

Abstract Numerical simulation of compressible two-component flows that consider different materials and physical properties is conducted. An explicit finite volume numerical framework based on an extended second order Godunov approach is developed and implemented to solve an Eulerian type mathematical model. This model consists of five partial differential equations in one space dimension and it is known as the transport reduced model. A fixed Eulerian mesh is considered and the hyperbolic problem is tackled using a robust and efficient HLL Riemann solver. The performance of the numerical solver is verified against a comprehensive suite of numerical and experimental case studies in multi-dimensional space. Computing the evolution of interfaces between two immiscible fluids is considered as a major challenge for the present model and the numerical technique. The achieved numerical results demonstrate a very good agreement with all reference data.

Keywords Compressible multi-component flows · Godunov approach · HLL Riemann solver · Interface evolution · Shock wave · Shock bubble interaction

A. Ballil (✉) · S. A. Jolgam · A. F. Nowakowski · F. C. G. A. Nicolleau
Sheffield Fluid Mechanics Group, Mechanical Engineering Department,
University of Sheffield, Sheffield S1 3JD, UK
e-mail: a.ballil@Sheffield.ac.uk

S. A. Jolgam
e-mail: mep08saj@Sheffield.ac.uk

A. F. Nowakowski
e-mail: a.f.nowakowski@Sheffield.ac.uk

F. C. G. A. Nicolleau
e-mail: f.nicolleau@Sheffield.ac.uk

1 Introduction

The numerical simulation of the creation and evolution of interfaces in compressible multi-component flows is a challenging research issue. Multi-component flows occur in several industries and engineering operations such as power generation, separation and mixing processes and the inertial confinement fusion [1]. Computation of this type of flow is complicated and causes some difficulties in various engineering applications such as safety of nuclear reactors [2]. Compressible multi-component flows can be represented numerically by two main approaches. These are: Sharp Interface Method (SIM) and Diffuse Interface Method (DIM). The main characteristic of the DIM is that it allows numerical diffusion at the interface. The DIM corresponds to different mathematical models and various successful numerical approaches: for instance, a seven equation model with two velocities and two pressures developed in [3]; a five equation model proposed in [4] known as the transport reduced model; a similar five equation model was derived from the generic seven equation model in [5] and other two reduced models derived in [6]. This paper introduces the development of the numerical algorithm which utilizes the mathematical model for compressible two-component flows first presented in [4]. The performance of the reduced mathematical models was investigated in [5] and [7] using classical benchmark test problems and Roe type solver. The performance of a numerical framework that has been developed based on this model using HLL and HLLC Riemann solvers has been examined in [8].

In this work an extension of our work in [8] has been made. Computation of compressible two-component flows with different materials and tracking the evolution of the interface between two immiscible fluids is the main aim of the present work. An extended numerical approach has been developed for tracking the interface evolution. The mathematical equations and the main procedures of the numerical framework have been stated for two-dimensional flow systems. The results have been re-demonstrated in the two dimensional case studies with more details. We also have extended the investigation of the performance of the developed numerical algorithm by computing a numerically challenging shock-bubble interaction problem and compare the results with available experimental data. Shock-bubble interaction is a well known multi-component flow phenomenon. It is common in many engineering applications; for example, during supersonic combustion in ramjet engine.

In the framework of multi-component flows with interface evolution and shock bubble interaction many interesting experiments have been carried out. For example, experiments to observe the interaction between a plane shock wave and various gas bubbles were presented in [9]. The deformation of a spherical bubble impacted by a plane shock wave via a multiple exposure shadow-graph diagnostic was examined in [10]. Quantitative comparisons between the experimental data and numerical results of shock-bubble interactions were made in [11]. On the other hand, many numerical simulations of compressible multi-component flows that consider the evolution of the interface have been made. For instance, a numerical method based on upwind schemes was introduced and applied to several two phase flows test problems in [12].

The interaction of the shock wave with various Mach numbers with a cylindrical bubble was investigated numerically in [13]. An interface interaction method for compressible multifluids was developed in [14]. An efficient method to simulate and track fluid interfaces called A front-tracking method was presented in [15]. A new finite-volume interface capturing method was introduced for simulation of multi-component compressible flows with high density ratios and strong shocks in [16].

This paper is organized as follows: The governing equations of the two component flow model are reviewed. The major steps of the numerical method are then described with the HLL Riemann solver. The obtained results are presented. Finally, the conclusion is made.

2 The Mathematical Equations

2.1 The Transport Reduced Model

The two-component flow model that has been considered in this work consists of six equations in 2d flows. It is structured as: Two continuity equations, two mixture momentum equations, a mixture energy equations augmented by a volume fraction equation.

Without mass and heat transfer the model can be written as follows:

$$\begin{aligned}
 \frac{\partial \alpha_1}{\partial t} + u \frac{\partial \alpha_1}{\partial x} + v \frac{\partial \alpha_1}{\partial y} &= 0, \\
 \frac{\partial \alpha_1 \rho_1}{\partial t} + \frac{\partial \alpha_1 \rho_1 u}{\partial x} + \frac{\partial \alpha_1 \rho_1 v}{\partial y} &= 0, \\
 \frac{\partial \alpha_2 \rho_2}{\partial t} + \frac{\partial \alpha_2 \rho_2 u}{\partial x} + \frac{\partial \alpha_2 \rho_2 v}{\partial y} &= 0, \\
 \frac{\partial \rho u}{\partial t} + \frac{\partial (\rho u^2 + P)}{\partial x} + \frac{\partial \rho u v}{\partial y} &= 0, \\
 \frac{\partial \rho v}{\partial t} + \frac{\partial \rho u v}{\partial x} + \frac{\partial (\rho v^2 + P)}{\partial y} &= 0, \\
 \frac{\partial \rho E}{\partial t} + \frac{\partial (u(\rho E + P))}{\partial x} + \frac{\partial (v(\rho E + P))}{\partial y} &= 0.
 \end{aligned} \tag{1}$$

The notations are conventional: α_k and ρ_k characterize the volume fraction and the density of the k th component of the flow, ρ , u , v , P and E represent the mixture density, the mixture velocity component in x -direction, the mixture velocity component in y -direction, the mixture pressure and the mixture total energy respectively.

The mixture variables can be defined as:

$$\begin{aligned}
 \rho &= \alpha_1 \rho_1 + \alpha_2 \rho_2 \\
 u &= (\alpha_1 \rho_1 u_1 + \alpha_2 \rho_2 u_2) / \rho \\
 v &= (\alpha_1 \rho_1 v_1 + \alpha_2 \rho_2 v_2) / \rho \\
 P &= \alpha_1 P_1 + \alpha_2 P_2 \\
 E &= (\alpha_1 \rho_1 E_1 + \alpha_2 \rho_2 E_2) / \rho.
 \end{aligned}$$

2.2 Equation of State (EOS)

In the present work, the isobaric closure is used with stiffened equation of state to close the model. The mixture stiffened (EOS) can be cast in the following form:

$$P = (\gamma - 1) \rho e - \gamma \pi, \quad (2)$$

where e is the internal energy, γ is the heat capacity ratio and π is the pressure constant.

The mixture equation of state parameters γ and π can be written as:

$$\frac{1}{\gamma - 1} = \sum_k \frac{\alpha_k}{\gamma_k - 1} \quad \text{and} \quad \gamma \pi = \frac{\sum_k \frac{\alpha_k \gamma_k \pi_k}{\gamma_k - 1}}{\sum_k \frac{\alpha_k}{\gamma_k - 1}},$$

where k refers to the k th component of the flow.

The internal energy can be expressed in terms of total energy as follows:

$$E = e + \frac{1}{2}u^2 + \frac{1}{2}v^2.$$

Finally, the mixture sound speed for isobaric closure has been calculated via:

$$c = \frac{\sum y_k \varepsilon_k c_k^2}{\varepsilon} \quad (3)$$

where, y_k is the mass fraction and it is given by $y_k = \alpha_k \rho_k / \rho$, c_k is the speed of sound of the k th fluid and $\varepsilon_k = 1/(\gamma_k - 1)$.

2.3 Quasi-Linear Equations of the Reduced Model

In two-dimensional flow with two fluids, the system of Eq. (1) can be re-written in quasi-linear form with primitive variables in the following compact form:

$$\frac{\partial W}{\partial t} + A(W) \frac{\partial W}{\partial x} + B(W) \frac{\partial W}{\partial y} = 0 \quad (4)$$

where the primitive vector W and the Jacobian matrices $A(W)$ and $B(W)$ for this system can be written as:

$$W = \begin{bmatrix} \alpha_1 \\ \rho_1 \\ \rho_2 \\ u \\ v \\ P \end{bmatrix}, \quad A(W) = \begin{bmatrix} u & 0 & 0 & 0 & 0 & 0 \\ 0 & u & 0 & \rho_1 & 0 & 0 \\ 0 & 0 & u & \rho_2 & 0 & 0 \\ 0 & 0 & 0 & u & 0 & 1/\rho \\ 0 & 0 & 0 & 0 & u & 0 \\ 0 & 0 & 0 & \rho c^2 & 0 & u \end{bmatrix}$$

and

$$B(W) = \begin{bmatrix} v & 0 & 0 & 0 & 0 & 0 \\ 0 & v & 0 & 0 & \rho_1 & 0 \\ 0 & 0 & v & 0 & \rho_2 & 0 \\ 0 & 0 & 0 & v & 0 & 0 \\ 0 & 0 & 0 & 0 & v & 1/\rho \\ 0 & 0 & 0 & 0 & \rho c^2 & v \end{bmatrix}.$$

The Jacobian matrix $A(W)$ provides the following eigenvalues: $u + c, u, u, u, u$ and $u - c$, whereas the Jacobian matrix $B(W)$ provides the following eigenvalues: $v + c, v, v, v, v$ and $v - c$, which represent the wave speeds of the system.

3 Numerical Method

The numerical algorithm for 2d problems is developed using an extended Godunov approach with the classical MUSCL scheme to achieve second order accuracy in space and time. The splitting scheme is applied for the discretization of the conservative vector in two time steps as follows:

$$U_{i,j}^{n+\frac{1}{2}} = U_{i,j}^n - \frac{\Delta t}{\Delta x} \left[F^n \left(U^* \left(U_{i+\frac{1}{2},j}^-, U_{i+\frac{1}{2},j}^+ \right) \right) - F^n \left(U^* \left(U_{i-\frac{1}{2},j}^-, U_{i-\frac{1}{2},j}^+ \right) \right) \right] \text{ and}$$

$$U_{i,j}^{n+1} = U_{i,j}^{n+\frac{1}{2}} - \frac{\Delta t}{\Delta y} \left[G^{n+\frac{1}{2}} \left(U^* \left(U_{i,j+\frac{1}{2}}^-, U_{i,j+\frac{1}{2}}^+ \right) \right) - G^{n+\frac{1}{2}} \left(U^* \left(U_{i,j-\frac{1}{2}}^-, U_{i,j-\frac{1}{2}}^+ \right) \right) \right].$$

The flux vectors in x -direction $F(U^*)$ and in y -direction $G(U^*)$ have been calculated using HLL Riemann solver, which was first presented in [17] and described in the context of the Riemann problem with details in [18].

Similarly, the discretization of the volume fraction equation with second order accuracy can be written as:

$$\alpha_{i,j}^{n+\frac{1}{2}} = \alpha_{i,j}^n - u \frac{\Delta t}{\Delta x} \left[\alpha^{*(n)} \left(\alpha_{i+\frac{1}{2},j}^-, \alpha_{i+\frac{1}{2},j}^+ \right) - \alpha^{*(n)} \left(\alpha_{i-\frac{1}{2},j}^-, \alpha_{i-\frac{1}{2},j}^+ \right) \right] \text{ and}$$

$$\alpha_{i,j}^{n+1} = \alpha_{i,j}^{n+\frac{1}{2}} - v \frac{\Delta t}{\Delta y} \left[\alpha^{*(n+\frac{1}{2})} \left(\alpha_{i,j+\frac{1}{2}}^-, \alpha_{i,j+\frac{1}{2}}^+ \right) - \alpha^{*(n+\frac{1}{2})} \left(\alpha_{i,j-\frac{1}{2}}^-, \alpha_{i,j-\frac{1}{2}}^+ \right) \right].$$

The numerical time step size Δt is chosen as in [18]:

$$\Delta t = CFL \times \min \left(\frac{\Delta x}{S_x}, \frac{\Delta y}{S_y} \right),$$

where CFL is the Courant-Friedrichs-Lewy number ($CFL \leq 1$, to insure the stability of the numerical method), S_x and S_y are the maximum wave speeds in the x and y directions respectively and they can be expressed as:

$$S_x = \max(0, u_{i\pm\frac{1}{2},j}^+ + c_{i\pm\frac{1}{2},j}^+, u_{i\pm\frac{1}{2},j}^- + c_{i\pm\frac{1}{2},j}^-),$$

$$S_y = \max(0, v_{i,j\pm\frac{1}{2}}^+ + c_{i,j\pm\frac{1}{2}}^+, v_{i,j\pm\frac{1}{2}}^- + c_{i,j\pm\frac{1}{2}}^-).$$

3.1 2D Form of the HLL Approximate Riemann Solver

With HLL Riemann solver, the numerical flux function at a cell boundary in x -direction can be written as follows:

$$F_{i+\frac{1}{2},j}^{HLL} = \begin{cases} F_{i,j} & \text{if } 0 \leq S_{XL}, \\ \frac{S_{i+\frac{1}{2},j}^+ F_{i,j} - S_{i+\frac{1}{2},j}^- F_{i+1,j} + S_{i+\frac{1}{2},j}^+ S_{i+\frac{1}{2},j}^- (U_{i+1,j} - U_{i,j})}{S_{i+\frac{1}{2},j}^+ - S_{i+\frac{1}{2},j}^-} & \text{if } S_{XL} \leq 0 \leq S_{XR}, \\ F_{i+1,j} & \text{if } 0 \geq S_{XR} \end{cases}$$

In the similar way the numerical flux function at a cell boundary in y -direction can be written as follows:

$$G_{i,j+\frac{1}{2}}^{HLL} = \begin{cases} G_{i,j} & \text{if } 0 \leq S_{YL}, \\ \frac{S_{i,j+\frac{1}{2}}^+ G_{i,j} - S_{i,j+\frac{1}{2}}^- G_{i,j+1} + S_{i,j+\frac{1}{2}}^+ S_{i,j+\frac{1}{2}}^- (U_{i,j+1} - U_{i,j})}{S_{i,j+\frac{1}{2}}^+ - S_{i,j+\frac{1}{2}}^-} & \text{if } S_{YL} \leq 0 \leq S_{YR}, \\ G_{i,j+1} & \text{if } 0 \geq S_{YR} \end{cases}$$

where subscripts S_{XR} and S_{XL} denotes to right and left wave speeds at each cell boundary in x -direction . Whereas, S_{YR} and S_{YL} denotes to right and left wave speeds at each cell boundary in y -direction.

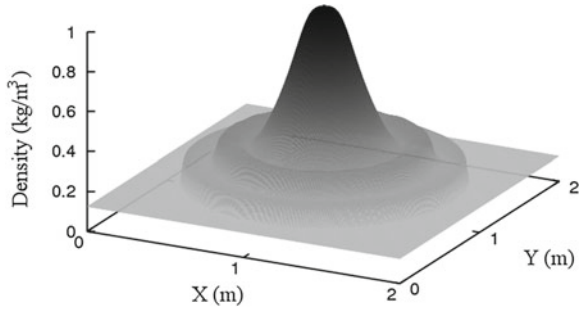
4 Test Problems

Four different test cases have been considered to observe the evolution of the interface and to assess the numerical algorithm that is developed in this work. These cases consider different initial states and physical properties, which provide different flow

Table 1 Initial conditions for the explosion test

Physical property	Bubble	Surrounding fluid
Density, kg/m^3	1	0.125
X-Velocity, m/s	0	0
Y- Velocity, m/s	0	0
Pressure, Pa	1	0.1
Heat capacity ratio, γ	1.4	1.4

Fig. 1 Evolution of density profile at $t = 0.25$ s for the explosion test



regimes. The results obtained from the first three cases have been compared with other numerical results which are generated using different models and numerical methods. In the fourth case the current results have been compared with available experimental data.

4.1 Explosion Test

This test is a single phase problem and the reduced model of the two-phase flows is applied for this test. In this test the two flow components stand for the same fluid which produce extreme conditions. The physical domain of this problem is a square of 2×2 m, which contains a circular bubble of 0.8 m in diameter located at the center of the domain. The initial condition is demonstrated in Table 1. The computation was made using 300×300 cells and the periodic boundary conditions (B.C) were considered.

The surface plots for density and pressure distribution at time $t = 0.25$ s are illustrated in Figs. 1 and 2 respectively. The current results are significantly close to the results that published in [18]. This confirms that the reduced model reproduces the physical behavior of the flow components with stiff initial conditions.

4.2 Interface Translation Test

The computational domain for this case study is a square of 1×1 m includes a circular interface of 0.32 m in diameter separates two fluids. The center of the bubble

Fig. 2 Evolution of pressure profile at $t = 0.25$ s for the explosion test

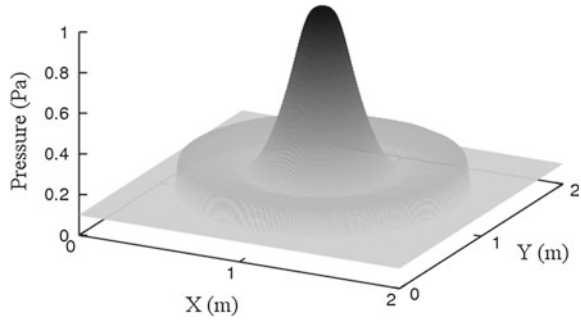
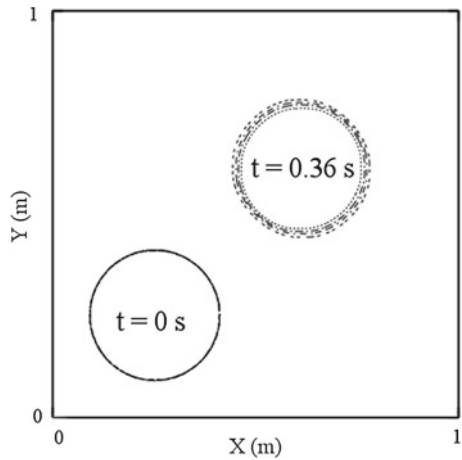


Table 2 Initial conditions for the interface test

Physical property	Bubble	Surrounding fluid
Density, kg/m^3	1	0.1
X-Velocity, m/s	1	1
Y- Velocity, m/s	1	1
Pressure, Pa	1	1
Heat capacity ratio, γ	1.4	1.6

Fig. 3 Volume fraction contours at the initial time $t = 0$ s and at time $t = 0.36$ s for the interface translation test



is located at 0.25, 0.25 m. The initial conditions for this test are stated in Table 2. The computation was made using 300×300 cells and periodic B.C.

The results are shown in Fig. 3 for volume fraction contours at the initial time $t = 0$ s and at the time $t = 0.36$ s. The results show the time interval during which the bubble has moved with a uniform velocity and pressure from its initial position to a new location where the center of the bubble has the coordinates (0.61, 0.61) m. The shape of the bubble can be compared with the numerical results presented in [19].

Table 3 Initial conditions for the under-water explosion test

Physical property	Bubble	Surrounding fluid
Density, kg/m^3	1.241	0.991
X-Velocity, m/s	0	0
Y- Velocity, m/s	0	0
Pressure, Pa	2.753	$3.059e^{-4}$
Heat capacity ratio, γ	1.4	5.5
Pressure constant, π	0	1.505

Fig. 4 Density evolution at time $t = 0.058$ s for the explosion under-water test

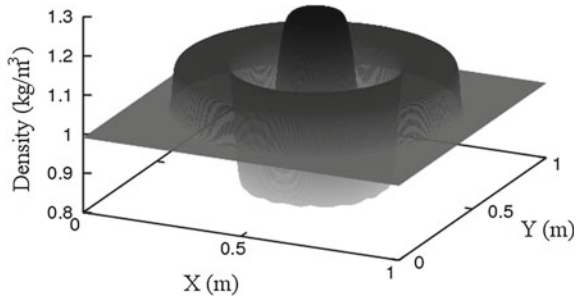
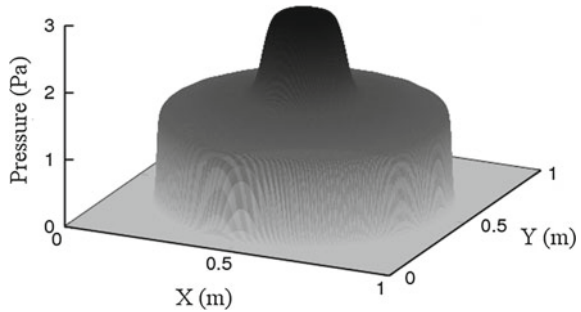


Fig. 5 Pressure distribution at time $t = 0.058$ s for the explosion under-water test



4.3 Bubble Explosion Under-Water Test

This test has been considered by many researchers, for example [15] and [19]. The computational domain of this test problem is a square of dimension 1×1 m, which including a bubble of 0.4 m in diameter located in the center of the domain. The initial state is shown in Table 3. The simulation was made using 300×300 cells and periodic B.C.

The surface plots for mixture density and pressure are presented in Figs. 4 and 5 respectively. The numerical results obtained are compared with the equivalent numerical results that published in [15] and [19]. The current results demonstrate a good compatibility with the other results. The numerical solutions obtained characterize and capture the physical behavior and the evolution of the interface correctly.

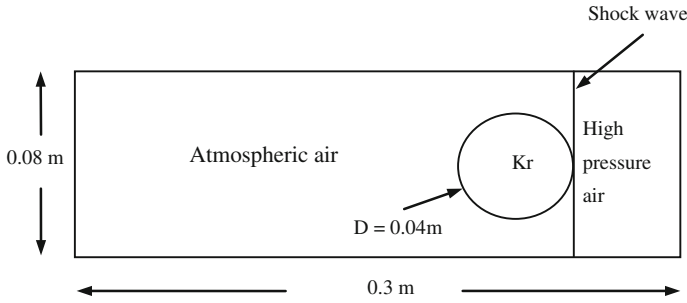


Fig. 6 Schematic diagram shows the physical domain of shock-bubble interaction test

Table 4 Initial conditions for shock-bubble interaction test

Physical property	Krypton bubble	Pre-shock air	Post-shock air
Density, kg/m^3	3.506	1.29	2.4021
X-Velocity, m/s	0	0	230.28
Y- Velocity, m/s	0	0	0
Pressure, Pa	101325	101325	249091
Heat capacity ratio, γ	1.67	1.4	1.4

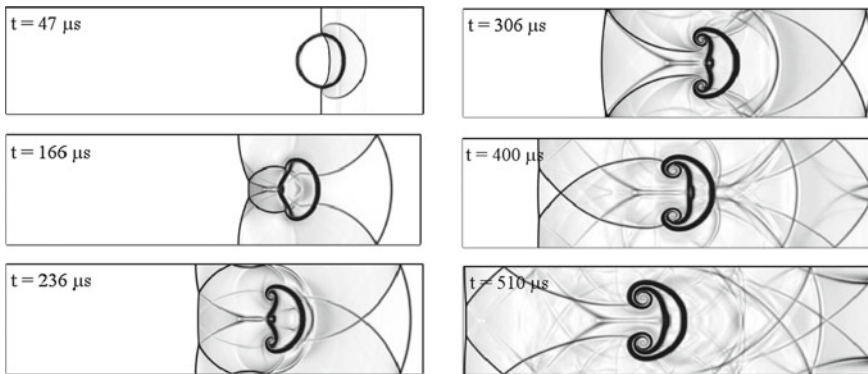


Fig. 7 The mixture density contours for the krypton bubble-air constitution at different times $t = 47\mu\text{s}$, $166\mu\text{s}$, $236\mu\text{s}$, $306\mu\text{s}$, $400\mu\text{s}$ and $510\mu\text{s}$

4.4 Validation of Shock-Bubble Interaction

Here the interaction between a moderate shock wave ($\text{Mach} = 1.5$) and Krypton gas bubble surrounded by air at atmospheric pressure has been simulated. The current numerical results have been compared with the experimental results in [11]. A schematic diagram of the initial physical state is illustrated in Fig. 6 and the initial conditions are shown in Table 4.

The results are demonstrated in Fig. 7. The evolution of the bubble contour with time due to the interaction with the shock wave is observable and it is in a good agreement with experimental results in [11]. It can be noticed that the present numerical method has the mechanism for tracking the physical phenomena that have been occurrence within the domain of the interaction. At the early stage of the interaction, one can notice a shock wave transmitting inside the gas bubble, incident shock outside the bubble and a reflection wave propagating backward to the right side. The deformation of the bubble is influenced by the differences in densities between the krypton and surrounded air especially in the early stages of the interaction. In the middle time a high speed penetrating jet, which moving towards the right side has generated on the line of symmetry of the bubble. At relatively later stages the effect of the vorticity on the interface deformation has appeared. One can observe the vortices that have been generated on the top and at the bottom of the bubble contour.

5 Conclusion

Numerical simulations of compressible flows between two immiscible fluids have been performed successfully. The numerical algorithm for these simulations has been developed based on an extended Godunov approach with HLL solver considering second order accuracy. The performance of the considered multi-component flow model and the numerical method has been verified effectively. This has been made using a set of carefully chosen case studies which are distinguished by a variety of compressible flow regimes. The obtained results show that the developed algorithm is able to reproduce the physical behavior of the flow components efficiently. Consequently, it could be applied to simulate a wide range of compressible multi-component flows with different materials and physical properties.

References

1. Lindl JD, McCrory RL, Campbell EM (1992) Progress toward ignition and burn propagation in inertial confinement fusion. *Phys Today* 45(9):32–40
2. Nowakowski A, Librovich B, Lue L (2004) Reactor safety analysis based on a developed two-phase compressible flow simulation. In: *Proceedings of the 7th Biennial conference on engineering systems design and analysis, ESDA 2004, Manchester, U.K., 19–22 July 2004, vol 1*, pp 929–936
3. Saurel R, Abgrall R (1999) A multiphase Godunov method for compressible multifluid and multiphase flows. *J Comput Phys* 150(2):425–467
4. Allaire G, Clerc S, Kokh S (2000) A five-equation model for the numerical simulation of interfaces in two-phase flows. *C. R. Acad Sci—Series I: Mathematics* 331(12):1017–1022
5. Murrone A, Guillard H (2005) A five-equation model for the simulation of interfaces between compressible fluids. *J Comput Phys* 202(2):664–698

6. Kapila AK, Menikoff R, Bdzil JB, Son SF, Stewart DS (2001) Two-phase modeling of deflagration to detonation transition in granular materials: reduced equations. *Phys Fluids* 13(10):3002–3024
7. Allaire G, Clerc S, Kokh S (2002) A five-equation model for the simulation of interfaces between compressible fluids. *J Comput Phys* 181(2):577–616
8. Ballil A, Jolgam S, Nowakowski AF, Nicoleau FCGA (2012) Numerical simulation of compressible two-phase flows using an Eulerian type reduced model. In: Proceedings of the world congress on engineering WCE 2012. Lecture Notes in Engineering and Computer Science. U.K, London, 4–6 July 2012, pp 1835–1840
9. Haas JF, Sturtevant B (1987) Interaction of weak shock waves with cylindrical and spherical gas inhomogeneities. *J Fluid Mech* 181:41–76
10. Layes G, Jourdan G, Houas L (2003) Distortion of a spherical gaseous interface accelerated by a plane shock wave. *Phys Rev Lett* 91(17):174502
11. Layes G, Le Métayer O (2007) Quantitative numerical and experimental studies of the shock accelerated heterogeneous bubbles motion. *Phys Fluids* 19:042105
12. Coquel F, Amine KE, Godlewski E, Perthame B, Rascle P (1997) A numerical method using upwind schemes for the resolution of two-phase flows. *J Comput Phys* 136:272–288
13. Bagabir A, Drikakis D (2001) Mach number effects on shock-bubble interaction. *Shock Waves* 11(3):209–218
14. Hu X, Khoo B (2004) An interface interaction method for compressible multifluids. *J Comput Phys* 198:35–64
15. Terashima H, Tryggvason G (2010) A front-tracking method with projected interface conditions for compressible multi-fluid flows. *Comput Fluids* 39:1804–1814
16. Shukla RK, Pantano C, Freund JB (2010) An interface capturing method for the simulation of multi-phase compressible flows. *J Comput Phys* 229:7411–7439
17. Harten A, Lax PD, Leer BV (1983) On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review* 25(1):35–61
18. Toro E (1999) Riemann solvers and numerical methods for fluid dynamics. Springer
19. Shyue K (1998) An efficient shock-capturing algorithm for compressible multi-component problems. *J Comput Phys* 142(1):208–242

Turbulent Boundary Layer Gas–Solid Flow Based on Two-Fluid Model

Hassan Basirat Tabrizi

Abstract Motion of Particles in a dilute turbulent boundary layer on a flat wall was simulated numerically. Eulerian-Eulerian two-way coupled model was used. Thermophoretic force and Brownian diffusion effects were investigated on the depositions of fine particles in a turbulent boundary layer. Turbulence closure was achieved by Prandtl's mixing length model. The set of equations was solved numerically by using finite difference method. Introduced particle diffusion term played a significant role in numerical convergence. The proposed two-fluid approach for evaluating the effect of different forces on small particle deposition from a turbulent flow over a flat plate produced similar finding compared to Lagrangian method and computationally less expensive.

Keywords Brownian · Deposition · Effective viscosity · Modeling · Particle diffusion · Simulation · Thermophoresis · Two-way coupling

1 Introduction

Prediction for deposition of droplet or particle is needed in many natural and industrial processes such as microchip manufacturing, chemical coating of metals, gas filtration, and heat exchangers. In most practical applications, transport occurs in the presence of turbulent flow and may be influenced by uncontrolled factors of temperature and humidity. To name a few example; Achebo [1] studied on computational analysis of erosion wear rate in a pipeline by using the drift flux model based on Eulerian continuum equations. Recently, Bernad et al. [2] investigated particle hemodynamics analysis after coronary angioplasty by using Lagrangian-Eulerian method. In a flow at presence of solid surface, a boundary layer will develop and the

H. Basirat Tabrizi (✉)
Department of Mechanical Engineering, Amirkabir University of Technology,
P.O. Box 15875-4413, 159163411 Tehran, Iran
e-mail: hbasirat@aut.ac.ir

energy and momentum transfer gives rise to temperature and velocity gradient. This convection by the bulk flow plays a significant role for particle deposition. Therefore, such flows should be model in the near wall region as accurately as possible. Generally, two basic approaches are used to predict gas-solid behavior. Lagrangian method treats trajectories of individual particles. Eulerian method, the dispersed phase is regarded as a continuum, and balances of mass, momentum and energy are written in differential form for both phase. Crow et al. [3] reported a comprehensive review of both modeling approaches when turbulence is important, the two-fluid model is computationally more efficient and this method is adopted here. The particles on entering close to wall, deposition will occur mostly due to inertial and sedimentary mechanism. The wall and boundary layer flow plays an important role on the deposition of particles. It slows particles due to the lift forces and in problems involving heat transfer in a layer due to thermophoresis force, which happens in a non-isothermal gas, in the opposite direction of the temperature gradient. Thus, particles tend to move toward cold areas and be repulsed from hot areas. This phenomenon, known as thermophoresis, is of considerable importance in particle deposition problems where it can play either beneficial or adverse roles.

Soo [4] discussed the case of laminar and turbulent boundary motion of gas-solid on a flat plate. Kallio et al. [5] presented a Lagrangian random-walk approach to modeling particle deposition in turbulent duct flows. Particle concentration profiles reveal that smaller particles tend to accumulate in the near-wall region due to the sudden damping of fluid turbulence. Eulerian model had been employed by Taniere et al. [6] to analyze the particle motion and mass flux distributions in the mixed regime in a turbulent boundary layer two-phase flow where saltation and turbulence effects were important. Recently, particle behavior in the turbulent boundary layer of a dilute gas-solid flow was studied experimentally and numerically using CFD software by Wang et al. [7]. Their results indicated a non-uniform distribution of particle concentration with the peak value inside the turbulent boundary layer on the flat plate. They related this phenomenon to the slip-shear lift force and particle-wall collision. They used Lagrangian numerical method and did not consider the non isothermal gas effect. The slip-shear lift force effect for larger particles was shown considerable ($d_p = 200 \mu\text{m}$).

Goren [8] gave a detailed theoretical analysis of thermophoretic deposition of aerosol particles in the laminar boundary layer on a flat plate. Jayaraj [9] studied the particle concentration boundary layer due to thermophoresis for a wide range of parameters. Greenfield et al. [10] simulated the particle deposition in a turbulent boundary layer in the presence of thermophoresis with a test case involving heated pipe using Lagrangian approach. Kroger et al. [11] studied the deposition of log-normally distributed particles in isothermal and heated turbulent boundary layer via Lagrangian random walk simulations. The velocity and temperature fields and thermophoretic force are considered Gaussian random fields. Their results showed that introduction of temperature gradient lead to a strong increase in particle deposition. Slater et al. [12] predicted the deposition rate of small particles for two-dimensional turbulent gas flows onto solid boundaries by employing a fully Eulerian two-fluid approach. The particle-density-weighted averaging of particle equation motion was

used which generate fewer turbulence correlations. He et al. [13] developed a numerical simulation procedure for studying deposition of aerosol particles in duct flows including the effect of thermal force under laminar and turbulent conditions. A semi empirical model had been presented which covers the thermophoretic effects for the entire range of Knudsen number. Thermally enhanced deposition velocities of particles in a turbulent duct flow of a hot gas had also been evaluated.

It can be noticed most researches adopted Lagrangian approach whereby the particle equations are integrated along particle path lines. However Shahriari and Basirat Tabrizi [14] studied two-fluid modeling of turbulent boundary layer of disperse medium on a flat plate by using the particulate phase diffusivity term with the Prandtl's mixing length theory. Further, two differential equation turbulent model was used by Gharraei et al. [15]. Small effect due to higher closure turbulence model and easy convergence due to the diffusion term for deposition of small particles was noticed. Further Shahriari and Basirat Tabrizi [16] studied the effect of different temperature gradients on particle deposition across the boundary layer where Brownian effects neglected due to large size particle. Similar approach is applied to predict the particle behavior in gas-solid turbulent boundary layer flow on a flat plate by taking into account Brownian diffusion and thermophoresis forces here. Prandtl's mixing length model is introduced for fluid turbulence.

2 Formulation and Solution

Two-dimensional turbulent boundary layer flow of fine spherical particles-fluid suspension past a horizontal flat plate is studied via coupled two-fluid model. The fluid gas is considered to be steady, incompressible, with constant properties. The volume fraction of particulate material is assumed to be small. Prandtl's mixing length model is used for turbulence closure model of the flowing gas. Effect of the aerodynamic lift and the rotation of particles on deposition are assumed small compare to the effect of the Brownian and thermophoretic force. The viscous, pressure terms and particle-particle collisions in the particulate phase momentum equation can be neglected. In addition, the particle diffusion term is considered in this modeling. Eulerian-Eulerian approach described by Soo [4] modified, leading to boundary layer approximations which are applied to fluid and solid phase. The governing equations of gas phase are:

$$\frac{\partial \bar{u}_f}{\partial x} + \frac{\partial \bar{v}_f}{\partial y} = 0 \quad (1)$$

$$\begin{aligned} \bar{u}_f \frac{\partial \bar{u}_f}{\partial x} + \bar{v}_f \frac{\partial \bar{u}_f}{\partial y} \\ = \frac{\partial}{\partial y} \left((v + v_t) \frac{\partial \bar{u}_f}{\partial y} \right) - \frac{\rho_p}{\rho_f} F (\bar{u}_f - u_p) \end{aligned} \quad (2)$$

$$\bar{u}_f \frac{\partial \bar{T}_f}{\partial x} + \bar{v}_f \frac{\partial \bar{T}_f}{\partial y} = \frac{\partial}{\partial y} \left(\left(\alpha + \frac{\nu_t}{Pr_t} \right) \frac{\partial \bar{T}_f}{\partial y} \right) \quad (3)$$

Here u, v bar stands for average velocity in x, y and T bar for average temperature. ν is kinematic viscosity, α is thermal diffusivity, Pr is Prandtl's number and subscript f, p, t stands for fluid, particle and turbulent respectively.

Solid phase flow governing equations are governed by:

$$\frac{\partial (\rho_p u_p)}{\partial x} + \frac{\partial (\rho_p v_p)}{\partial y} = (D_B + D_p) \frac{\partial^2 \rho_p}{\partial y^2} \quad (4)$$

$$u_p \frac{\partial u_p}{\partial x} + v_p \frac{\partial u_p}{\partial y} = \frac{D_p}{\rho_p} \left(\frac{\partial}{\partial y} \left(\rho_p \frac{\partial u_p}{\partial y} \right) \right) + \frac{F(\bar{u}_f - u_p)}{C_c} + F_{B,x} \quad (5)$$

$$u_p \frac{\partial v_p}{\partial x} + v_p \frac{\partial v_p}{\partial y} = \frac{D_p}{\rho_p} \left(\frac{\partial}{\partial x} \left(\rho_p \frac{\partial u_p}{\partial y} \right) + 2 \frac{\partial}{\partial y} \left(\rho_p \frac{\partial v_p}{\partial y} \right) \right) + \frac{F(\bar{v}_f - v_p)}{C_c} + F_{Th} + F_{B,y} \quad (6)$$

The thermophoretic force per unit particle mass can be written as [17]:

$$F_{Th} = - \frac{18\mu_f \nu_f \kappa_T}{d_p^2 \rho_{mp}} \frac{1}{T_f} \frac{\partial T_f}{\partial y} \quad (7)$$

where κ_T is a coefficient, which depends on the ratio of the gas and particle thermal conductivities and Knudsen number (Kn).

$$\kappa_T = \frac{2C_s(k_g/k_p + C_t Kn)}{(1 + 3C_m Kn)(1 + 2k_g/k_p + 2C_t Kn)} \quad (8)$$

Here $Kn = \frac{2\lambda}{d_p}$, $C_s = 1.17$, $C_t = 2.18$, $C_m = 1.14$, k thermal conductivity, λ is mean free path, μ is dynamic viscosity, ρ density, d_p particle diameter and subscripts g, p, mp stands for gas, particle and particle mass respectively.

Soo [4] gives the Brownian force per unit particle mass:

$$F_{B,x} = -F \frac{D_B}{\rho_p} \frac{\partial \rho_p}{\partial x}; \quad F_{B,y} = -F \frac{D_B}{\rho_p} \frac{\partial \rho_p}{\partial y} \quad (9)$$

where D_B , is the Brownian diffusion coefficient and equal to $K_B T / 3\pi \mu d_p$ and Cunningham correction factor, C_c follows (see Schlichting [18]):

$$C_c = 1 + Kn [1.257 + 0.4 \exp(-1.1/Kn)] \quad (10)$$

For simulation convenience, Basirat Tabrizi and Kermani [19] introduced dimensionless quantities.

Turbulent or eddy viscosity, ν_t is calculated by employing Prandtl's mixing length model (see Boothroyd [20]):

$$\nu_t = LU_\infty \sqrt{\text{Re}_L} (\bar{l}_m)^2 \left| \frac{\partial \bar{U}_f}{\partial \bar{Y}} \right| \quad (11)$$

$$\bar{l}_m = 0.41 \frac{\bar{Y}}{\sqrt{\text{Re}_L}} \left[1 - \exp\left(-\frac{\bar{Y}L\sqrt{\tau_w/(\rho_f \text{Re}_L)}}{26\nu}\right) \right] \quad (12)$$

where τ_w the shear stress at the wall, U_∞ free stream velocity, L length of plate, Re is Reynolds number.

The skin friction coefficient defined:

$$C_f = \frac{\tau_w}{\frac{1}{2}\rho_f U_\infty^2} \quad (13)$$

τ_w is wall shear stress.

Moreover, the displacement thickness follows:

$$\bar{\delta}_f^* = \int_0^\infty (1 - \bar{U}_f) d\bar{Y} \quad (14)$$

Here subscript w stands for wall. The boundary conditions in dimensionless form are:

$$\begin{aligned} \bar{Y} = 0 : \bar{U}_f = \bar{V}_f = \bar{V}_p = 0, \bar{T}_f = 1 \\ \bar{U}_p \frac{\partial \bar{U}_p}{\partial \bar{X}} = \frac{D_p}{\bar{\rho}_p \nu} \left(\frac{\partial}{\partial \bar{Y}} (\bar{\rho}_p \frac{\partial \bar{U}_p}{\partial \bar{Y}}) \right) \\ - \frac{\bar{U}_p \bar{F} \bar{\rho}_f}{C_c} + \bar{F}_{B,x} \\ \frac{\partial (\bar{\rho}_p \bar{U}_p)}{\partial \bar{X}} = \frac{(D_B + D_p)}{\nu} \frac{\partial^2 \bar{\rho}_p}{\partial \bar{Y}^2} \end{aligned} \quad (15a)$$

$$\begin{aligned} \bar{Y} \rightarrow \infty : \bar{U}_f = \bar{U}_p = \bar{\rho}_p = 1, \\ \frac{\partial \bar{V}_f}{\partial \bar{Y}} = \frac{\partial \bar{V}_p}{\partial \bar{Y}} = \frac{\partial \bar{T}_f}{\partial \bar{Y}} = 0 \end{aligned} \quad (15b)$$

$$\bar{X} = 0 : \bar{U}_f = \bar{U}_p = \bar{\rho}_p = \bar{T}_f = 1, \bar{V}_f = \bar{V}_p = 0 \quad (15c)$$

$$\bar{X} \rightarrow \infty :$$

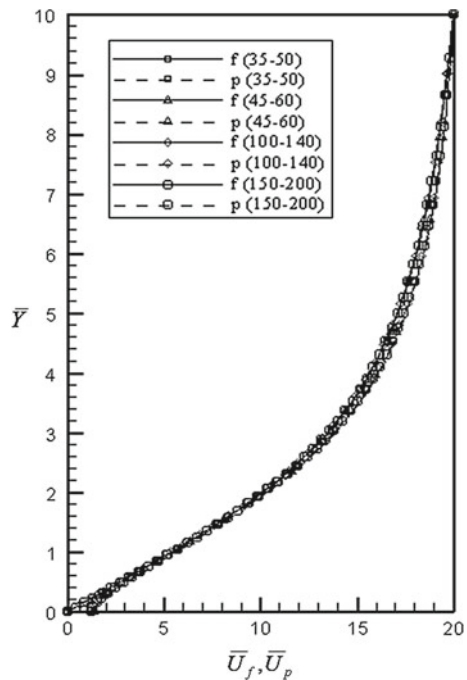
$$\frac{\partial \bar{U}_f}{\partial \bar{X}} = \frac{\partial \bar{U}_p}{\partial \bar{X}} = \frac{\partial \bar{V}_f}{\partial \bar{X}} = \frac{\partial \bar{V}_p}{\partial \bar{X}} = \frac{\partial \bar{T}_f}{\partial \bar{X}} = \frac{\partial \bar{\rho}_p}{\partial \bar{X}} = 0 \quad (15d)$$

Term, D_p shows particle flux due to particle diffusion and can be estimated from the turbulent Schmidt number $Sc_t = \nu_t / D_t$ which might take values close to unity or else. The effect of particle diffusion term was discussed without thermophoresis and Brownian force elsewhere (see Shahriari and Basirat Tabrizi [16]).

3 Numerical Procedure

The hydrodynamic and thermal boundary layer equations and their boundary conditions are solved numerically using finite difference scheme. A two-dimensional logarithmic mesh form is superimposed on the flow field. Different mesh sizes are compared to obtain grid-independent results for inlet velocity of 20 m/s and 2.5 μm particle sizes. It is shown in Fig. 1 maximum error due to different computational mesh sizes reveals in the near-wall region. A 100×140 mesh is selected as the

Fig. 1 Mesh comparison [19]



optimum mesh regarding to the computational time. The minimum value of Δy has set to be 0.03. Coarse mesh sizes lead to unphysical results and convergence problems.

Gas and solid continuity equations and boundary conditions for particle density along the plate are solved implicitly while other governing equations and boundary conditions are implied explicitly. Computation proceeds up to the steady state state that is considered to reach for a relative change of field variables less than 10^{-6} .

4 Results and Discussion

Since, there exist no coupled two-fluid approach for this kind of modeling even without the diffusion part, so the model predictions without the thermophoresis effect have been compared to the only available experimental and theoretical results by Taniere et al. [6]. Here, the physical property of those reported is used for the model simulation. They used a two-fluid model and neglected the effect of particle diffusion term; instead, a turbulence source term was used. Inlet velocity of 10.6 m/s, particle size $60\ \mu\text{m}$ with density of $2500\ \text{Kg/m}^3$, loading ratio $\beta = 0.1$ at location of 1.89 m. Figure 2 indicates this comparison. The particle diffusion term used in the present model is taken to be equal to the thermal diffusivity of particle material (for glass, $D_p/\nu = 0.05$). The velocities ($u^+ = \bar{u}/u_\tau$) are expressed in wall distance unit ($y^+ = yu_\tau/\nu$). The model prediction seems to be in a closer agreement to their results. In addition, it can be seen the influence of particle diffusion is qualitatively well predicted.

Further comparisons were shown in our previous studied [14, 15]. They indicated that the diffusion term and Prandtl’s mixing length theory are well suited with the experimental observation.

Fig. 2 Model comparison of gas and particle mean velocity profiles with Taniere et al. [6]

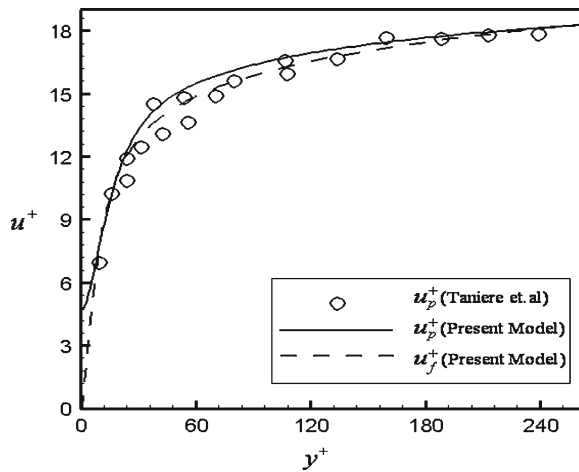


Fig. 3 Solid line gas velocity, dash line particle velocity, square pure gas velocity, black dot particle density on the surface, upper figure at $X/L = 0.5$ and lower figure at $X/L = 5$

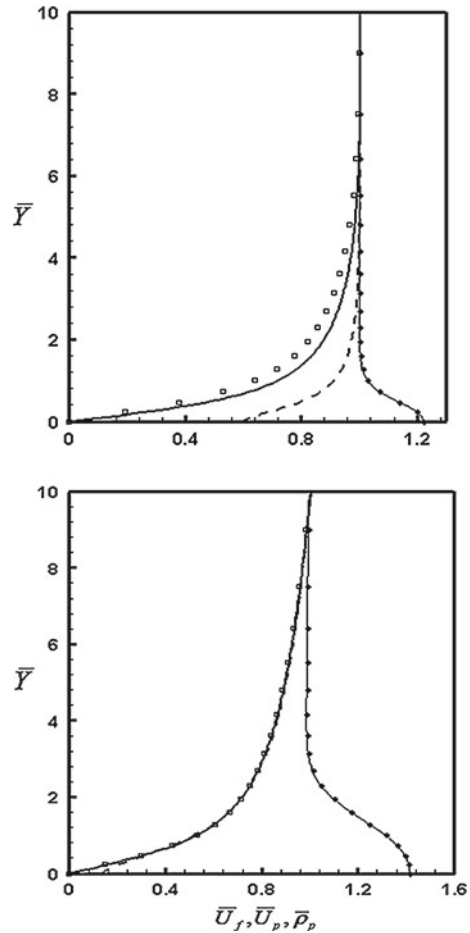


Figure 3 shows typical gas and solid velocity and particle density along the wall at different locations without temperature gradient. This is for inlet velocity of 30 m/s, loading ratio of 1.0 and particle density and size of 1608 Kg/m³, 10 μm respectively. Three different regions can be recognized accordingly. For $X/L \ll 1$ particle velocity profile differs greatly from fluid because of particle inertia. Along the plate, this difference is reducing due to fluid-particle interactions. At $X/L \gg 1$ both phases are approaching to each other. The particle wall density is also compared in this figure. Particle density adopts a maximum at $X/L \approx 1$.

This point on the numerical simulation is carried with $D_p/\nu = 0.05$, loading ratio of 1.0, and particle size 2.5 μm with density of 2500 Kg/m³ unless state else.

Figure 4 shows effect of D_p/ν on particle density profiles without temperature difference. Small D_p/ν corresponds to almost no particle diffusion. In addition, approaching D_p/ν to zero leads to incapability of numerical method to solve flow

Fig. 4 Effect of particle density on the surface

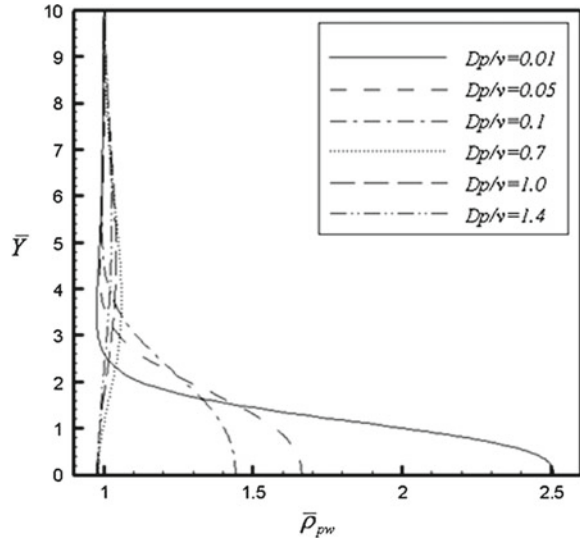
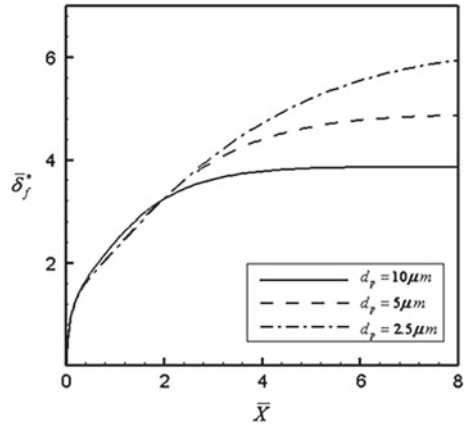


Fig. 5 Effect of particle size on the fluid displacement thickness



governing equations and needs to modify numerical scheme for better convergence. However, indeed experiments should be conducted in the near-wall region to gain a proper value for D_p/ν .

Varying parameters such as loading ratio, particle size and particle density on the surface were discussed in previous studied [14–16, 19]. Here, mainly is focused on particle size effect on the skin friction coefficient and displacement thickness, which is important for design criteria.

Figures 5 and 6 shows the effect of particle size on the fluid displacement thickness and the skin friction factor due to the thermophorsis force for temperature difference of 50 K. It indicates the smaller size particle has lower displacement thickness and

Fig. 6 Particle size effect on the skin friction coefficient

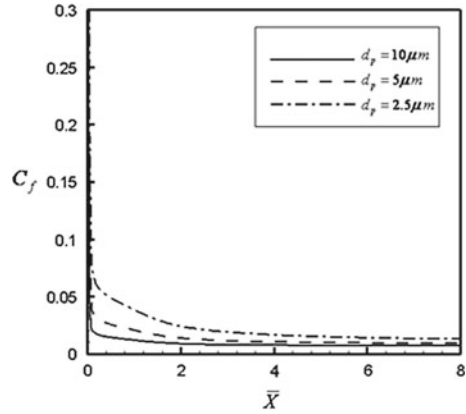
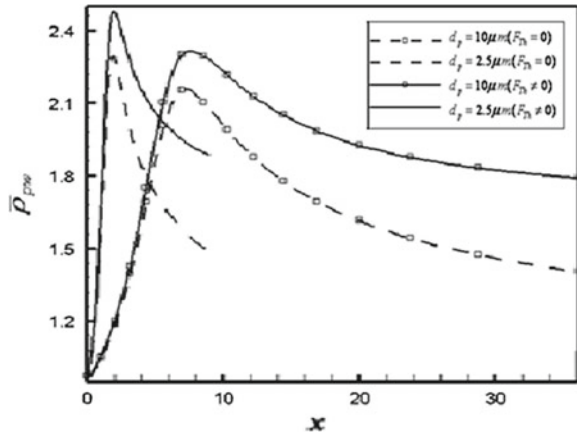


Fig. 7 Surface particle density



adjusts itself to the main gas-solid stream rapidly. In addition, for smaller size particle due to higher deposition, the skin friction factor is higher than larger size particle.

Further effect of some parameters, such as the skin friction coefficient, thermal displacement thickness of gas and solid flow with and without thermophoresis or temperature variations’ were discussed in Ref. [19].

To conclude the above study, one should investigate the surface particle density distribution affects with and without thermophoresis force. Figure 7 shows higher deposition with the thermophoresis force at temperature difference of 300 K. In addition, it indicates higher value for the smaller size particle and closer to the entrance region.

Further, the kinetic theory of granular flow was applied to the transport equations of the solid-phase. In addition, one-equation model for turbulent kinetic energy was used for gas-phase turbulence. Inter-particle and particle-wall inelastic collisions were modeled by means of restitution coefficients; e and e_w respectively

(see Dehghan and Basirat Tabrizi [21]). Ignoring the thermophoresis force and considering particles viscosity and the accuracy of viscosity simulation have some influence on particulate-phase velocity especially near the wall. Further research needs to incorporate the thermophoresis force and its effect.

5 Conclusion and Future Work

Turbulent gas-solid flow on a horizontal flat plate by using the two-way coupling Eulerian–Eulerian approach was discussed. Prandtl’s mixing length theory was used as closure model for gas turbulence. In addition, for the particle deposition at the wall the particle diffusion term was introduced. Because of the thermophoretic force, particle deposition increased which was predicted by using other method of approaches such as Eulerian-Lagrangian [10]. Presenting the dispersed phase within the concept of Eulerian approach and adding the particle diffusion term for the dispersed phase, the problem of convergence was solved much easier.

Changing the temperature difference between plate surface and free stream had great influence on particle deposition on the flat plate. The larger particle delayed the deposition peak further distance away from the plate front edge. The skin friction coefficient increased and the displacement thickness decreased. This can be considered in design criteria.

In addition, the current approach on small particle deposition from a turbulent flow over a flat plate uses less computing time and memories.

Recently, the kinetic theory of granular flow and one-equation model for turbulent kinetic energy for gas-phase turbulence was used by ignoring the thermophoresis force (see Dehghan and Basirat Tabrizi [21]). It should be more precise for predicting dispersion and deposition on the wall. However, needs more verifications and modeling to be done on the granular part and especially on the effect of thermophoresis force.

Acknowledgments The author would like to thank his graduate students specially Ms. G. Kermani, Mr. Sh. Shahriari, Mr. R. Gharraei, Mr. M. Dehghan and many others who worked through this research.

References

1. Achebo JI (2009) Computational analysis of erosion wear rate in a pipe line using the drift flux models based on Eulerian continuum equations. In: Proceedings of the world congress on engineering 2009, vol I, WCE, London, UK, pp 719–721, 1–3 July 2009
2. Bernad SI, Totoream AF, Vinatu VF, Susan-Resiga RF (2010) Particle hemodynamics analysis after coronary angioplasty. In: Proceedings of the world congress on engineering 2010, vol I, WCE, London, UK, pp 583–588, 30 June–2 July 2010
3. Crowe CT, Trout JN (1996) Numerical models for two-phase turbulent flows. *Ann Rev Fluid Mech* 28:11–43

4. Soo SL (1989) Particulates and continuum: multiphase fluid dynamics, HPC
5. Kallio GA, Reeks MW (1989) A numerical simulation of particle deposition in turbulent boundary layers. *Int J Multiph Flow* 15(3):433–446
6. Taniere A, Oesterle B, Foucaut JM (1997) Modeling of saltation and turbulence effects in a horizontal gas-solid boundary layer. *Part Sci Technol* 14:337–350
7. Wang J, Levy EK (2006) Particle behavior in the turbulent boundary layer of a dilute gas-particle flow past a flat plate. *Exp Therm Fluid Sci* 30:473–483
8. Goren SL (1977) Thermophoresis of aerosol particles in the laminar boundary layer on a flat surface. *J Colloid Interface Sci* 61:77–85
9. Jayaraj S (1995) Thermophoresis in laminar flow over cold inclined plates with variable properties. *Heat Mass Transf* 30:167–173
10. Greenfield C, Quarini G (1998) A Lagrangian simulation of particle deposition in a turbulent boundary layer in the presence of thermophoresis. *Appl Math Model* 22:759–771
11. Kroger C, Drossinos Y (2000) A random-walk simulation of thermophoretic particle deposition in a turbulent boundary layer. *Int J Multiph Flow* 26:1325–1350
12. Slater SA, Leeming AD, Young JB (2003) Particle deposition from two-dimensional turbulent gas flows. *Int Multiph Flow* 29:721–750
13. He C, Ahmadi G (1998) Particle deposition with thermophoresis in laminar and turbulent duct flows. *Aerosol Sci Tech* 29:525–546
14. Shahriari Sh, Basirat Tabrizi H (2003) Modeling of turbulent boundary layer of a disperse medium on a flat plate. In: *Proceedings of the IASTED2003, international conference, applied simulation and modeling*, ACTA Press, 2003, pp 7–12
15. Gharraei R, Basirat Tabrizi H, Esmailzadeh E (2004) Prediction of turbulent boundary layer flow of a particulate suspension using $\kappa - \tau$ model. In: *Proceedings of canadian society of mechanical engineers (CSME) 2004 Forum*, pp 29–35
16. Shahriari Sh, Basirat Tabrizi H (2007) Two-fluid model simulation of thermophoretic deposition for fine particles in a turbulent boundary layer. *ASME-JSME2007 thermal engineering summer heat transfer conference, HT2007-32117*, vol 2, 2007, pp 849–854
17. Talbot L, Cheng RK, Schefer RW, Willis DR (1980) Thermophoresis of particles in a heated boundary layer. *J Fluid Mech* 101:737–758
18. Schlichting H (1979) *Boundary layer theory*. McGraw-Hill, New York
19. Basirat Tabrizi H, Kermani G (2012) Thermophoresis of particles over flat plate turbulent flow based on two-fluid modeling. In: *Proceedings of the world congress on engineering 2012*, vol III, WCE, London, UK, pp 1675–1680, 4–6 July 2012
20. Boothroyd RG (1971) *Flowing gas-solids suspensions*. Chapman and Hall, London
21. Dehghan M, Basirat Tabrizi H (2012) On near-wall behavior of particles in a dilute turbulent gas-solid flow using kinetic theory of granular flow. *Powder Technol* 224:273–280

Molten Carbonate Fuel Cell as a Reducer of CO₂ Emissions from Gas Turbine Power Plants

Jaroslav Milewski, Rafal Bernat and Janusz Lewandowski

Abstract A Molten Carbonate Fuel Cell (MCFC) is shown to reduce CO₂ emissions from a Gas Turbine Power Plant (GTTP). The MCFC is placed in the flue gas stream of the gas turbine. The main advantages of this solution are: higher total electricity generated by a hybrid system and reduced CO₂ emissions with power generation efficiency remained the same. The model of the MCFC is given and described. The results obtained show that use of an MCFC could reduce CO₂ emissions by 73.

Keywords CO₂ sequestration · Emissions · Engineering · Fuel cells · Gas turbine · Modeling · Molten carbonate fuel cell · Optimization · Power plants

1 Introduction

Fuel cells are considered to be one of the most prospective electricity sources. They are thought to replace mobile phones and computer batteries, become eventual drives for cars, and produce electricity in distributed power plants of larger scale. Various types of fuel cells may be distinguished by different catalysts, different ions being the proton carriers, different operational temperatures and different fuels that may be used. In general, we may recognize low- (e.g. Polymer Exchange Fuel Cells [13]) and high-temperature fuel cells [1] and among the latter Solid Oxide Fuel Cells (SOFC) [9, 15] and Molten Carbonate Fuel Cells (MCFC) [3, 4]. They are both of high

J. Milewski (✉) · R. Bernat · J. Lewandowski
Institute of Heat Engineering, Warsaw University of Technology, ul. Nowowiejska 21/25,
00-665 Warszawa, Poland
e-mail: jaroslaw.milewski@itc.pw.edu.pl

R. Bernat
e-mail: rafal.bernat@itc.pw.edu.pl

J. Lewandowski
e-mail: janusz.lewandowski@itc.pw.edu.pl

efficiencies and have the priceless feature of methane utilization—already available fuel [7] including bio-fuels [10]. Others need hydrogen feeding whose production requires significant amount of energy. Additionally, high temperature fuel cell can be coupled to gas turbine for ultra-efficient power generation. Alternatively, the small units based on fuel cells can be utilized as power sources in a Distributed Generation system.

Furthermore, MCFCs enable to concentrate carbon dioxide [11, 17], e.g. from coal [16] or gas fired power plants, and might become a part of a Carbon Capture and Storage system [8]. Operation of MCFC requires flow of CO_3^{2-} as the proton carrier through the electrolyte. This is achieved by feeding CO_2 to the cathode, where it reacts and flows as CO_3^{2-} to the anode. There, after another reaction, it becomes carbon dioxide again and, after removing water vapor, may be transported as pure gas to the storage point. One may say that MCFCs work as a filter allowing exclusive flow of CO_2 (Fig. 1).

The European Union has placed limits on CO_2 emissions by Member States as a part of its Emission Trading Scheme [2]. This impacts fossil fuel power plants to a significant degree as their emissions are governed by the number of emission allowances they receive from the Member State allocation. Excess CO_2 emissions have to be covered by purchasing extra allowances, which is in effect a penalty. According to the European Energy Exchange, on the 11 of October 2012 it was 7,59 and 7,76 euro/tonne CO_2 for the primary and secondary market respectively. In contrast, undershooting emission limits enables the emitter to sell CO_2 allowances. This is possible to the end of 2012. Then it will be even more strict. From 2013 all emitters will be forced to buy emission allowances from the pool granted to the EU Member. This forces fossil based economies to develop technology adapted to the political situation. CCS is an option. However, one has to consider that carbon dioxide sequestration by, for example wet amine scrubbing requires additional energy. This results in efficiency decrease of the whole system, so in order to produce the same net amount of energy more fuel has to be used. On the contrary, Molten Carbonate Fuel Cells not only separate the gas, but also, simultaneously produce heat and electricity contributing to the total energy generation of the system. They may even increase the efficiency of the whole system.

Keeping in mind that they may as well use many fuels like hydrogen, natural gas, methanol or bio-gas it seems that it is currently feasible to apply them and consider them as extremely competitive.

It is, of course possible to combine a gas turbine with MCFC what will result in a hybrid system (HS) with increased efficiency and decreased carbon dioxide emission. The exhaust gases of a gas turbine power plant consist mainly of nitrogen, oxygen, steam and carbon dioxide. This mixture can be used as the oxidant in the MCFC (cathode feeding).

Negative ions are transferred through the molten electrolyte. Each ion is composed of one molecule of carbon dioxide, one atom of oxygen and two electrons. This means that an adequate ratio of carbon dioxide to oxygen is 2.75 (mass based) or 2.0 (mole based).

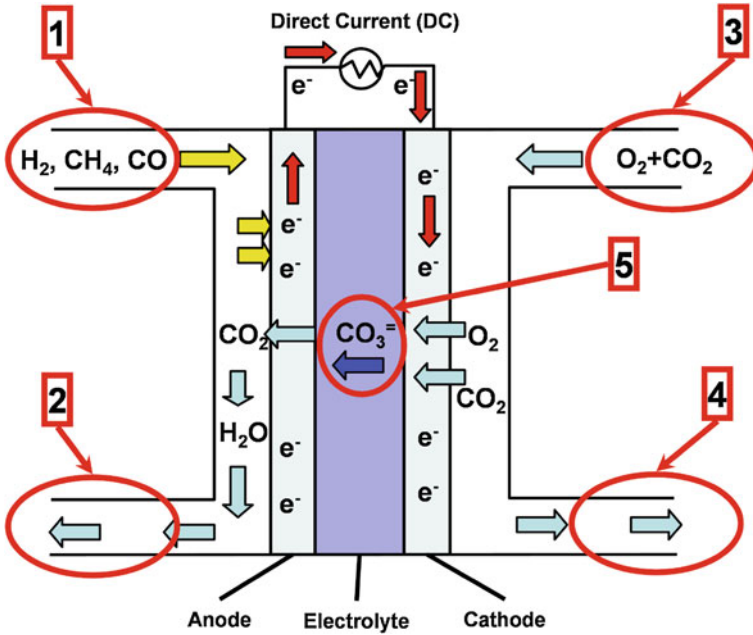


Fig. 1 Working principles of MCFC; 1 fuel input, 2 mixture of CO₂, H₂ and H₂O, 3 oxidant input, 4 exhaust, 5 ions of CO₃²⁻

Table 1 Exhaust gas composition

Component	Mass fraction (%)	Mole fraction (%)
CO ₂	5.2	3.4
H ₂ O	4.1	6.6
O ₂	15.3	13.6
N ₂	74.0	75.4
Ar	1.4	1.0
CO ₂ /O ₂	0.34	0.25

The typical gas turbine flue gas composition is shown in Table 1. The ratio of carbon dioxide to oxygen is hence 0.25 (mole based) and 0.34 (mass based). This means that flue gas contains an insufficient quantity of oxygen to trap all CO₂.

2 Mathematical Model and Optimization

As for many other engineering applications, mathematical modeling is the basic method for analyzing fuel cells systems. In order to model system elements correctly, a zero-dimensional approach was used. The parameters, that are considered to be the most significant ones in the modeling process, are briefly presented below. The model

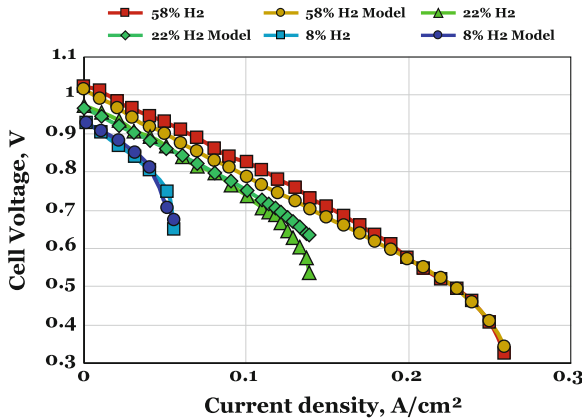
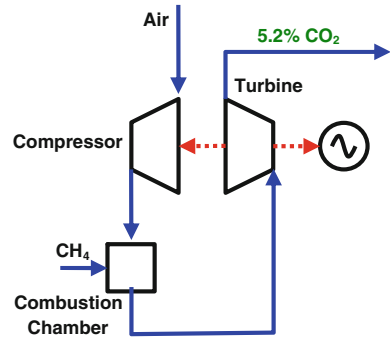


Fig. 2 Experimental and simulations data at different H₂ molar fractions, experimental data taken from [14]

Fig. 3 GTTP system



that used here is a new conceptual, alternative mathematical model of an MCFC [12]. The discrepancy between the model and experimental data given by Morita and alia are presented in the Fig. 2. It is conspicuous that it is valid for different hydrogen compositions—the error is marginal.

However, the MCFC is only a part of the plant that's emissions are to be reduced. The analyzed, sole gas turbine system is presented in the Fig. 3 (Table 2).

In the system the compressed air is delivered to the combustion chamber where fuel (natural gas) is combusted. Hot gas expands in the gas turbine and is rejected to the atmosphere. The mathematical model of the GTTP was created based on three main assumptions:

- air compressor isentropic efficiency: 79 %
- gas turbine isentropic efficiency: 88 %
- no pressure drops across the combustion chamber.

Table 2 Nominal parameters of GT power plant [6]

Name	Value
Air compressor inlet pressure (MPa)	0.1
Air compressor inlet temperature (°C)	15
Pressure ratio	17.1
Fuel	Natural gas
Fuel mass flow (kg/s)	4.0
Turbine inlet temperature (°C)	1210
Exhaust gas mass flow (kg/s)	213
Turbine outlet temperature (°C)	587
GT Power (MW)	65
GT Efficiency (LHV) (%)	33
CO ₂ annually emission (Gg/a)	250
Relative emission of CO ₂ (kg/MWh)	609
CO ₂ mass flow (kg/s)	11

In order to make the model more real a commercial gas turbine unit was chosen to analyze [6]. Nominal parameters of the GTPP and exhaust gas composition are shown in Tables 1 and 3, respectively.

To compose a CO₃²⁻ ion, it is necessary to split a half mole of O₂ with one mole of CO₂. Adequate mass and molar ratios of CO₂ to O₂ (for capture all carbon dioxide) are 1.38 and 2, respectively. However, from data given in Table 1 it seems that, theoretically, all CO₂ could be captured (some of the oxygen will be simply rejected to the atmosphere).

All analyzed cases were optimized with the objective function being total power generation efficiency. Nevertheless, there is room for discussion as to the choice of this as the objective function of the optimizing process [19]. While the main task of an MCFC is to capture CO₂ from flue gas, it also increases total power generation due to its higher efficiency compared with that of the steam cycle (44 vs. 30 %).

The size of the MCFC installed at the flue gas rejection pipelines can be varied in wide range. From the other hand the same fuel utilization ratio can be realized by fuel cells of different size. There are three main parameters which determine the MCFC size: fuel utilization factor, maximum current density and inlet fuel flow. At least two from these three parameters determine the size of the MCFC. The stack fuel utilization factor was chosen at constant level of 90 %. The maximum current density and fuel mass flow were taken as primary variables of the optimizing process.

Optimized parameters:

- MCFC fuel mass flow
- The value of i_{max} in the range 0.06–0.3 A/cm²
- Heat Exchanger efficiency in the range 0–85 %.

The optimizing process was carried out with the temperature inside the stack below 750 °C.

3 Gas Turbine Power Plant with MCFC

Two cases of gas turbine power plant with the MCFC were investigated. Case 1 concerns a situation when there is no intervention in GTTP cycle. It means that MCFC is added at GTTP outlet stream. Case 2 concerns the situation when heat exchangers before combustion chamber are added to the GTTP. These heat exchangers are fed by MCFC exhaust streams. This case, however, seems to be very difficult to apply in reality. It is obvious, that due to the design of gas turbines, there is not enough space to fix heat exchangers just after the compressor. We also have to consider that relatively low CO₂ content in flue gas results in low MCFC efficiency (about 34 % (based on Lower Heating Value, LHV)). Therefore, another option to compose low efficient MCFC with high efficient Combined Cycle Gas Turbine (with efficiency about 55 %) seems to be unreasonable and this case was not investigated.

The CO₂ reduction emission factor is defined as follows:

$$\eta_{CO_2} = 1 - \frac{\dot{m}_{CO_2,out}}{\dot{m}_{CO_2,in}} \quad (1)$$

where: \dot{m} —mass flow, kg/s; *out*—MCFC outlet cathode stream; *in*—MCFC inlet cathode stream.

What should be the objective of the optimizing process is not obvious. The MCFC is installed to capture the CO₂, from this point of view the quantity of captured CO₂ should be maximized. But on the other side, the MCFC utilizes the same fuel as the gas turbine and produces electricity and heat. From that reason both analyzed cases were optimized to obtain maximum system efficiency.

We have to remember that the fuel cells system itself is not the only equipment that has to be installed in order to separate CO₂. Additionally, we need a CO₂ separator, a catalytic burner, and a DC/AC converter. The CO₂ separator is a water cooled heat exchanger that cools down the gases that are rejected from the fuel cell. The condensate is then taken away purifying the flue gases stream to carbon dioxide only. The catalytic burner is fed by pure oxygen to utilize the rest of methane, hydrogen and carbon oxide. Naturally, oxygen extraction (e.g. from air) requires energy. The production of one kilogram of oxygen at atmospheric pressure requires from 200 to 300 kJ. The mean value of 250 kJ was taken into calculations, what is included in the model and decreases the system efficiency depending on the amount of consumed oxygen.

One may not forget that installation of MCFC at gas turbine outlet results in back pressure drop of about 1 %. It decreases the efficiency of the GTTP from 33 to 32 % (Fig. 4).

Parameters obtained during the optimizing process are given in Table 3. It is easily visible that in both cases quite high values of CO₂ reduction are achievable.

The MCFC-GTTP in Case 2 was created by adding two heat exchangers (see Fig. 5). The heat exchangers have a role to recover exhaust heat from MCFC outlet

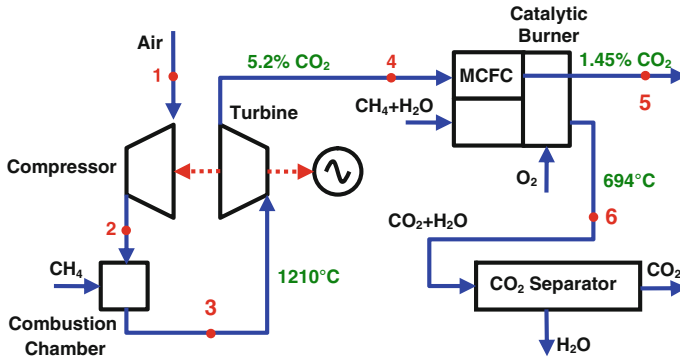


Fig. 4 GTPP-MCFC system—case 1

Table 3 Nominal parameters of GTPP-MCFC

Name	Case 1	Case 2
GTTP-MCFC power (total power) (MW)	80	77
GTTP power/total power (%)	81	82
MCFC power/total power (%)	19	18
GTTP-MCFC efficiency (LHV) (%)	33	40
CO ₂ emission reduction factor (%)	73	91
Annual CO ₂ emission (Gg/a)	67	18
Relative CO ₂ emission (kg/MWh)	132	37
MCFC efficiency (LHV) (%)	34	36
GTTP efficiency (LHV) (%)	32	41
Fuel utilization factor (%)	90	90
Average cell voltage (mV)	513	486
Current density (mA/cm ²)	29.5	29.6
Oxygen mass flow (kg/s)	0.2	0.2
MCFC/GTTP fuel ratio	0.52	0.65

streams. Note that the GTTP efficiency would increase with a recuperative heat exchanger when no MCFC is installed as well.

The system was optimized with the same conditions like Case 1. Nominal parameters of Case 2 of GTTP-MCFC system are given in Table.

GTTP-MCFC Case 2 generates slightly less power in comparison with Case 1. During the simulations a constant value of Turbine Inlet Temperature (TIT) was assumed. The implementation of heat exchangers means lower fuel mass flow demanded by the combustion chamber.

A reduction of the CO₂ emission of 91% is obtained. Simultaneously, electric efficiency is increased to 40% (LHV) what gives the relative emission of CO₂ of 37 kg/MWh.

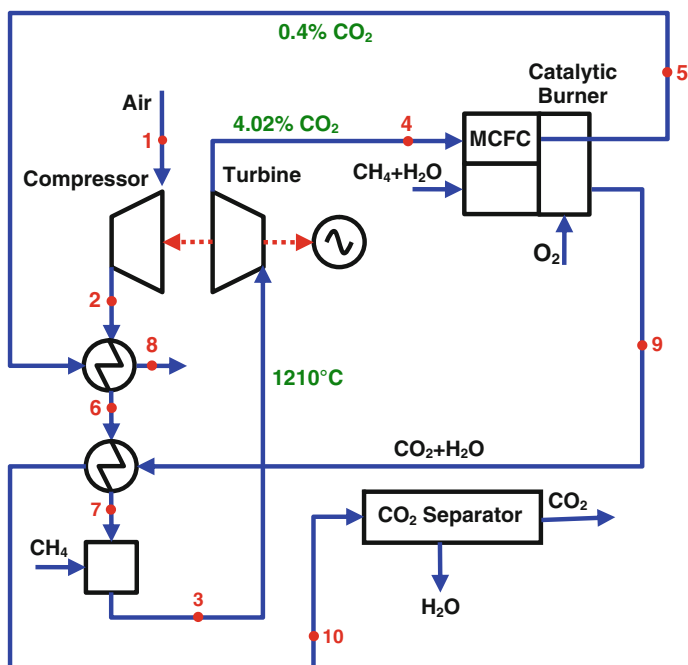


Fig. 5 GTPP-MCFC system—case 2

4 Experimental Investigation

After the mathematical model was created, there arose the need to check the possibility of CO_2 separation in practice. The laboratory of Institute of Heat Engineering at Warsaw University of Technology holds proper apparatus necessary for such investigation. Similar investigations are also being proceed in other research laboratories (e.g. [5, 18]). The MCFC tested has a planar area of 100 cm^2 where the anode was a porous (55%) Ni structure with thickness of 0.76 mm, the cathode a porous (60%) nickel oxide structure with thickness of 0.7 mm and the electrolyte a lithium carbonate and potassium carbonate— $(\text{Li}_2\text{CO}_3)_{0.62}(\text{K}_2\text{CO}_3)_{0.38}$ —mixture (three matrices of 0.3 mm each, in total 0.9 mm). The electrode-electrolyte matrix is sandwiched between the two opposing separator plates and the fuel and oxidant flow concurrently in opposing channels. The cathodic and anodic current collectors were made in the form of stainless steel embossed sheets. The cell was tested in the experimental facility, in which the cell was held in a vessel and it was possible to set up and control each operating parameter. The vaporizer system provides measurement and control of the water vapor directed to the anode and cathode. The temperature of the system is subject to external control, and local temperatures collected by thermocouples, present in various positions on the cell, are processed by the PC board. Proprietary software allows one to easily view and manipulate the

Table 4 Main parameters of the experiments performed

Variant	Fed to	Flow (ml/min/cm ²)	H ₂ (%)	H ₂ O (%)	N ₂ (%)	Air (%)	CO ₂ (%)
GT flue gases	Anode	2.29	70	13	–	–	17
+ reference point	Cathode	36.14	–	6.7	24.5	65.4	3.4
GT flue gases	Anode	1.97	70	13	–	–	17
+ reference point (optimized)	Cathode	38.22	–	6.7	24.5	65.4	3.4

condition and performance of the system. The cell fabrication was concluded by the first heat up (“cell conditioning”) during which the components assumed their final form. The cells were operated at atmospheric pressure with the same reference point gas compositions and flow rates. The fuel gas (H₂)₈₀(CO₂)₂₀ and the oxidant gas (Air)₇₀(CO₂)₃₀ were fed. The cell voltage is directly measured at the two electrodes and its value is processed by a National Instruments board. The cell resistance was measured by using a HIOKI 3560 AC mW HiTESTER (four wires, 1 kHz). The gas composition and flow rates were controlled by a set of mass flow controllers. The gases and water flow rate are measured and controlled by Brooks 5850E Digital Mass Flow Controllers, chosen for their high accuracy and for their ability to be managed by software through serial PC ports. For load demand DC electronic load (SAE Electronic Conversion SRL) was used. Three series of tests were performed at 650 °C, the temperature kept constant on the cell plane using heating plates equipped with three electric heater each. A first one aimed to analyze the simplest case, in which just flue gases are fed into cathode side. The laboratory tests were conducted at operational temperature of 650 °C and the necessary parameters, i.e. CO₂ and H₂ mass flows, voltage and current, were measured among others (Table 4).

Two cases were analyzed—one where the flue gases were directly fed on the fuel cell working on nominal parameters and the other, where the MCFC was optimized to obtain higher efficiency and CO₂ reduction rate. The voltage/current characteristic for the investigated fuel cell is presented in the Fig. 6a.

The results proved that concentration of carbon dioxide for its further separation from the flue gases is possible.

The graph shows typical performance of a Molten Carbonate Fuel Cell. However, the most important issue for this chapter is the possibility to extract carbon dioxide from the flue gases of a gas turbine power plant. The research carried out in the laboratory proved this to be possible. As one may observe in the Fig. 6b reduction rate of CO₂ from the gas turbine exhaust gases of more than 60 % was achieved. This, obviously, varies with the load that the fuel cell is subject to.

Naturally, the laboratory stand may not be compared with industrial, large-scale installations. Its operation parameters differ greatly, when compared with pilot-scale installations. For example, the efficiency of the MW-scale fuel cell systems is much greater than the one of the single, investigated cell (Fig. 7). The difference is equal to, more or less, 15 % points. Due to the small size of the equipment used it was

not possible to clearly determine the feasibility and economic profitability of such solutions. Nevertheless, it is certain that it is possible to concentrate and extract quite pure carbon dioxide from GTPP flue gases using Molten Carbonate Fuel Cells. Moreover, if a larger-scale installation would be used for this case, it could increase electricity and heat production of the whole system. This is a priceless advantage comparing to other ways of CO₂ capture.

5 Final Remarks

The CO₂ emission reduction factor and CO₂ relative emission were used to compare the systems. These values for all analyzed cases are given in the Table 5. The MCFC could reduce the CO₂ emission from gas turbine power plant exhaust by more than 70%. The relative CO₂ emission decreases more significantly because the MCFC produces additional power.

Relatively low efficiency of the MCFC is caused by low CO₂ content at gas turbine exhaust, which gives low maximum cell voltage. The combination of MCFC with GTPP requires higher investment costs. However, common CO₂ separation methods also require capital investment and, instead of producing energy they consume it.

Moreover, application of the MCFC in a Gas Turbine Power Plant gives a relatively high reduction in CO₂ emissions. The relative CO₂ emission of the GTPP is estimated at 609 kgCO₂/MWh while in contrast the MCFC-GTPP hybrid system has an emission rate of 135 kgCO₂/MWh. The quantity of CO₂ emitted by the MCFC-GTPP is 73% lower than is the case with the GTPP.

As mentioned earlier, all cases were optimized to achieve maximum power generation efficiency. However, this may be open to challenge if it is accepted that the main task of the MCFC is to limit CO₂ emissions, which would result in the CO₂ emission reduction factor being used as the objective function of the optimizing process. If this factor is optimized the cell voltage at last cell can fall below zero and the MCFC will work as a CO₂ concentrator only. At the very least, the MCFC

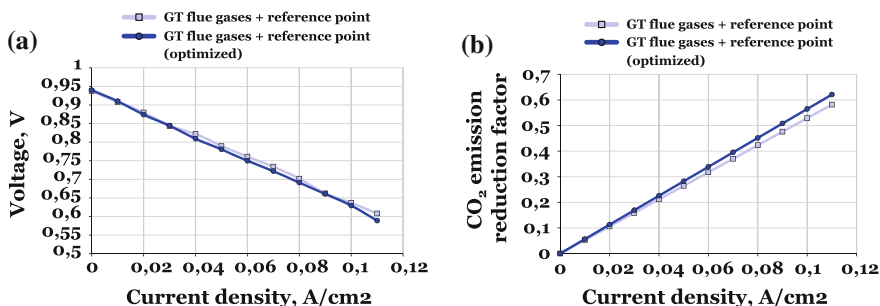


Fig. 6 Results of the experiments. a Voltage to current density for the investigated fuel cell. b CO₂ reduction rate

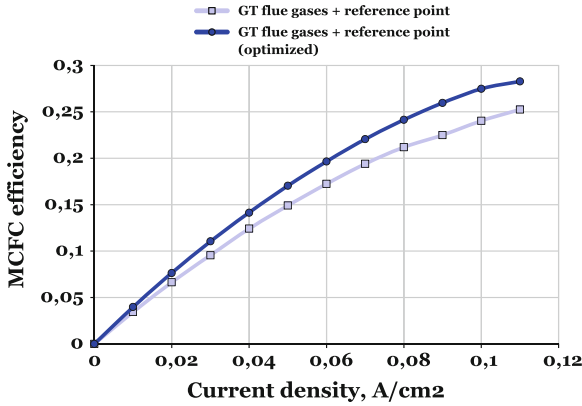


Fig. 7 Efficiency of the MCFC

Table 5 Main parameters of analyzed systems

Name	GTPP	Case 1	Case 2
MCFC _{fuel flow} /GT _{fuel flow} (%)	0	29	45
System efficiency (%)	33	33	40
CO ₂ emission reduction factor (%)	0	73	91
MCFC/GTPP fuel ratio	0	0.52	0.65
Relative CO ₂ emission (kg/MWh)	609	135	37
Annual CO ₂ emission (Gg/a)	250	68	18
Total system power (MW)	65	80	77

would generate no power, and might even consume some. However, the main task of a power plant is power generation; hence hybrid system efficiency was chosen as the objective function for optimization.

It should be borne in mind that prices of tradeable CO₂ allowances are relatively constant at present, which affords opportunity to realize profits from carbon trading.

Important technical issues such as sulfur or dust resistances of the MCFC fell outside the remit of this paper, although they can evidently limit the application of MCFCs in gas turbine power plants.

MCFCs could be profitably used in existing power plants which have been given CO₂ limits. MCFCs could potentially decrease CO₂ emissions, leaving the power generation capacity of the system at least the same, if not greater.

Acknowledgments The results presented in this paper were obtained from research work co-financed by the National Centre of Research and Development in the framework of Contract SP/E/1/67484/10 - Strategic Research Programme - Advanced technologies for energy generation: Development of a technology for highly efficiency zero-emission coal-fired power units integrated with CO₂ capture. Special acknowledgments for Prof. Lee Choong Gon from Hanbat National University, South Korea for his help and support with the experimental part of the work done.

References

1. Blum L, Deja R, Peters R, Stolten D (2011) Comparison of efficiencies of low, mean and high temperature fuel cell systems. *Int J Hydrogen Energy* 36(17):11056–11067
2. Budzianowski W (2010) An oxy-fuel mass-recirculating process for H₂ production with CO₂ capture by autothermal catalytic oxyforming of methane. *Int J Hydrogen Energy* 35(14):7454–7469
3. De Lorenzo G, Fragiaco P (2012) Electrical and electrical-thermal power plants with molten carbonate fuel cell/gas turbine-integrated systems. *Int J Energy Res* 36(2):153–165
4. De Lorenzo G, Fragiaco P (2012) A methodology for improving the performance of molten carbonate fuel cell/gas turbine hybrid systems. *Int J Energy Res* 36(1):96–110
5. Discepoli G, Cinti G, Desideri U, Penchini D, Proietti S (2012) Carbon capture with molten carbonate fuel cells: experimental tests and fuel cell performance assessment. *Int J Greenh Gas Control* 9:372–384
6. Granser D, Rocca F (1996) New high-efficiency 70 mw heavy-duty gas turbine. In: *Proceedings of Power-Gen conference, New Delhi, India*
7. Jeong H, Cho S, Kim D, Pyun H, Ha D, Han C, Kang M, Jeong M, Lee S (2012) A heuristic method of variable selection based on principal component analysis and factor analysis for monitoring in a 300 kw mcfc power plant. *Int J Hydrogen Energy* 37(15):11394–11400
8. Kotowicz J, Bartela T (2012) Optimisation of the connection of membrane CCS installation with a supercritical coal-fired power plant. *Energy* 38(1):118–127
9. Kupecki J, Badyda K (2011) SOFC-based micro-CHP system as an example of efficient power generation unit. *Arch Thermodyn* 32(3):33–43
10. Lanzini A, Santarelli M, Orsello G (2010) Residential solid oxide fuel cell generator fuelled by ethanol: cell, stack and system modelling with a preliminary experiment. *Fuel Cells* 10(4):654–675
11. Milewski J, Bernat R, Lewandowski J (2012) Reducing CO₂ emissions from a gas turbine power plant by using a molten carbonate fuel cell. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, London, UK, pp 1773–1778, 4–6 July 2012*
12. Milewski J, Wołowicz M, Badyda K, Misztal Z (2011) Operational characteristics of 36 kW PEMFC-CHP unit. *Rynek Energii* 92(1):150–156
13. Milewski J, Wołowicz M, Miller A (2012) An alternative model of molten carbonate fuel cell: a proposal. In: *EmHyTeC 2012, number P.1-34, pp 130–132*
14. Morita H, Komoda M, Mugikura Y, Izaki Y, Watanabe T, Masuda Y, Matsuyama T (2002) Performance analysis of molten carbonate fuel cell using a Li/Na electrolyte. *J Power Sour* 112(2):509–518
15. Mueller F, Gaynor R, Auld AE, Brouwer J, Jabbari F, Samuelsen GS (2008) Synergistic integration of a gas turbine and solid oxide fuel cell for improved transient capability. *J Power Sour* 176(1):229–239
16. Nomura R, Iki N, Kurata O, Kawabata M, Tsutsumi A, Koda E, Furutani H (2011) System analysis of IGFC with exergy recuperation utilizing low-grade coal, vol 4, pp 243–251
17. Wee J-H (2011) Molten carbonate fuel cell and gas turbine hybrid systems as distributed energy resources. *Appl Energy* 88(12):4252–4263
18. Wu W, Luo J-J (2010) Nonlinear feedback control of a preheater-integrated molten carbonate fuel cell system. *J Process Control* 20(7):860–868
19. Zhang H, Lin G, Chen J (2011) Performance analysis and multi-objective optimization of a new molten carbonate fuel cell system. *Int J Hydrogen Energy* 36(6):4015–4021

Computational Contact Modelling of Hip Resurfacing Devices

Murat Ali and Ken Mao

Abstract A combination of computational models and theoretical methods have been used and developed to study the contact of hip resurfacing devices under normal and edge loading conditions. Techniques were developed and the solutions based on using the finite element method. It was found that the study of hip joint modelling, numerical methodologies of mechanical wear simulations and shakedown analysis can be developed to study the contact mechanics and biotribology of hip resurfacing devices under central and edge loading conditions. Each method developed in this study provides a unique platform to study these problems.

Keywords Biotribology · Contact · Finite element analysis · Microseparation · Shakedown · Wear

1 Introduction

Contact mechanics, wear and surface damage of hip resurfacing devices are subjects which have been studied since very early implantations and the longevity of the devices are becoming increasingly important. The wear and surface damage of these bearing surfaces occur through normal gait loading conditions, however, another problem is the stripe wear patterns observed on metal-on-metal patient retrievals [1] and assessed devices following hip simulator studies [2]. It has been claimed that edge loading occurs during the walking cycle of the patient; therefore ‘microseparation’ is simulated into each cycle during experimental wear testing [3]. The laxity of the hip joint is understood to lead to microseparation during the gait

M. Ali (✉) · K. Mao
School of Engineering, University of Warwick, Coventry CV4 7AL, UK
e-mail: murat.ali@warwick.ac.uk

K. Mao
e-mail: k.mao@warwick.ac.uk

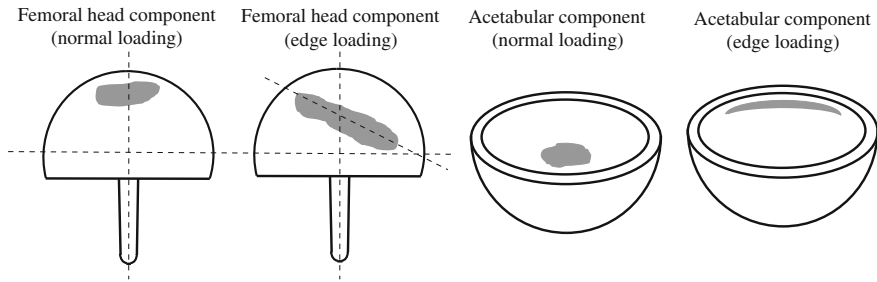


Fig. 1 Hip resurfacing device with normal and edge loading wear patterns

cycle, and fluoroscopy studies have revealed how edge loading of the hip joint occurs due to lateral sliding of the femoral component during gait [3]. The differences between wear patterns observed during normal and edge loading conditions is shown schematically in Fig. 1. This study expands on the research conducted by Ali and Mao [4] to further develop techniques in assessing both the contact mechanics for wear modelling and the application shakedown theory to cyclically loaded hip resurfacing devices, particularly those under normal and microseparation conditions leading to edge loaded hip resurfaced bearings.

2 Contact, Wear and Shakedown Theory

Contact mechanics forms an integral part to predicting the contact stresses and mechanical wear associated with hip resurfacing devices. For studying the wear of orthopaedic devices, the Archard wear model [5] has been used with finite element analysis techniques. Although the Archard wear model [6] appears in many forms, the form most appropriate to be used within the finite element method has been described (1) where h is the linear wear depth, k_w is the dimensional wear coefficient, p is the contact pressure and s is the sliding distance.

$$h = k_w p s \quad (1)$$

Along with mechanical wear under cyclic contact, residual stresses can act to protect the component from plastic deformation by ensuring purely elastic material behaviour is reached in the longer term. Shakedown theory can be applied to assess the repetitive rolling and sliding contacts of elastic-perfectly plastic materials [7]. The transition from elastic to perfectly plastic occurs at the yield point of the stress-strain curve and this assumes that the material does not harden under loading conditions. Shakedown theory is based on Koiter's and Melan's theorems. Where Koiter's theorem defines the upper shakedown limit and kinematic shakedown theorem, and Melan's theorem provides the lower shakedown limit and static shakedown

theorem [8]. Under normal cyclic walking and edge loading of the hip joint, rolling and sliding contact is present. This is another indication that shakedown theory can be applicable to hip joints studies, as the theory was originally used to study rolling and sliding contact of elastic bodies. For the shakedown theory to be valid then plastic deformation must occur to initiate residual stresses leading to purely elastic steady state cycles, or in other words the yield strength must be exceeded for the residual stresses to be present following the load removal.

Edge loading has been assessed using experimental simulators under cyclic loading considering the rotation of the hip [9]. The significance of mild and severe microseparation conditions were shown, also in a separate study the kinematics and motions had a significant effect on the contact mechanics and wear rates of devices [10].

3 Methods and Materials

Computational and numerical methods have been used to investigate the mechanical contact of hip resurfacing devices under normal and edge loading conditions. A technique has been developed to take patient bone scans and build finite element (FE) contact models as described in Fig. 2. The acetabular cup and femoral head components were modelled using SolidWorks. These orthopaedic models were combined with pelvis and femur models in an assembly. The associative interface between the computer aided design model and finite element model allowed for geometrical modifications to be made to the orthopaedic devices. The analysis was conducted using ABAQUS (version 6.10-1) in combination with user defined subroutines and custom programming.

Common to all of the finite element models, the hip resurfacing device had a bearing diameter \varnothing_f of 50 mm and diametral clearance \varnothing_c of $80\ \mu\text{m}$ [11]. A simple contact model was used to carry out specific comparison studies, therefore the hip resurfacing components were backed and fully tied to rigid parts which were referred to as model 1 (shown in Fig. 3). The elasticity of attached bone was considered for the simulation of models 2–4. For initial conditions the cup and femoral head bearing centre's coincided, and all model boundary conditions were subsequently applied within specified time steps. The assembly of model 2 has been provided in Fig. 4.

A number of vertical loads were considered including: a 3900 N load which was based on the peak load F_y expected during the walking cycle and an ISO (International Organization of Standardization) load F_I of 3000 N. A stumbling load F_s of 11000 N was also applied as these high vertical loads have been highlighted to occur during patient stumbling [12]. For model 1 and model 2 the microseparation was modelled by translating the cup bearing centre in the lateral direction (i.e. along the anatomical lateral-medial axis) as used in experimental testing methods [9] and a finite element study of edge loading [13]. In addition to this method, 'pure' microseparation was also simulated, which more closely replicates the theoretical microseparation model proposed by Mak et al. [14].

Fig. 2 Bone scans to FE contact models

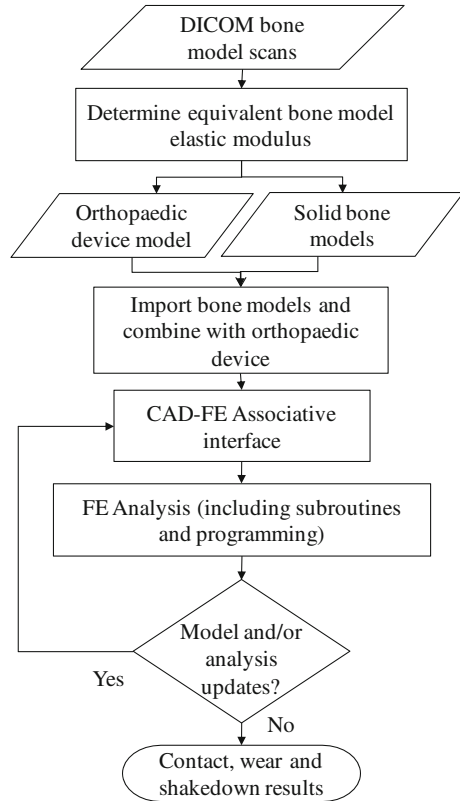
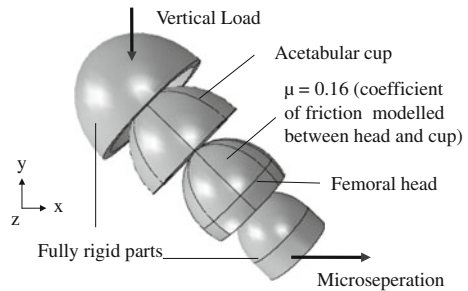
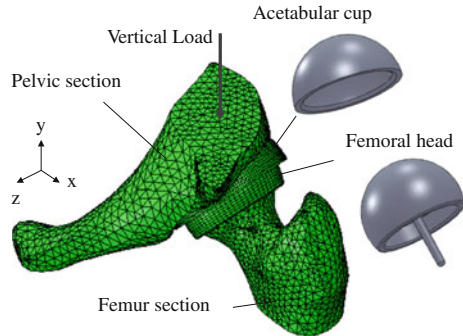


Fig. 3 Assembly of rigid backed components (model 1)



The coefficient of friction (μ) between the head and cup was defined as 0.16 based on the friction factor of CoCrMo on CoCrMo (cobalt chromium molybdenum) in both bovine serum and synovial fluid [15]. The coefficient of friction value modelled in finite element analysis was shown to have a negligible effect on the contact stress [11], however, as the surface friction coefficient increases during the life of the component the subsurface stresses will also increase [16]. Therefore it has been considered in

Fig. 4 Segmented hip model (model 2)



this study as the long term wear simulations can then take into account the increase in surface friction and surface roughness during the cyclic life of the component.

For this analysis, normal “Hard” contact behaviour was modelled and the material properties were obtained from literature and are summarised in Table 1 [17, 18]. For the application of shakedown theory, elastic-perfectly plastic ASTM F75 CoCrMo ‘as cast’ material properties were used [19]. For the bone model material an assessment was conducted to find an equivalent bone elastic modulus for the femur B_{EF} and pelvis B_{EP} to provide a simplified material model for the contact analysis. These values were determined by comparing the model stiffness of a CT (computed tomography) scanned femur and pelvis loaded in all three anatomical directions (x, y, z) using the finite element method. A sensitivity analysis was carried out on the bone material model to compare the elastic modulus values between 3 and 25 GPa applied to the pelvis and femur.

A full hip finite element model (Fig. 5) were developed to provide validation for using a segmented model. Except for modelling the full femur and pelvic model, this geometrically matches that of the segmented model (model 2) for ease of comparison. By carrying out the finite element discretisation within the finite element analysis package meant that all the advanced tools within this environment could be utilized. A 2D axis symmetric model (Fig. 6) was developed following the techniques described by Udofia et al. [11] as a model to conduct a cyclic shakedown analysis and assess the subsurface stresses under different vertical loading conditions. Standard ISO loading and angular displacement data was used where flexion-extension and internal-external rotation was simulated for the wear analysis, which is where the majority of the sliding distance between the bearing devices would occur from.

Table 1 Material properties

Material	Elastic modulus (GPa)	Poisson’s ratio	Density (kg/m ³)
CrCoMo	230	0.3	8270
B_{EF}	12.3	0.3	1900
B_{EP}	6.1	0.3	1900

Fig. 5 Full hip joint model (model 3)

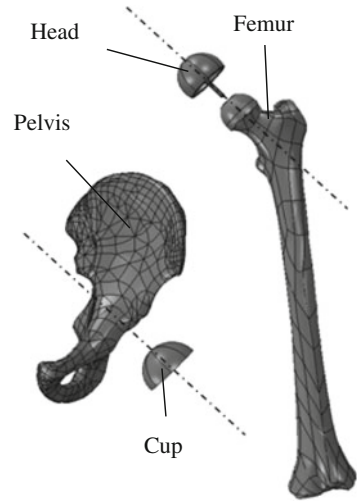
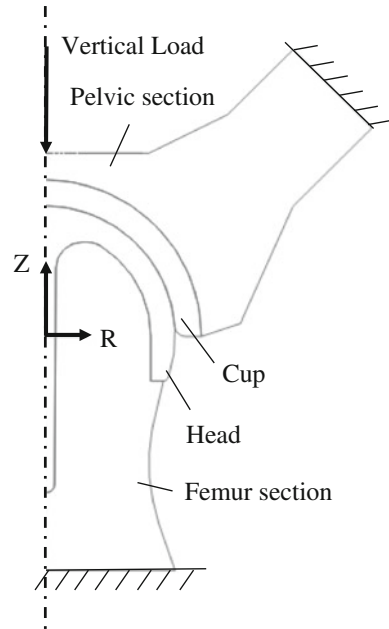


Fig. 6 2D Axis-symmetric model (model 4)



By studying the kinematics of the hip joint, it is claimed that microseparation occurs during the swing phase of gait [20, 21]. The swing phase occurs between 60 and 100% of the gait cycle, where the head and cup relocate fully during heel strike and edge loading occurs. As the frequency of the walking cycle ranges from 0.4–2.2 Hz [22], it is expected that edge loading could occur over a time period of 0.5 s.

The wear simulations extract data from finite element analysis and the data was used to calculate the sliding distance and wear depth over a number of cycles. At predefined cyclic intervals the mesh is then updated without the need to use an adaptive meshing algorithm. For each identified number of cycles the total wear depth was calculated at each node on the bearing surface as shown in (2), where h_I is the total wear depth calculated over the total number of increments n for the analysis at each node of the bearing surface. The total volumetric wear over the testing period is given by h_T as shown in (3), where m is the total number of finite element mesh update increments. The dimensional wear coefficients are based on values provided in literature [23].

$$h_I = \sum_{i=1}^n kp_i(s_i - s_{i-1}) \quad (2)$$

$$h_T = \sum_{i=1}^m h_I \quad (3)$$

4 Results

For model 1 and based on the walking gait peak vertical load of 3900 N, the maximum contact stress was 101 MPa without consideration of microseparation. The contact pressure increased to a maximum of 1284 MPa along with 675 MPa von Mises stress when 250 μm of lateral displacement was applied in combination with peak vertical loading conditions. By considering a lateral reaction force of 500 N (in line with experimental simulator test methods) without any vertical load led to a maximum contact stress of 564 MPa, von Mises stress of 456 MPa and maximum principal stress of 431 MPa. The simulation conducted on this model considered one cycle of edge loading and when the edge load was removed (i.e. contact removed) plastic strain was predicted to be less than 0.03 %. Through the assessment of edge loading due to ‘pure’ microseparation, the contact pressure between the head and cup again was predicted (Fig. 7). A symmetrical profile occurred about the centre of contact and the magnitude decreased as the distance from the centre of contact increased. The maximum contact stress did not occur directly on the rim radius of the cup, but above the rim radius. By using a spherical coordinate system to define the position of results, the contact stress magnitude decreased as the azimuthal angle ϕ moved away from the centre of contact. All of the contact was observed below 7° of the polar angle θ . By taking advantage of the customisable rigid backed model and through an efficient parametric study, the affect of cup inclination angle under ‘pure’ microseparation conditions have been observed (Fig. 8). The contact pressure increases as the cup inclination angle increased between 30 and 60°.

For model 2, a 250 μm translation in the lateral direction led to an edge loading reaction force of 907 N. Based on the walking gait 3900 N peak vertical load the

Fig. 7 Edge loading from 'pure' microseparation

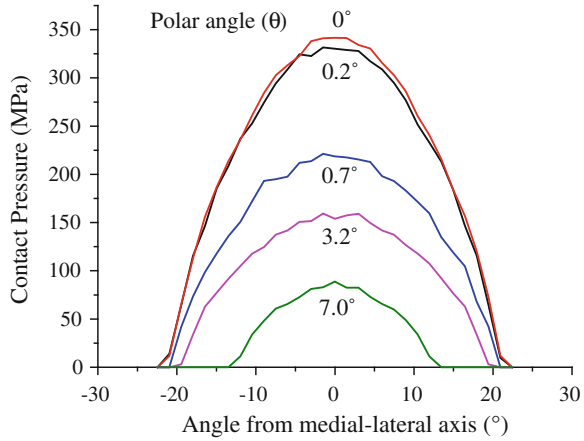
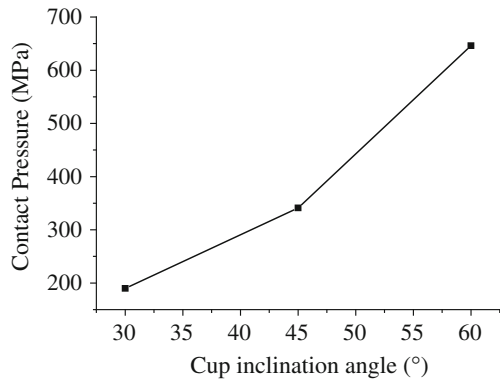


Fig. 8 Variation of contact stress against cup inclination angle



maximum contact pressure was 18 MPa without consideration of microseparation (Fig. 9). The contact pressure and von Mises stress increased to a maximum of 142 and 141 MPa respectively when microseparation conditions were applied in combination with the peak vertical gait load (Fig. 10). When modelling an ISO gait loading profile in combination with a lateral sliding edge load the contact pressure profile is observed to be elliptical with the maximum contact pressure of 85 MPa occurring in the centre of contact.

The contact patch for edge loaded acetabular cups and femoral heads were noted to be elliptical (with a high b/a ratio) compared with a circular contact area during normal loading conditions. The total contact area for normal loading contact and edge loading contact is provided in Fig. 11, where the maximum contact area for normal loading (N_l and N_p) and edge loading (M_l and M_p) is highlighted.

By considering the affect of anteversion the segmented hip joint assembly was modified with a cup anteversion angle of 15° , the maximum contact stress occurred in the region where the cup was backed by a stiffer region of the pelvis. The results

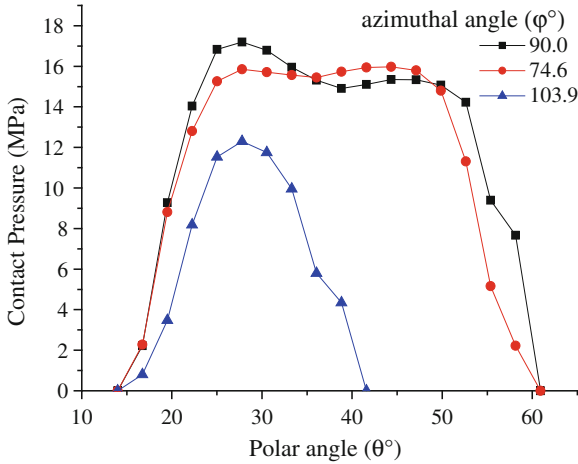


Fig. 9 Contact pressure distribution during normal loading

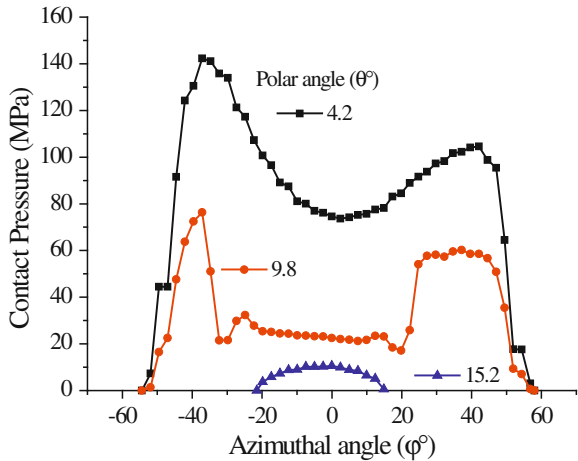


Fig. 10 Contact pressure distribution during edge loading

obtained from the full hip model (model 3), showed maximum contact stress under normal loading conditions to be 17 MPa. The sensitivity analysis of bone material elasticity modulus is shown in Fig. 12.

Based on a representation of shakedown maps for line and circular contact [7, 24] and a friction coefficient of 0.16, the component will remain in an elastic state under contact loading as long as the load intensity P_o/k does not exceed 3 (Fig. 13), where P_o and k are the maximum contact stress and material shear yield strength respectively. Based on theoretical shakedown maps and considering the maximum contact stress observed, the load intensity of the hip resurfacing device P_o/k is predicted to

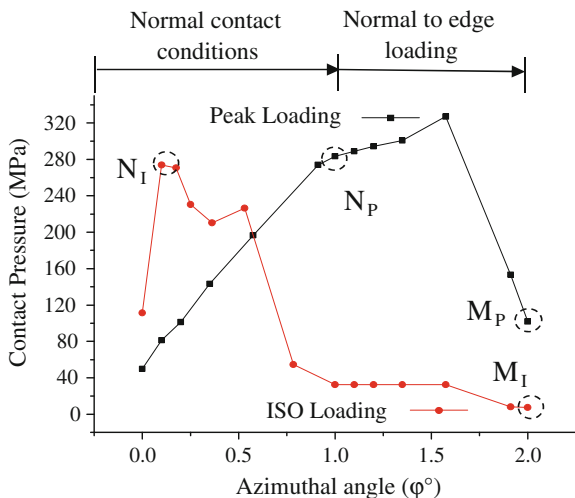


Fig. 11 Total contact area between the femoral head and acetabular cup

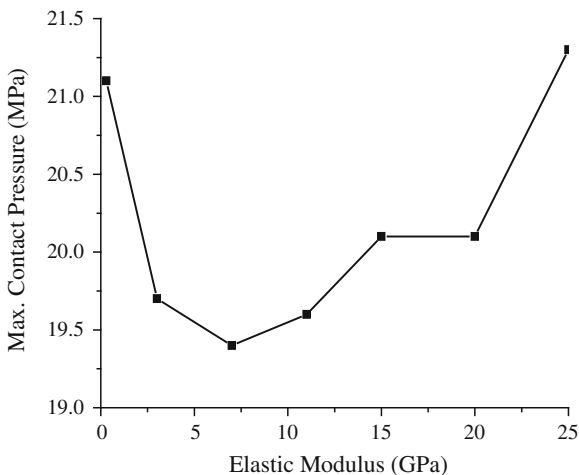


Fig. 12 Variation of contact pressure against bone elastic modulus

lie within the elastic region of the shakedown map and no elastic shakedown is predicted to occur. Following on from this result, by conducting the 2D axis-symmetric cyclic analysis using model 4, the stress-strain curve (Fig. 14) also predicted the hip resurfacing device material to remain within the elastic region under normal loading conditions even when high stumbling loads were considered.

The maximum von Mises stresses under the vertical loads F_I , F_y and F_s are provided in Table 2. All the maximum von Mises stresses occurred below the surface of contact, it was only when a stumbling load F_s that the maximum von Mises

Fig. 13 Shakedown map representation for line contact

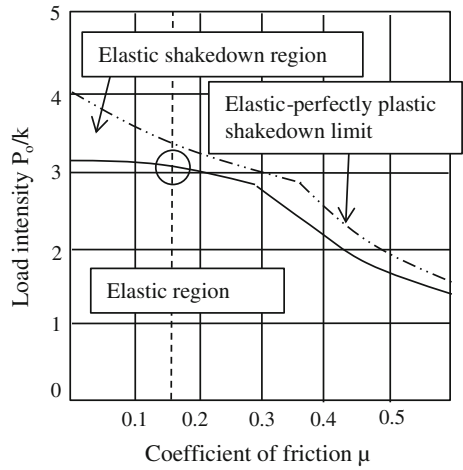


Fig. 14 2D-Axis symmetric cyclic stress-strain curve during normal loading.

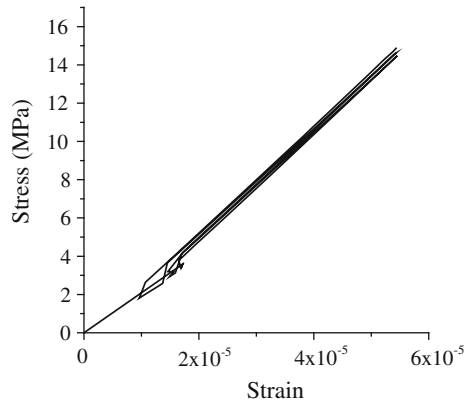


Table 2 Maximum von Mises stress under vertical loads

Load	Acetabular cup (max. stress MPa)	Femoral head (max. stress MPa)
F_l	34	15
F_y	45	21
F_s	125	66 ^a

^a 107 MPa predicted at the base of the femoral head stem

stress occurred at the at the bottom base of the head component. The mechanical wear prediction for the volumetric material loss due to mechanical wear of femoral head under flexion-extension rotation, internal-external rotation and ISO gait loading conditions was $82 \text{ mm}^3/\text{mc}$ (per million cycles).

5 Discussion

By comparing the results obtained for all computational models the effect of bone elasticity on the contact stress and von Mises stress distributions were shown. Any asymmetrical contact and stress distributions were predicted to be caused by the unsymmetrical geometry of the human anatomy, but more importantly the vertical loading and microseparation conditions. When edge loading occurred through a lateral displacement of the femoral head with an ISO loading profile the contact pressure profile was symmetrical about the centre of contact. When modelling microseparation conditions, it was observed that the maximum contact pressure and von Mises stress occurred towards the anterior end of the acetabular cup and femoral head. For all three dimensional models, the plastic strains and stress were predicted to occur above the rim radius of the cup which matches the inspections from patient retrievals and bearing components following experimental simulator testing with microseparation. The corresponding contact profile on the femoral head component was also dependent upon the anteversion angle of the implanted cup. Due to the geometric nature of the femoral head, the anteversion of the cup would not have any effect upon the contact pressure profile and magnitude on the acetabular cup. The contact pressures were also found to be insensitive to bone elastic modulus, even though a large range of E values were modeled as a form of methodology verification.

The magnitude of stresses and contact pressures may appear large for model 1 however, the rigidity of backing components have shown to increase the results by at least a factor of 5 over the results obtained using models 2–4. These high levels of contact pressures and stresses have also been observed by Mak et al. [13]. The total contact area under edge loading conditions was at least 2.7 times less than under central or normal contact conditions. This is an important finding as the contact patch dimensions directly affects the linear wear as does the contact pressure according to the Archard wear model used to study wear of the bearing surfaces. Following on from the predictions and study of contact pressures, the mechanical wear simulations provided a numerical method for predicting the gravimetric volume loss of material following finite element ablation and comparing the total element volumes before and after the cyclic wear process. The results were dependent upon the specific boundary conditions, dimensional wear coefficient, contact pressure and sliding distance. The wear loss increased linearly which also agrees with linear wear observed from device tested using experimental simulators.

When considering both cyclic gait loading and high stumbling loads no plasticity was observed in models 2–4, therefore, in reality it is predicted that material plasticity is not predicted to occur under normal, edge loading or even extreme stumbling load conditions. Although no fatigue assessments were carried out in this study, it is an important consideration for any cyclically loaded component. Throughout this study it is deemed that fatigue strength along with fracture toughness of Cobalt Chromium Molybdenum are significantly larger than bone. The fracture toughness of CoCrMo is many times greater than for bone. This high fracture toughness would much sooner

cause femoral neck fracture, before fracture or fatigue failure of the metal-on-metal device.

The microseparation distance of $250\ \mu\text{m}$ was equivalent to a force greater than that considered in experimental simulator studies which is typically $200\text{--}500\ \text{N}$ in magnitude. It was possible to assess the reaction forces in the edge loaded regions to determine the contact stress results at specific loading magnitudes. This observation also explains the high values of edge loading contact stress observed in model 1. Based on the maximum contact pressure and calculated value of k , a low value of load intensity, suggests that the component under central and edge loading conditions would remain within the elastic region of a contact shakedown map, which is a 'safe' region for the component to be operating in under rolling and sliding contact. Therefore, in terms of the hip resurfacing devices response to loading, elastic shakedown, plastic shakedown or ratcheting behaviour is unlikely to be observed, during normal contact conditions, edge loading or stumbling load conditions.

By assessing affect of cup inclination angle under 'pure' microseparation and relocation, the increases in contact pressure above a 45° cup inclination agrees with the increased wear rates from patients with implanted hip resurfacing devices [25], however, it should be noted that this was conducted without any anteversion of the acetabular cup.

The mechanical wear simulations provided comparative results against the results obtained from experimental simulator studies considering the vast variation in methodologies and assumptions made between the numerical and experimental strategies.

6 Conclusion and Further Work

A combination of computational, numerical and theoretical techniques have been used and developed, which formed the basis of studying the contact mechanics, wear and shakedown of hip resurfacing device. The finite element method was used to build contact models, develop numerical mechanical wear techniques from previous studies and assess the application of shakedown theory to normal and edge loaded hip joint resurfacing devices under different loading conditions. The severity of edge loading contact was observed along with the significance and sensitivity of results based on the bone backed anatomical geometry and component assembly. From the assumptions made in this study and the modelling conditions to simulate normal and edge loading for hip joint resurfacing devices, predictions have shown that although cyclic loading is present during the operation of the hip resurfacing devices, elastic shakedown, plastic shakedown or ratcheting is not predicted to occur. The resurfacing device material is predicted to remain operating within the elastic region. It should be noted that this conclusion is drawn without the direct assessment of asperity shakedown, which will be considered in future studies.

The scope for studying the contact mechanics and wear of hip resurfacing devices within its designed applications of being implanted into patients is possible without

the need for complex density based material models. During this study it was found that an equivalent bone modulus can be used without the need for refinement as the affect on producing varied contact pressures was negligible.

The modelling of microseparation was carried out in two distinct and separate ways by applying both lateral sliding and ‘pure’ microseparation. Laxity of the joint is simulated based on a theoretical microseparation model which provides further explanation of the increasing wear rates observed by in-vitro studies and patient retrievals. Both microseparation simulation models showed an increase of contact pressure by at least a factor of 2 over normal or centrally loaded hip resurfacing devices depending on a number of factors, including the anteversion of the acetabular cup and loading conditions. This level of contact stress increase agrees with the level of wear rate increase from in-vitro experimental simulator studies of standard testing including microseparation.

The Archard wear model in combination with the FE solver, provided a basis for predicting the wear of the hip bearing surfaces. The methodological approach adopted in this study meant that numerical and process checks could be performed at every step to ensure that the developed simulations provided understandable results. Further work is required to reduce the cyclic block increments to update the finite element mesh more regularly, this will in turn allow for a contact pressure distribution which is dependant upon the worn surface geometry. The wear simulations should also consider the variation in dimensional wear coefficient throughout the cyclic life of the bearing components.

Modelling verification and comparative solutions to other studies and theoretical models have been developed for centered contact conditions; however, further work is required to develop theoretical and computational models to more accurately simulate and assess the effects of edge loading and microseparation on hip resurfacing devices. The kinematics of these conditions during human joint motion should be considered in more depth if simulations are to more accurately model these problems. Overall, using a combination of techniques and theoretical models has shown to be beneficial in developing the simulations to hip resurfacing devices under specific conditions.

Acknowledgments This work was fully supported and funded by the EPSRC (Engineering and Physical Sciences Research Council).

References

1. Bowsher JG, Donaldson TK, Williams PA, Clarke IC (2008) Surface damage after multiple dislocations of a 38-mm-diameter, metal-on-metal hip prosthesis. *J Arthroplast* 23:1090–1096
2. Williams S, Stewart TD, Ingham E, Stone MH, Fisher J (2004) Metal-on-metal bearing wear with different swing phase loads. *J Biomed Mater Res Part B-Appl Biomater* 70B:233–239
3. Leslie IJ, Williams S, Isaac G, Ingham E, Fisher J (2009) High cup angle and microseparation increase the wear of hip surface replacements. *Clin Orthop Relat Res* 467:2259–2265
4. Ali M, Mao K (2012) Modelling of hip resurfacing device contact under central and edge loading conditions. In: *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012*, 4–6 July. London, U.K, pp 2054–2059

5. Archard JF (1953) Contact and rubbing of flat surfaces. *J Appl Phys* 24:981–988
6. Maxian TA, Brown TD, Pedersen DR, Callaghan JJ (1996) A sliding-distance-coupled finite element formulation for polyethylene wear in total hip arthroplasty. *J Biomech* 29:687–692
7. Ponter ARS, Chen HF, Ciavarella M, Specchia G (2006) Shakedown analyses for rolling and sliding contact problems. *Int J Solids Struct* 43:4201–4219
8. Williams JA, Dyson IN, Kapoor A (Apr 1999) Repeated loading, residual stresses, shakedown, and tribology. *J Mater Res* 14:1548–1559
9. Stewart T, Tipper J, Streicher R, Ingham E, Fisher J (2001) Long-term wear of HIPed alumina on alumina bearings for THR under microseparation conditions. *J Mater Sci-Mater Med* 12:1053–1056
10. Firkins PJ, Tipper JL, Ingham E, Stone MH, Farrar R, Fisher J (2001) Influence of simulator kinematics on the wear of metal-on-metal hip prostheses. *Proc Inst Mech Eng Part H-J Eng Med* 215:119–121
11. Udofia IJ, Yew A, Jin ZM (2004) Contact mechanics analysis of metal-on-metal hip resurfacing prostheses. *Proc Inst Mech Eng Part H-J Eng Med* 218:293–305
12. Bergmann G, Graichen F, Rohlmann A, Bender A, Heinlein B, Duda GN et al (2010) Realistic loads for testing hip implants. *Bio-med Mater Eng* 20:65–75
13. Mak M, Jin Z, Fisher J, Stewart TD (2011) Influence of acetabular cup rim design on the contact stress during edge loading in ceramic-on-ceramic hip prostheses. *J Arthroplast* 26:131–136
14. Mak MM, Besong AA, Jin ZM, Fisher J (2002) Effect of microseparation on contact mechanics in ceramic-on-ceramic hip joint replacements. *Proc Inst Mech Eng Part H-J Eng Med* 216:403–408
15. Scholes SC, Unsworth A, Goldsmith AAJ (2000) A frictional study of total hip joint replacements. *Phys Med Biol* 45:3721–3735
16. Farley J (2008) Development of a computational method of low cycle fatigue prediction for multi-layer surfaces under rolling/sliding contact conditions. Ph.D. dissertation, School of Engineering and Design, Brunel University
17. Hodgskinson R, Currey JD (1992) Young modulus, density and material properties in cancellous bone over a large density range. *J Mater Sci-Mater in Med* 3:377–381
18. Dalstra M, Huijskes R, Vanerling L (1995) Development and validation of a 3-dimensional finite-element model of the pelvis bone. *J Biomech Eng-Trans Asme* 117:272–278
19. ASM International (2009) *Materials and coatings for medical devices: cardiovascular*. ASM International, Cleveland
20. Lombardi AV, Mallory TH, Dennis DA, Komistek RD, Fada RA, Northcut EJ (2000) An in vivo determination of total hip arthroplasty pistoning during activity. *J Arthroplast* 15:702–709
21. Dennis DA, Komistek RD, Northcut EJ, Ochoa JA, Ritchie A (2001) In vivo determination of hip joint separation and the forces generated due to impact loading conditions. *J Biomech* 34:623–629
22. Affatato S, Spinelli A, Zavalloni M, Mazzega-Fabbro C, Viceconti A (Dec 2008) Tribology and total hip joint replacement: current concepts in mechanical simulation. *Med Eng Phys* 30:1305–1317
23. Liu F, Leslie I, Williams S, Fisher J, Jin Z (2008) Development of computational wear simulation of metal-on-metal hip resurfacing replacements. *J Biomech* 41:686–694
24. Williams JA (2005) The influence of repeated loading, residual stresses and shakedown on the behaviour of tribological contacts. *Tribol Int* 38:786–797
25. Hart AJ, Buddhdev P, Winship P, Faria N, Powell JJ, Skinner JA (2008) Cup inclination angle of greater than 50 degrees increases whole blood concentrations of cobalt and chromium ions after metal-on-metal hip resurfacing. *Hip Int* 18:212–9

Transport Phenomena in Engineering Problems: CFD-Based Computational Modeling

Maksim Mezhericher

Abstract Computational Fluid Dynamics is a popular modeling approach which utilizes numerical methods and computer simulations to solve and analyze problems that involve transport phenomena in fluid flows. CFD-based models demonstrate high versatility and capability of dealing with a wide range of engineering problems. This chapter presents two examples of CFD-based computational modeling successfully applied for different fields of engineering: particle engineering by drying processes and thermal management.

Keywords Computational fluid dynamics · Design · Modeling · Numerical simulations · Particle engineering · Thermal management · Transport phenomena

1 Introduction

Many contemporary engineering problems involve flows of liquid and/or gas, transport of heat by conduction, convection and radiation mechanisms, mass transfer by diffusion and convection, flows of bubbles, drops or particles, combustion etc. These complex problems require fundamental understanding that cannot be provided only by available experimental techniques, and therefore theoretical and numerical modeling are essential.

Recent progress in computer industry stimulated fast development of computational approaches, among them Computational Fluid Dynamics (CFD). This is a wide spread modeling approach which utilizes numerical methods and computer simulations to solve and analyze problems that involve transport phenomena in fluid flows. Lots of commercial and open computer codes are implementing CFD tech-

M. Mezhericher (✉)

Department of Mechanical Engineering, Shamoon College of Engineering,
Bialik/Basel Sts., Beer Sheva 84100, Israel
e-mail: maksime@sce.ac.il

nique: ANSYS FLUENT, ANSYS CFX, FLOW-3D, STAR-CD, COMSOL CFD, OpenFOAM, OpenFVM and many others.

This book chapter presents two examples of CFD-based computational modeling successfully applied for different fields of engineering: particle engineering by drying processes and thermal management of a car compartment. In spite of apparent differences, these two models have common roots in description of transport phenomena of the fluid phase.

2 Particle Engineering by Drying Processes

2.1 *Spray Drying*

Spray drying is a widely applied technology utilized to transform solutions, emulsions or suspensions into dry granules, and particle agglomerates, by feeding the liquid mixture as a spray of droplets into a medium with a hot drying agent. Because spray drying can be used either as a preservation method or simply as a fast drying technique, this process is utilized in many industries, such as food manufactures, pharmaceutical, chemical and biochemical industries. Spray drying is a rapid process (up to several seconds) compared to other methods of drying (e.g., pulse combustion drying, drum drying, freeze drying) due to the small spray droplet sizes and their large specific surface areas that maximize rates of heat and mass transfer. Therefore, this technique is the preferred drying method for many thermally-sensitive materials. Spray drying also turns a raw material into a dried powder in a single step, which can be advantageous for profit maximization and process simplification. Along with other drying techniques, spray drying also provides the advantage of weight and volume reduction. Dyestuffs, paint pigments, plastics, resins, catalysts, ceramic materials, washing powders, pesticides, fertilizers, organic and inorganic chemicals, skim and whole milk, baby foods, instant coffee and tea, dried fruits, juices, eggs, spices, cereal, enzymes, vitamins, flavors, antibiotics, medical ingredients, additives, animal feeds, biomass—this list of the spray-dried products is far from being complete.

A typical spray drying tower includes an atomizer, which transforms the supplied liquid feed into a spray of droplets, and a contact mixing zone, where the spray interacts with a hot drying agent. As a result of this interaction, the moisture content of the drying agent increases due to liquid evaporation from the droplets. In turn, due to evaporation, the spray droplets shrink and turn into solid particles. The drying process proceeds until the dried particles with the desired moisture content are obtained and then the final product is recovered from the drying chamber. Depending on the production requirements, droplet sizes from 1 to 1000 μm can be achieved using either nozzle or rotating disk atomizers; for the most common applications, average spray droplet diameters are between 100 and 200 μm . Air under atmospheric pressure, steam and inert gases like nitrogen are popular drying agents used nowadays. It is

noted that inert gases are applied for drying of flammable, toxic or oxide-sensitive materials.

2.2 Pneumatic Drying

Pneumatic (flash) drying is another example of extensively used technology in food, chemical, agricultural and pharmaceutical industries. The main advantages of this process are fast elimination of free moisture from pre-prepared feed of wet particles and operation in continuous mode. Typically, the feed is introduced into the drying column by a screw via a Venturi pipe. The particles dry out in seconds as they are conveyed by hot gas (air) stream. Then, the product is separated using cyclones which are usually followed by scrubbers or bag filters for final cleaning of the exhaust gases. In spite of its apparent simplicity, the process of pneumatic drying is a complex multi-scale multi-phase transport phenomenon involving turbulent mixing of humid gas and multi-component wet particles, heat and mass transfer interaction between the drying gas and dispersed phase, and internal heat and moisture transport within each conveyed wet particle.

3 Thermal Management of Car Compartment

One of the most energy-expensive units in contemporary vehicles is the air conditioning system (ACS). On an average, such systems consume up to 17 % of the overall power produced by vehicle engines of the world, depending on the cooling regime and environment thermal load [1]. It is remarkable that air-conditioning units in cars and light commercial vehicles burn more than 5 % of the vehicle fuel consumed annually throughout the European Union [2]. For instance, the United Kingdom emits about 3 million tons of CO₂ each year simply from powering the air-conditioning systems in vehicles. In South European and Mediterranean countries the problem of air pollution increase by ACS powering is even more acute.

Extensive theoretical and experimental studies have been performed throughout the recent years, aimed to reduce the fuel consumption and environmental pollution due to vehicle air conditioning. Proper thermal management based on the gas dynamics inside the vehicle passenger compartment is crucial for air conditioning and heating systems performance as well as for the comfort of passengers. On the other hand, the nature of the flow, namely the velocity field in combination with the temperature distribution, has a strong influence on the human sensation of thermal comfort [3].

In the present research it is proposed to develop a three-dimensional theoretical model of transport phenomena in car compartment. This model is based on Eulerian approach for the gas flow and takes into account thermal energy transfer by simultaneous conduction, convection and radiation mechanisms within the compartment

as well as outside the vehicle. The model is able to predict steady-state and transient profiles of air velocity, density, pressure, temperature and humidity for various regimes of the air conditioning, vehicle driving modes, compartment configurations and ambient conditions.

To assess the passenger level of thermal comfort, a methodology based on published studies [4, 5] may be developed and the corresponding equations may be coupled to the developed computational model. Moreover, the effect of the passengers themselves on their thermal comfort (e.g., heat emission by human body, air inhalation and gas mixture exhalation etc.) might be considered.

The CFD-based model of thermal management can be utilized as a tool for the following parametric investigations: enhancement of natural convection within a compartment, influence of thermal insulation and trim materials on the passenger thermal comfort as well as effect of introduction of innovating passive cooling techniques on energy consumption by ACS.

4 Theoretical Modeling

4.1 Spray and Pneumatic Drying

Transport phenomena in drying processes is subdivided into *external* (gas-particles mixing) and *internal* (within dispersed droplets/particles).

The fluid dynamics of continuous phase of drying gas is treated by an Eulerian approach and a standard k- ϵ model is utilized for turbulence description. The utilized three-dimensional conservation equations of continuity, momentum, energy, species, turbulent kinetic energy and dissipation rate of turbulence kinetic energy are as follows ($i, j = 1, 2, 3$):

– continuity

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j} (\rho u_j) = S_c, \quad (1)$$

where ρ and u are drying gas density and velocity, and S_c is mass source term.

– momentum conservation

$$\begin{aligned} \frac{\partial}{\partial t} (\rho u_i) + \frac{\partial}{\partial x_j} (\rho u_i u_j) = & -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \left[\mu_e \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] \\ & + \Delta \rho g_i + U_{pi} S_c + \sum F_{gp}, \end{aligned} \quad (2)$$

where p and μ_e are drying gas pressure and effective viscosity, U_p is particle velocity and $\sum F_{gp}$ is sum of the forces exerted by particles on the gas phase.

– energy conservation

$$\frac{\partial}{\partial t} (\rho h) + \frac{\partial}{\partial x_j} (\rho u_j h) = \frac{\partial}{\partial x_j} \left(\frac{\mu_e}{\sigma_h} \frac{\partial h}{\partial x_j} \right) - q_r + S_h, \quad (3)$$

where h is specific enthalpy, q_r and S_h are thermal radiation and energy source terms, respectively.

– species conservation

$$\frac{\partial}{\partial t} (\rho Y_v) + \frac{\partial}{\partial x_j} (\rho u_j Y_v) = \frac{\partial}{\partial x_j} \left(\frac{\mu_e}{\sigma_Y} \frac{\partial Y_v}{\partial x_j} \right) + S_c, \quad (4)$$

where Y_v is mass fraction of vapour in humid gas.

– turbulence kinetic energy

$$\frac{\partial}{\partial t} (\rho k) + \frac{\partial}{\partial x_j} (\rho u_j k) = \frac{\partial}{\partial x_j} \left(\frac{\mu_e}{\sigma_k} \frac{\partial k}{\partial x_j} \right) + G_k + G_b - \rho \varepsilon, \quad (5)$$

where k is turbulent kinetic energy and ε is dissipation rate of turbulent kinetic energy.

– dissipation rate of turbulence kinetic energy

$$\frac{\partial}{\partial t} (\rho \varepsilon) + \frac{\partial}{\partial x_j} (\rho u_j \varepsilon) = \frac{\partial}{\partial x_j} \left(\frac{\mu_e}{\sigma_\varepsilon} \frac{\partial \varepsilon}{\partial x_j} \right) + \frac{\varepsilon}{k} (C_1 G_k - C_2 \rho \varepsilon). \quad (6)$$

The production of turbulence kinetic energy due to mean velocity gradients is equal to:

$$G_k = \mu_T \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j}. \quad (7)$$

The production of turbulence kinetic energy due to buoyancy is given by:

$$G_b = -\beta g_j \frac{\mu_T}{\sigma_T} \frac{\partial T}{\partial x_j}, \quad (8)$$

where T is gas temperature and β is coefficient of gas thermal expansion:

$$\beta = -\frac{1}{\rho} \left(\frac{\partial \rho}{\partial T} \right)_p. \quad (9)$$

The utilized constants are $C_1 = 1.44$, $C_2 = 1.92$, and the Prandtl numbers are equal to $\sigma_k = \sigma_h = \sigma_Y = \sigma_T = 0.9$ and $\sigma_\varepsilon = 1.3$. The effective viscosity, μ_e , is defined as:

$$\mu_e = \mu + \mu_T, \quad (10)$$

where μ_T is turbulent viscosity

$$\mu_T = C_\mu \rho \frac{k^2}{\varepsilon}. \quad (11)$$

In the above expression $C_\mu = 0.09$.

The constitutive relationship between air temperature, pressure and density is given by the ideal gas law (such model is sufficient because of small humidity of the gas involved in the considered multiphase flow):

$$p = \frac{\rho}{M} \Re T, \quad (12)$$

where \Re is universal gas constant and M is molecular weight of the gas phase.

To track the trajectories and other valuable parameters of spray of droplets and particles, a Discrete Phase Model (DPM) based on Lagrangian formulation is utilized. The motion of the droplets/particles is described by Newton's Second Law:

$$\frac{d\vec{U}_p}{dt} = \vec{g} + \frac{\sum \vec{F}_p}{m_p}. \quad (13)$$

Here $\sum \vec{F}_p$ is sum of the forces exerted on given spray droplet/particle by the gas phase, by other particles and walls of the spray drying chamber; \vec{g} is gravity acceleration, and \vec{U}_p and m_p are droplet/particle velocity and mass, respectively. In general, the acting forces on spray droplet/particle are as follows:

$$\sum \vec{F}_p = \vec{F}_D + \vec{F}_B + \vec{F}_A + \vec{F}_{PG} + \vec{F}_C + \vec{F}_{other}, \quad (14)$$

where \vec{F}_D is drag force, \vec{F}_B is buoyancy force, \vec{F}_A is added mass force, \vec{F}_{PG} is pressure gradient force and \vec{F}_C is contact force (neglected in the present work). The term \vec{F}_{other} represents other forces, usually important for submicron particles and/or at specific conditions, e.g., phoretic, Basset, Saffman, Magnus forces etc., and neglected in the present work for simplicity.

The drag force is determined by the expression:

$$\vec{F}_D = \frac{\pi d_p^2}{8} \rho C_D \left| \vec{u} - \vec{U}_p \right| \left(\vec{u} - \vec{U}_p \right), \quad (15)$$

where d_p is droplet/particle diameter and \vec{u} is velocity of gas phase. The drag coefficient, C_D , is calculated according to well-known empirical correlations for spherical particles.

The buoyancy force opposes gravity and in the present study it is much smaller than the latter, because the densities of spray droplets and drying gas differ more than thousand times. For this reason the buoyancy of droplets/particles is currently neglected.

The added mass (“virtual-mass”) force, required to accelerate the gas surrounding the droplet/particle, is given by:

$$\vec{F}_A = \rho \frac{\pi d_p^3}{12} \frac{d}{dt} (\vec{u} - \vec{U}_p). \quad (16)$$

For flow in dryers, this force may be important in the droplets/particles entrance region, where the velocities of injected dispersed phase and drying gas are substantially different and their intensive mixing leads to high change rates of the relative velocities.

The pressure gradient in the gas phase additionally accelerates droplets/particles and results in the following force:

$$\vec{F}_{GP} = -\frac{\pi d_p^3}{6} \vec{\nabla} p. \quad (17)$$

This force can be essential and worth consideration in the dryer regions with fast pressure changes like inlet, outlet and swirling zones of the gas phase.

In the present work the *internal transport phenomena* within each droplet/particle are described with the help of previously developed and validated two-stage drying kinetics model, see [6]. The drying process of droplet containing solids is divided in two drying stages. In the first stage of drying, an excess of moisture forms a liquid envelope around the droplet solid fraction, and unhindered drying similar to pure liquid droplet evaporation results in the shrinkage of the droplet diameter. At a certain moment, the moisture excess is completely evaporated, droplet turns into a wet particle and the second stage of a hindered drying begins. In this second drying stage, two regions of the wet particle can be identified: layer of dry porous crust and internal wet core. The drying rate is controlled by the rate of moisture diffusion from the particle wet core through the crust pores towards the particle outer surface. As a result of the hindered drying, the particle wet core shrinks and the thickness of the crust region increases. The particle outer diameter is assumed to remain unchanged during the second drying stage. After the point when the particle moisture content decreases to a minimal possible value (determined either as an equilibrium moisture content or as a bounded moisture that cannot be removed by drying), the particle is treated as a dry non-evaporating solid sphere. It is worth noting that all droplets and particles are assumed to be spherical and a full radial symmetry of inter-droplet physical parameters (temperature, moisture content etc.) is believed. The concept of two-stage droplet drying kinetics is illustrated by Fig. 1.

4.2 Thermal Management of Car Compartment

The conservation equations of the above 3D model of transport phenomena in particle engineering processes can also be applied to describe the gas flow and heat

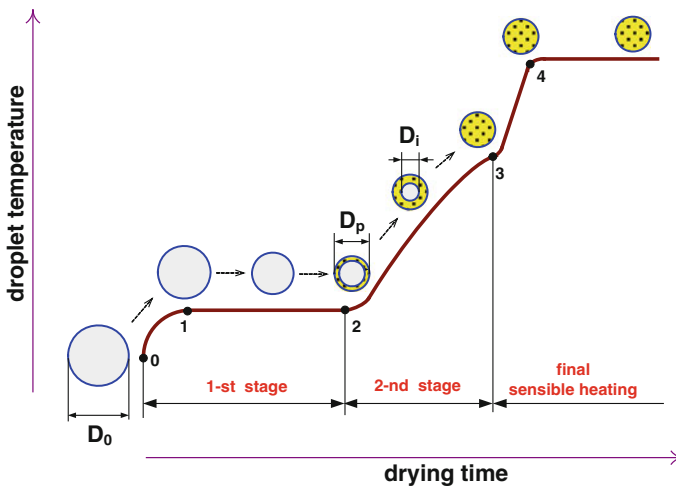


Fig. 1 The concept of two-stage droplet drying kinetics

and mass transfer within a car passenger compartment and in its surroundings. Particularly, in the present study 3D compressible Reynolds Averaged Navier-Stokes (RANS) equations including $k-\varepsilon$ turbulence formulation (1–12) are utilized to predict flow patterns of the gas mixture (air, vapor, carbon dioxide) inside a passenger compartment and air flow outside the cabin. These equations are solved by applying a Finite Volumes Method (FVM). The computational model formulated in such way is capable of predicting air velocity, temperature and species flow patterns inside the compartment under different ambient conditions. Moreover, the model can be used to prognosticate energy consumption of ACS that is necessary for providing passengers comfort at given outside thermal load. In addition, numerical simulations with this 3D model may facilitate revealing weak design points and optimization of existing/newly designed car air conditioning systems.

5 Numerical Simulations

5.1 Spray Drying

A cylinder-on-cone spray dryer with co-current flow of drying air and spray of droplets (Fig. 2) is adopted from the literature [6, 7]. Preheated atmospheric air at temperature 468 K and absolute humidity of 0.009 kg H₂O/kg dry air is supplied into the drying chamber through a central round inlet of the flat horizontal ceiling, without swirling and at angle of 35°, with respect to the vertical axis. The air inlet velocity is 9.08 m/s, whereas its turbulence kinetic energy is equal to 0.027 m²/s²

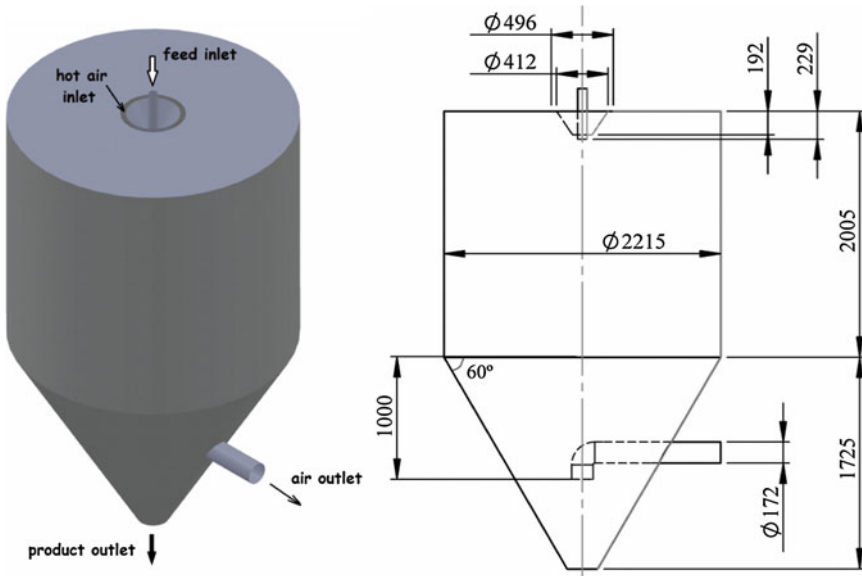


Fig. 2 Sketch of spray drying tower

and turbulence energy dissipation rate is $0.37 \text{ m}^2/\text{s}^3$. The spray of liquid droplets is obtained by atomizing the liquid feed in a pressure nozzle. The spray cone angle is assumed to be 76° , the droplet velocities at the nozzle exit are assigned to 59 m/s , and the temperature of the feed is set at 300 K . The distribution of droplet diameters in the spray is assumed to obey the Rosin-Rammler distribution function, where the mean droplet diameter is assumed to be $70.5 \text{ }\mu\text{m}$, the spread parameter is set at 2.09 , and the corresponding minimum and maximum droplet diameters are taken as $10.0 \text{ }\mu\text{m}$ and $138.0 \text{ }\mu\text{m}$, respectively. The overall spray mass flow rate is equal to 0.0139 kg/s (50 kg/hr). The walls of drying chamber are assumed to be made of 2 mm stainless steel, and the coefficient of heat transfer through the walls is set to zero as though there is a perfect thermal insulation of the chamber. The gage air pressure in the outlet pipe of the drying chamber is set to -100 Pa .

The numerical solution and simulations have been performed by utilizing a 3D pressure-based solver incorporated in CFD package ANSYS FLUENT 13. The solver is based on the finite volumes technique and enables two-way coupled Euler-DPM algorithm for treatment of the continuous and discrete phases. The chamber geometry has been meshed by $619,711$ unstructured grid cells of tetrahedral and polyhedral shape with the various mesh sizes.

The 3D spray from pressure nozzle is modeled by 20 spatial droplet streams. In turn, each droplet stream is represented by 10 injections of different droplet diameters: minimum and maximum diameters are $10.0 \text{ }\mu\text{m}$ and $138.0 \text{ }\mu\text{m}$, whereas the intermediate droplet sizes are calculated by applying Rosin-Rammler distribution

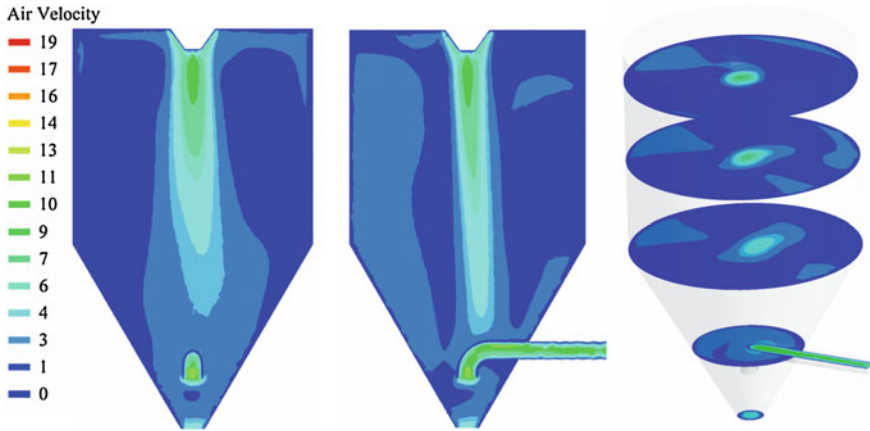


Fig. 3 Flow fields of air velocity (m/s) in spray dryer: frontal (*left*), side (*middle*) and isometric (*right*) cuts

function with $70.5 \mu\text{m}$ of droplet average size. In this way, totally 200 different droplet injections have been introduced into the computational domain.

All the numerical simulations of spray drying process have been performed in steady-state two-way coupling mode of calculations. For continuous phase, the spatial discretization was performed by upwind scheme of second order for all conservation equations (except pressure solved by PRESTO! procedure) and SIMPLE scheme was used for coupling between the pressure and velocity. For the dispersed phase, the tracking scheme was automatically selected between low order implicit and high order trapezoidal schemes based on the solution stability. DPM sources were updated every iteration of the continuous phase. The overall steady-state numerical formulation was of the second order of accuracy.

The computation of internal transport phenomena for the discrete phase was accomplished using the concept of user defined functions (UDF). The numerical solution was implemented as a subroutine and linked to the ANSYS FLUENT solver via a set of the original UDFs. The results of simulations are shown in Figs. 3, 4, and 5.

5.2 Pneumatic Drying

For the purposes of the theoretical study the geometry of Baeyens et al. [8] experimental set-up was adopted. Hot dry air and wet particles are supplied to the bottom of vertical pneumatic dryer with 1.25 m internal diameter and 25 m height (see Fig. 6). The developed theoretical model was numerically solved with the help of Finite Volume Method and 3D simulations of pneumatic drying were performed using the CFD package ANSYS FLUENT. To this end, the 3D numerical grid with 9078 distributed



Fig. 4 Flow patterns of air temperature (Kelvin) in spray dryer: frontal (*left*), side (*middle*) and isometric (*right*) cuts

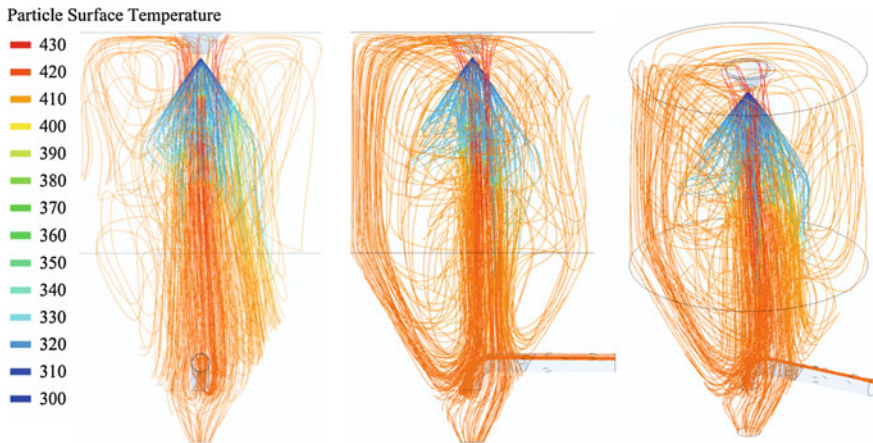


Fig. 5 Particle trajectories in spray dryer colored by particle surface temperature (Kelvin): frontal (*left*), side (*middle*) and isometric (*right*) views

hexahedral/ wedge cell volumes was generated in GAMBIT 2.2.30 using the Cooper scheme.

The flow of wet PVC particles in the pneumatic dryer was simulated through 89 injections of spherical particles. Each injection began on the bottom of the dryer at the centroid of one of the 89 bottom plane mesh elements. The particle injections were normal to the dryer bottom plane and parallel to each other.

The numerical simulations were performed in the following way. First, the flow of drying air was simulated without the discrete phase until converged solution was obtained. At the next step, wet particles were injected into the domain and two-way coupled simulations were performed until convergence.

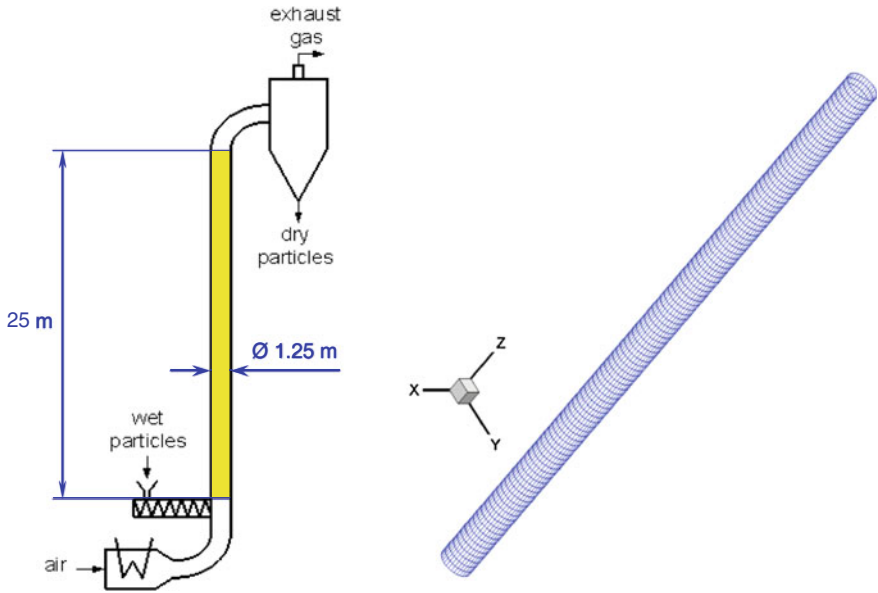


Fig. 6 Schematic sketch of pneumatic dryer [8] (left) and corresponding numerical grid (right)

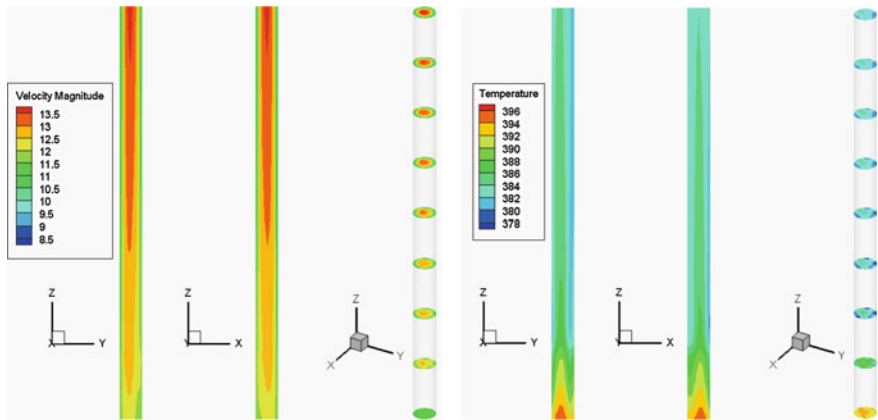


Fig. 7 Calculated flow patterns for drying of PVC wet particles in pneumatic dryer(adiabatic flow). *Left*—air velocity magnitude, m/s, *right*—air temperature, K

The convergence of the numerical simulations was determined by means of residuals of the transport equations. Particularly, the converged values of the scaled residuals were ensured to be lower than 10^{-6} for the energy equation and 10^{-3} for the rest of equations. The convergence was also verified by negligibly small values of the global mass and energy imbalances. The predicted flow patterns are given in Fig. 7.

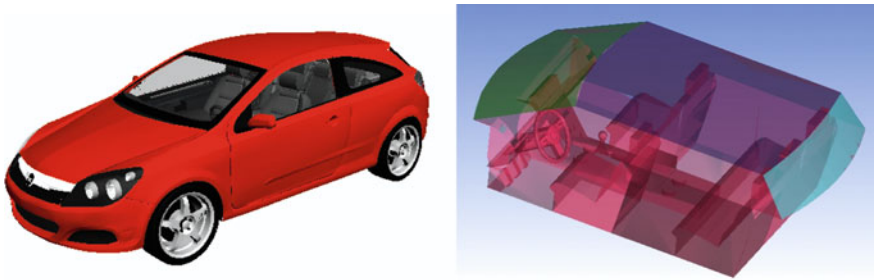


Fig. 8 Car model in 3D Studio Max software (*left*) and extracted compartment geometry (*right*)

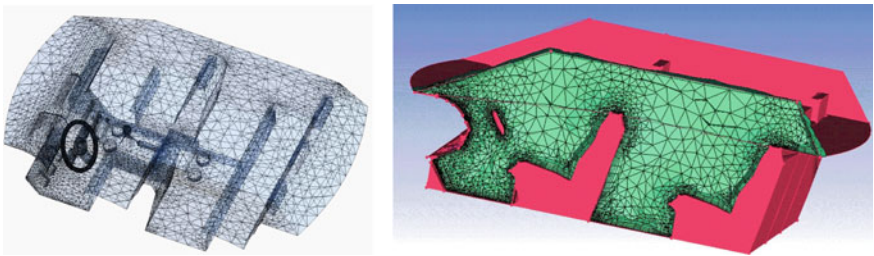


Fig. 9 Numerical grid of car passenger compartment

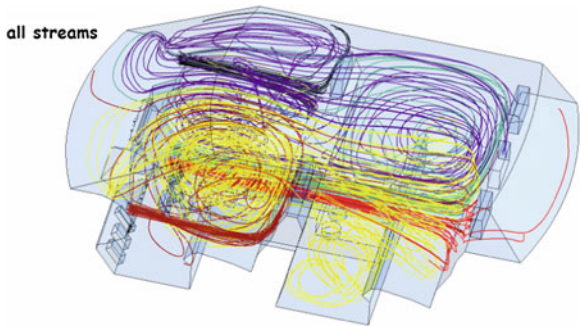


Fig. 10 Calculated streamlines of air velocity in car passenger compartment

5.3 Air Flow Patterns in Car Passenger Compartment

A simplified model of conventional car passenger compartment was adopted for numerical simulations. The original car model was created in 3D Studio Max software, whereas the extracted compartment geometry was processed in ANSYS ICEM CFD program and meshed with 518,705 polyhedral (3–6 faces) grid cells (Figs. 8 and 9). Then, steady state 3D numerical simulations using ANSYS FLUENT code were performed.

Figure 10 demonstrates the results of numerical simulations of air flow patterns in the compartment. Four air inlets with velocity 1 m/s and temperature 12 °C and two air outlets with -150 Pa gage pressure were established. The ambient temperature was taken 37 °C and heat transfer coefficient was set to $10 \text{ W}/(\text{m}^2 \cdot \text{K})$, assuming parked car.

6 Conclusion

The computational CFD-based modeling is a powerful tool for description, simulation and analysis of variety of engineering problems involving macroscopic transport phenomena (fluid/gas dynamics, turbulence, heat transfer and mass transfer). CFD-based models demonstrate high versatility and capability of dealing with a wide range of engineering problems. Basic knowledge in CFD-based modeling and ability to work with CFD software are becoming to be necessary skills for contemporary engineers and researchers.

References

1. Lambert MA, Jones BJ (2006) Automotive adsorption air conditioner powered by exhaust heat. part 1: conceptual and embodiment design. *J Automobile Eng* 220(7):959–972
2. Sorption energy seeking to commercialize waste heat-driven adsorption heat pump technology for vehicle air conditioning. Green car congress. <http://www.greencarcongress.com>, 24 April 2010
3. Application briefs from FLUENT. EX170. Vehicle ventilation system. http://www.fluent.com/solutions/automotive/ex170_vehicle_ventilation_system.pdf
4. Huang L, Han TA (2005) Case study of occupant thermal comfort in a cabin using virtual thermal comfort engineering. In: Proceedings of EACC 2005–2nd european automotive CFD conference, Frankfurt, Germany, 29–30 June 2005
5. Lombardi G, Maganzi M, Cannizzo F, Solinas G (2007) The Use of CFD to improve the thermal comfort in the automotive field. In: Proceedings of EACC 2007–3rd european automotive CFD conference, Frankfurt, Germany, 5–6 July 2007
6. Mezhericher M (2011) Theoretical modeling of spray drying processes. Drying kinetics, two and three dimensional CFD modeling. vol 1. LAP Lambert Academic Publishing, Saarbrücken, Germany. ISBN 978-3-8443-9959-2
7. Mezhericher M (2012) CFD-based modeling of transport phenomena for engineering problems, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012, WCE 2012, 4–6 July London, UK, pp 1803–1808
8. Baeyens J, van Gauwbergen D, Vinckier I (1995) Pneumatic drying: the use of large-scale experimental data in a design procedure. *Powder Technol* 83(2):139–148

Investigating the Effects on the Low Speed Response of a Pressure Charged IC Engine Through the Application of a Twin-Entry Turbine Housing

Alex Kuzstelan, Denis Marchant, Yufeng Yao and Yawei Wang

Abstract In this study, one-dimensional analysis using AVL Boost software has been carried out on a series of compression and spark ignition engines utilizing a manufacturer fitted single-entry turbocharger and a modified twin-entry unit, the latter adopting two symmetrical turbine housing inlet ports. The model reconstruction using AVL Boost considers parameters that accurately represent the physical engine conditions including manifold geometry, turbocharger flow maps and combustion chamber characteristics. Model validations have been made for a standard single-entry turbocharger configuration to predict the maximum engine power and torque, in comparison with available manufacturer data and analytical calculations. Further studies concentrate on engine performance comparisons between single- and twin-entry turbochargers at low engine speed conditions, typically in a range of 1000–3000 RPM. Improvements in turbine shaft speed, engine power and torque have been achieved, thus implying improved low speed engine response. This study reveals the potential commercial benefits of adopting a twin-entry turbocharger and contribution to the academic community through this additional research.

Keywords Engine simulation · Internal combustion engine · Pressure charging · Turbine housing · Turbocharging · Twin-entry turbochargers

A. Kuzstelan (✉) · Y. Wang
Kingston University London, Friars Avenue, London SW15 3DW, UK
e-mail: K0306918@kingston.ac.uk

Y. Wang
e-mail: y.wang@kingston.ac.uk

D. Marchant
Kingston University London, Penhyn Road, Kingston Upon Thames, Surrey KT1 2EE, UK
e-mail: D.Marchant@kingston.ac.uk

Y. Yao
University of the West of England, Coldharbour Lane, Bristol BS16 1QY, UK
e-mail: Yufeng.Yao@uwe.ac.uk

1 Introduction

Turbochargers have been extensively used for “engine downsizing” practices as they can largely enhance the engines power and torque output without the need of increasing the swept volume of each cylinder. However, for turbocharged downsized diesel engines, the slower response of the turbine at low engine speeds, typically in a range of 1000–3000 RPM, appears to be a common problem. Various solutions have been proposed and studied, including variable geometry turbochargers (VGT), two-stage turbocharger and turbo-compounding methods. Both Arnold [1] and Hawley [2] observed that adopting a narrow vane angle within a VGT turbine housing at low engine speeds increases exhaust flow to the impeller, thus improving the boost performance of the compressor. Recently Chadwell and Walls [3] suggested a new technology known as a SuperTurbo to overcome the slow response of a turbocharger at low engine speeds. This type of turbocharger can be coupled to a continuously variable transmission (CVT) which is directly run via the crankshaft of the engine, thus allowing the turbocharger to act as a supercharger boosting device at lower engine speeds. Similar increases in performance using turbo-compounding methods are observed by Ishii [4] and Petitjean et al. [5]. Two-stage turbocharging as discussed by Watel et al. [6] uses high and low pressure turbochargers working in series to overcome the effects of reduced exhaust pressure encountered at low engine speeds. One method which has not been fully researched is the application of a twin-entry turbocharger with two turbine inlet ports. This arrangement may lead to an improved engine response at lower engine speeds, primarily due to the separated inlet port arrangement, thus avoiding the interactions between the differently pulsed exhaust gases inside the manifold, and enhancing the energy transfer from exhaust gas to the turbine impeller. In contrast to a single-entry turbocharger, a twin-entry turbine housing (as shown in Fig. 1) will better utilize the energy of the pulsating exhaust gas to boost the turbine performance which directly increases the rotational speed of the compressor impeller. For example a four-cylinder engine with a 1-3-4-2 firing order equipped with a single-entry turbocharger and 4 into 1 exhaust manifold will produce the following conditions: at the end of the exhaust stroke in cylinder 1 (i.e. when the piston is approaching top dead centre (TDC)), the momentum of the exhaust gas flowing into the manifold will scavenge the burnt gas out of the cylinder. In the meantime in cylinder 2, the exhaust valve is already open allowing for exhaust gas to enter the manifold as well. This means that the exhaust gas from cylinder 2 will interact with the flow of exhaust gas from cylinder 1, thus affecting the energy transfer to the turbine [7]. One solution to this problem is to adopt a twin-entry turbocharger with a split-pulse manifold that keeps the differently pulsed exhaust gasses separate, thus allowing the majority of the pulsating energy of the exhaust gas to be used by the impeller. This is not only more practical and economical but also provides a potential for improvement in the reduction of gaseous emissions. Twin-entry turbochargers have been widely used in industry for large-size engines, but limited research has been undertaken for medium-sized engines. Therefore more

Fig. 1 Turbocharger cut-away highlighting the twin-entry volute geometry [8]



studies are necessary to provide further insight into the key benefits, or otherwise, of adopting a twin-entry turbocharger as shown in this study.

2 Analysis of Experimental Engine Models

2.1 Engine Model

A commercially available downsized four-cylinder Renault 1.5L compression ignition (DCi) engine is used as a base engine for the 1-D simulation. The engine is fitted with a single-entry turbocharger as part of its standard specification. This factor is beneficial as a crucial aspect of the experimental criteria involves an analysis of a standard engine and the same engine equipped with a twin-entry turbine housing utilizing the same trim and area ratio. Table 1 gives the key parameters of the model required by the AVL Boost code [9]. It is worthwhile to point out that the purpose

Table 1 The key engine parameters as required by the AVL Boost simulation code

Parameters	1.5L DCi	2.0L CI	1.8L SI
Bore	76 mm	85 mm	81 mm
Stroke	80.5 mm	88 mm	86 mm
Exhaust valve lift	8.6 mm	5.0 mm	9.3 mm
Inlet valve lift	8.0 mm	4.6 mm	7.67 mm
Compression ratio	17.9:1	18:1	9.5:1
No. of cylinders	4	4	4
Valves per cyl.	2	2	5

of choosing this type of engine is to fulfill the current trend of engine downsizing as frequently cited in engine technology international [10]. Two further engines are also modeled using the AVL Boost code to evaluate whether the same observations could be met due to the application of a twin-entry turbocharger. These include a Peugeot 2.0L CI engine and Audi 1.8L SI engine respectively.

2.2 Engine Boundary Conditions

For the purpose of this study, the engine modeling is based on 100 simulation cycles using variable operating conditions of engine speed ranging between 1000 and 5000 RPM. Both the exhaust and the inlet valve lift profiles and dimensions are also defined using original data from the manufacturer to provide realistic operating conditions of the combustion cycle. Furthermore, identical compressor geometry and flow maps are used for both single- and twin-entry turbocharger configurations. This provides more accurate boundary conditions as the flow characteristics of the compressor will only be affected by the differences in turbine inlet and exhaust manifold geometry. It is essential to maintain the same compressor housing in order to derive accurate and convincing conclusions. In order to model engine operating conditions, the intake and the exhaust piping lengths and diameters of the physical engine are directly measured and replicated within the software. In conjunction with the Vibe combustion model, the Woschni heat transfer model and the Patton et al. friction model [11] are used to define the heat transfer conditions within the combustion chamber for each simulated engine RPM stage, which allows the AVL Boost code to accurately replicate a realistic compression ignition combustion cycle within a simulation environment.

2.3 Single and Twin-Entry Turbocharger Models

Figure 2 shows the complete simulation model of the Renault four-cylinder 1.5L DCi engine with a standard single-entry turbocharger configuration. The exhaust manifold has a 4 into 1 geometry which will result in strong flow interactions and turbulent flow mixing of the pulsating exhaust gases [12]. This implies that the energy transfer from the exhaust gases to the impeller of the turbine are not optimized, thus not realizing the full potential of the engine outputs, particularly power and torque.

In order to implement a twin-entry turbine housing in Fig. 2, a modified manifold configuration is introduced with a split-pulse design. By using the known firing order (1-3-4-2) of the original engine, the 4 into 1 manifold has been changed to allow for the exhaust gases from cylinders 1 and 4 and 2 and 3 to remain separate as highlighted in Fig. 3. Therefore the software will recognize the number of exhaust to turbine housing inputs being changed to a corresponding twin-entry configuration.

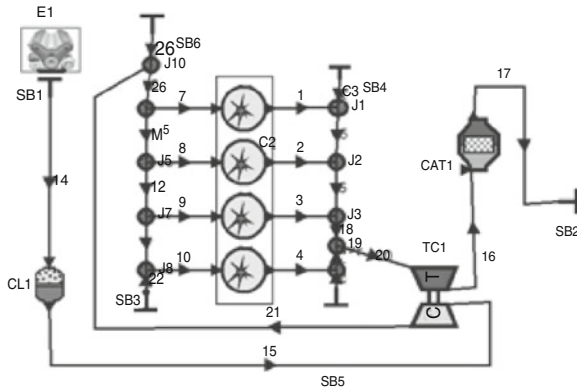


Fig. 2 AVL Boost model of the single-entry turbocharger configuration for the Renault 1.5L DCi engine

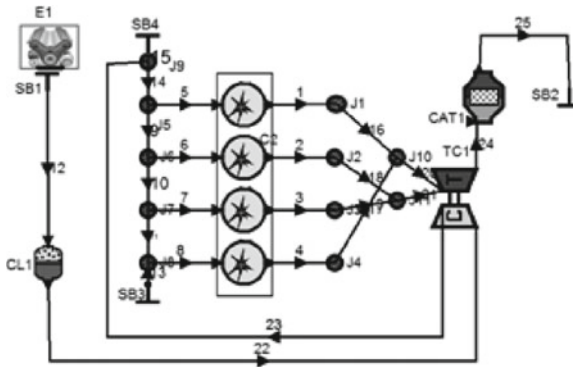


Fig. 3 AVL Boost model of the twin-entry turbocharger configuration for the Renault 1.5L DCi engine

3 AVL Boost Model Validation

Model validation has been performed using parameters of a standard Renault 1.5L DCi engine with a single-entry turbocharger and the results will be compared to those provided by the manufacturer. Data such as peak engine power of 50kW at 4000RPM with the BorgWarner KP35 single-entry turbocharger, and other key engine parameters as shown in Table 1, will be used.

3.1 Maximum Engine Power and Torque

The engine model was run for 100 simulation cycles using the parameters described above. To validate the model engines the performance results are compared to those

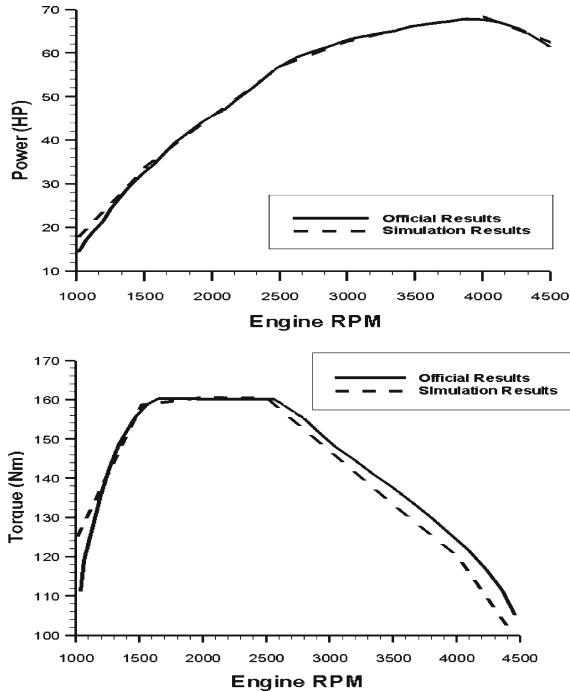


Fig. 4 AVL Boost simulated power and torque results in comparison with published manufacturing data of the 1.5L DCi Renault engine

published by the engines manufacturer as shown for the Renault 1.5L DCi engine in Fig. 4 displaying power and torque output results as a function of engine speed in a range of 1000–4500 RPM. It is clear that the simulated model Renault engine has produced very accurate predictions of peak power and torque values at an engine speed of 4000 and 1500 RPM respectively, when compared to the manufacturer's data. Similar model validations were performed for the Peugeot 2.0L CI (see Table 2) and Audi 1.8L SI engines (see Fig. 5). For the 1.8L SI engine however the torque results exhibit some discrepancies between the simulation and the manufacturer data, particularly the torque curves. This is likely to be attributed due to the inaccuracies of the combustion shape parameter which specifies the combustion characteristics within each cylinder in the AVL Boost simulation code. These characteristics are constantly changing within a running physical engine which means that a fixed number, as specified with the software cannot accurately represent a complete combustion definition.

Table 2 shows that there is only a 1.4% and 2.3% increase in peak power and torque results for the simulated 2.0L CI Peugeot engine indicating that the AVL Boost code has accurately re-produced the operational condition of the 2.0L compression ignition engine.

Table 2 Manufacturer and simulated data acquired by the Boost code for the Peugeot 2.0L compression ignition engine [13]

	Simulation results	Official data	Error in %
Max power	68 kW@4000 RPM	67 kW@4000 RPM	1 kW [1.4 % error]
Max torque	220 Nm@2000 RPM	215 Nm@2000 RPM	5 Nm [2.3 % error]

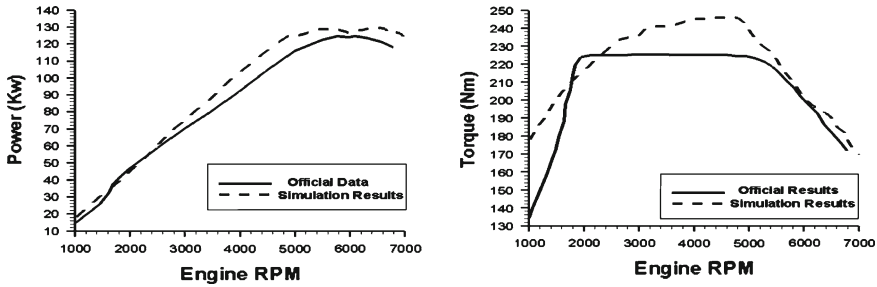


Fig. 5 Comparison between simulated engine and torque results for the Audi 1.8L SI engine to the data specified by the engine manufacturer [13]

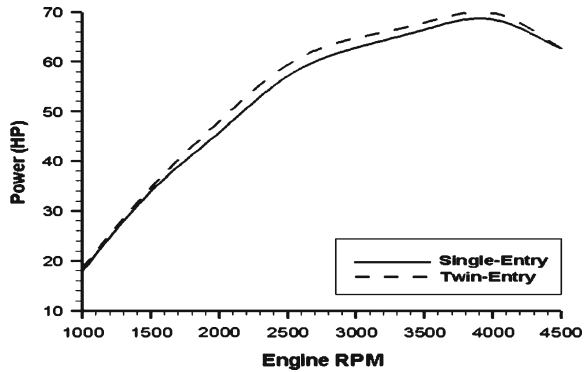
The validation results acquired from the three engines clearly indicate that the AVL Boost 1-D simulations have achieved reliable results considering that combustion, thermodynamic and heat transfer models are used to simulate viable engine operation.

Based on the above validations, it was concluded that the simulated model engines equipped with the standard single-entry turbocharger are working correctly. The models can therefore be subsequently adapted to a twin-entry turbine housing as described in Sect. 2.3. A direct comparison analysis between the single- and twin-entry turbocharger configurations will be used to conduct comprehensive studies concentrating on potential changes in engine performance due to the adoption of a twin-entry turbocharger geometry. These will include engine torque, power and brake mean effective pressure (BMEP).

4 Simulation Results

The engine response at low engine speeds is an area of primary interest when analyzing the application of twin-entry turbochargers for downsized engines. A common problem for turbochargers is the response time that the turbine needs to reach sufficient impeller speeds often known as “spooling time”, in order for the compressor to work effectively, i.e. to produce sufficient boost. Having a long “spooling time” means the engine is susceptible to a long time delay in responsiveness, so-called ‘turbo-lag’, before the effect of the turbocharger becomes effective. It was therefore decided that the engine characteristics in a range of 1000–3500 RPM would be

Fig. 6 Increased power output of the Renault 1.5L DCi engine in 1000–4000 RPM engine speed range using a twin-entry turbocharger



closely investigated as this is the range where the ‘spooling time’ and the ‘turbo-lag’ have the greatest effect. It is expected that the adoption of a twin-entry turbocharger could reduce these undesirable characteristics.

4.1 Power and Torque Outputs

The main benefit of increasing the spooling time of the turbocharger during low engine crankshaft speeds is the improvement in time required for the compressor to reach its optimum boost output. This implies the increase of engine power and torque.

Figure 6 shows the comparison of power output from both single- and twin-entry turbochargers for the Renault 1.5 DCi engine. The greatest gain in power output for the twin-entry configuration was observed at 2500 RPM, providing approximately 2.5 HP of extra power output. This power gain is resultant from the increase in compressor performance due to the improved energy transfer from the exhaust gases to the turbine impeller. When calculated over the complete RPM range (i.e. 1000–4500 RPM) the twin-entry configuration produces 3.27% greater power when compared to the benchmark data. Figure 7 shows the comparison of engine torque acquired from the simulation of the Renault 1.5L DCi engine. Adopting a twin-entry turbine housing has clearly improved the torque characteristics of the engine. For example at 2000 RPM the torque has increased from approximately 160–170 Nm. This increase of 5.55% @ 2000 RPM is highly favorable as the engine response performance will have noticeably improved during the low engine speed range of 1000–3500 RPM.

Similar trends in an increase in engine power output have been revealed by the AVL Boost simulation code where a single- and a twin-entry turbocharger comparative analysis was performed on the Audi 1.8L SI engine as shown in Fig. 8.

The third simulation using a Peugeot 2.0L CI engine was performed using the AVL Boost code to further illustrate the effect of a twin-entry turbine housing on

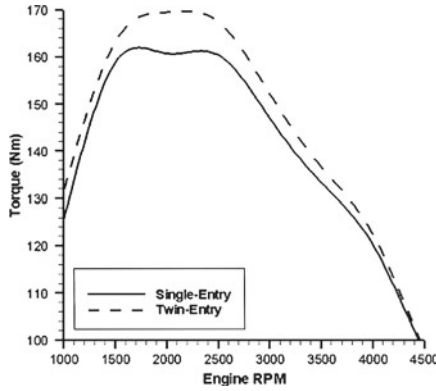


Fig. 7 Increased engine torque output due to the adoption of a twin-entry turbocharger on the Renault 1.5L DCi engine

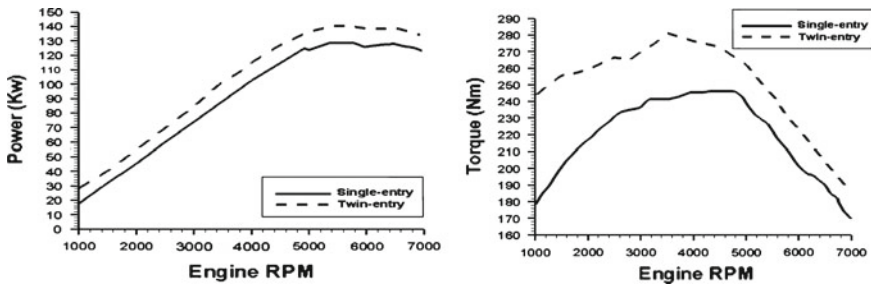


Fig. 8 Increased power and torque outputs for the Audi 1.8L SI engine using a twin-entry turbocharger [13]

the output performance characteristics of the engine. An average increase in power (7.76 %) and torque (7.52 %) calculated from 1000 to 3000 RPM are shown in Fig. 9.

4.2 BMEP Improvement

The additional air flow rate due to the twin-entry turbine configurations also causes an increase in compressor efficiency i.e. compressor discharge pressure, which not only improves the volumetric efficiency (VE) of the engine but also the Break Mean Effective Pressure (BMEP). BMEP is another important parameter used to characterize the performance of engine output and is related to torque as shown in Eq. 1.

$$Torque = \frac{BMEP \times Swept\ Engine\ Volume}{2\pi} \tag{1}$$

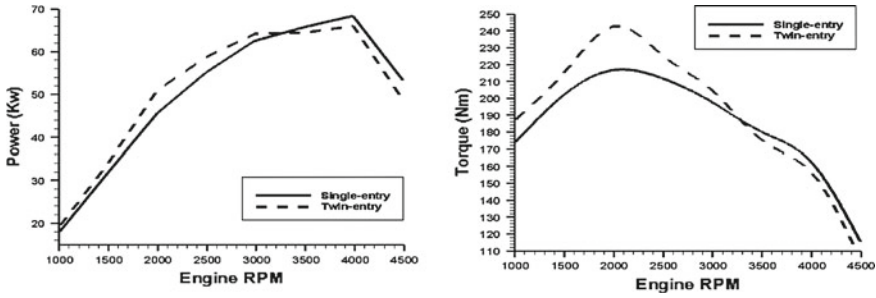
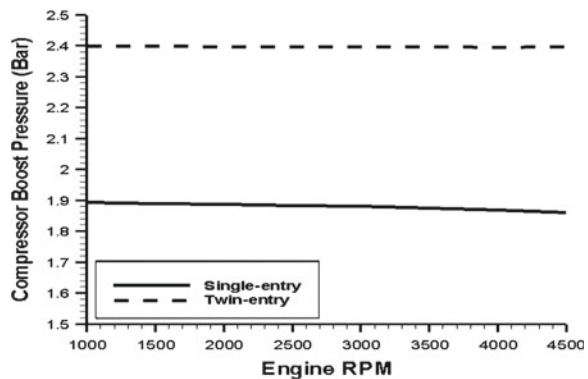


Fig. 9 Increased engine power and torque output for the Peugeot 2.0L CI engine using a twin-entry turbocharger [13]

Fig. 10 Compressor “boost” performance gain increasing the manifold pressure and therefore BMEP



It is clear from the equation that increasing the BMEP of an engine also results in increased torque characteristics, as previously shown in Figs. 8 and 9.

Figure 10 shows a comparison of compressor discharge pressure variations for both single- and twin-entry turbocharger configurations used on the 1.5L DCi engine. A clear increase in engine boost due to the twin-entry turbine housing is illustrated. The increase in pressure, although relatively small (0.5 Bar at 2500RPM) is more favorable as any large rise in discharge boost pressure may lead to a surge condition which could deem it inappropriate for practical application. It is apparent from the discussed results that small gains in the performance of the compressor will provide an improved overall engine performance output, e.g. the increase in compressor air flow resulting in a consequently larger VE and therefore BMEP as shown in Fig. 11. A maximum BMEP increase of approximately 1 Bar is observed at 2000 RPM.

Overall, there is a noticeable increase (4.03 %) in BMEP over an engine speed RPM range of 1000–3500 RPM which is crucial for performance and response of the engine in urban driving environments. It is evident that with only a 0.5 Bar increase in compressor boost pressure the twin-entry configured engine can achieve a 1 Bar

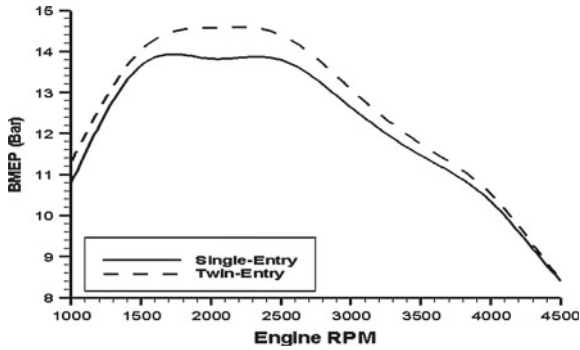


Fig. 11 Comparison between single and twin-entry engine BMEP results of Renault 1.5L DCi engine

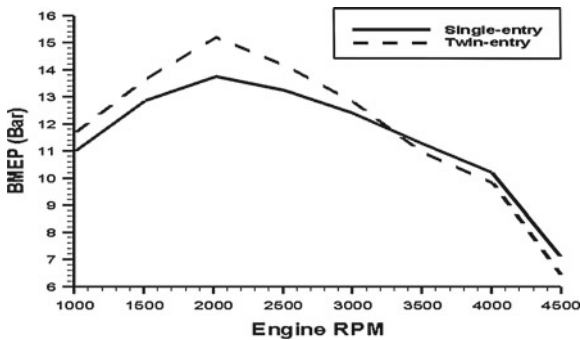


Fig. 12 A 11% improvement at 2000RPM in BMEP for the Peugeot 2.0L CI engine [13]

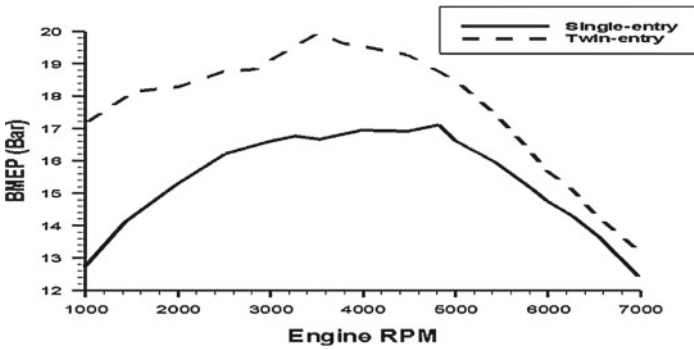


Fig. 13 A 22% improvement in BMEP exhibited by the Audi 1.8L SI engine [13]

increase in the BMEP. This therefore clearly shows that the adoption of a twin-entry turbine housing is more beneficial than a single-entry one.

An improvement in BMEP is also noted from the Peugeot 2.0L CI engine as shown in Fig. 12. There is an overall increase in BMEP from 13.75 to 15.25 Bar at 2000 RPM

Table 3 Average increase in engine performance calculated during 1000–3500 RPM due to the application of a twin-entry turbocharger

	1.5L DCi engine (%)	2.0L CI engine (%)	1.8L SI engine (%)
Power	3.99	7.76	26.46
Torque	4.04	7.52	22.23
BMEP	4.03	6.81	22.43

which equates to a 10.91 % improvement due to the addition of a twin-entry turbine housing. The simulation results acquired from the Audi 1.8L SI engine Fig. 13 also showed that the BMEP increased by 22.43 % during 1000–3500 RPM after the engine model was modified to accept the twin-entry turbocharger configuration.

A summary of the improvements exhibited by the AVL Boost simulation for all three engines is shown in Table 3.

5 Conclusions

The AVL Boost engine simulation code has demonstrated potential performance improvements on a variety of engines due to the adoption of a twin-entry turbocharger with a corresponding split-pulse manifold. The results for the Renault 1.5L DCi engine show that the application of a symmetrical twin-entry volute design enhances the performance of the engine when operating during low RPM conditions, the most effectiveness being observed from 1500 to 3000 RPM showing a maximum 4.03 % increase in BMEP. Both engine torque and power performance also increased by approximately 5.5 % at 2000 RPM resulting in an average performance increase of 4 % within the 1000–3500 engine RPM range. The addition of the extra torque and power is more beneficial during low engine speeds as the turbocharger delay time will be reduced making the engine more responsive to driver input. The “drivability” of the vehicle has therefore also improved.

References

1. Arnold D (2004) Turbocharging technologies to meet critical performance demands of ultra-low emissions diesel engines. SAE International Technical Paper Series, 2004-01-1359
2. Hawley J, Wallace F, Cox A, Horrocks R, Bird G (1999) Variable geometry turbocharging for lower emissions and improved torque characteristics. Proc Inst Mech Eng Part D: J Automobile Eng 213(2):145–159
3. Chadwell CJ, Walls M (2010) Analysis of a superturbocharged downsized engine using 1-D CFD simulation. SAE International Technical Paper Series, 2010-01-1231
4. Ishii M (2009) System optimization of a turbo-compound engine. SAE International Technical Paper Series, 2009-01-1940

5. Petitjean D, Bernardini L, Middlemass C, Shahed SM (2004) Advanced gasoline engine turbocharging technology for fuel economy improvements. SAE International Technical Paper Series, 2004-01-0988
6. Watel E, Pagot A, Pacaud P, Schmitt J (2010) Matching and evaluating methods for Euro 6 and efficient two-stage turbocharging diesel engine. SAE International Technical Paper Series, 2010-01-1229
7. Aghaali H, Hajilouy-Benisi A (2008) Experimental modelling of twin-entry radial turbine. Iran J Sci Technol Trans B Eng 32(B6):571–584
8. Baines N (2011) Concepts NREC, Trends in off-highway turbocharging. SAE Vehicle Engineering
9. AVL Boost V. 2010 Theory manual. Edition 11/2010 Document number 01.0114.2010
10. Weissbaeck M (2011) Diesel downsizing. Engine Technol Int January:26–28
11. AVL Boost V. 2010.1 User guide. Edition 03/2011 Document number 01.0104.2010.1
12. Hillier V, Coombes P (2004) Fundamentals of motor vehicle technology, 5th edn. Nelson Thornes Ltd., Cheltenham
13. Kusztelan A, Marchant D, Yao Y, Wang Y, Selcuk S, Gaikwad A (2012) Increases in low speed response of an IC engine using a twin-entry turbocharger. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE 2012, 4–6 July 2012, London, UK

A Data Mining Approach to Recognize Objects in Satellite Images to Predict Natural Resources

Muhammad Shahbaz, Aziz Guergachi, Aneela Noreen
and Muhammad Shaheen

Abstract This paper presents an approach for the classification of satellite images by recognizing various objects in them. Satellite images are rich in geographical information that can be used in a number of useful ways. The proposed system classifies satellite images by extracting different objects from the images. Our object recognition mechanism extracts attributes from satellite images under two domains namely: color pixels' organization and pixel intensity. The extracted attributes aid in the identification of objects lying inside the satellite images. Once we are able to identify objects, we proceeded further to classify satellite images with the help of decision trees. The system has been tested for a number of satellite images acquired from around the globe. The objects in the images have been further subdivided into different sub categories to improve the classification and prediction process. This is a novel approach which is not using any image processing techniques but is utilizing the extracted features to identify objects and then using these objects to classify the satellite images.

M. Shahbaz (✉)
Department of Computer Science and Engineering,
University of Engineering and Technology, Lahore, Pakistan
e-mail: Muhammad.Shahbaz@gmail.com

A. Guergachi
Information Technology Management, Ted Rogers School,
Ryerson University, Toronto, ON, Canada
e-mail: a2guerga@ryerson.ca

A. Noreen
Department of Computer Science, University of Engineering and Technology,
Lahore, Pakistan
e-mail: aneela.uet@gmail.com

M. Shaheen
Department of Computer Science, FAST NU, Peshawar, Pakistan
e-mail: shaheentanoli@gmail.com

Keywords Classification of images · Data mining · Decision tree · Machine learning · Object recognition · Satellite images

1 Introduction

Remotely sensed images are rich in geographical information by capturing various geographical objects. Geographical information can be useful for different sectors like government, business, science, engineering and research institutes. Geographical information can be used for planning, extraction and analysis of natural resources and help improve the vegetation of an area. These are few examples but there can be gazillions of its advantages.

Remotely sensed images that we can acquire through satellites, sensors and radars are very large in numbers and in size as well. With the advancement of technologies, such as image digitization and storage, quantity of images has also elevated [1]. Each image has huge information residing inside it in the form of objects. It is difficult for humans to go through each image and extract patterns from such images. With the help of state of the art image storage, data analysis and classification techniques it is possible to automate who process to understand hidden patterns and help improve the prediction, in various domains.

A number of techniques have been proposed for object recognition. Adnan A. Y. Mustafa [2] proposed an object recognition scheme by matching boundary signature. For this purpose, he introduced four boundary signatures: the Curvature Boundary Signature (SCB), the Direction Boundary Signature (SAB), the Distance Boundary Signature (SDB) and the Parameter Boundary Signature (SPB). These signatures can be constructed by using boundary's local and global geometric shape attributes. This approach is based on the shape of the object. For shape of object, boundary of object is required to compute. But some objects don't have distinct boundaries. Soil, ecological zone are good examples that don't have distinct sharp boundaries.

Ender Ozcan et al. [3] proposed a methodology for partial shape matching by using genetic algorithms. For this purpose, they described features of model shapes in form of line segments and angles. For recognition purpose, they matched input shape with model shapes. In search, GAs uses a population of individuals of fixed size. New solutions can be produced by using operators.

Farzin Mokhtarian et al. [4] proposed a method that was based on the maxima of curvature zero crossing contours. They used it for representing it as feature vector of Curvature Scale Space (CSS) image that can be used for describing the shapes of object boundary counters. In their proposed matching algorithm, they compared two sets of maxima. Then they assigned a matching value as a measure of similarity.

Kyoung Sig Roh et al. [5] recognize Object by using invariant counter descriptor and projective refinement methodology. For this purpose, they proposed a contour descriptor that consists of a series of geometric invariance of five equally spaced coplanar points on the contour. This descriptor is basically used for indexing a hash table that helps for recognizing 2D curved objects. For verification, they used pro-

jective refinement. They repeatedly compute a projective transformation between model and scene contours. They used resulting pixel error after projective refinement to prove whether a generated hypothesis is true or false.

For recognizing objects, Mariana Tsaneva et al. [6] used transform methods e.g. Gabor filter and wavelet transform for texture analysis. The drawback of this technique is that wavelet performance is not satisfactory in image processing when bandwidth is not high and images are in compressed form.

Jean-Baptiste Bordes et al. [7] prescribed an algorithm that integrates an angular local descriptor and radial one. It also used the co-efficient that are produced as a result of the Fourier analysis. Their proposed algorithm work for roundabout detection but it is not for general object detection. Secondly proposed methodology is not fast enough.

Talibi-Alaoui and Sbihi [8] used unsupervised classification approach of neural networks to classify the textual images. Jiebo Luo et al. [9] proposed a methodology for characterizing satellite images that was based on the color- and structure-based vocabularies. The drawback of this methodology is usage of larger vocabulary. Large vocabulary provides only small gain in accuracy but it requires a greater computation cost.

Most of the object recognition algorithms don't produce satisfactory results when boundaries of the objects are not constant, clear and sharp enough. Desert, sea and ecological zone are well known examples of such objects that don't have distinct boundaries [1]. We propose a methodology that is independent of the shape and boundary of the object. In our proposed object recognition methodology, we shall focus on two major domains that are: pixel intensity and organization of color pixels [10]. We shall use these two techniques to calculate the attributes from the image. Decision tree shall be helpful for taking decision by classifying the objects by using attributes that are extracted from the image for object recognition. The attributes shall be stored in a database that shall serve as knowledge base which will be useful for classification of satellite images based on similarities [11]. The process is initiated with the broader domain of objects like greenery or water but later such objects will be further subdivided into various types such as different kinds of greenery like grass, bushes, trees etc. for greenery and water into sea, pond, river or freshwater etc.

2 System Mode Details and its Description

Our system has two modes, one of them is the operational or testing mode and another one is training mode which is working on testing and training data sets respectively.

The whole data which is collected from Google Earth is divided into two portions, training data and testing data. In training mode, we used half of sample images for each object. After acquiring the images, we employed the pixel intensity and organization of color pixels to each image. As a result of employment of pixel intensity and organization of color pixels on each image, we extracted attributes from each image and then these attributes were stored in the database. Based on the

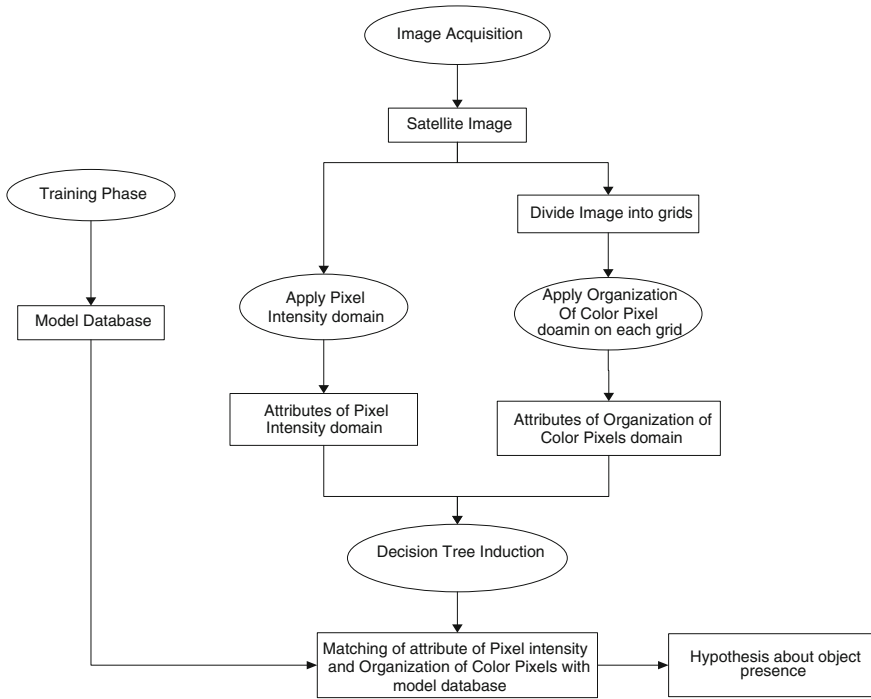


Fig. 1 Block diagram of object recognition system

result of Pixel Intensity and Organization of Color Pixels, we build a decision tree that shall be helpful for object recognition from new incoming images.

In testing mode, we provide a satellite image that has some objects. We employ the pixel intensity and organization of color pixels to incoming satellite image and extract attributes from the image. Afterward the extracted attributes are given to decision tree classifier for making decision about the presence of the object.

Figure 1 shows a block diagram of the object recognition system. The input to the system is the satellite image. For recognizing object from satellite image, first we employed Pixel intensity to extract attributes from the input image. We then employed Organization of color pixels on the image to extract attributes. In organization of color pixels, we divide the image into grids and for each grid we computed attributes of organization of color pixels. On the basis of the attributes extracted under the Pixel intensity domain and organization of color pixel, we have drawn ID3 decision tree with the help of database established in the training mode in which an hypothesis is made about presence of objects in image.

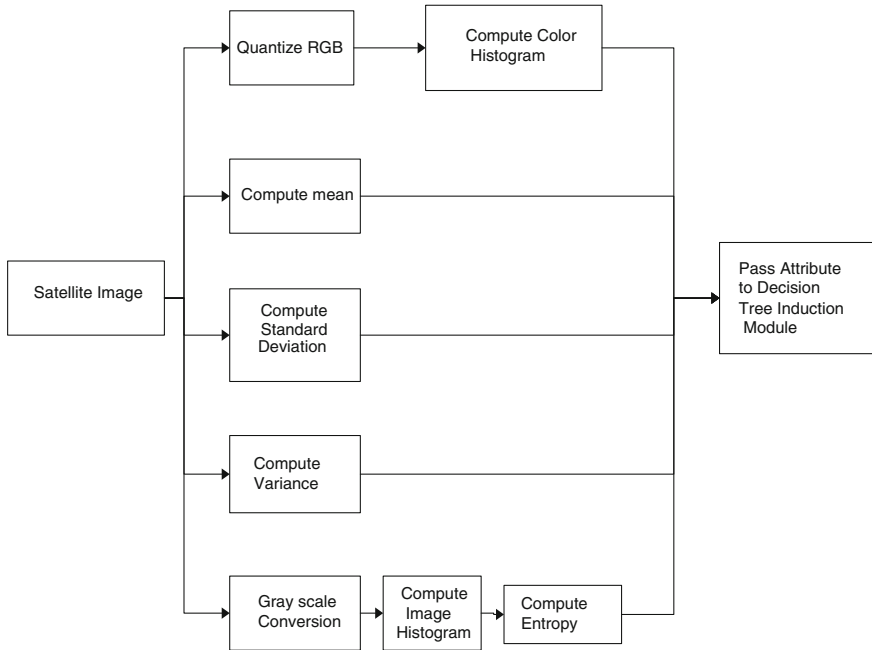


Fig. 2 Block diagram of pixel intensity

2.1 Object Recognition Scheme

An object can be recognized by its form, texture and color. If we deal with the form of the object then boundary of an object is very important. In satellite image one of the major issues is the boundary of the object as some objects like soil don't have distinct boundary [1]. To deal with such a problem, we developed a technique that is independent of the object boundary. Color and texture can play prominent role in object recognition because color is an important factor in object recognition rather than intensity [12]. The proposed object recognition algorithm shall get cue from color and texture of the object for making decision about the presence of the object. The algorithm is also independent of rotational, size and translational invariance.

i. Pixel intensity

In pixel intensity, we deal with the intensity of image. We focus on color histogram and other statistical attributes that are calculated from RGB color space of an image and gray scale image which are the contributors for recognizing objects from the image. We compute a number of attributes under pixel intensity domain that are color histogram of image, mean, standard deviation and variance of RGB color space and entropy of gray scale image.

Figure 2 shows all the attributes of the pixel intensity domain that are computed from satellite image that are acquired through the Google Earth.

Color histogram represents the distribution of color pixels inside the image. In simple words, we can say that the Color histogram is the counting of each color inside an image. Color histogram refers to the joint probabilities of intensities of three color channels. Color histogram can be defined as [13]:

$$h_{R,G,B}[r, b, g] = N \cdot \text{Prob} \{R = r, G = g, B = b\}$$

where R, G and B refer to three color channels and N refers to number of pixels in the image. Color histogram can be computed by counting the number of pixels of each color from an image. It is more convenient approach to convert three channel histogram into a single variable histogram. This transformation can be achieved by using the following equation [13]:

$$m = r + N_r g + N_r N_g b$$

where N_r and N_g are the number of bins for colors red and green respectively and m shall be the single variable histogram.

Histogram can be computed by using the equation [14]:

$$h(m_k) = n_k$$

where m_k is k th single value color, n_k is the number of pixels in the image having single value color m_k and $h(m_k)$ is the single value color histogram .

If we deal an image that lies in RGB color space then the computation complexity of image shall be $256 \times 256 \times 256 = 16,777,216$. For reducing the computation complexity of the color space, we shall prefer to quantize the color space. We can discretize the color of the image into the finite number of bins. In our case we divide the RGB color space into 3 bins. As a result of three bins, the color of image 16,777,216 is quantized to the $3 \times 3 \times 3 = 27$ colors which deteriorate the quality of the results. The process can obviously be improved with a slight decrease in the efficiency by increasing the number of bins from 3 to 4 and so on until we reach at the full intensity of the image.

The range of bin1 lies between 0 and 85 intensities, range of bin2 lies between 86 and 172 intensities and bin 3 lies between 173 and 255 intensities. After quantization of RGB level, we convert three channels color into a single variable. Afterward we compute the color histogram of single color variable.

Texture of object is also useful in object recognition process. For extracting the texture information of the image, we computed the entropy of the image. Entropy is another attribute of our pixel intensity domain. Entropy is a statistical measure of randomness. Entropy can be useful for characterizing the texture of the input image. We can utilize entropy for measuring the information that lies in the signal. Entropy typically represents the uniformity [15, 16].

For computing the entropy of the image, we transform image into the gray scale. After getting the grayscale image, we shall compute the histogram of gray scale image. After computing the histogram, entropy of grayscale image can be computed by using the following equation [17]:

$$Entropy = - \sum p_i \log_2 p_i$$

Another attribute of the pixel intensity domain is the mean of RGB image. Mean is the statistical measure that is used for determining the central value from the distribution of data set. Mean can be computed by the dividing sum of all elements of data set to the number of data set. We can compute the mean of RGB image by using the following equation [18]:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

Another attribute of the pixel intensity domain is the standard deviation of RGB image. Standard deviation is a statistical measure that can be used for measuring the dispersion inside the data set. If we take the square root of the standard deviation then we shall acquire the variance. We can compute the standard deviation of RGB image by using the following equation [18]:

$$S_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

where μ is the mean of all elements of the data set

The last attribute of the pixel intensity domain is the variance of RGB image. Variance is a statistical measure that can be used to measure how far value lies from mean. It is used for measuring dispersion. Variance can be measured by using the following equation [18]:

$$S_N^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

After extracting all of the attributes of the pixel intensity domain, the attributes handover to the decision tree induction module that make a hypothesis about the presence of the object.

ii. *Organization of color pixel*

For improving our object recognition method, we introduced another domain that is Organization of color pixels. In Organization of color pixels, we divide our image into the number of equal size grids and extracts useful attributes from each grid.

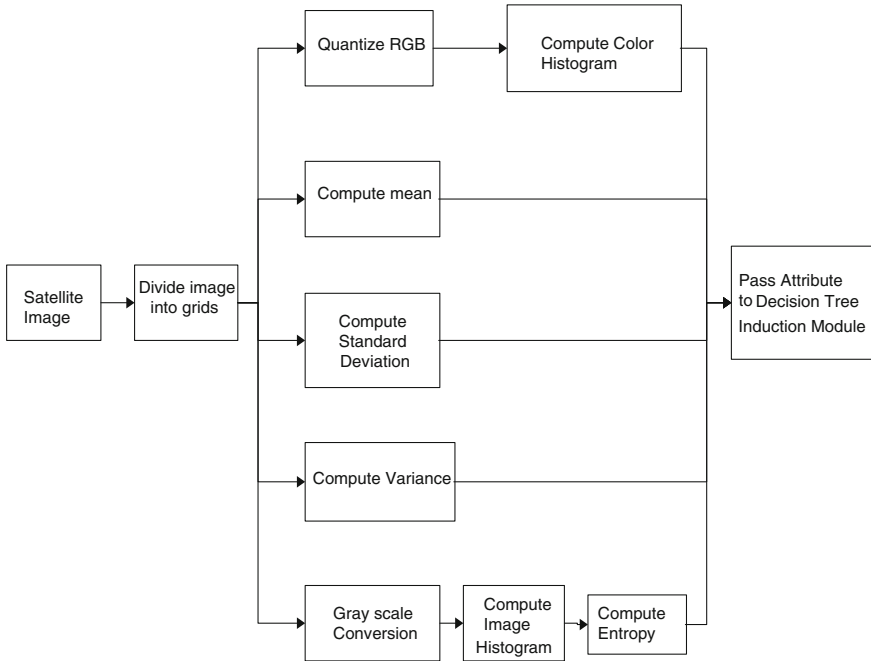


Fig. 3 Block diagram of organization of color pixels

From each grid, we computed a number of useful attributes that might help in recognizing an object from the image. These attributes are color histogram of image, mean, standard deviation and variance of RGB color space and entropy of gray scale image.

Figure 3 is showing that when satellite image is provided to the system for computing the attributes of organization of color pixels then first it is divided into number of grids and afterward all the attributes of the organization of color pixels are computed from each grid. For each grid, we shall compute quantized color histogram, mean, standard deviation, variance and entropy.

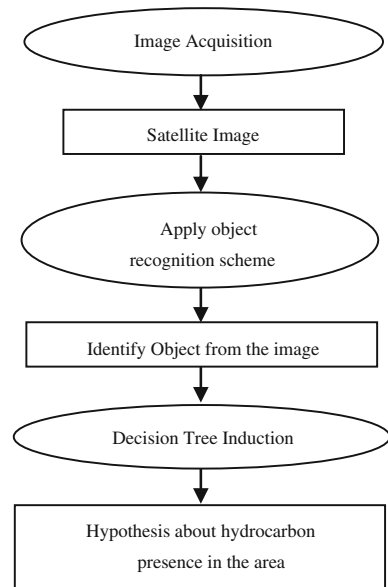
After extracting all of the attributes of organization of color pixel from each grid of an image, the attributes are mapped in a decision tree induction module that makes a hypothesis about the presence of the object.

iii. *Decision tree induction*

This module is helpful in making decision about the object inside an image. We stored the evaluated attributes in a database.

In training phase, a number of objects for each class are taken. For each object belonging to a certain class we extracted the attributes under pixel intensity domain and organization of color pixels. These all attributes are helpful for recognizing objects from satellite image. After acquiring attributes from training phase, we make a decision tree that shall be helpful for making decision about incoming object into the system.

Fig. 4 Block diagram for classification of images



2.2 Classification of Images

The images are extracted from Google Earth and rectified for classification purposes. The objects from the image are recognized on the basis of algorithm proposed in Sect. 3. Afterward recognized objects are passed to a decision tree that makes a decision in which class satellite image lies. We generalized our image classification scheme to determine if hydrocarbons are present in a particular area or not.

Figure 4 shows the block diagram of the image classification scheme. First of all we shall recognize objects from a satellite image then passed identified object to a decision tree that make a decision either hydrocarbons is present in that area or not.

2.3 Summary of Algorithm

We can summarize our proposed algorithm as:

1. Acquire the satellite image from the Google earth.
2. Apply Pixel intensity domain on satellite image for calculating attributes under Pixel Intensity domain.
3. Apply Organization of Color Pixel domain on satellite image. For this purpose, we divide our satellite image into a number of grids and calculate attribute of each grid.
4. Pass calculated attributes from Step 2 and 3 to the decision tree.

5. Decision tree makes a decision about the presence of the objects inside the image.
6. After identification of the object, information is passed over to the classification module.
7. Classification module has a decision tree that makes a decision about the presence of the hydrocarbon inside an area.

3 Verification of the Proposed Scheme

For verifying our purposed scheme, we focused on five main classes of objects that are tree, greenery, water, soil and rocks. Later on these classes are further divided into sub class to refine the results.

Figure 5 shows the mean, standard deviation, variance and entropy of RGB of the five classes. As you can see in the figure that value of each attribute of each class is different from another class that is helpful for differentiate one class from another classes.

For performing experiment, we took a number of satellite images from Google Earth. We took a number of images of each class. For object recognition from satellite images, we used our proposed object recognition scheme as we described in Sect. 3. For each image of each class, we applied the Pixel intensity and Organization of Color Pixel domain. As a result of this we acquired the attribute of both domains.

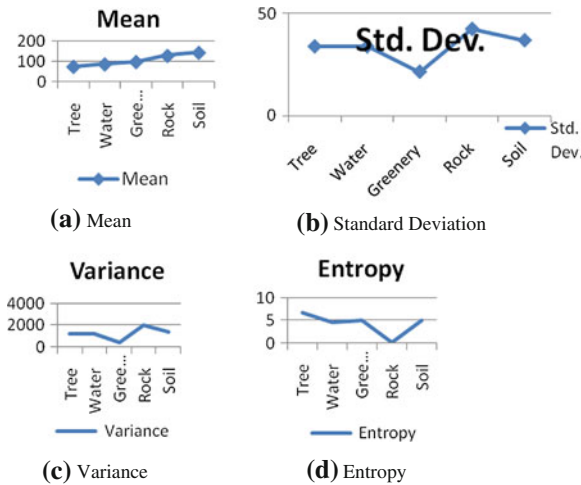


Fig. 5 Attributes of Pixel intensity domain (a) Mean of RGB of the five classes (*Tree, Water, Greenary, Rock and Soil*) (b) Standard deviation of RGB of the five classes (c) Variance of RGB of the five classes and (d) Entropy of RGB of the five classes

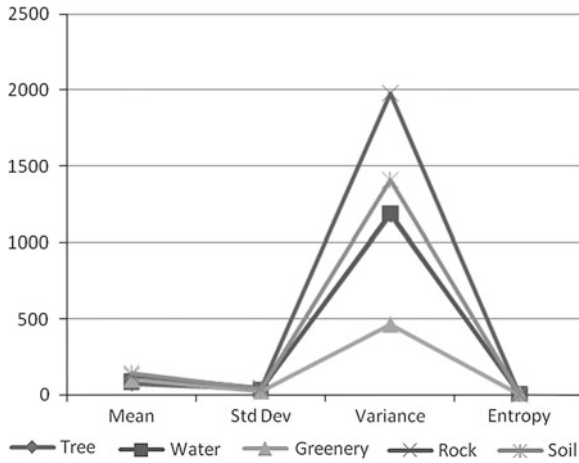


Fig. 6 Mean, Standard deviation, Entropy and Variance of the five classes

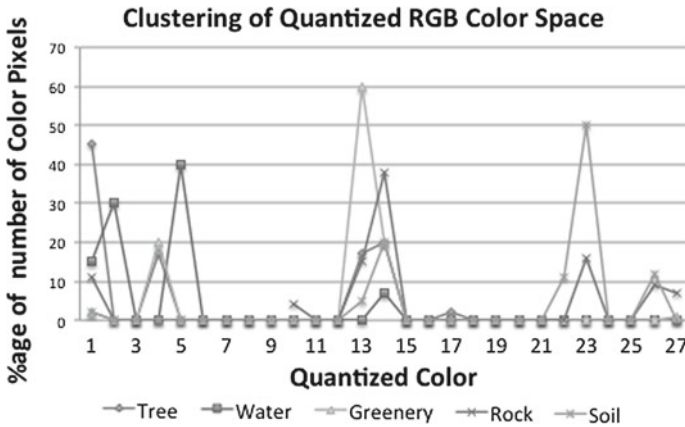


Fig. 7 Quantized color histogram of the tree, water, greenery, rock and soil

Figure 6 shows the mean, standard deviation, entropy and variance of the five classes. These four attributes play a vital role in making a decision about object recognition.

Figure 7 shows the quantized color histogram of the tree, water, greenery, rock and soil. As you can see in the figure that peak of histogram varies from class to class that is helpful for differentiate each class from another class.

Figure 7a shows quantized color histogram of tree. As you can see in the figure that peak of histogram is remarkably high for the color set 1, 4, 13 and 14. These colors play an important role in differentiating tree from other objects. Figure 7b shows the quantized color histogram of the water. As you can see, peak of histogram is remarkably high for the color set 1, 2, 5 and 14. Figure 7c shows the quantized

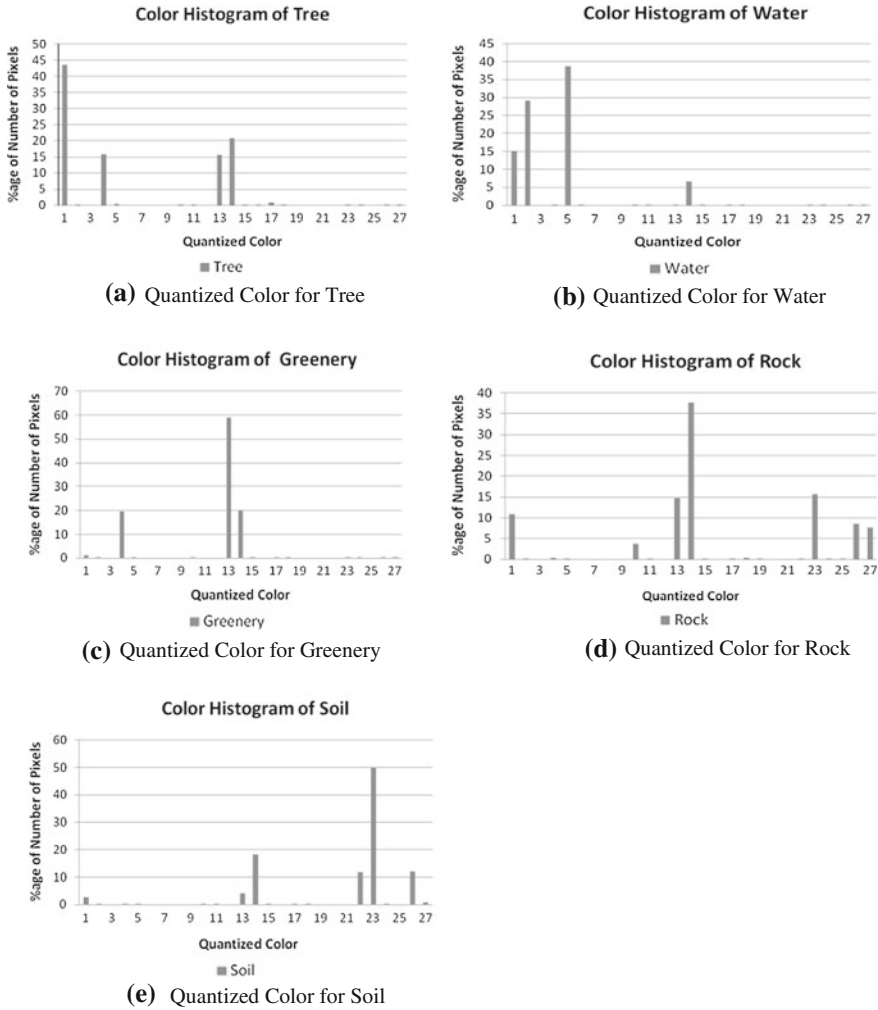


Fig. 8 Attribute of pixel intensity domain: Color histogram: **a** Color Histogram of *Tree* **b** Color histogram of *Water* **c** Color histogram of *Greenery* **d** Color histogram of *Rock* and **e** Color histogram of *Soil*

color histogram of the greenery. As you can see for greenery class that peak of histogram is remarkably high for the color set 4, 13 and 14. Figure 8d shows the quantized color histogram of the rock. As you can see in the figure that peak of histogram is remarkably high for the color set 1, 13, 14, 23, 26 and 27. Figure 7e shows the quantized color histogram of the soil. As you can see in the figure that peak of histogram is remarkably high for the color set 13, 14, 22, 23 and 26.

In order to improve the object refinement the objects can further be subdivided into sub classes for all the main five classes used for this experimentation. Rock can further

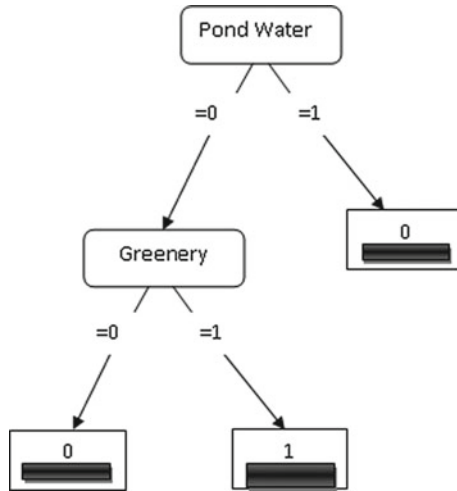


Fig. 9 Decision tree for classification of hydrocarbon

subdivided into Andesite, Boninite, Pegmatite etc. Similarly water can be of various types such as pond water, river water, sea water etc. These subclasses of objects can remarkably increase the classification results after refine object identification.

In our experiment we divided water into further three subclasses i.e. pond, river and sea water. Figure 8b shows the color quantized histogram of water. The water subclass’s quantized histograms are shown in Fig. 10.

Figure 10 shows the quantized color histogram of the Pond River and Sea water. As you can see in the figure that peak of Pond histogram is remarkably high for the color set 1 and 14. These colors play an important role in differentiating pond from other subclasses. For River Water the peak of histogram is remarkably high for the color set 2 and 5. These colors differentiate river from other subclasses. Similarly peak of histogram for sea water is remarkably high for the color set 5 and 6. These colors play an important role in differentiating Sea from other subclasses.

In order to verify the working of the algorithm for the recognition of various water types, the results are tabulated in the form of confusion matrix show in Table 1.

River water result is not 100 percent accurate. Samples of river water are sometimes identified as both river and sea water. This problem can be further solved if we incorporate higher number of RGB color bins.

After extracting attributes from the pixel intensity and organization of color pixel, these attributes are hand over to the decision tree induction module for making hypothesis about the presence of the object. For identifying objects from an image, this information was passed over to the image classification module that decides either hydrocarbon is present in the area of given image or not.

In our experiments, we consider only five types of the classes: tree, water, greenery, rock and soil. These classes can be useful for finding the presence of the hydrocarbon in an area. We took total 42 samples of different area from which some of area has hydrocarbon and others don’t have. We make a database that has the information of

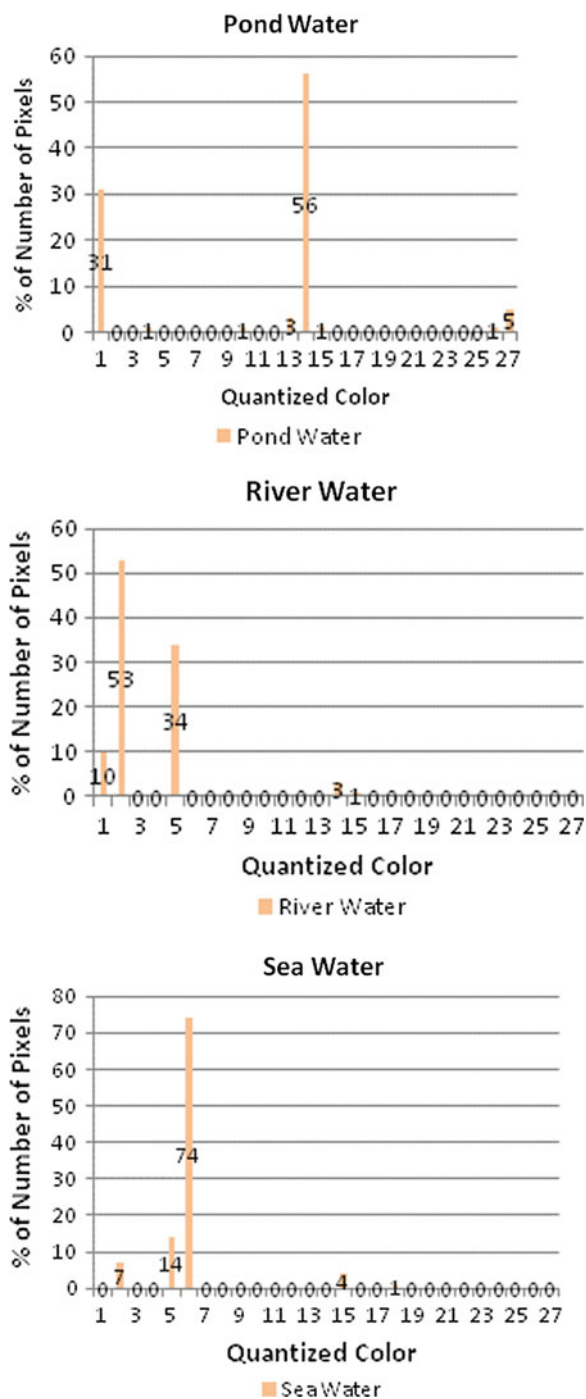


Fig. 10 Color Histogram of water subclasses *Pond*, *River* and *Sea water*

Table 1 Confusion matrix of subclasses of water

	Pond water	River water	Sea water
Pond water	3	0	0
River water	0	2	2
Sea water	0	0	6

the five classes and presence of hydrocarbons. From this data, we make decision tree on the base of presence of hydrocarbons and our five classes: tree, water, greenery, rock and soil. Rapid miner is used to build the decision tree and the result is shown in Fig. 9.

The figure shows the decision tree for classification of hydrocarbon. In Fig. 9, 1 indicates the presence and 0 indicates the absence. From decision tree, we can conclude that water and greenery play an important role in making decision about the presence of hydrocarbons in an area. From decision tree, we can easily infer that if there is water in an area then there can be chance of presence of hydrocarbons in that area. If water is not present but greenery is present then there can also be chance of hydrocarbons in that region.

4 Conclusion

In this research paper, we have proposed a method for classification of satellite images by using data mining technique. Major concern in classification of images is the object recognition from satellite image. Our object recognition technique is based on two domains that are Pixel intensity and organization of color pixels. All of the attributes that are extracted under Pixel Intensity and Organization of Color Pixel domain are helpful for making hypothesis about object's identification. Our purposed object recognition scheme is also independent of rotational, size and translational invariance. It is also computationally efficient. After recognition of objects from an image, information is handed over to decision tree that makes decision about the presence of hydrocarbon inside that area.

Our proposed methodology has been tested for 42 satellite images. The experimental results are satisfactory and show that the system's accuracy is around 80%. The proposed model shows a higher degree of robustness and accuracy for the object recognition process.

We focused on only five classes: tree, water, greenery, rock and soil. Our image classification scheme can be improved by adding more classes. Addition of sub classes of these classes (tree, water, greenery, rock and soil) can also boost the performance of our proposed methodology. For example: sub class of water can be lake water, sea water, river water and water covered with greenery etc.

For hydrocarbon classification we just consider the classes that were related to natural phenomena. Tree, water, greenery, rock and soil are part of nature. Similar classification scheme can be employed for enemy attack detection but the classes used for this activity should be related to enemy attack. Unnecessary human movement, presence of tanks, unwanted bushes and sand blocks etc. can be an indication of enemy attack.

References

1. Aghbari ZA (2009) Effective image mining by representing color histograms as time series. *JACIII* 13(2):109–114
2. Mustafa AAY, Shapiro LG, Ganter MA (1996) 3D object recognition from color intensity images. In: 13th International Conference on Pattern Recognition, Vienna, Austria, pp 25–30
3. Ozcan E, Mohan CK (1997) Partial shape matching using genetic algorithms. *Pattern Recogn Lett Elsevier Sci* 18:987–992
4. Farzin M, Abbasi S, Kittler J (1996) Robust and efficient shape indexing through curvature scale space. *British Machine Vision Conference*, Edinburgh, pp 53–62
5. Roh KS, Kweon IS (1998) 2-D object recognition using invariant contour descriptor and projective refinement. *Pattern Recognition*, Elsevier V31(4):441–455
6. Tsaneva M, Petkov D (2007) Recognition of objects on the Earth's surface through texture analysis of satellite images. In: *Proceeding of the third scientific conference with international participation Space, Ecology, Nanotechnology, Safety Varna, Bulgaria*, pp 27–29
7. Bordes J-B, Roux M (2006) Detection of roundabouts in satellite image. *ISPRS*, Ankara (Turkey)
8. Talibi-Alaoui M, Sbihi A (2012) Application of a mathematical morphological process and neural network for unsupervised texture image classification with fractal features. *IAENG Int J Comput Sci* 39(3):286–294
9. Luo J, Hao W, McIntyre D, Joshi D, Yu J (2008) Recognizing picture-taking environment from satellite images: a feasibility study. *Pattern Recognition*, ICPR
10. Shahbaz M et al. (2012) Classification by object recognition in satellite images by using data mining. In: *Lecture notes in engineering and computer science: proceedings of the world congress on engineering*, WCE 2012, 4–6 July, London, UK, pp 406–416
11. Lorrenz P (2010) Classification of incomplete data by observation. *Engineering Letters*, ISSN: 1816-0948 (online) 1816-093x (print), 18(4):1–10
12. Seinstra FJ, Geusebroek J-M (2006) ECCV Workshop on computation Intensive Mthods for Computer Vision, 9th European Conference on Computer Vision, Graz, Austria, 7–13 May, 2006
13. Smith JR, Chang S-F (1996) Tools and techniques for color image retrieval. In: *Symposium on electronic imaging: science and technology—storage and retrieval for image and video databases IV*, vol. 2670, IS&T/SPIE, San Jose, USA
14. Gonzalez RC, Woods RE (2001) *Digital image processing*. 2nd edn. Prentice Hall, pp 120–139. ISBN:0201180758
15. Petrou M, Sevilla PG (2006) *Image processing dealing with texture*. Wiley, pp 282–529. ISBN: 047002628
16. *Image Processing Toolbox*, For use with MATLAB, The Maths Work Inc, 2001
17. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423, 623–656
18. Kenney JF, Keeping ES (1962) *Mathematics of statistics*. Pt. 1, 3rd edn. Princeton, Van Nostrand

Handling the Data Growth with Privacy Preservation in Collaborative Filtering

Xiwei Wang and Jun Zhang

Abstract The emergence of electric business facilitates people in purchasing merchandises over the Internet. To sell the products better, online service providers use recommender systems to provide recommendations to customers. Most recommender systems are based on collaborative filtering (CF) technique. This technique provides recommendations based on users' transaction history. Due to the technical limitations, many online merchants ask a third party to help develop and maintain recommender systems instead of doing that themselves. Therefore, they need to share their data with these third parties and users' private information is prone to leaking. Furthermore, the fast data growth should be handled by the data owner efficiently without sacrificing privacy. In this chapter, we propose a privacy preserving data updating scheme for collaborative filtering purpose and study its performance on two different datasets. The experimental results show that the proposed scheme does not degrade recommendation accuracy and can preserve a satisfactory level of privacy while updating the data efficiently.

Keywords Collaborative filtering · Data growth · Missing value imputation · Non-negative matrix factorization · Privacy preservation · Singular value decomposition · Updating

1 Introduction

A recommender system is a program that utilizes algorithms to predict users' purchase interests by profiling their shopping patterns. With the help of recommender systems, online merchants (also referred to as data owners) could better sell their

X. Wang (✉) · J. Zhang

Department of Computer Science, University of Kentucky, Lexington, KY40506-0633, USA
e-mail: xiwei@cs.uky.edu

J. Zhang

e-mail: jzhang@cs.uky.edu

products to the returning customers. Most recommender systems are based on collaborative filtering (CF) techniques, e.g., item/user correlation based CF [9], SVD (singular value decomposition) based latent factor CF [10]. Due to the technical limitations, many online merchants buy services from professional third parties to help build their recommender systems. In this scenario, merchants need to share their commercial data with the third party which has the potential for privacy leakage of user information. Typical private information in transaction data includes, but is not limited to, the ratings of a user left on particular items and items that this user has rated. People would not like others (except for the website where they purchased the products) to know this information. Thus privacy preserving collaborative filtering algorithms [4, 8] were proposed to tackle the problem. It is obvious that the data should be perturbed by online merchants before they release it to the third party.

Besides the privacy issue, data owner has to take care of the fast growing data as well. Once new data arrives, e.g., new items or new users' transaction records, it should be appended to the existing data. To protect privacy, data owner needs to do the perturbation. If he just simply redoes the perturbation on the whole data, he needs to resend the full perturbed data to the third party. This process not only takes a large amount of time to perturb data but also requires the model to be rebuilt on the site of the third party. It is infeasible to provide fast real-time recommendations.

In this chapter, we propose a privacy preserving data updating scheme in collaborative filtering. This scheme is based on truncated SVD updating algorithms [2, 7] which can provide privacy protection when incorporating new data into the original one in an efficient way. We start with the pre-computed SVD of the original data matrix. New rows/columns are then built into the existing factor matrices. We try to preserve users' privacy by truncating new matrix together with randomization and post-processing. Two missing value imputation methods during the updating process are studied as well. Results of the experiments that are conducted on MovieLens dataset [10] and Jester dataset [6] show that our scheme can handle data growth efficiently and keep a low level of privacy loss. The prediction quality is still at a good level compared with most published results.

The remainder of this chapter is organized as follows. Section 2 outlines the related work. Section 3 defines the problem and related notations. Section 4 describes the main idea of the proposed scheme. Section 5 presents the experiments on two datasets and discusses the results. Some concluding remarks and future work are given in Sect. 6.

2 Related Work

Most CF models suffer from the privacy leakage issue. Users' private information is fully accessible without any disguise to the provider of recommender systems. Canny [4] first proposed the privacy preserving collaborative filtering (PPCF) that deals with the privacy leakage in the CF process. In his distributed PPCF model, users could control all of their data. A community of users can compute a public

“aggregate” of their data that does not expose individual users’ data. Each user then uses local computation to get personalized recommendations. Nevertheless, most popular collaborative filtering techniques are based on a central server. In this scheme, users send their data to a server and they do not participate in the CF process; only the server needs to conduct the CF. Polat and Du [8] adopted randomized perturbation for both correlation-based CF and SVD-based CF to provide privacy protection. They use uniform distribution and Gaussian distribution to generate random noise and apply the noise to the original data. Differing from Canny, Polat and Du, we focus on the framework in which data owner has all original user data but he needs to do perturbation before releasing it to a third party [15].

In this framework, data owner should be able to handle the fast data growth without leaking the privacy. Among all data perturbation methods, SVD is acknowledged as a simple and effective data perturbation technique. Stewart [12] surveyed the perturbation theory of singular value decomposition and its application in signal processing. A recent work by Tougas and Spiteri [13] demonstrated a partial SVD updating scheme that requires one QR factorization and one SVD (on small intermediate matrices and thus not expensive in computation) per update. Based on their work, Wang et al. [14] presented an improved SVD-based data value hiding method and tested it with clustering algorithm on both synthetic data sets and real data sets. Their experimental results indicate that the introduction of the incremental matrix decomposition produces a significant increase in speed for the SVD-based data value hiding model. Our scheme is similar to this model but we have modified the SVD updating algorithm with missing value imputation and post-processing so that it can be incorporated into collaborative filtering process smoothly.

3 Problem Description

Suppose the data owner has a sparse user-item rating matrix, denoted by $R \in \mathbb{R}^{m \times n}$, where there are m users and n item. When new users’ transactions are available, the new rows, denoted by $T \in \mathbb{R}^{p \times n}$, should be appended to the original matrix R , as

$$\begin{bmatrix} R \\ T \end{bmatrix} \rightarrow R' \tag{1}$$

Similarly, when new items are collected, the new columns, denoted by $F \in \mathbb{R}^{m \times q}$, should be appended to the original matrix R , as

$$\begin{bmatrix} R & F \end{bmatrix} \rightarrow R'' \tag{2}$$

In both cases, data owner could not simply release T or F because they contain the real user ratings. He can not directly release R' or R'' either due to the scalability and privacy issues. An ideal case is, suppose a perturbed version (SVD-based) of

R with privacy protection has been released, data owner only releases the perturbed incremental data, say \tilde{T} and \tilde{F} which preserves users' privacy and does not degrade the recommendation quality.

4 Privacy Preserving Data Updating Scheme

In this section, we will present the data updating scheme that could preserve the privacy during the whole process. We try to protect users' privacy in three aspects, i.e., missing value imputation, randomization-based perturbation and SVD truncation. The imputation step can preserve the private information—"which items that a user has rated". Meanwhile, a second phase perturbation done by randomization and truncated SVD techniques solves another problem—"what are the actual ratings that a user left on particular items". On one hand, random noise can alter the rating values a bit while leaving the distribution unchanged. On the other hand, truncated SVD is a naturally ideal choice for data perturbation. It captures the latent properties of a matrix and removes the useless noise.

In (1), we see that T is added to R as a series of rows. The new matrix R' has a dimension of $(m + p) \times n$. Assuming the truncated rank- k SVD of R has been computed previously,

$$R_k = U_k \Sigma_k V_k^T \quad (3)$$

where $U_k \in \mathbb{R}^{m \times k}$ and $V_k \in \mathbb{R}^{n \times k}$ are two orthogonal matrices; $\Sigma_k \in \mathbb{R}^{k \times k}$ is a diagonal matrix with the largest k singular values on its diagonal.

Since SVD cannot work on an incomplete matrix, we should impute the missing values in advance. We would like to use two different imputation methods: mean value imputation and WNMTF (weighted non-negative matrix tri-factorization) imputation.

In mean value imputation [10], we calculate each column mean and impute all the missing values in every column with its mean value.

In WNMTF [5, 16] imputation, we use WNMTF to factorize an incomplete matrix $A \in \mathbb{R}^{m \times n}$ into three factor matrices, i.e., $W \in \mathbb{R}^{m \times l}$, $G \in \mathbb{R}^{l \times t}$, and $H \in \mathbb{R}^{n \times t}$, where l and t are the column ranks of W and H . The objective function of WNMTF is

$$\min_{W \geq 0, G \geq 0, H \geq 0} f(A, Y, W, G, H) = \|Y \circ (A - WGH^T)\|_F^2 \quad (4)$$

where $Y \in \mathbb{R}^{p \times n}$ is the weight matrix that indicates the value existence in the matrix A .

The update formula corresponds to (4) is

$$W_{ij} = W_{ij} \frac{[(Y \circ A)HG^T]_{ij}}{\{[Y \circ (WGH^T)]HG^T\}_{ij}} \quad (5)$$

$$G_{ij} = G_{ij} \frac{[W^T(Y \circ A)H]_{ij}}{\{W^T[Y \circ (WGH^T)]H\}_{ij}} \quad (6)$$

$$H_{ij} = H_{ij} \frac{[(Y \circ A)^T WG]_{ij}}{\{[Y \circ (WGH^T)]^T WG\}_{ij}} \quad (7)$$

With either of the above two imputation methods, we obtain the imputed matrices: \hat{R} (with its factor matrices \hat{U}_k , $\hat{\Sigma}_k$ and \hat{V}_k) and \hat{T} . Now the problem space has been converted from (1) to (8):

$$\begin{bmatrix} \hat{R} \\ \hat{T} \end{bmatrix} \rightarrow \hat{R}' \quad (8)$$

After imputation, random noise that obeys Gaussian distribution is added to the new data \hat{T} , yielding \dot{T} . Then we follow the procedure in Tougas and Spiteri [13] to update the matrix. First, a QR factorization is performed on $\ddot{T} = (I_n - \hat{V}_k \hat{V}_k^T) \dot{T}^T$, where I_n is an $n \times n$ identity matrix. Thus we have $Q_T S_T = \ddot{T}$, in which $Q_T \in \mathbb{R}^{n \times p}$ is an orthonormal matrix and $S_T \in \mathbb{R}^{p \times p}$ is an upper triangular matrix. Then

$$\begin{aligned} \hat{R}' &= \begin{bmatrix} \hat{R} \\ \hat{T} \end{bmatrix} \approx \begin{bmatrix} \hat{R}_k \\ \hat{T} \end{bmatrix} \approx \begin{bmatrix} \hat{R}_k \\ \dot{T} \end{bmatrix} \\ &= \begin{bmatrix} \hat{U}_k & 0 \\ 0 & I_p \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_k & 0 \\ \dot{T} \hat{V}_k & S_T^T \end{bmatrix} [\hat{V}_k \ Q_T]^T \end{aligned} \quad (9)$$

Next, we compute the rank- k SVD on the middle matrix, i.e.,

$$\begin{bmatrix} \hat{\Sigma}_k & 0 \\ \dot{T} \hat{V}_k & S_T^T \end{bmatrix}_{(k+p) \times (k+p)} \approx U'_k \Sigma'_k V'_k{}^T \quad (10)$$

Since $(k + p)$ is typically small, the computation of SVD should be very fast. Same as in Wang et al. [14], we compute the truncated rank- k SVD of \hat{R}' instead of a complete one,

$$\hat{R}'_k = \left(\begin{bmatrix} \hat{U}_k & 0 \\ 0 & I_p \end{bmatrix} U'_k \right) \Sigma'_k ([\hat{V}_k \ Q_T] V'_k)^T \quad (11)$$

In CF context, the value of all entries should be in a valid range. For example, a valid value r in MovieLens should be $0 < r \leq 5$. Therefore, after obtaining the truncated new matrix \hat{R}'_k , a post-processing step is applied to it so that all invalid values will be replaced with reasonable ones. The processed version of \hat{R}'_k is denoted by $\Delta \hat{R}'_k$.

In our scheme, we assume the third party owns \hat{R}_k so we only send ΔT ($\Delta T = \Delta \hat{R}'_k(m + 1 : m + p, :) \in \mathbb{R}^{p \times n}$)¹ to them.

The following algorithm summarizes the SVD-based row/user updating.

¹ $\Delta \hat{R}'_k(m + 1 : m + p, :)$ is a Matlab notation that means the last p rows of $\Delta \hat{R}'_k$

Privacy Preserving Row Updating Algorithm {**INPUT:**Pre-computed rank- k SVD of \hat{R} : \hat{U}_k , $\hat{\Sigma}_k$ and \hat{V}_k ;New data $T \in \mathbb{R}^{p \times n}$;Parameters for Gaussian random noise: μ and σ ;**OUTPUT:**SVD for updated full matrix: \hat{U}'_k , $\hat{\Sigma}'_k$ and \hat{V}'_k ;Perturbed new data: ΔT ;Impute the missing values in $T \rightarrow \hat{T}$;Apply random noise $X(X \sim N(\mu, \sigma))$ to $\hat{T} \rightarrow \hat{\hat{T}}$;Perform QR factorization on $\hat{\hat{T}} = (I_n - \hat{V}_k \hat{V}_k^T) \hat{\hat{T}} \rightarrow Q_T S_T$;Perform SVD on $\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_k & 0 \\ \hat{T} & S_T \end{bmatrix} \rightarrow \hat{\Sigma} \approx U'_k \hat{\Sigma}'_k V'^T_k$;Compute $\left(\begin{bmatrix} \hat{U}_k & 0 \\ 0 & I_p \end{bmatrix} \cdot U'_k \right) \rightarrow \hat{U}'_k$ Compute $\left(\begin{bmatrix} \hat{V}_k & Q_T \end{bmatrix} \cdot V'_k \right) \rightarrow \hat{V}'_k$ Compute the rank- k approximation of $\hat{R}' \rightarrow \hat{R}'_k = \hat{U}'_k \hat{\Sigma}'_k \hat{V}'_k{}^T$;Process the invalid values $\rightarrow \Delta \hat{R}'_k$; $\Delta \hat{R}'_k(m+1 : m+p, :) \rightarrow \Delta T$;Return \hat{U}'_k , $\hat{\Sigma}'_k$, \hat{V}'_k , and ΔT .

}

Like row updating, the column/item updating algorithm is presented as follows.

Privacy Preserving Column Updating Algorithm {**INPUT:**Pre-computed rank- k SVD of \hat{R} : \hat{U}_k , $\hat{\Sigma}_k$ and \hat{V}_k ;New data $F \in \mathbb{R}^{m \times q}$;Parameters for Gaussian random noise: μ and σ ;**OUTPUT:**SVD for updated full matrix: \hat{U}''_k , $\hat{\Sigma}''_k$ and \hat{V}''_k ;Perturbed new data: ΔF ;Impute the missing values in $F \rightarrow \hat{F}$;Apply random noise $X(X \sim N(\mu, \sigma))$ to $\hat{F} \rightarrow \hat{\hat{F}}$;Perform QR factorization on $\hat{\hat{F}} = (I_m - \hat{U}_k \hat{U}_k^T) \hat{\hat{F}} \rightarrow Q_F S_F$;Perform SVD on $\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_k & \hat{U}_k^T \hat{\hat{F}} \\ 0 & R_F \end{bmatrix} \rightarrow \hat{\Sigma} \approx U''_k \hat{\Sigma}''_k V''_k{}^T$;Compute $\left(\begin{bmatrix} \hat{U}_k & Q_F \end{bmatrix} U''_k \right) \rightarrow \hat{U}''_k$ Compute $\left(\begin{bmatrix} \hat{V}_k & 0 \\ 0 & I_q \end{bmatrix} V''_k \right) \rightarrow \hat{V}''_k$ Compute the rank- k approximation of $\hat{R}'' \rightarrow \hat{R}''_k = \hat{U}''_k \hat{\Sigma}''_k \hat{V}''_k{}^T$;Process the invalid values $\rightarrow \Delta \hat{R}''_k$; $\Delta \hat{R}''_k(:, n+1 : n+q) \rightarrow \Delta F$;Return \hat{U}''_k , $\hat{\Sigma}''_k$, \hat{V}''_k and ΔF .

}

Data owner should keep the updated SVD of new user-item rating matrix (\hat{U}'_k, Σ'_k and \hat{V}'_k for row updating, \hat{U}''_k, Σ''_k and \hat{V}''_k for column updating) for future update and the perturbed new data matrix (ΔT for row updating, ΔF for column updating) as a reference.

5 Experimental Study

In this section, we discuss the test dataset, prediction model, evaluation strategy and experimental results.

5.1 Data Description

As most research papers in collaborative filtering, we adopt both 100K MovieLens [10] and Jester [6] datasets as the test data. The 100K MovieLens dataset has 943 users and 1,682 items. The 100,000 ratings, ranging from 1 to 5, were divided into two parts: the training set (80,000 ratings) and the test set (20,000 ratings). The Jester datasets has 24,983 users and 100 jokes with 1,810,455 ratings ranging from -10 to $+10$. In our experiment, we pick up 5,000 users together with their ratings and randomly select 80% ratings as the training set while use the rest as test set.

5.2 Prediction Model and Error Measurement

In our experiments, we build the prediction model by SVD-based CF algorithm [10]. For a dense matrix A , its rank- k SVD is $A \approx \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^T$. Compute the user factor matrix ($UF \in \mathbb{R}^{m \times k}$) and item factor matrix ($IF \in \mathbb{R}^{n \times k}$):

$$UF = \tilde{U}_k \sqrt{\tilde{\Sigma}_k}, \quad IF = \tilde{V}_k \sqrt{\tilde{\Sigma}_k} \quad (12)$$

The predicted rating for user i left on item j is computed by taking the inner product of the i th row of UF and the j th row of IF :

$$p_{ij} = (\tilde{U}_k \sqrt{\tilde{\Sigma}_k})_i (\tilde{V}_k \sqrt{\tilde{\Sigma}_k})_j^T \quad (13)$$

When testing the prediction accuracy, we build the SVD model on training set; then for every rating in the test set, we compute the corresponding predicted value and measure the difference. We use mean absolute error (MAE) [3, 11] as the specific error metric (the lower the better).

5.3 Privacy Measurement

When we measure the privacy, we mean to what extent the original data could be estimated if given the perturbed data. In this chapter, we use the privacy measure that was first proposed by Agrawal and Aggarwal [1] and was applied to measure the privacy loss in collaborative filtering by Polat and Du [8].

In [1], they proposed $\Pi(Y) = 2^{h(Y)}$ as the privacy inherent in a random variable Y with $h(Y)$ as its differential entropy. Thus given a perturbed version of Y , denoted by X , the average conditional privacy(also referred to as Privacy Level) of Y given X is $\Pi(Y|X) = 2^{h(Y|X)}$.

Similar with Polat and Du's work, we take $\Pi(Y|X)$ as privacy measure to quantify the privacy in the experiments. Note that in this chapter, Y corresponds to all the existing values in the training set, meaning that we do not consider the missing values (treated as zeros in this case). This is slightly different from [15].

5.4 Evaluation Strategy

The proposed scheme is tested in several aspects: the prediction accuracy in recommendation, the privacy protection level, when to recompute SVD, and randomization degree with its impact in perturbation, etc.

To test when to recompute SVD, data in training set, which is viewed as a rating matrix, is split into two sub sections with a particular ratio ρ . The first ρ data is assumed to be held by the third party; the remaining data will be updated into it. For instance, when a row split is performed with $\rho = 40\%$, the first 40% rows in training set is treated as R in (1). An imputation process should be done on this data without the knowledge from the remaining 60% data, yielding \hat{R} in (8). Then a rank- k SVD is computed for this matrix. We call the rank- k approximation of \hat{R} the starting matrix. These data structures are utilized as the input of row updating algorithm. Results are expected to be different with varying split ratio in the training data. If the result is too far from the predefined threshold or the results starts to degrade at some point, a re-computation should be performed.

However, we don't update the remaining 60% rows in training set in one round since data in real world application grows in small amount compared with the existing one. In our updating experiments, the 60% rows are repetitively added to the starting matrix in several times, with 1/9 of the rows in each time [15].

We evaluate the algorithms on both MovieLens and Jester datasets by investigating the time cost of updating, prediction error and privacy measure on the final matrix. The machine we use is equipped with Intel® Core™ i5-2405S processor, 8GB RAM and is installed with UNIX operating system.

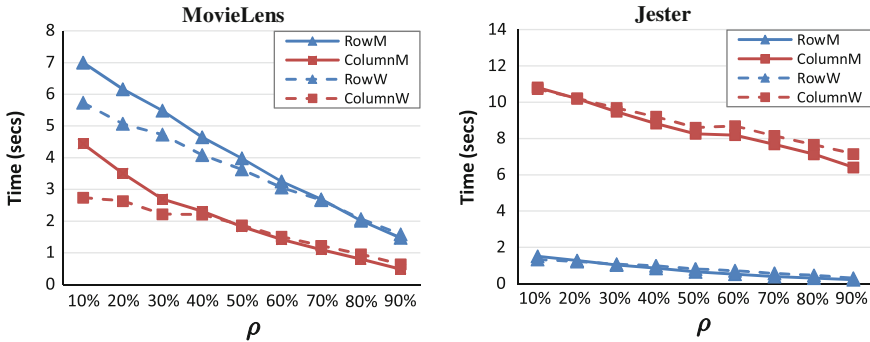


Fig. 1 Time cost variation with split ratio ρ

5.5 Results and Discussion

5.5.1 Split Ratio Study

Owing to the inherent property of SVD updating algorithms, errors are generated in each run. The data owner should be aware of the suitable time to recompute SVD for whole data so that the quality of the data can be kept. We study this problem by experimenting with the split ratio ρ . Note we use 13 and 11 as the truncation rank for MovieLens and Jester datasets. To be consistent, in WNMTF imputation, both l and t are set to 13 for MovieLens and 11 for Jester. As a reference, the maximum number of iterations in WNMTF is set to 10.

The time cost for updating new data with varying ρ is plotted in Fig. 1, where “RowM” and “ColumnM” represent row update and column update with mean value imputation while “RowW” and “ColumnW” represent the updates with WNMTF imputation. It is expected that updating fewer rows/columns takes less time.

Furthermore, the figure indicates the relation between the time cost of row and column updating—it depends on dimensionality of row and column. For example, the MovieLens data has more columns (1,682 items) than rows (943 users) while the Jester data has much fewer columns (100 items) than rows (24,983 users). We observed each step of both row and column updating algorithms and found that, when the number of columns is greater than the number of rows, steps 1 and 3 in row updating algorithm need more time than that in column updating algorithm due to higher dimensionality and vice versa. Compared with the time cost for imputing and computing SVD on the complete raw training set (i.e., the non-incremental case), our scheme runs much faster on both datasets (See Table 1). As for the imputation methods in this scenario, it is intuitive that WNMTF takes much shorter time than mean value imputation. Nevertheless, the difference is not that apparent in total time cost of row/column updating algorithms(i.e., the incremental case). This is because imputation time is a smaller part of the total time cost than the time for computing incremental SVD.

Table 1 Time cost of prediction on raw training data

Dataset	Imputation method	Imputation time	SVD time	Total time
MovieLens	Mean value	18.2617	0.3562	18.6179
	WNMTF	0.8764	0.3483	1.2247
Jester	Mean value	16.7078	0.1105	16.8183
	WNMTF	1.2392	0.1196	1.3588

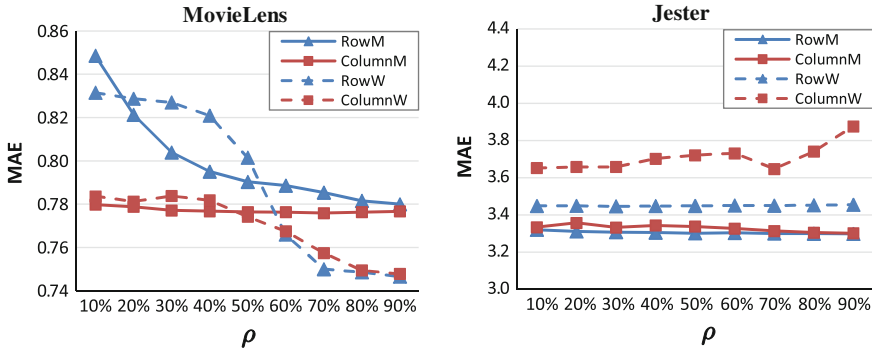


Fig. 2 MAE variation with split ratio ρ_1

Figure 2 shows the mean average error. For MovieLens dataset, the column update with mean value imputation almost stays at the same errors while the row update has a descending trend. With the WNMTF imputation, both updates produce smaller errors with rising split ratio. It is interesting that after some points, WNMTF provides lower prediction errors than mean value imputation. Thing is different for Jester data where MAE keeps stable in all updates and mean value imputation performs better. We get this difference because Jester has denser rating matrix than MovieLens meaning the former relies less on the new data than the latter.

The privacy level with varying split ratio is displayed in Fig. 3. As mentioned before, we only consider the existing ratings in the training set and eliminating the missing values when calculating the privacy level. This makes the results different from [15]. The privacy level without taking into account missing values does not change much with increasing split ratio. The curves look fluctuating because we reduced the interval of Y-axis. In this experiment, we notice that WNMTF imputation produces higher privacy level than mean value imputation. It can be attributed to more changes on ratings done by WNMTF. Since mean value imputation does not alter the existing values whereas WNMTF recompute the whole data, the latter perturbs the data to a deeper extent.

Now we can decide when to recompute SVD for the whole data according to Figs. 2 and 3. For mean value imputation, the MAEs on both datasets drop more slowly after $\rho \geq 50\%$ and there is no apparent variation of the slope for privacy measure curves, the re-computation can be performed when ρ reaches 50%. For

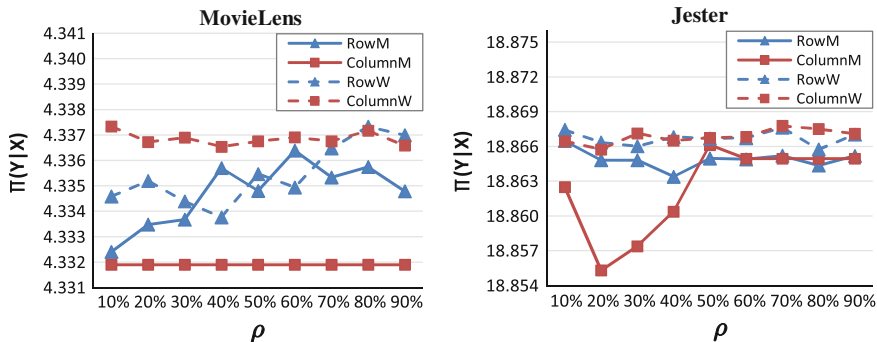


Fig. 3 Privacy level variation with split ratio ρ_1

WNMTF imputation, the MAEs keep decreasing all the way in MovieLens dataset, so the re-computation is not needed so far; the MAE for column update in Jester data increases when $\rho \geq 70\%$, meaning that the pre-computed matrices of SVD starts to degrade and thus a re-computation is helpful.

5.5.2 Role of Randomization in Data Updating

So far, we have not applied randomization technique to our data updating scheme. In this section, we study the role of randomization (Gaussian noise with μ and σ as its parameters in both row and column updating algorithms) in both data quality and privacy preservation. In the following experiments, ρ is fixed to 40% and we use the WNMTF to impute the missing values. We probe μ in $\{0, 1\}$ and σ in $\{0.1, 1\}$ for both datasets. Table 2 collects the statistics of the test.

In this table, the results of row and column updating with randomization is compared with the non-randomized version. As can be seen, after applying random noise to new data before updating it, the privacy level in all cases improves to a certain extent. Nevertheless, we lost some utility of the data which results in greater MAEs at the same time. Hence, the parameters should be carefully chosen to deal with the

Table 2 Randomization in data updating on MovieLens dataset

Parameters		Row updating		Column updating	
μ	σ	MAE	$\Pi(Y X)$	MAE	$\Pi(Y X)$
0	0	0.8209	4.3338	0.7818	4.3365
0	0.1	0.8210	4.3341	0.7818	4.3368
0	1	0.8292	4.3355	0.7980	4.3390
1	0.1	1.0584	4.3342	1.0447	4.3372
1	1	1.0158	4.3361	0.9738	4.3389

trade-off between data utility and data privacy. Moreover, the results indicate that the expectation μ affects the results more than the standard deviation σ . We suggest data owner determine μ first and then tweak σ .

As a summary, randomization technique can be used as an auxiliary step in SVD-based data updating scheme to provide better privacy protection. It brings in randomness that perturbs the data before SVD updating. Therefore, data will be perturbed twice (randomization + SVD) in addition to imputation during the updating process and thus can achieve a higher privacy level. However, with the latent factors captured by SVD, most critical information can be retained which ensures the data quality for recommendation.

6 Conclusion and Future Work

In this chapter, we present a privacy preserving data updating scheme for collaborative filtering purpose. It is an incremental SVD based scheme with randomization technique and could be utilized in updating user-item matrix and preserving the privacy at the same time. We try to protect users' privacy in three aspects, i.e., missing value imputation, randomization-based perturbation and SVD truncation. The experimental results on MovieLens and Jester datasets show that our proposed scheme could update new data into the existing (processed) data very fast. It can also provide high quality data for accurate recommendation while keep the privacy.

Future work will take into account users' latent factor to obtain a more reasonable imputation strategy in updating the data. Other matrix factorization techniques, e.g., non-negative matrix factorization, will be considered to explore an alternative way of updating the new data with privacy preservation so that a possible better scheme can be provided.

References

1. Agrawal D, Aggarwal, CC (2001) On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, PODS '01, ACM, pp 247–255
2. Brand Matthew (2006) Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra Appl* 415(1):20–30
3. Breese JS, Heckerman D, Kadie CM (1998) Empirical analysis of predictive algorithms for collaborative filtering. Technical report, Microsoft Research, Microsoft Corporation
4. Canny J (2002) Collaborative filtering with privacy. In: Proceedings of the 2002 IEEE symposium on security and privacy, IEEE Computer Society, pp 45–57
5. Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix trifactorizations for clustering. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD '06, ACM, pp 126–135
6. Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: a constant time collaborative filtering algorithm. *Inf Retr* 4(2):133–151

7. Koch O, Lubich C (2007) Dynamical low-rank approximation. *SIAM J Matrix Anal Appl* 29(2):434–454
8. Polat H, Du W (2005) Privacy-preserving collaborative filtering. *Int J Electron Commer* 9(4):9–35
9. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: *Proceedings of ACM 1994 conference on computer supported cooperative work*, ACM, pp 175–186
10. Sarwar BM, Karypis G, Konstan JA, Riedl JT (2000) Application of dimensionality reduction in recommender systems—a case study. In: *Proceedings of ACM WebKDD workshop*, ACM
11. Shardanand U, Maes P (1995) Social information filtering: algorithms for automating “word of mouth”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '95*, ACM Press/Addison-Wesley Publishing Co, pp 210–217
12. Stewart GW (1990) Perturbation theory for the singular value decomposition. Technical report, Computer Science Department, University of Maryland
13. Tougas J, Spiteri RJ (2007) Updating the partial singular value decomposition in latent semantic indexing. *Comput Stat Data Anal* 52:174–183
14. Wang J, Zhan J, Zhang J (2008) Towards real-time performance of data value hiding for frequent data updates. In: *Proceedings of the IEEE international conference on granular computing*, IEEE Computer Society, pp 606–611
15. Wang X, Zhang J (2012) SVD-based privacy preserving data updating in collaborative filtering. In: *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, IAENG*, pp 377–284
16. Zhang S, Wang W, Ford J, Makedon F (2006) Learning from incomplete ratings using non-negative matrix factorization. In: *Proceedings of the sixth SIAM international conference on data mining*, SIAM, pp 548–552

Machine Learning-Based Missing Value Imputation Method for Clinical Datasets

M. Mostafizur Rahman and D. N. Davis

Abstract Missing value imputation is one of the biggest tasks of data pre-processing when performing data mining. Most medical datasets are usually incomplete. Simply removing the incomplete cases from the original datasets can bring more problems than solutions. A suitable method for missing value imputation can help to produce good quality datasets for better analysing clinical trials. In this paper we explore the use of a machine learning technique as a missing value imputation method for incomplete cardiovascular data. Mean/mode imputation, fuzzy unordered rule induction algorithm imputation, decision tree imputation and other machine learning algorithms are used as missing value imputation and the final datasets are classified using decision tree, fuzzy unordered rule induction, KNN and K-Mean clustering. The experiment shows that final classifier performance is improved when the fuzzy unordered rule induction algorithm is used to predict missing attribute values for K-Mean clustering and in most cases, the machine learning techniques were found to perform better than the standard mean imputation technique.

Keywords Cardiovascular · FURIA · Fuzzy rules · J48 · K-Mean · Missing value

M. M. Rahman (✉) · D. N. Davis
Department of Computer Science, University of Hull, Hull, UK
e-mail: mmbappy@gmail.com

D. N. Davis
e-mail: D.N.Davis@hull.ac.uk

M. M. Rahman
Department of Computer Science, Eastern University Dhaka, Dhaka, Bangladesh
e-mail: M.M.Rahman@2009.hull.ac.uk

1 Introduction

Many researchers have identified several important and challenging issues [1–3] for clinical decision support. In “Grand challenges for decision support” Sittig et al. [1] set out ten critical problems for “designing, developing, presenting, implementing, evaluating, and maintaining all types of clinical decision support capabilities for clinicians, patients and consumers”. However Sittig et al.’s identification does cover little about data preprocessing. Sometimes, improved data quality is itself the goal of the analysis, usually to improve processes in a production database [4] and designing of decision support.

Two types of databases are available in medical domain [5]. The first is the dataset acquired by medical experts, which are collected for a special research topic where data collection is triggered by the generated hypothesis of a clinical trial. The other type is a huge dataset retrieved from hospital information systems. These data are stored in a database automatically without any specific research purpose. These data records are often used for further analysis and building clinical decision support system. These types of datasets are very complex where the numbers of records are very huge, with a large number of attributes for each record; many missing values and typically the datasets are mostly imbalanced with regard to their class label. In this paper we will be addressing the issue of missing value in clinical (cardiovascular) datasets.

Many real-life data sets are incomplete. The problem with missing attribute values is a very important issue in Data Mining. In medical data mining the problem with the missing values has become a challenging issue. In many clinical trials, the medical report pro-forma allow some attributes to be left blank, because they are inappropriate for some class of illness or the person providing the information feels that it is not appropriate to record the values for some attributes [6].

Typically there are two types of missing data [7]; one is called missing completely at random or MCAR. Data is MCAR when the response indicator variables R are independent of the data variables X and the latent variables Z . The MCAR condition can be succinctly expressed by the relation $P(R|X, Z, \mu) = P(R|\mu)$. The second category of missing data is called missing at random or MAR. The MAR condition is frequently written as $P(R = r|X = x, Z = z, \mu) = P(R = r|X^\circ = x^\circ, \mu)$ for all x^μ , z and μ [8, 9].

In general, methods to handle missing values belong either to sequential methods like leastwise deletion, assigning most common values, arithmetic mean for the numeric attribute etc. or parallel methods where rule induction algorithm are used to predict missing attribute values [10]. There are reasons for which sequential leastwise deletion is considered to be a good method [7], but several works [6, 7, 11] have shown that the application of this method on the original data can corrupt the interpretation of the data and mislead the subsequent analysis through the introduction of bias.

While several techniques for missing value imputation are employed by researchers, most of the techniques are single imputation approaches [12]. The most traditional missing value imputation techniques are deleting case records, mean value imputation, maximum likelihood and other statistical methods [12]. In recent

years, research has explored the use of machine learning techniques as a method for missing values imputation in several clinical and other incomplete datasets [13]. Machine learning algorithm such as multilayer perceptron (MLP), self-organising maps (SOM), decision tree (DT) and k-nearest neighbours (KNN) have been used as missing value imputation methods in different domains [11, 14–21]. Machine learning methods like MLP, SOM, KNN and decisions tree have been found to perform better than the traditional statistical methods [11, 22].

In this paper we examine the use of Machine Learning techniques as a missing values imputation method for real life incomplete cardiovascular datasets. Where, we have used classifier to predict the value for a missing field and impute the predicted value to make the dataset complete. In order to compare the performance we have used four classifiers, Decision Tree [10], KNN [32], SVM [35] and FURIA [23] to predict the missing values. The datasets are later classified using Decision Tree, KNN, FURIA and K-Means Clustering; the results are compared with commonly used mean-mode imputation methods.

2 Overview of FURIA

Fuzzy Unordered Rule Induction Algorithm (FURIA) is a fuzzy rule-based classification method, which is a modification and extension of the state-of-the-art rule learner RIPPER. Fuzzy rules are obtained through replacing intervals by fuzzy intervals with trapezoidal membership functions [23]:

$$I^F(v) \stackrel{\text{df}}{=} \begin{cases} 1 & \phi^{c,L} \leq v \leq \phi^{c,U} \\ \frac{v-\phi^{s,L}}{\phi^{c,L}-\phi^{s,L}} & \phi^{s,L} \leq v \leq \phi^{c,L} \\ \frac{\phi^{s,U}-v}{\phi^{s,U}-\phi^{c,U}} & \phi^{c,U} \leq v \leq \phi^{s,U} \\ 0 & \text{else} \end{cases} \quad (1)$$

where $\phi^{c,L}$ and $\phi^{c,U}$ are the lower and upper bound of the membership of the fuzzy sets. For an instance $x = (x_1, \dots, x_n)$ the degree of the fuzzy membership can be found using the formula [23]:

$$\mu_{r^F}(x) = \prod_{i=1..k} i_i^F(x_i) \quad (2)$$

For fuzzification of a single antecedent only relevant training data is D_T^i considered and data are partitioned into two subsets and rule purity is used to measure the quality of the fuzzification [23]:

$$D_T^i = \{x = (x_1, \dots, x_k) \in D_T \mid I_j^F(x_j) > 0 \text{ for all } j \neq i\} \subseteq D_T \quad (3)$$

$$P_{ur} = \frac{P_i}{p_i + n_i} \tag{4}$$

where

$$p_i \stackrel{\text{def}}{=} \sum_{x \in D_{T+}^i} \mu_{A_i}(A)$$

$$n_i \stackrel{\text{def}}{=} \sum_{x \in D_{T-}^i} \mu_{A_i}(A)$$

The fuzzy rules $r_1^{(j)} \dots r_k^{(j)}$ have been learned for the class λ_j , the support of this class is defined by [23]:

$$s_j(x) \stackrel{\text{df}}{=} \sum_{i=1 \dots k} \mu_{r_i^{(j)}}(x) \cdot CF(r_i^{(j)}) \tag{5}$$

where, the certainty factor of the rule is defined as

$$CF(r_i^{(j)}) = \frac{2 \frac{|D_T^{(j)}|}{|D_T|} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)}{2 + \sum_{x \in D_T} \mu_{r_i^{(j)}}(x)} \tag{6}$$

The use of the algorithm in of data mining can be found in [23–25].

3 Decision Tree

The decision tree classifier is one of the most widely used supervised learning methods. A decision tree is expressed as a recursive partition of the instance space. It consists of a directed tree with a “root” node with no incoming edges and all the other nodes have exactly one incoming edge. [10]. Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions [26].

Ross Quinlan introduced a decision tree algorithm (known as Iterative Dichotomiser (ID 3)) in 1979. C4.5, as a successor of ID3, is the most widely-used decision tree algorithm [27]. The major advantage to the use of decision trees is the class-focused visualization of data. This visualization is useful in that it allows users to readily understand the overall structure of data in terms of which attribute mostly affects the class (the root node is always the most significant attribute to the class). Typically the goal is to find the optimal decision tree by minimizing the generalization error [28]. The algorithms introduced by Quinlan [29, 30] has proved to

be an effective and popular method for finding a decision tree to express information contained implicitly in a data set. WEKA [31] makes use of an implementation of C4.5 algorithm called J48 which has been used for all of our experiments.

4 K-Nearest Neighbour Algorithm

K-Nearest Neighbor Algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space (defined using for example the Similarity measure). K-NN is a type of instance-based learning [32] or lazy learning where the function is only approximated locally and all computation is deferred until classification.

$$\text{Similarity}(\mathbf{x}, \mathbf{y}) = -\sqrt{\sum_{i=1}^n \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i)} \quad (7)$$

The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms where an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbour.

5 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms proposed by Macqueen in 1967, which has been used by many researchers to solve some well-known clustering problems [10]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The algorithm first randomly initializes the clusters center. The next step is to calculate the distance (discussed in the above section) between an object and the centroid of each cluster then take each point belonging to a given data set and associate it to the nearest centre and re-calculate the cluster centres. The process is repeated with the aim of minimizing an objective function known as squared error function given by:

$$\mathbf{J}(\mathbf{v}) = \sum_{i=1}^C \sum_{j=1}^{C_i} (\|\mathbf{x}_i - \mathbf{v}_j\|)^2 \quad (8)$$

where, $\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j , c_i is the number of data points in i^{th} cluster and c is the number of cluster centers.

6 Cardiovascular Data

We have used two datasets from Hull and Dundee clinical sites. The Hull site data includes 98 attributes and 498 cases of cardiovascular patients and the Dundee site data includes 57 attributes, and 341 cases from cardiovascular patients. After combining the data from the two sites, 26 matched attributes are left.

Missing values: After combining the data and removing redundant attributes we found that out of 26 attributes 18 attributes have a missing value frequency from 1 to 30% and out of 832 records 613 records have 4 to 56% missing values in their attributes.

From these two data sets, we prepared a combined dataset having 26 attributes with 823 records. Out of 823 records 605 records have missing values and 218 records do not have any missing values. Among all the records 120 patients are alive and 703 patients are dead. For this experiment according to clinical risk prediction model (CM1) [33], patients with status “Alive” are consider to be “Low Risk” and patients with status “Dead” are consider to be “High Risk”.

7 Mean and Mode Imputation

This is one of the most frequently used methods. It consists of replacing the unknown value for a given attribute by the mean (\bar{x}) (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute [21].

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \quad (9)$$

It replaces all missing records with a single and unique value \bar{x} , which is the mean value of that attribute.

8 Proposed Missing Value Imputation Process

The original data set is first portioned in to groups. The records having missing values in their attributes are in one group (the *complete data set*) and the records without any missing values are placed in a separate group. The classifier is trained with the complete data sets, and later the incomplete data is given to the model for predicting the missing attribute values. The process is repeated for the entire set of attributes that have missing values. At the end of training, this training dataset and missing value imputed datasets are combined to make the finalised data. The final dataset is then fed to the selected classifier for classification (as shown in Fig. 1).

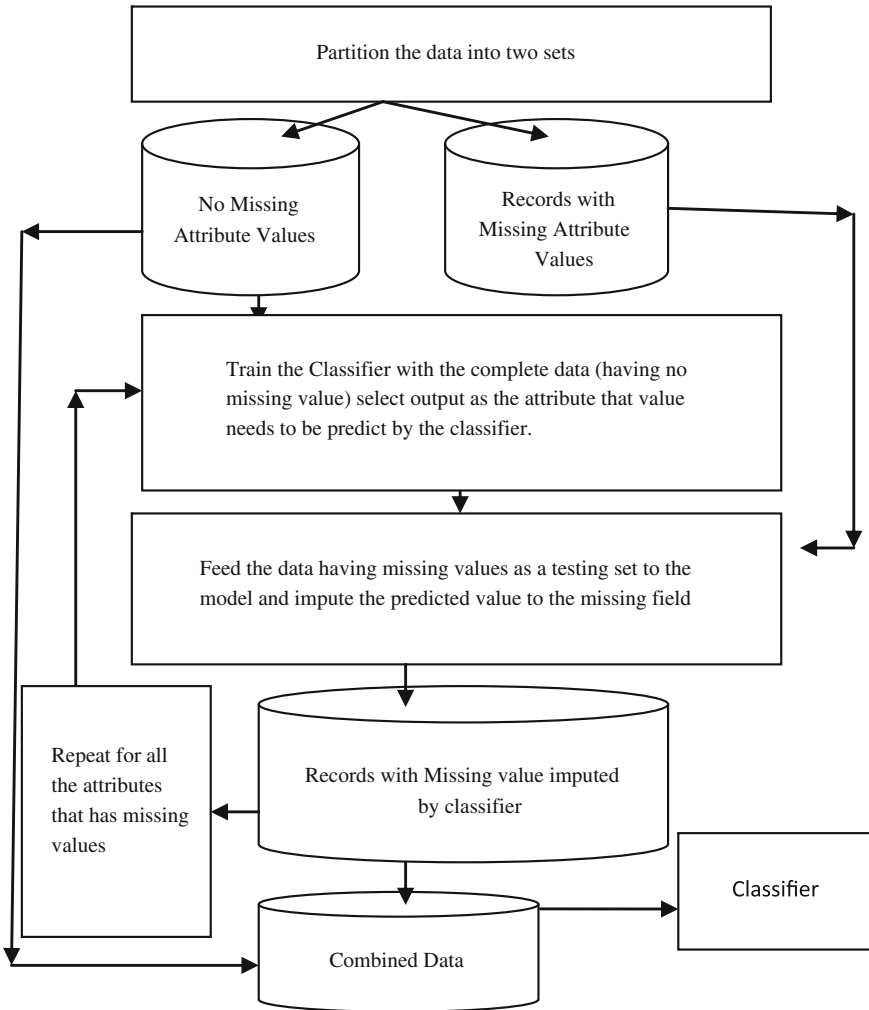


Fig. 1 Missing value imputation process

9 Results

We have experimented with a number of machine learning algorithms as missing value imputation mechanisms; such as FURIA, decision tree [34], and SVM [35]. The performance is compared with the most commonly used missing imputation statistical method mean-mode. The results are also compared with the previously published results of the same experimental dataset with mean-mode imputation for K-Mix clustering [36].

Table 1 Different missing imputation methods with k-mean clustering

Missing imputation methods	Confusion matrix							
	Risk	Classified high risk	Classified low risk	ACC	SEN	SPEC	PPV	NPV
Decision tree (J48)	High	36	84	0.64	0.30	0.70	0.15	0.85
	Low	212	491					
FURIA	High	52	68	0.58	0.43	0.60	0.16	0.86
	Low	281	422					
SVM	High	36	84	0.62	0.30	0.67	0.14	0.85
	Low	229	474					
Mean and mode	High	35	85	0.63	0.29	0.69	0.14	0.85
	Low	219	484					

From the Table 1 one can see that for K-mean clustering, decision tree imputation method shows accuracy of 64 % (slightly better than the other methods) but the sensitivity is 30 % which is almost as poor as the mean/mode imputation. SVM and mean/mode mutation show very similar performance with accuracy of 62–63 % and sensitivity of 29–32 %. On the other hand, fuzzy unordered rule induction algorithm as a missing value imputation method shows sensitivity of 43 % with accuracy of 58 %. Table 2 shows the comparison results of previously published results of K-Mix [37] clustering algorithm with mean mode imputation and simple K-mean clustering with FURIA missing value imputation. The result shows that the K-mean with FURIA as missing value imputation has higher sensitivity (43 %) than the K-mix with conventional mean/mode imputation method (0.25 %).

The datasets prepared by different imputation methods are also classified using well known classifier decision tree (J48), KNN and also with FURIA. The classification outcomes are presented in Tables 3, 4, 5. Table 6 presents the highest sensitivity value found of all the datasets prepared by different imputation methods and miss-

Table 2 Comparison results with k-mix clustering

Classifier with different missing imputation methods	Confusion matrix				
	Risk	Classified high risk	Classified low risk	SEN	SPEC
K-Mix (with mean mode imputation)	High	35	21	0.25	0.89
	Low	107	177		
K-Mean with Fuzzy unordered rule induction algorithm used as missing value imputation method	High	52	68	0.43	0.60
	Low	281	422		

Table 3 Different missing imputation methods with J48 classification

Missing imputation methods	Confusion matrix		ACC	SEN	SPEC	PPV	NPV	
	Actual risk ↓	Classified risk						
		High						Low
Decision tree (J48)	High	27	93	0.80	0.23	0.90	0.27	0.87
	Low	72	631					
K-NN	High	20	100	0.80	0.17	0.90	0.23	0.86
	Low	68	635					
FURIA	High	24	96	0.80	0.20	0.90	0.25	0.87
	Low	72	631					
SVM	High	18	102	0.78	0.15	0.89	0.19	0.86
	Low	79	624					
Mean	High	13	107	0.80	0.11	0.92	0.19	0.86
	Low	56	647					

Table 4 Different missing imputation methods with K-NN classification

Missing imputation methods	Confusion matrix		ACC	SEN	SPEC	PPV	NPV	
	Actual risk	Classified risk						
		High						Low
Decision tree (J48)	High	24	96	0.71	0.20	0.80	0.15	0.85
	Low	140	563					
K-NN	High	29	91	0.81	0.24	0.91	0.32	0.88
	Low	63	640					
FURIA	High	25	95	0.79	0.21	0.89	0.24	0.87
	Low	79	624					
SVM	High	24	96	0.71	0.20	0.80	0.15	0.85
	Low	140	563					
Mean	High	25	95	0.77	0.21	0.87	0.21	0.87
	Low	92	611					

ing value imputation using FURIA shows the sensitivity 43.3% which is the highest among all the machine learning methods and statistical method explored in this paper.

For clinical data analysis it is important to evaluate the classifier based on how well the classifier is performing to predict the “High Risk” patients. As indicated earlier the dataset shows an imbalance on patient’s status. Only 120 records, out of 832 records, are of “High Risk” (14.3% of the total records). A classifier may give very high accuracy if it can correctly classify the “Low Risk” patients but is of limited use if it does not correctly classify the “High Risk” patients. For our analysis we gave more importance to Sensitivity and Specificity then Accuracy to compare the classification outcome.

If we analyse the ROC [38] space for all the imputation methods classified with three classifiers mentioned earlier and one clustering algorithm plotted in Fig. 2, we will find that most the machine learning methods are above the random line and most of the cases better than the statistical mean/mode imputation.

Table 5 Different missing imputation methods with Fuzzy Rule Induction Algorithm classification

Missing imputation methods	Confusion matrix		ACC	SEN	SPEC	PPV	NPV	
	Actual risk	Classified risk						
		High						Low
Decision tree (J48)	High	48	72	0.63	0.40	0.67	0.17	0.87
	Low	230	473					
K-NN	High	36	84	0.67	0.30	0.73	0.16	0.86
	Low	190	513					
FURIA	High	36	84	0.67	0.30	0.73	0.16	0.86
	Low	190	513					
SVM	High	22	98	0.74	0.18	0.83	0.16	0.86
	Low	117	586					
Mean	High	27	93	0.72	0.23	0.80	0.16	0.86
	Low	140	563					

Table 6 Highest sensitivity value found with each of the imputation method

Missing imputation methods	Highest sensitivity (%)	With the accuracy (%)	The classifier used to classify
FURIA	43.3	58	K-Mean
K-NN	42.5	51	K-Mean
J48	40	63	FURIA
SVM	30	62	K-Mean
Mean	29	63	K-Mean

If we evaluate the missing imputation based on the sensitivity than we can see the FURIA missing value imputation outperformed all the other machine learning and traditional mean/mode approaches to missing value imputation methods that we have examined in this work.

10 The Complexity of the Proposed Method

The complexity of the proposed method is related with the complexity of the classifier is used for the missing value imputation. If we use FURIA, than the fuzzy unordered rule induction algorithm can be analysed by considering the complexity of the rule fuzzification procedure, rule stretching and re-evaluating the rules. For $|D_T|$ training data and n numbers of attribute the complexity of the fuzzification procedure is $O(|D_T| n^2)$ [23], with $|RS|$ numbers of rules and $|D_T|$ training data the complexity of rule stretching is $O(|D_T| n^2)$ [23], and rule r with antecedent set $A(r)$ the complexity for the rule re-evaluating is $O(|A(r)|)$. For the experimental data of 823 records with 23 attributes on an average it took 0.69s to build the model for each attribute of missing values.

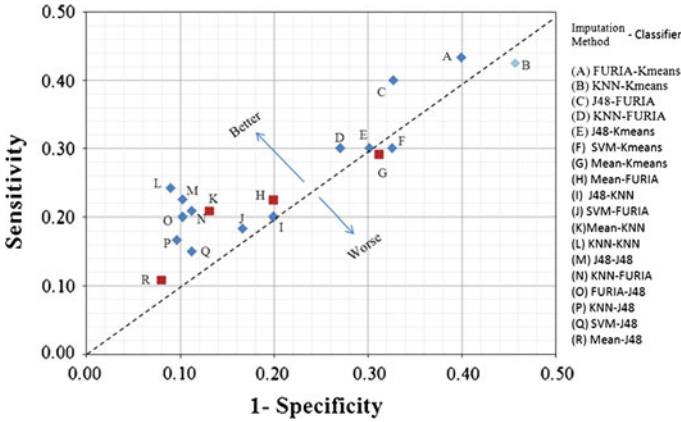


Fig. 2 The ROC space and plots of the different imputation methods classified with J48, FURIA, KNN and K-Means.

11 Conclusion

Missing attribute values are common in real life datasets, which causes many problems in pattern recognition and classification. Researchers are working towards a suitable missing value imputation solution which can show adequate improvement in the classification performance. Medical data are usually found to be incomplete as in many cases on medical reports some attributes can be left blank, because they are inappropriate for some class of illness or the person providing the information feels that it is not appropriate to record the values. In this work we examined the performance of machine learning techniques as missing value imputation. The results are compared with traditional mean/mode imputation. Experimental results show that all the machine learning methods which we explored outperformed the statistical method (Mean/Mode), based on sensitivity and some cases accuracy.

The process of missing imputation with our proposed method can be computationally expensive for large numbers of attribute having missing values in their attributes. However, we know that data cleaning is part of data pre-processing task of data mining which is not a real time task and neither a continuous process. Missing value imputation is a onetime task. With this extra effort we can obtain a good quality data for better classification and decision support.

We can conclude that machine learning techniques may be the best approach to imputing missing values for better classification outcome.

References

1. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS et al (2008) Grand challenges in clinical decision support. *J Biomed Inform* 41:387–392
2. Fox J, Glasspool D, Patkar V, Austin M, Black L, South M et al (2010) Delivering clinical decision support services: there is nothing as practical as a good theory. *J Biomed Inform* 43:831–843
3. Bellazzi R, Zupan B (2008) Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 77:81–97
4. Dasu T, Johnson T (2003) *Exploratory data mining and data cleaning*. Wiley-Interscience, New York
5. Tsumoto S (2000) Problems with mining medical data. In: *Computer software and applications conference, COMPSAC*, pp 467–468
6. Almeida RJ, Kaymak U, Sousa JMC (2010) A new approach to dealing with missing values in data-driven fuzzy modelling. *IEEE International Conference on Fuzzy Systems (FUZZ)*, Barcelona
7. Roderick JAL, Donald BR (2002) *Statistical analysis with missing data*, 2nd edn. Wiley, New York
8. Marlin BM (2008) *Missing data problems in machine learning*. Doctor of Philosophy, Graduate Department of Computer Science, University of Toronto, Toronto, Canada
9. Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. *J Sch Psychol* 48:5–37
10. Maimon O, Rokach L (2010) *Data mining and knowledge discovery handbook*. Springer, Berlin
11. Jerez JM, Molina I, Garcí'a-Laencina JP, Alba E, Nuria R, Miguel Mn et al (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med* 50:105–115
12. Peugh JL, Enders CK (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. *Rev Educ Res* 74:525–556
13. Rahman MM, Davis DN (2012) Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-Mean clustering on real cardiovascular data. *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering* (2012) London, UK, pp 391–394
14. Esther-Lydia S-RR, Pino-Mejias M, Lopez-Coello M-D, Cubiles-de-la-Vega (2011) Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks* 24:1
15. Weiss SM, Indurkha N (2000) Decision-rule solutions for data mining with missing values. In: *IBERAMIA-SBIA*, pp 1–10
16. Pawan L, Ming Z, Satish S (2008) *Evolutionary regression and neural imputations of missing values*. Springer, London
17. Setiawan NA, Venkatachalam P, Hani AFM (2008) Missing attribute value prediction based on artificial neural network and rough set theory. In: *Proceedings of the international conference on biomedical engineering and informatics, BMEI 2008*, p 306–310
18. Yun-fei Q, Xin-yan Z, Xue L, Liang-shan S (2010) Research on the missing attribute value data-oriented for decision tree. *2nd International conference on signal processing systems (ICSPS) 2010*
19. Meesad P, Hengprapromh K (2008) Combination of KNN-based feature selection and KNN based missing-value imputation of microarray data. In: *Proceedings of the 3rd international conference on innovative computing information and control, ICICIC '08*
20. Wang L, Fu D-M (2009) Estimation of missing values using a weighted K-nearest neighbors algorithm. In: *Proceedings of the international conference on environmental science and information application technology*, pp 660–663
21. Garcí'a-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M (2009) K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neuro Comput* 72:1483–1493

22. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. *Atmos Environ* 38:1352–2310
23. Hühn J, Hüllermeier E (2009) Fuzzy unordered rules induction algorithm. *Data Min Knowl Disc* 19:293–319
24. Lotte F, Lecuyer A, Arnaldi B (2007) FuRIA: A novel feature extraction algorithm for brain-computer interfaces using inverse models and Fuzzy regions of interest. In: *Proceedings of the 3rd international IEEE/EMBS conference on neural engineering, CNE '07*
25. Lotte F, Lecuyer A, Arnaldi B (2009) FURIA: An inverse solution based feature extraction algorithm using Fuzzy set theory for brain-computer interfaces. *IEEE Trans Signal Process* 57:3253–3263
26. Barros RC, Basgalupp MP, de Carvalho ACPLF, Freitas AA (2012) A survey of evolutionary algorithms for decision-tree induction. *IEEE Trans Syst Man Cybern Part C Appl Rev* 42:291–312
27. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF et al (Aug 2012) Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst* 36:2431–48
28. Maimon O, Rokach L (2010) *Data mining and knowledge discovery handbook*. Springer, Berlin
29. Quinlan JR (1985) *Induction of decision trees*. School of Computing Sciences, Broadway, N.S.W., Australia: New South Wales Institute of Technology
30. Quinlan JR (1993) *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann
31. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P et al (2010) WEKA-Experiences with a Java open-source project. *J Mach Learn Res* 11:2533–2541
32. Aha DW, Kibler D, Albert MK (Jan 1991) Instance-based learning algorithms. *Mach Learn* 6:37–66
33. Davis DN, Nguyen TTT (2008) Generating and verifying risk prediction models using data mining (A case study from cardiovascular medicine). Presented at the European society for cardiovascular surgery, 57th Annual congress of ESCVS, Barcelona Spain, 2008
34. Marsala C (2009) A fuzzy decision tree based approach to characterize medical data. In: *Proceedings of the IEEE International Conference on Fuzzy Systems, 2009*
35. Devendran V, Hemalatha T, Amitabh W (2008) Texture based scene categorization using artificial neural networks and support vector machines: a comparative study. *ICGST-GVIP*, vol 8, 2008
36. Nguyen TTT (2009) Predicting cardiovascular risks using pattern recognition and data mining. Ph.D., Department of Computer Science, The University of Hull, Hull, UK
37. Nguyen TTT, Davis DN (2007) A clustering algorithm for predicting cardiovascular risk. Presented at the international conference of data mining and knowledge engineering, London, 2007
38. Landgrebe TCW, Duin RPW (2008) Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Trans Pattern Anal Mach Intell* 30:810–822

Opto-Electronic Hybrid Integrated Platform for High-Speed Telecom/Datacom Applications: Microwave Design and Optimization

Wei Han, Marc Rensing, Xin Wang, Peter O'Brien and Frank H. Peters

Abstract An opto-electronic hybrid integrated platform was developed to enable the fabrication of broadband, low-cost, and compact transceivers for telecommunications. On this platform, an opto-electronic device such as a high-speed laser or a photodetector chip is integrated with a RF driver or an amplifier IC. A Kovar heatsink with multistep structure is designed for ease of optical coupling using a laser welding process. In order to control the high frequency resonances and improve the signal integrity, AIN based subcircuits are designed to feed the RF and DC signals separately. The interconnection networks between the IC and the opto-electronic device and also between the chips and high-speed transmission lines are carefully investigated to optimize the microwave performances. The influence of the packaging for this opto-electronic integration platform on the microwave performance is also analyzed in detail. The simulation results obtained and successful fabrication of a transmitter module demonstrate that the proposed platform can meet the requirements for high-speed WDM or TDM systems.

Keywords Driver IC · Microwave transmission · Opto-electronic hybrid integration · RF resonance · Transceiver module · Transmission line

W. Han (✉) · M. Rensing · X. Wang · P. O'Brien · F.H. Peters
Integrated Photonics Group, Tyndall National Institute, Cork, Ireland
e-mail: wei.han@tyndall.ie

M. Rensing
e-mail: marc.rensing@tyndall.ie

X. Wang
e-mail: xin.wang@tyndall.ie

P. O'Brien
e-mail: peter.obrien@tyndall.ie

F.H. Peters
Department of Physics, University College Cork Co., Cork, Ireland
e-mail: frank.peters@tyndall.ie

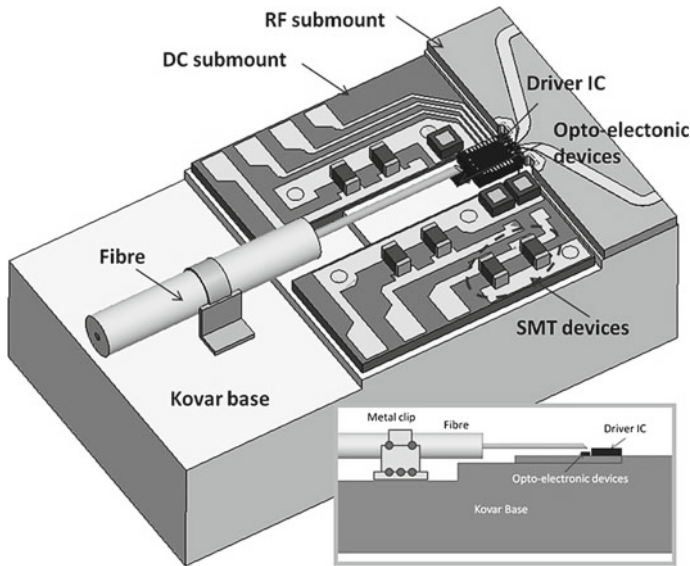


Fig. 1 The proposed opto-electronic hybrid integrated platform, insertion show the multistep structure of the Kovar base

1 Introduction

As telecommunications demand higher speeds and larger capacities, transmission networks such as time division multiplexing (TDM) and wavelength division multiplexing (WDM) have been studied extensively. Such networks require broadband, compact and low-cost opto-electronic modules such as high-speed modulators and photodetectors. Recent progress in long haul communications and short distance access network systems has already increased the transmission rates to several ten gigabits per second per channel [1–3]. To meet these requirements, it is important to develop hybrid integrated opto-electronic modules in which semiconductor devices (e.g. amplifiers and driver ICs), opto-electronic devices (e.g. lasers, modulators and PDs) and passive components (e.g. lens, fibers and waveguides) are assembled with RF or DC substrates. Such hybrid integrated platforms are often packaged and encapsulated in metal-ceramic or epoxy resin packages to fulfil different functions such as O/E and E/O conversions, all-optical 3R regenerations, optical routers and all-optical wavelength assignments [2–5]. The high-level integration will result in a relatively strong signal crosstalk or RF attenuation when the bit rates exceed 10 Gb/s. Especially at frequencies above 40 GHz, high order modes can be easily excited in microwave waveguides and thus generate RF resonances and rapid signal drop-off within the expecting bandwidth.

In this chapter, we present an opto-electronic hybrid integrated platform to enable the fabrication of broadband, low-cost, and compact transceivers for telecommuni-

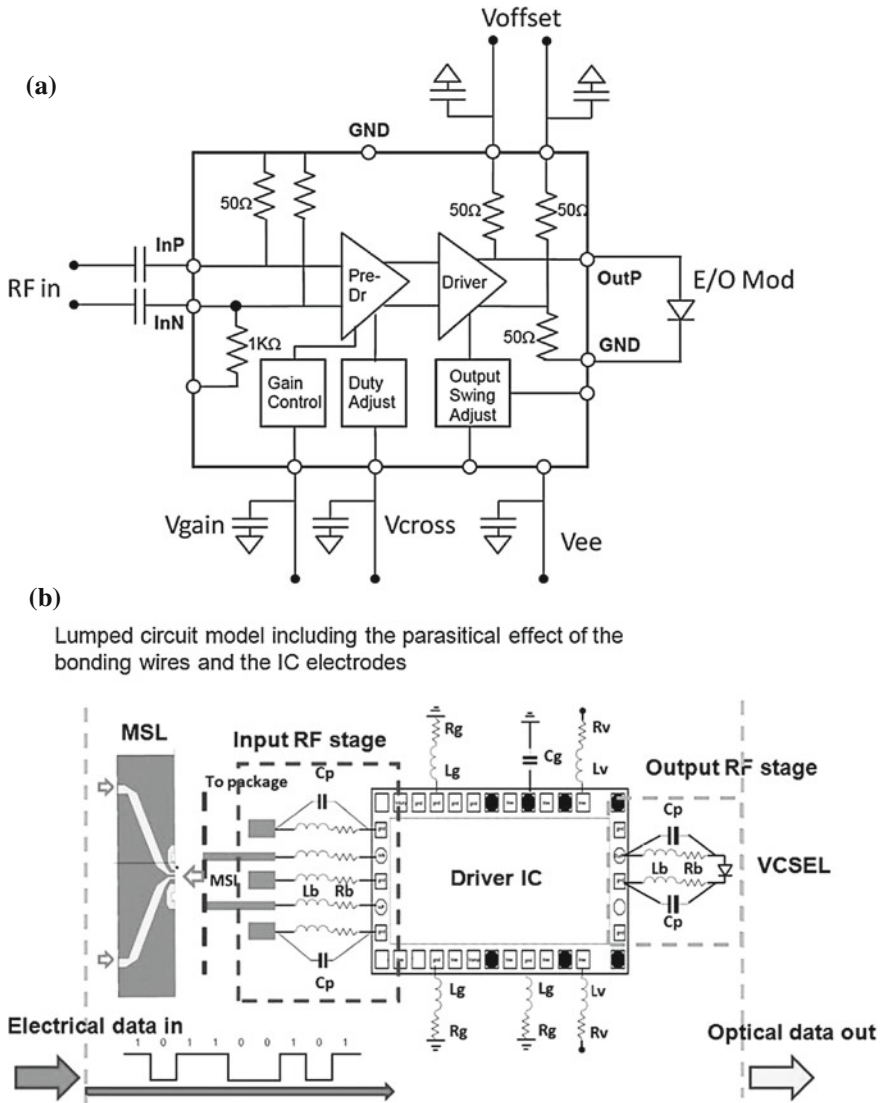


Fig. 2 Schematic diagram of the platform circuit (a), circuit model including the parasitical effect of the bonding wires and the IC electrodes (b)

cations applications. The proposed system-in package scheme is designed to support differential or dual way RF inputs with the transmission capability of 40GHz. By building the equivalent circuit model and distributed electromagnetic model of this platform, the residual parameters caused by the bonding wire, IC electrodes, and microwave transmission lines are investigated carefully to restrain the signal attenuations and improve the signal integrities. Likewise the signal crosstalk and high

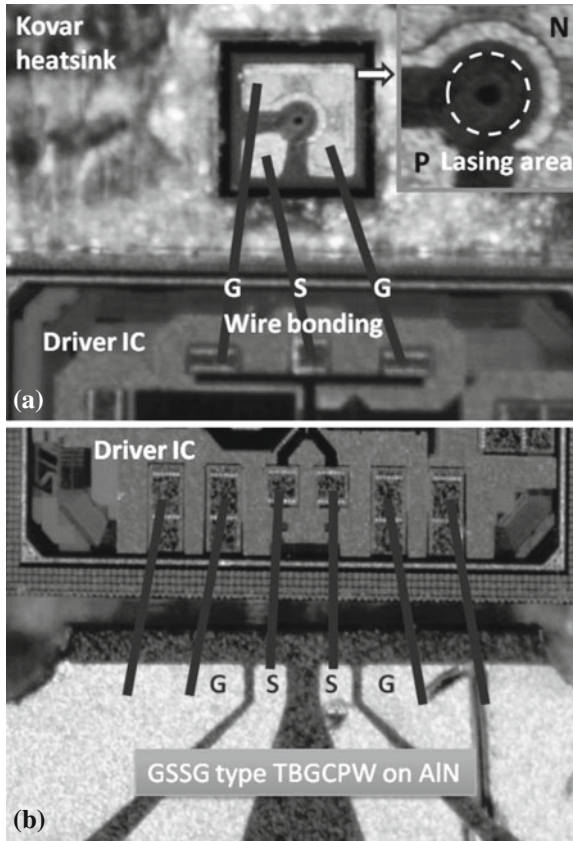


Fig. 3 Photograph of the electrical connection between the LD chip and driver IC (a), and between driver IC and RF subcircuits (b)

frequency resonances are extensively studied in RF and DC sub-circuit designs. By assembling this hybrid integrated platform into a butterfly package (BTF), a transceiver module can be constructed for various applications in WDM and TDM systems.

2 EM Analysis of the Opto-Electronic Integrated Platform

2.1 Schematic Diagram and Circuit Model of the Platform

The proposed opto-electronic hybrid integrated platform is shown in Fig. 1 including: an opto-electronic device such as a high-speed laser or a photodetector chip, a RF driver or an amplifier IC, a Kovar heatsink with multistep structure (as shown in the

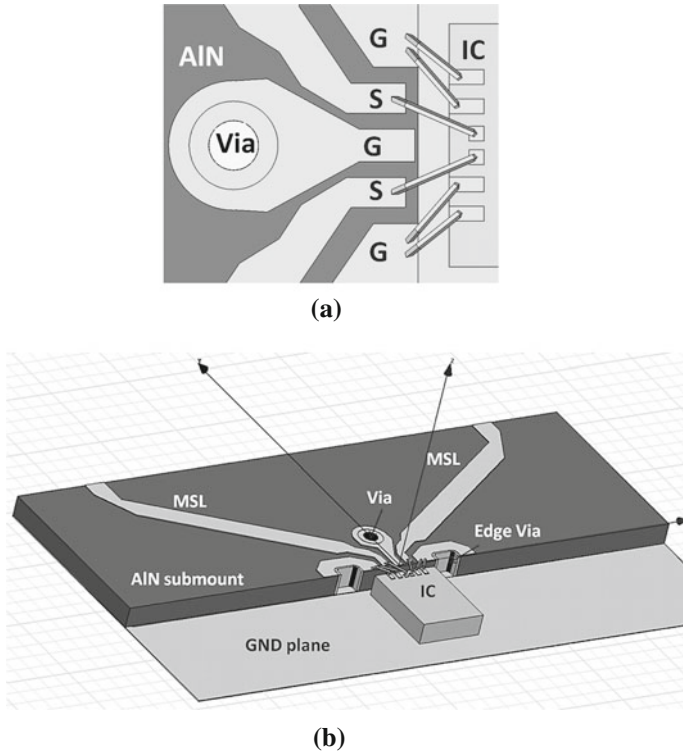


Fig. 4 Pads configuration of the AIN circuit and IC (a); model of the electrical connection between IC and RF circuit (b)

insert of Fig. 1) for easy of optical coupling and laser welding, optical fibre to deliver the optical signals in and out of the module, three sub-circuits to feed the DC and RF signals, and passive elements such as SMT capacitors and resistors. The active chips are fixed on the Kovar heatsink using solder or silver epoxy, and the electronic connections are realized by wire and ribbon bonding. The block diagram of the platform is shown in Fig. 2a, in which an E/O modulator and modulator driver IC are used in an optical transmitter. The modulator driver IC has an AC coupling differential input stage. The gain control, crossing point adjustment, and output offset control are available. Single-ended output is provided to drive an external E/O modulator. DC decoupling networks are also included in this model, and the undivided stable ground is provided by the Kovar heatsink.

Heat generated in the driver IC is a significant problem in high-speed opto-electronic integrated platforms because the power consumption of the IC will increase drastically with the operating bit rates. In this chapter, the driver IC is assembled with face-up electrical connection for better thermal control. However, the face-up connections of the IC requires wire bonding or ribbon bonding which will bring unwanted parasitic parameters at high frequencies. In Fig. 2b, the parasitical effects

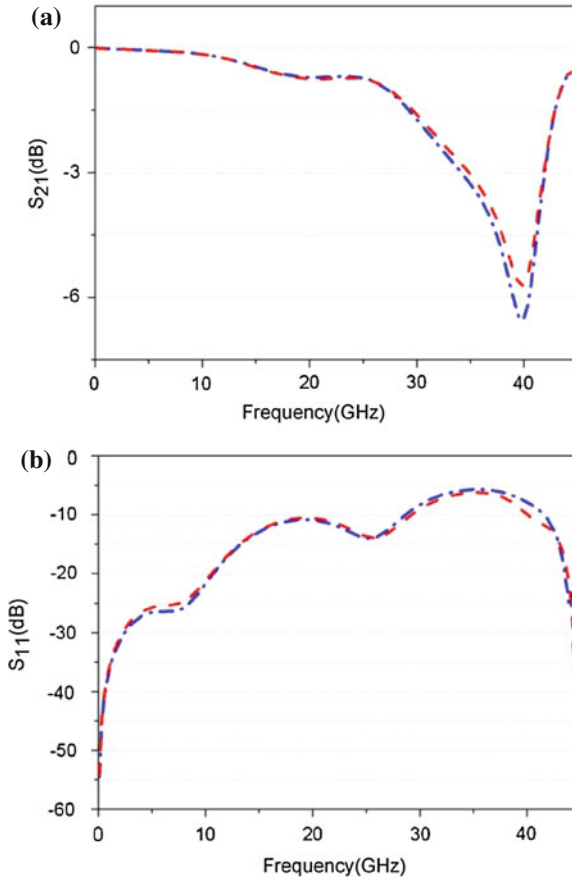


Fig. 5 Frequency responses of the interconnection and the assembly of the AlN circuit and driver IC, S_{21} (a) and S_{11} (b); dashed line positive input trace; dash-dot line negative input trace

of the bonding wires and the IC electrodes are added and compared with the circuit scheme in Fig. 2a. In the diagram, L_b and R_b are the inductance and resistance caused by the ribbon bonding in the RF path; R_g and L_g are due to the bonding wire in ground path; R_v and L_v are from the power supply path; C_p and C_g are the parasitic capacitance between the IC electrodes.

2.2 Distributed EM Model of the Platform

The electrical connection between the chips and circuit are described in Fig. 3. In the photograph, a high-speed vertical-cavity surface-emitting laser (VCSEL) is used as an E/O modulator, and the insertion of Fig. 1a shows the mesa structure and lasing

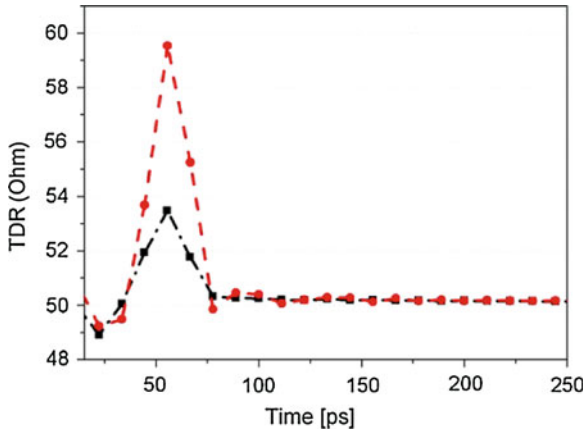


Fig. 6 simulation results on TDR Z_{in} ; *dash-dot line* input to the MSL; *dashed line* input to CPW

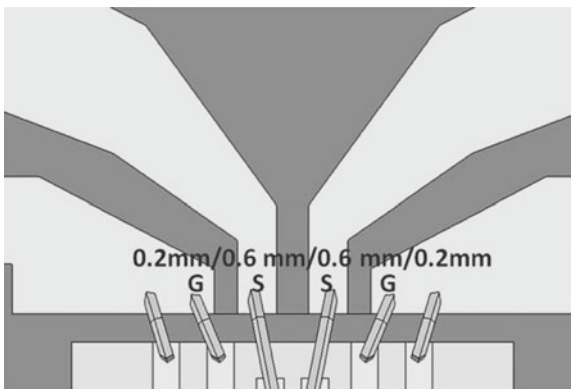
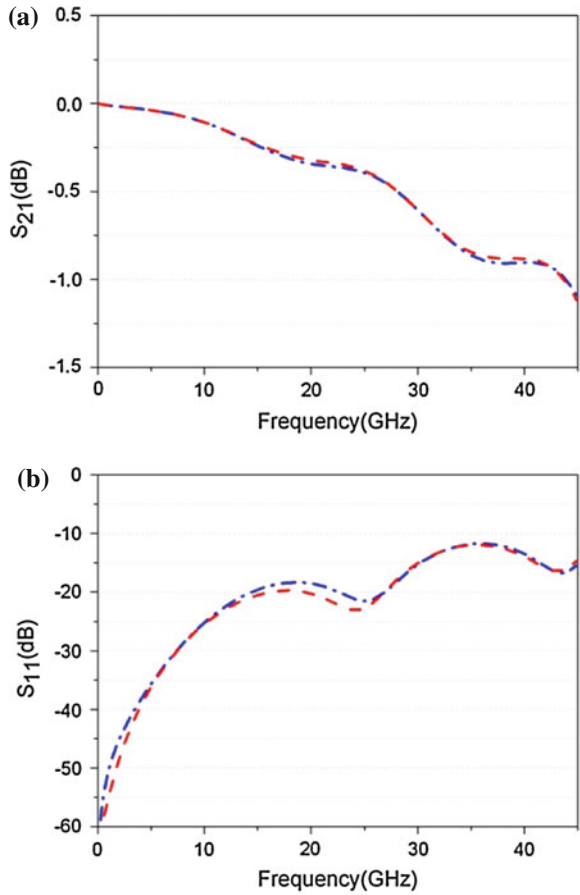


Fig. 7 GSSG electrical connection to improve high-frequency performances

area of the laser. A GSG type wire connection is employed between the LD and driver IC. Figure 3b shows the pad configuration of the driver IC, and an AlN based RF interface circuit with tapered and grounded coplanar waveguide (TBGCPW) is designed to meet the connection requirement of the IC. In order to optimize the RF subcircuit and the interconnection between the IC and the circuit, a distributed electromagnetic (EM) model is built using microwave simulator Ansoft HFSS. As shown in Fig. 4, the given RF circuit is based on AlN ceramic with a dielectric constant of 8.7 and loss tangent of 0.002. The RF transmission path is consisted of a microstrip line (MSL) for the connection with the RF connectors, and a coplanar waveguide (CPW) structure for the GSSG pad configuration of the IC.

On the AlN circuit, a through via and two edge vias are employed to form the ground planes. In Fig. 4a, a centre ground plane is added between the signal paths to make two GSG pad structures which correspond to the positive and negative data

Fig. 8 Frequency responses of the interconnection and the assembly of the AlN circuit and driver IC, S₂₁ (a) and S₁₁ (b); *dashed line* positive input trace; *dash-dot line* negative input trace



signal inputs of the driver IC. The ratio of the width and the gap of the CPW are set to 0.1/0.05 mm, and the spacing between the IC and circuit is 0.1 mm. The length of the bonding wire is 0.250 mm for the signal paths and 0.2 mm for the ground paths.

2.3 EM Simulation and Analysis on the Platform

Figure 5 gives the simulated frequency response of the interconnection and the assembly of the AlN circuit and driver IC. Strong resonance is observed at 40 GHz which brings a signal attenuation of approximately 6 dB at 40 GHz and thus limits the -3 dB bandwidth to 34 GHz. In addition, the RF reflection is above -10 dB at 28 GHz which will bring more RF return loss due to the impedance mismatch. Furthermore, to investigate the input impedance of the transmission line, time domain

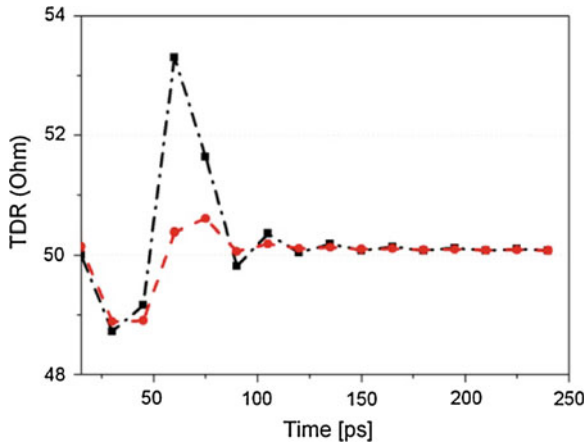


Fig. 9 Simulation results on TDR Z_{in} ; *dashed line* input to the MSL; *dash-dot line* input to CPW

impedance analyses are performed. As shown in Fig. 4, the RF feed path consists of a MSL and a CPW structure. Two wave ports are assigned at both ends of the transmission line.

In Fig. 6, the simulated input impedance shows that on the MSL side, the impedance can be controlled to around 53 Ohm. However, on CPW side, the impedance will increase to 59 Ohm which means a strong impedance mismatch generated in the GSGSG region.

In order to suppress the resonance and increase the bandwidth of the platform, an optimized GSSG type transmission line is designed. In Fig. 7, the centre ground plane is removed and the ratio of the width/gap is changed to 0.6/0.2 mm for ground and signal paths and 0.6/0.6 mm between signal paths. Moreover, the length of the bonding wire is decreased to 0.2 mm for the signal paths and to 0.15 mm for the ground paths.

In Fig. 8, the simulated frequency responses show a bandwidth increase as well as RF reflection suppression. The signal attenuation decreases to 0.68 dB at 40 GHz along with a S_{11} of -14 dB throughout the useful bandwidth. The simulated input impedance in Fig. 9 shows an obvious impedance improvement in the GSSG region where it decreases to 54 Ohm.

The comparison of the E-field distributions between the GSGSG structure and the GSSG structure is shown in Fig. 10. With respect to GSGSG structure, it can be seen from the plot that a strong resonance is generated between the centre ground plane and the signal path. The symmetric and narrow GSG structure and the impedance mismatch due to the conversion from MSL to CPW induce strong RF coupling and excites a standing wave in this region. In Fig. 10b, the optimized GSSG structure shows relatively low E-field intensity and the coupling between signal lines is designed to be minimized in this region.

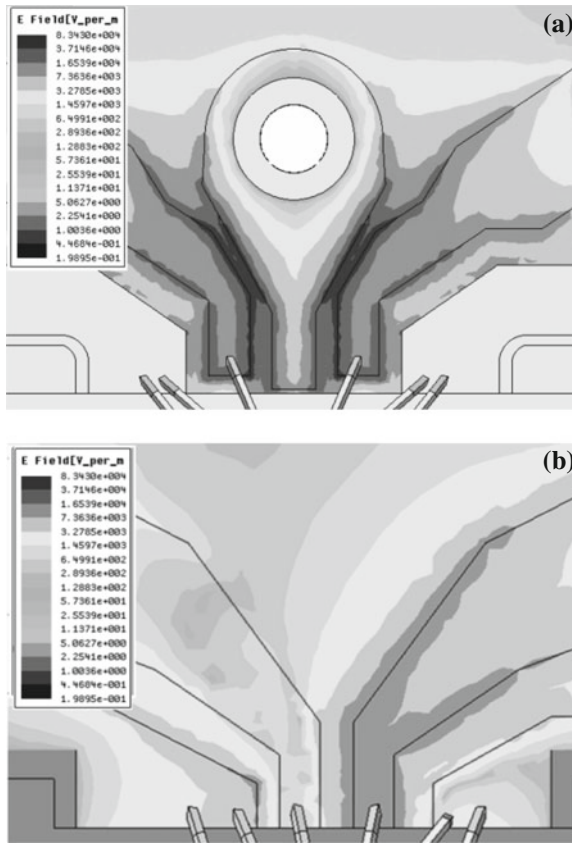


Fig. 10 E-field distributions at 40 GHz in GSGSG (a) and GSSG (b).

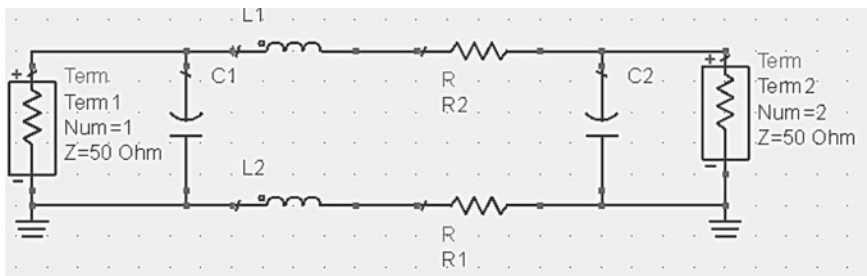


Fig. 11 Equivalent circuit model of the GSSG type bonding wires

In Fig. 7, the GSSG type bonding pads are designed to provide the electrical connection between the IC and RF substrate. As discussed in Sect. 2.1, the extra bonding wires will bring parasitical parameters, such as inductance between wires

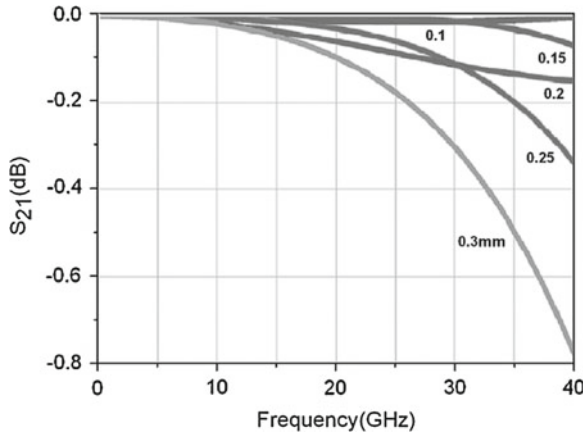


Fig. 12 Simulated S₂₁ with different wire length

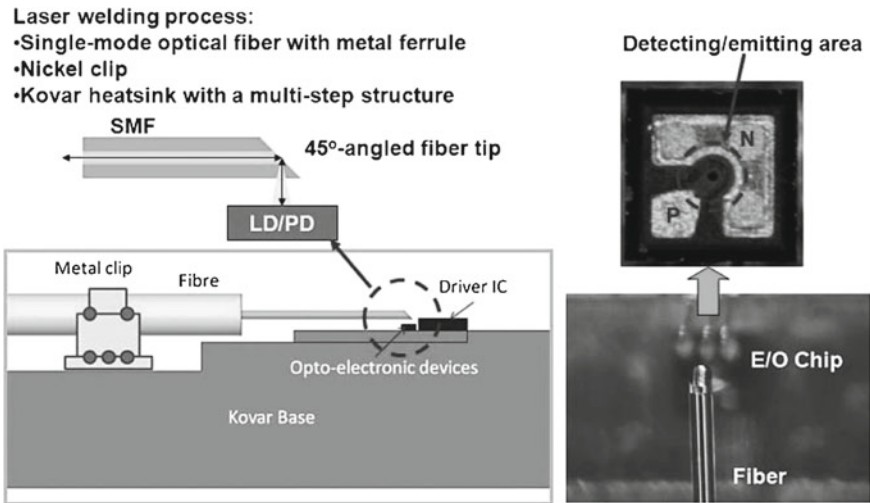


Fig. 13 Optical coupling scheme for proposed opto-electronic integrated platform. An angled fibre is used for the face-up optical coupling

and capacitance between bonding pads. In Fig. 11, an equivalent is built to examine the influence of the gold wires on transmission performance.

In Fig. 12, the transmission performance of bonding wire is investigated using the circuit model shown in Fig. 11. In terms of bonding wire inductance, a typical value of 1 nH/mm is used in the simulation. It is clear to see that when increasing the wire length from 0.1 to 0.3 mm, the RF insertion loss at 40 GHz will increase from 0.02 to 0.8 dB. It is mean that the bonding wire length should be kept as short as possible.

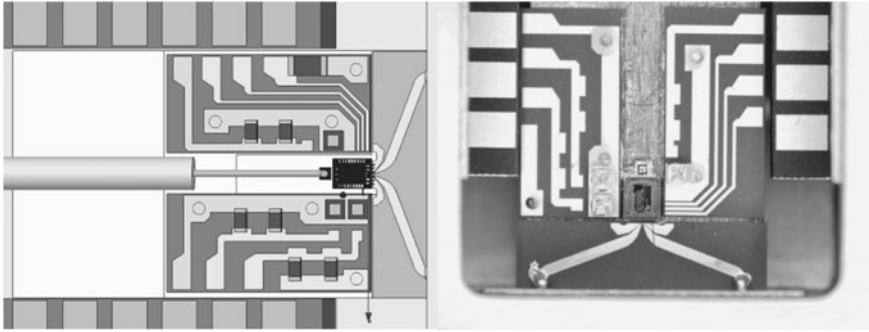


Fig. 14 Schematic diagram and photography of the proposed opto-electronic integrated platform packaged into a BTF to construct a high-speed transmitter module

In the real packaging, the wire length of the signal path is controlled below 0.25 mm and multi-wire bonding is used in the GND pads to decrease the inductance.

3 Conclusion

In this chapter, we presented an opto-electronic hybrid integrated platform to enable the fabrication of broadband, low-cost, and compact transceivers for telecommunications. The residual parameters caused by the bonding wire, IC electrodes, and microwave transmission lines in this platform were investigated carefully to minimize the signal attenuation and improve the signal integrities. In order to meet the desired bandwidth of 40 GHz, an optimized RF design was developed to minimize potential frequency decay and RF resonances. The EM simulation results obtained indicated that the RF signal transmission interface including the RF circuit and the bonding wires has an insertion loss of only 0.68 dB at 40 GHz along with a low RF reflection below -14 dB. By assembling this hybrid integrated platform into a butterfly package (BTF) as shown in Figs. 13 and 14, a transceiver module can be constructed for various applications in WDM and TDM systems.

References

1. Han W, Rensing M, Wang X, O'Brien P, Peters FH (2012) Investigation on the microwave performance of a high-speed opto-electronic hybrid integrated platform. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, London, UK, pp 1048–1052
2. Sakai K, Aruga H, Takagi SI, Kawano M, Negishi M, Kondoh Y, Kaneko S-I (2004) 1.3 μm uncooled DFB laser-diode module with a coupled differential feed for 10-Gb/s ethernet application. *J Lightwave Technol* 22(2):574–581

3. Frederique D, Catherine A, Fabrice B et al (2011) New nonlinear electrical modeling of high-speed electroabsorption modulators for 40 Gb/s optical networks. *J Lightwave Technol* 29(6):880–887
4. Han W, Zhu NH, Liang X, Ren M, Sun K, Zhang BH, Li L, Zhang HG (2009) Injection locked Fabry-Perot laser diodes for WDM passive optical network spare function. *Opt Commun* 282(17):3553–3557
5. Peschke M, Saravanan B, Hanke C, Knodl T, Stegmuller B (2004) Investigation of the capacitance of integrated DFB-EAMs with shared active layer for 40 GHz bandwidth. *Lasers Electro-Opt Soc (LEOS)* 2:673–674

Direct Torque Control of In-Wheel BLDC Motor Used in Electric Vehicle

Alireza Tashakori Abkenar and Mehran Motamed Ektesabi

Abstract Zero running emission, sustainability and efficiency of Electric Vehicle (EV) make it appropriate option for future transportation. In-wheel propulsion system of electric vehicles has been one of the main research concentrations in past decades. Brushless DC (BLDC) motor is the most suitable in-wheel motor because of its high efficiency, torque/speed characteristics, high power to size ratio, high operating life and noiseless operation. In this chapter direct torque control (DTC) switching technique with digital pulse width modulation (PWM) speed controller of BLDC motor for drive train system of EV has been reported. Effectiveness of the proposed BLDC motor drive is investigated through simulation and experiment. Obtained results show effective control of torque and remarkable reduction of torque ripple amplitude compare to conventional reported switching techniques. Improvements of in-wheel motor's torque controllability result to more efficient and safer electric vehicles.

Keywords BLDC motor · Direct torque control (DTC) · Electric vehicle · In-wheel motors · PWM technique · Torque ripple

1 Introduction

Idea of using electricity instead of fossil fuels for propulsion system of vehicles is not new. Scientists and manufacturers have attempted to design or improve electric vehicles from long time ago. Rodert Anderson built the first electric carriage in 1839

A. Tashakori Abkenar (✉) · M. Motamed Ektesabi
Faculty of Engineering and Industrial Science, Swinburne University of Technology,
PO Box 218, Hawthorn, VIC 3122, Australia
e-mail: atashakoriabkenar@swin.edu.au

M. Motamed Ektesabi
e-mail: mektesabi@swin.edu.au

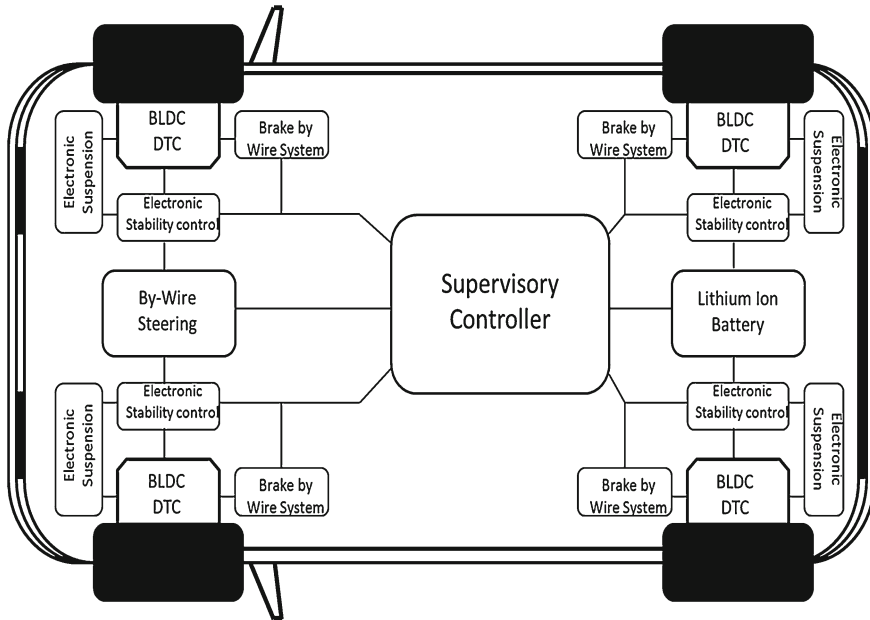


Fig. 1 Four-wheel drive train system of an IFECV

and David Salomon developed an electric car with a light electric motor in 1870, however batteries were heavy at that time and performance was poor [1]. Nowadays hybrid vehicles are more popular than electric vehicles due to better mileage and absence of enough infrastructures for charging battery of electric vehicles. Using in-wheel technology, by wire technology and intelligent control systems instead of conventional hydraulic or pneumatic control systems result to an Intelligent Fully Electronically Controlled Vehicle (IFECV) [2]. Schematic diagram of four wheel drive train of an IFECV is shown in Fig. 1.

Improving performance of in-wheel motor and its controller can increase efficiency, controllability and safety of electric vehicles. Various electrical motors have been used by manufacturers in last decades. Brushed DC, induction, switched reluctance and BLDC motors have been compared and BLDC has been recommended for high performance electric vehicle [2]. High efficiency, high speed ranges and high dynamic response due to permanent magnet (low inertia) rotor are immediate advantages of BLDC motor over brushed DC and induction motors for electric vehicle application. BLDC motor is a type of DC motor which commutation is done electronically. Therefore it has more complex control algorithm compare to other motor types. Commutation is done by knowing exact position of permanent magnet rotor. Typically there are two algorithms for rotor position detection. One uses usually sensors (Hall Effect) and the other does not which is called sensorless. Hall Effect sensors are mounted inside motor in 120 electrical degrees to detect rotor position. Optical

encoders are used for high resolution applications [3]. Back-EMF monitoring, flux linkage-based technique and free-wheeling diode conduction are some of sensorless control methods that can be used to commutate BLDC motor instead of using sensors [4]. Reducing complexity of motor construction, cost and maintenance are obvious advantages of sensorless control techniques but sensing back-EMF at low speeds, transient time and discontinuous response due to high commutation rates are its disadvantages.

Priceless researches and works have been discussed for developing different control algorithms of BLDC motor [4–8]. Sensorless control technique with a new flux linkage function has been reported for BLDC motors [4]. This method improves problem of sensorless control techniques at low speeds. A speed-independent position function named “ $G(\theta)$ ” has been defined with respect to mechanical angle of rotor. This technique is able to detect position of rotor at around 1.5 % of nominal speed. It is suitable for in-wheel application because we need to control the motor from stall position. Four-switch converter with the current controlled PWM control technique has been proposed for BLDC [5]. Difficulties in generating 120° conducting current profiles in three phase winding of BLDC with four-switched converter and current distortion in two phase cause by back-EMF of silent phase are main problems of proposed technique. A new power supply, DSP-controlled PWM chopper and C-dump converter has been reported for BLDC motor drive [6]. A dual speed and current closed-loop control is used to keep ratio of voltage to frequency constant to have constant torque operation of motor. Forced commutation RC circuits and snubber circuits to control commutation and dv/dt rating on switches have been discussed. Simulation results show number of current spikes which cause increase on torque ripple of BLDC motor which is not suitable for in-wheel application. An adaptive fuzzy control scheme via parallel distributed compensation has been applied to control velocity of small BLDC motors [7]. Simulation results show stable velocity control of BLDC motor in case of any parameter perturbation. Although stability and smoothness of torque is essential in high performance EV's, but it would be constructive for other applications like scooters and electric bikes. A digital controller of BLDC motor with two mode of operation, namely conduction angle control and current mode control has been introduced [8]. Torque is directly proportional to current in BLDC, thus current control results to torque control of motor. Speed ripple of BLDC is reduced via proposed digital controller up to maximum of 3.4%. This method could be more suitable for EV application if torque ripple reduction also has been considered [9].

DTC technique is a sensorless control technique. It does not use any Hall Effect sensor for detecting rotor position. Output torque of an electric motor is directly proportional to output power of the motor. Therefore torque control is one of important factors in drive train of electric vehicle. Reduction of torque ripples cause to deliver smoother power to the wheel. Delivering as minimum as possible ripple free torque with desired value to the wheels in various conditions, essentially increase safety and efficiency of electric vehicle. Therefore DTC switching technique is a suitable choice for high performance electric vehicles [9].

In this chapter direct torque control switching technique and PWM speed controller of BLDC motor reported in [9] for drive train of electric vehicles are described in more details. Simulation and experimental results of proposed BLDC drive are presented and discussed.

2 Direct Torque Control of BLDC Using Three Phase Conduction Mode

Direct torque control (DTC) technique of induction motors has been proposed for the first time by Takahashi and Noguchi in 1986 [10] and Depenbrock in 1988 [11]. Recently, many researchers worked on DTC of BLDC motor for applications which need precise torque control [12–17]. DTC of BLDC for drive train of hybrid electric vehicle is presented [16]. Schematic diagram of DTC of BLDC motor is shown in Fig. 2.

Torque error, stator flux error and stator flux angle are regularly used to select proper voltage space vector for switching in DTC technique. Here flux linkage error is eliminated Because of variations of stator flux magnitude regarding changes in resistance, current and voltage and specifically sharp dips at every commutation [12]. BLDC operates in both constant torque region and constant power region. Back-EMF

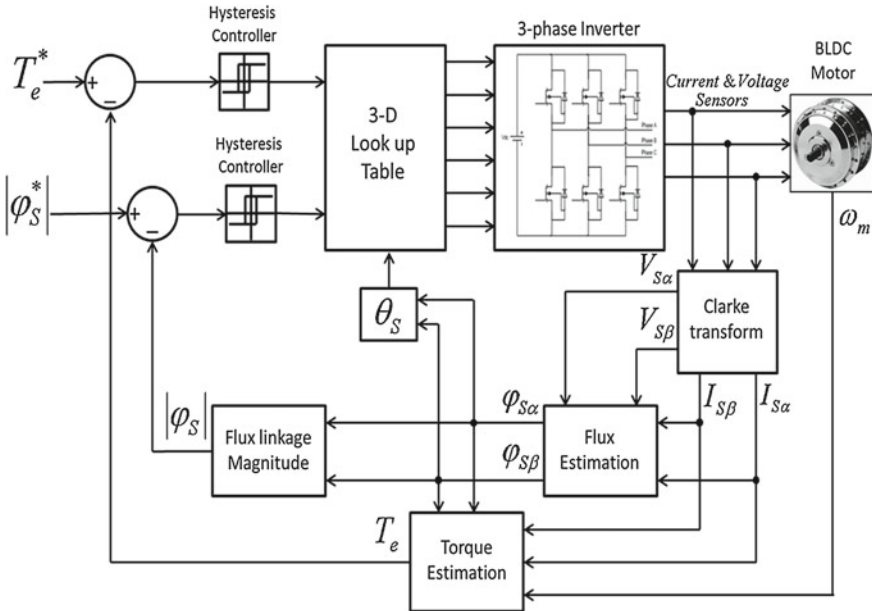


Fig. 2 Schematic diagram of DTC for BLDC motor

of motor is below DC voltage source of inverter in constant torque region (below base speed) and is increased more than DC voltage value above nominal speed. Stator inductance avoids abrupt increase of phase current and deteriorates output torque of motor. Therefore in this paper operation of BLDC motor is considered at constant torque region. Accurate estimation of flux linkage magnitude and torque is required for DTC of in-wheel motors. In some techniques, Current sensors have been used to determine flux linkage and estimate voltage from DC bus of inverter [14, 16, 17]. This method is too sensitive to voltage errors caused by dead-time effects of inverter switches, voltage drop of power electronic devices and fluctuation of DC link voltage [18]. In this chapter, both current and voltage sensors are used for accurate estimation of flux linkage magnitude and torque [15].

Precise estimation are mainly depends on accurate sensing of currents and voltages. Variations of stator resistance due to changes of temperature cause error in stator flux estimation. Pure analogue integrator also produces DC offset in signal. It is considered that BLDC motor is working in constant torque region below rated speed. Therefore there is no need of flux magnitude change during operation. Second algorithm proposed in [19] with limiting level of $2K_L\pi/(3\sqrt{3})$ (K_L is flux linkage) is used to solve analogue integrator DC drift error.

Clarke transformation is used to convert the balanced three phase system (voltages and currents) to the $\alpha\beta$ -axis references. Stator flux linkage magnitude, stator flux angle and electrical torque of motor can be estimated by [9],

$$\varphi_{S\alpha} = \int (V_{S\alpha} - Ri_{S\alpha}).dt \quad (1)$$

$$\varphi_{S\beta} = \int (V_{S\beta} - Ri_{S\beta}).dt \quad (2)$$

$$\varphi_{r\alpha} = \varphi_{S\alpha} - Li_{S\alpha} \quad (3)$$

$$\varphi_{r\beta} = \varphi_{S\beta} - Li_{S\beta} \quad (4)$$

$$|\varphi_S| = \sqrt{\varphi_{S\alpha}^2 + \varphi_{S\beta}^2} \quad (5)$$

$$\theta_S = \tan^{-1} \left(\frac{\varphi_{S\beta}}{\varphi_{S\alpha}} \right) \quad (6)$$

$$e_\alpha = \frac{d\varphi_{r\alpha}}{dt} \quad (7)$$

$$e_\beta = \frac{d\varphi_{r\beta}}{dt} \quad (8)$$

$$T_e = \frac{3}{2} \frac{P}{2} \left[\frac{e_\alpha}{\omega} i_{S\alpha} + \frac{e_\beta}{\omega} i_{S\beta} \right] \quad (9)$$

Table 1 Three phase conduction switching for DTC of BLDC

Torque error	Flux angle sectors					
	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6
1	V_6 (101)	V_1 (100)	V_2 (110)	V_3 (010)	V_4 (011)	V_5 (001)
0	V_2 (110)	V_3 (010)	V_4 (011)	V_5 (001)	V_6 (101)	V_1 (100)

where R is stator resistance, L is stator inductance, V_α is α -axis voltage vector, V_β is β -axis voltage vector, i_α is α -axis current vector, i_β is β -axis current vector, $\varphi_{S\alpha}$ is α -axis stator flux vector, $\varphi_{S\beta}$ is β -axis stator flux vector, $\varphi_{S\alpha}$ is stator flux magnitude, θ_S is stator flux angle, e_α is α -axis back-EMF, e_β is β -axis back-EMF, T_e is electric torque. After finding values of stator flux linkage in the stationary α β -axis by (1) and (2), flux linkage magnitude and angle are calculated from (5) and (6). By deriving rotor flux linkage from (3) and (4) torque of BLDC is evaluated by (9).

Hysteresis controller will generate square wave pulse with respect to torque error. Hysteresis controller output is '1' if the actual value of torque produced by motor is more than reference torque value of controller and is '0' if actual torque value is less than reference value. In this paper hysteresis band limits has been set 1, 0.1 and 0.01 to show torque ripple reduction capability of controller. The maximum switching frequency for minimum value of hysteresis band limits is near 10 KHz.

Three phase conduction mode is used for switching of VSI of BLDC motor. Six none zero voltage vectors that have been used to switch VSI are V_1 (100), V_2 (110), V_3 (010), V_4 (011), V_5 (001), V_6 (101). Voltage space vectors are chosen with respect to output of torque hysteresis controller and stator flux angle of motor [20]. Estimated stator flux angle of BLDC motor has been divided to six equal sectors. Sector one is starting from -30 degrees to 30 degrees and so on to complete one full rotation of flux linkage. Switching table of inverter to choose space vectors in each sector is shown in Table 1.

3 Digital PWM Speed Controller of BLDC

Various algorithms have been used for speed control of BLDC motor. Hysteresis current control and pulse width modulation (PWM) control are the most widely used BLDC motor control techniques [21]. Speed of BLDC motor is directly proportional to its terminal voltages. A three phase voltage source is used to supply voltage to the BLDC motor. There are two methods to adjusting the average output voltage of VSI, the variable DC link inverter and PWM switching technique. In variable DC link technique, DC bus voltage of inverter is adjusted to get the desired speed of motor. In PWM technique, a duty cycle controlled high frequency signal is multiplied to either control switching signals of upper switch of each leg, or lower switch of each leg, or all six switches of VSI [3].

Table 2 BLDC motor specification for simulation

Description	Value	Unit
DC voltage	400	V
Phase resistance (R)	2.875	Ω
Phase inductance (L)	0.8	m-H
Inertia (J)	0.8e-3	KG-m ²
Damping ratio (β)	0.001	N-s/m ²
Flux linkage	0.175	Wb
Pole pairs	4	–

In the proposed drive of BLDC motor, PWM technique is employed for speed control. PWM technique reported by Sathyan et al. [21], two predetermined states of duty cycle values (high, DH, and low, DL, PWM duty cycles) have been chosen to control the speed of BLDC motor. In this chapter a proportional and integral (PI) controller is utilized to choose a duty cycle of high frequency signal with respect to speed error. Therefore, ideally one duty cycle is chosen by controller for any particular reference speed. However, practically the controller continuously changes duty cycle value in boundaries around ideal duty cycle instead of having two predefined values [22]. Regarding safety issue which is vital factor in electric vehicles, high frequency signal is added to all six switches of inverter [3].

4 Simulation Results and Discussion

The proposed direct torque controlled drive of BLDC motor for in-wheel application is simulated by MATLAB/SIMULINK. Specification and parameters of BLDC used in simulation model are listed in Table 2. Simulation results are compared with conventional Hall Effect switching method. It is shown that controller is able to estimate torque, flux linkage magnitude and angle of BLDC precisely. By applying various hysteresis band limits, electrical torque variation is controlled in desired limit.

Speed response and torque response of DTC of BLDC for 1500 rpm speed reference under 10 N.m torque load is shown in Fig. 3. Sampling time is 5 μ s and Hysteresis band limit is set to 0.01. Model is tested for different hysteresis bands and torque responses are shown in Fig. 4. Simulation results of Fig. 4 show that width of Hysteresis band limit effects amplitude of torque ripple of BLDC motor. Peak to peak amplitude of torque ripple is reduced up to four percent of reference torque (0.4 N.m) in simulation model of proposed DTC. It is 10 times lesser than conventional Hall Effect sensors control technique. Proposed BLDC drive has a better torque response compare to presented model in [16].

Although reducing hysteresis band limits result in smoother torque but increase switching frequency of VSI. Switching frequency directly affects switching loss

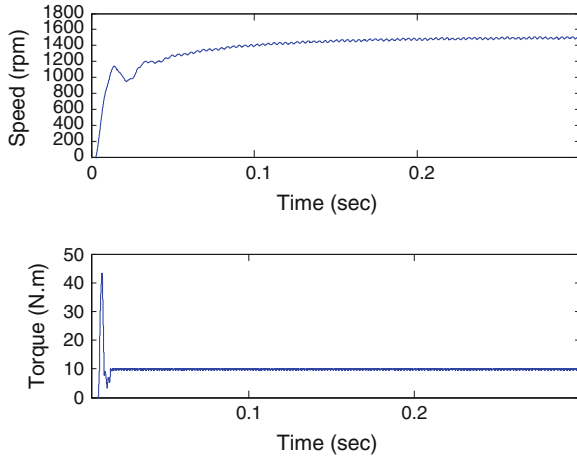


Fig. 3 Speed and torque response of BLDC motor

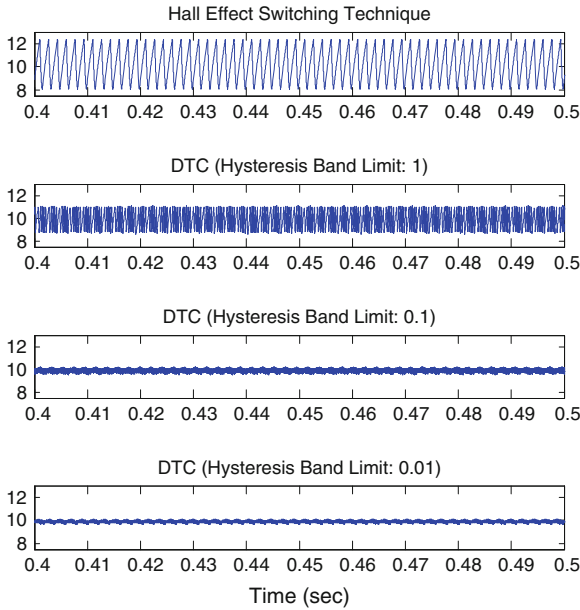


Fig. 4 Pulsating torque of BLDC for different hysteresis band limits

of inverter and practically it may not possible to have high switching frequencies. Therefore hysteresis band limits cannot be less than a particular threshold practically.

The stator flux magnitude and flux angle of BLDC motor are shown in Fig. 5. It can be observed that the stator flux magnitude in constant torque region is oscillating

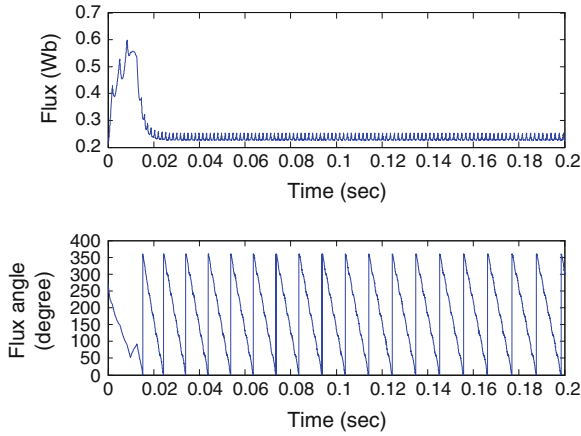


Fig. 5 Estimated stator flux linkage magnitude and angle of BLDC

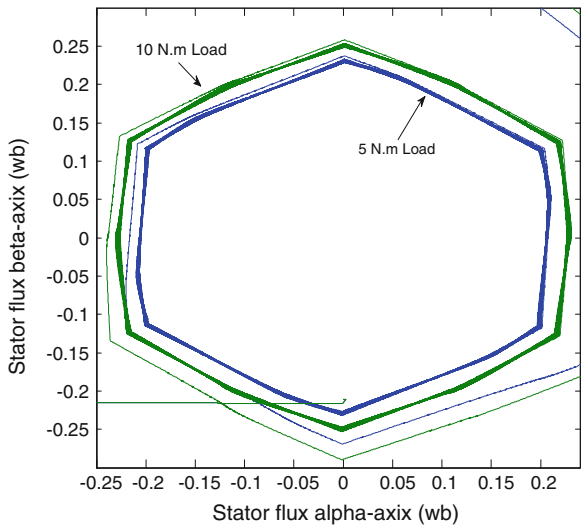


Fig. 6 Stator flux linkage trajectory for 5 N.m and 10 N.m loads

around 0.22 Wb. The flux magnitude value is almost the same as limiting level of integration algorithm.

Stator flux linkage locus of BLDC for different load torques (5 N.m and 10 N.m) is shown in Fig. 6.

The flux linkage angle sectors are clearly visible in Figure. As it can be seen increasing of torque load cause increase of flux magnitude and more sharp changes are observed [12].

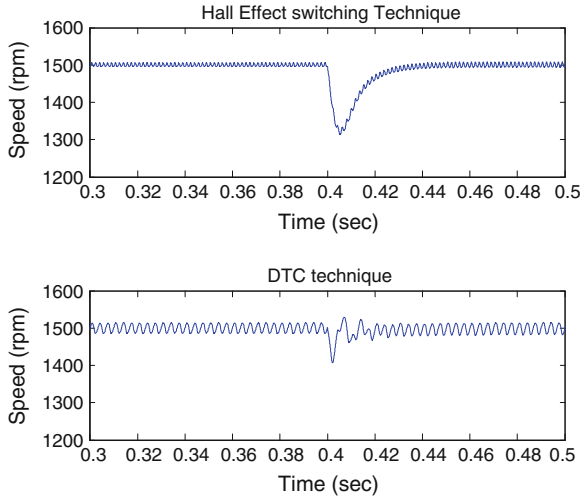


Fig. 7 Speed characteristics under same mechanical shock

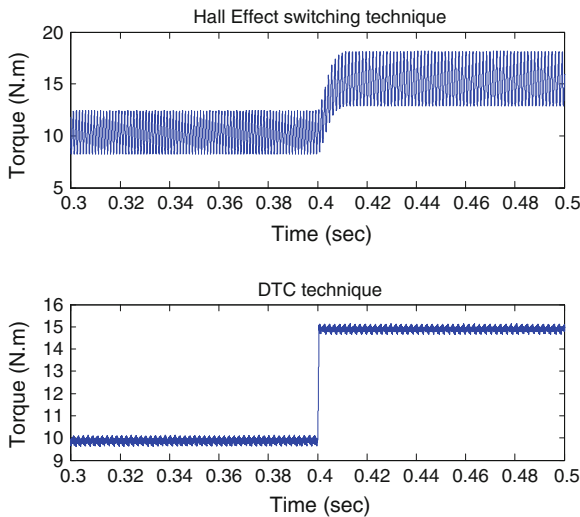


Fig. 8 Torque characteristics under same mechanical shock

According to in-wheel motor requirements, robustness of motor and controller is critical in safety point of view. Behavior of BLDC motor for proposed DTC drive and conventional Hall Effect switching technique is compared under same mechanical shocks. Sudden fifty percent change of torque load (from 10 N.m to 15 N.m) is applied to BLDC model at 1500 rpm reference speed. Speed response and torque response of BLDC motor under torque load change are shown in Figs. 7 and 8, respectively.

Abrupt change of torque load is applied at time 0.4 second. As it can be seen in Fig. 7, the speed response of DTC technique is almost fifteen times faster than con-

ventional switching technique under mechanical shock. Speed fluctuation of DTC is more than conventional switching technique, however its torque response is smoother. Figure 8 shows that dynamic torque response of DTC is much faster than Hall Effects witching technique.

5 Experimental Results

Effectiveness of proposed DTC switching technique and PWM speed controller for BLDC motor is investigated by experiment. Experimental setup of BLDC motor is shown in Fig. 9. A low voltage BLDC motor is tested as prototype. PIC18F4231 microcontroller on development board of microchip is programmed to implement proposed drive of BLDC motor. A three phase voltage source inverter with MOS-FET switches is used to supply BLDC motor. The BLDC motor specification for experiment is given in Table 3.

Direct torque control switching technique of BLDC motor is tested for 0.1 N.m reference torque and 0.01 hysteresis band limits. Reference speed of PWM speed controller is set to 2000 RPM (below nominal speed of motor). Torque response of BLDC is shown in Fig. 10.

High frequency controlled PWM signal is applied to all six switches of inverter. The line voltage of phase 'A' of BLDC motor with respect to negative terminal of DC source of inverter is shown in Fig. 11. Experimental results of proposed drive are satisfactory and show correct performance of BLDC motor.

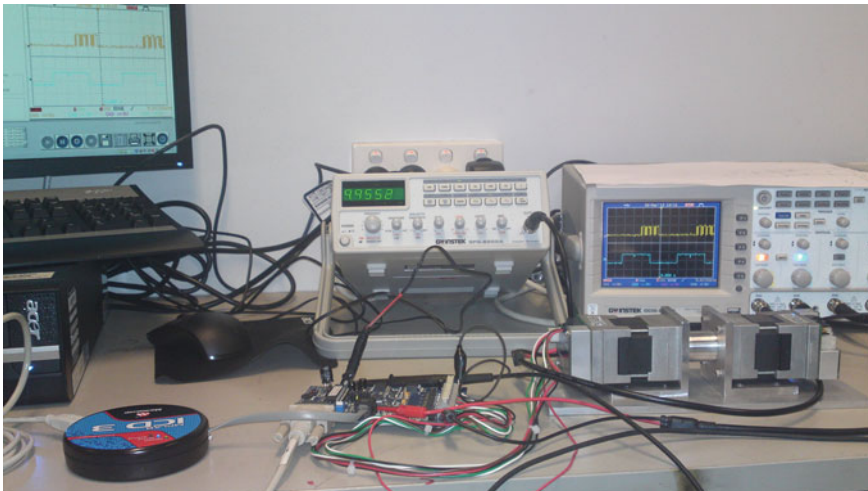


Fig. 9 Experimental setup

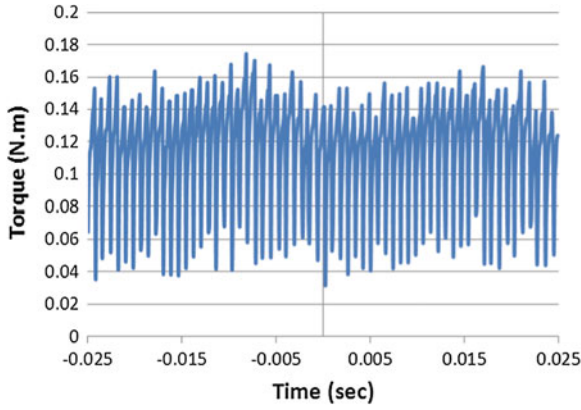


Fig. 10 Experimental torque characteristics

Fig. 11 Experimental line voltage of phase A

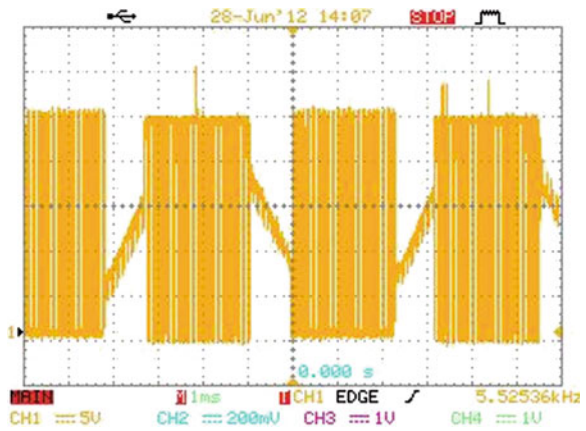


Table 3 BLDC motor specification for experiment

Description	Value	Unit
DC voltage	24	V
Rated speed	3000	RPM
Rated torque	0.28	N.m
Phase resistance (R)	2	Ω
Phase inductance (L)	4.6	m-H
Inertia (J)	4.43e-6	KG- m ²
Pole pairs	10	-

6 Conclusion

There is a growing attention to the electric vehicles in automotive industry due to control of emission of greenhouses gases in atmosphere. In-wheel technology is an advanced propulsion system in electric vehicles. BLDC motors are in interest of many manufacturers for in-wheel application. In this chapter, direct torque control switching technique of BLDC motor has been proposed as a suitable choice for drive train of electric vehicles. Simplified proposed DTC model of BLDC motor without flux observation for constant torque region has been simulated simultaneously with a digital PWM speed control technique. Proposed model has also been practically tested. The obtained results show that estimated torque calculated with state observer is very close approximation of actual output torque of motor. In this case, it has been possible to successfully control torque ripple amplitude by adjusting of Hysteresis band limit. In this model, torque ripple amplitude has reduced effectively. Hence, the developed DTC switching technique is capable of minimizing the pulsating torque of BLDC motor to deliver smoother power to the wheels. Consequently safety and efficiency of electric vehicle has been improved.

References

1. Bergsson K (2005) Hybrid vehicle history more than a century of evolution and refinement, <http://www.hybrid-vehicle.org/hybrid-vehicle-history.html>, Accessed 18 Sep 2012
2. Tashakori A, Ektesabi M, Hosseinzadeh N (2011) Characteristics of suitable drive train for electric vehicle. In: Proceedings of international conference on instrumentation, measurement, circuits and systems (ICIMCS 2011), vol 2. ASME, pp 51–57
3. Tashakori A, Ektesabi M (2012) Comparison of different PWM switching modes of BLDC motor as drive train of electric vehicles. (World Academy of Science) J Eng Technol 67:719–725
4. Tae-Hyung K, Ehsani M (2004) Sensorless control of the BLDC motors from near-zero to high speeds. IEEE Trans Power Electron ISSN 0885–8993, p1635
5. Lee BK, Ehsani M (2001) Advanced BLDC motor drive for low cost and high performance propulsion system in electric and hybrid vehicles, Texas A&M University, Dept. of Electrical Engineering, College Station, TX 77843–3128, USA
6. Luo FL, Yeo HG (2000) Advanced PM brushless DC motor control and system for electric vehicles. In: Proceedings of IEEE industry applications conference 2000, vol 2. Singapore, pp 1336–1343
7. Choi J, Park C, Rhyu S, Sung H (2004) Development and control of BLDC Motor using fuzzy models. In: Proceedings of IEEE conference on robotics, automation and mechatronics 2004, vol 2. Singapore, pp 1180–1185
8. Rodriguez F, Emadi A (Oct. 2007) A novel digital control technique for brushless dc motor drives. IEEE Trans on Industrial Electronics 54:2365–2373
9. Tashakori A, Ektesabi M (2012) Direct torque controlled drive train for electric vehicle. In : Proceeding of the world congress on engineering, WCE 2012. Lecturer notes in engineering and computer science, London, UK, 4–6 July 2012, pp 948–952
10. Takahashi I, Noguchi T (1986) A new quick-response and high-efficiency control strategies of an induction motor. IEEE Trans Ind Appl 22(5):820–827
11. Depenbrock M (Oct. 1988) Direct self-control of inverter-fed induction machine. IEEE Trans on Power Electronics 3(4):420–429

12. Ozturk SB, Toliyat HA (2007) Direct torque control of brushless dc motor with non-sinusoidal back-EMF. In: Proceedings of IEEE international conference on electric machines & drives, vol 1. IEMDC, pp 165–171
13. Oz turk SB, Toliyat HA (2008) Sensorless direct torque and indirect flux control of brushless dc motor with non-sinusoidal back-EMF. In: proceedings of 34th annual conference of IEEE on industrial electronics IECON 2008, pp 1373–1378
14. Yang J, Hu Y, Huang W, Chu J, Gao J (2009) Direct torque control of Brushless DC motor without flux linkage observation. In: proceedings of IEEE 6th international power electronics and motion control conference IPEMC 2009, pp 1934–1937
15. Wei-Sheng Y, Hai L, Hong L, Wei Y (2009) Sensorless direct torque controlled drive of brushless DC motor based on fuzzy logic. In: Proceedings of 4th IEEE conference on industrial electronics and applications ICIEA 2009, pp 3411–3416
16. Gupta A, Taehyung K, Taesik P, Cheol L (2009) Intelligent direct torque control of brushless dc motors for hybrid electric vehicles. In: Proceedings of IEEE conference on vehicle power and propulsion VPPC, pp 116–120
17. Ozturk SB, Toliyat HA (April 2011) Direct torque and indirect flux control of brushless dc motor. IEEE/ASME Trans Mechatron 16:351–360
18. Vas P (1998) Sensorless vector and direct torque control, Oxford University Press, 1998
19. Hu J, Wu B (1998) New integration algorithms for estimating motor flux over a wide speed range. IEEE Trans Power Electron 13(5):969–977
20. Zhong L, Rahman MF, Hu WY, Lim KW (1997) Analysis of direct torque control in permanent magnet synchronous motor drives. IEEE Trans Power Electron 12(3):528–536
21. Sathyan A, Milivojevic N, Lee Y, Krishnamurthy M, Emadi A (2009) An FPGA-based novel digital PWM control scheme for BLDC motor drives, IEEE Trans Ind Electron 56(8): 3040–3049
22. Tashakori A, Ektesabi M (2012) Stability analysis of sensorless BLDC motor drive using digital PWM technique for electric vehicles. In: Proceedings of 38th annual conference of IEEE on industrial electronics, (IECON 2012) (accepted), Montreal, Canada, 25–28 Oct 2012

Different Detection Schemes Using Enhanced Double Weight Code for OCDMA Systems

Feras N. Hasoon, Mohammed H. Al-Mansoori and Sahbudin Shaari

Abstract This chapter investigates the performance of enhanced double weight (EDW) code for spectral-amplitude-coding optical code division multiple access (SAC-OCDMA) system using different detection techniques. EDW code possess ideal cross-correlation properties such as the maximum cross correlation of one that are important characteristics in the optical CDMA systems since these can eliminate multiple access interference and reduce noise. The EDW code has numerous advantages including the efficient and easy code construction, simple encoder/decoder design, existence for every natural number n , and weight, which can be any odd number greater than one. The experimental simulation results as well as the transmission performances are presented in this chapter.

Keywords Cross-correlation · Detection technique · Enhanced double weight · Multiple access techniques · OCDMA · Optical coding

1 Introduction

Optical code-division multiple-access (OCDMA) systems are attracting more and more attention in the field of all-optical communications. With such systems, adapted from successful implementations in wireless networks, multiple users can access

F. N. Hasoon (✉)

Faculty of Computing and IT, Sohar University, Sohar, Sultanate of Oman
e-mail: fhasoon@soharuni.edu.om

M. H. Al-Mansoori

Faculty of Engineering, Sohar University, Sohar, Sultanate of Oman
e-mail: mmansoori@soharuni.edu.om

S. Shaari

Institute of Micro Engineering and Nanoelectronics, Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Malaysia
e-mail: sahbudin@vlsi.eng.ukm.my

the network asynchronously and simultaneously with a high level of transmission security [1, 2]. In Optical spectrum CDMA (OSCDMA) systems, each user is assigned a sequence code that serves as its address by slicing the spectrum; a CDMA user modulates its code (or address) with each data bit and asynchronously initiates transmission. This modifies the spectrum appearance, in a way recognizable only by the intended receiver. Otherwise, only noise-like bursts are observed [3]. The advantages of OSCDMA technique over other multiplexing techniques such as Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) are numerous [4, 5].

Optical CDMA communication systems require neither the time nor frequency management systems of previous techniques. The most important consideration in OCDMA is the code design. Many codes have been proposed for OCDMA, such as Hadamard [6–8], optical orthogonal (OOC) [9, 10], Prime [11], Modified Frequency-Hopping (MFH) [12–14], and Modified Double Weight [15, 16] codes. However, these all suffer from limitations—in one way or another. The codes are either too long (OOC and Prime), the constructions are complicated (OOC and MFH), or the cross-correlation are not ideal (Prime and Hadamard).

In OCDMA systems, the existence of multiple users accessing the same medium, at the same time and frequencies to transmit their data streams concurrently produce MAI. MAI is the dominant source of deterioration in an OCDMA system; therefore, a good design of the code sequences and detection scheme is important to reduce the affect of MAI [17].

In an effort to reduce the effect of MAI, Lei et al. [18] and Jen-Fa et al. [19] suggest that MAI can be minimized using the subtraction techniques, which can be done at the detection system. Several detection techniques were proposed by many researchers [20–28], most of these researches [20–25] uses complementary subtraction technique, AND subtraction technique [26], the spectral direct detection technique (SDD) [27] and XOR subtraction detection [28]. In this chapter, we investigate the performance of EDW code [29] for OSCDMA system. We compare the complementary subtraction detection technique with SDD technique; we discuss in details the effect of distance, data rate, and input power on the performance of the OSCDMA system. At the same data rate, the performance of the OSCDMA system using SDD technique is better in term of bit-error rate (BER) and optical signal-to-noise ration as compared to that of complementary subtraction technique.

The organization of this chapter is as follows. Optical code division multiple access is given in Sect. 2. In Sect. 3, we present the code structure for the spectral amplitude coding optical CDMA system, based on the double weight (DW) code families. Simulation set up and performance analyses are given in Sect. 4. Finally, a conclusion is drawn in Sect. 5.

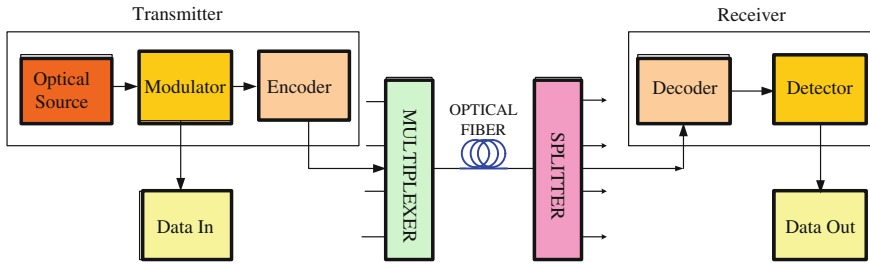


Fig. 1 Optical CDMA system diagram

2 Optical Code Division Multiple Access

To enable a large number of users to asynchronously access the network, the code division multiple access technique is used in the field of wireless communications. In this type of access, every user in the local area network is given a unique code sequence, permitted by the technology that is extended to optical fiber communication systems. This code (either temporal, spectral, or two-dimensional code) serves as an identification for any data bit sent to each user. Therefore, only the permitted or intended user will be able to retrieve the code using the encoded data stream.

Salehi [1] and Wei et al. [7] described that an encoder device in the optical CDMA network, is employed in every transmitter to encode every modulated data bit with the preset code to each intended user as shown in Fig. 1. A broadcast architecture, using a passive star coupler, is used to connect all users' to all the receivers. Then, the star coupler combines the signals from all users and these are sent to all of the matched filter receivers in the network. When a particular signal reaches the receiver, a decoder will match the code used in an individual data bit with the receiver code. The fundamental task of the decoder is to retrieve the data bits encoded with the local code and discard the rest of the signals. In point of fact, these encoded bits which do not match the codes are not totally rejected by the encoder and some noises or also known as interference may also pass through the decoder. When several active users simultaneously access the network, this scenario causes the small single user interference to add up to generate multiple access interference which directly confines the operation of the optical CDMA system. Based on this, interference is the main obstacle which needs to be dealt with in the optical CDMA network.

3 OCDMA Codes

The signals in the OCDMA systems have been encoded using various types of codes, of which these are discrete codes with every bit encoded into a sequence of small chips that signify the targeted location of the signal.

Various work done on the incoherent optical CDMA schemes sparked the creation of several major code groups like the optical orthogonal codes, or OOC [1, 2] and the prime sequence codes [30, 31] which according to Prucnal et al. [5] are well-organized in the optical delay line networks. The codes have been designed with very long code lengths and small code weights to reduce crosstalk, and lower the temporal overlap between the pulses from various users at the intensity correlator output. The length of the code is just about the square of the number of codes reasonably supported by the system.

The optical CDMA systems with delay line loops are found to be spectrally inefficient because of the long code length and small code weight used. Therefore, a sub-picosecond-pulsed optical source, with a pulse width much smaller than the bit duration, is needed. Co-channel user interference that is caused by the non-orthogonality leads to a very poor performance penalty even when the codes are carefully designed. Park et al. [32, 33] suggest that the bit error rate, or BER is commonly rather high and with a limited number co-active users allowed.

Therefore, Griffin et al. [34–36] proposed the use of a cascaded ladder encoder to reduce the splitting loss due to encoding and decoding process. According to Yang et al. [37, 38] the proposed codes to date are designed based on the modification of Prime codes. Nevertheless, the modified Prime codes offer lesser codes as compared to the original Prime codes, which directly causes the number of network subscribers to be further limited. This clearly indicates that in term of the allowed number of users, the ladder network-based systems perform even worse than many other systems which employ the use of delay line encoders.

Meghavoryan et al. [39] state that the Hadamard codes are more suitable to be used because the density of “1”s and “0” is more homogeneous, while at the same time, every user has almost the same average power. Lam et al. [6] reveal that one of the disadvantages of these codes is that they do not support many channels in a single transmission. To add to this, Aljunid et al. [40] stated that the bit error rate is rather high and that only two co-active users are allowed at a time.

Above all, it can be concluded that these codes are still restricted by various limitations in many ways. Among these, the constructions of the codes are either complicated as in the OOC and Modified Frequency Hopping (MFH) codes, the cross-correlation are not ideal, as in the Hadamard and Prime codes, or the length of the code length is too long as in the OOC and Prime codes. Due to the preference and requirement for either very wide band sources or very narrow filter bandwidths, long code lengths are regarded as a drawback in its implementation. This section includes a discussion of double weight (DW) code families for the OCDMA systems.

3.1 DW Code Constructions

In Aljunid et al. [15], the DW code can be constructed using the following steps:

Step 1:

The DW code can be represented using the $K \times N$ matrix. In the DW codes structures, the matrix K rows and N columns will represent the number of users and the minimum code length, respectively. A basic DW code is given by a 2×3 matrix, as shown below:

$$H_1 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{1}$$

Notice that H_1 has a chip combination sequence of 1, 2, 1 for the three columns (i.e., $0 + 1, 1 + 1, 1 + 0$).

Step 2:

A simple mapping technique is used to increase the number of codes as shown below:

$$H_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & H_1 \\ H_1 & 0 \end{bmatrix} \tag{2}$$

Note that as the number of users, K increases, the code length, N also increases. The relationship between the two parameters, K and N is given by Equation below:

$$N = \frac{3K}{2} + \frac{1}{2} \left[\sin\left(\frac{K\pi}{2}\right) \right]^2 \tag{3}$$

3.2 MDW Code Construction

MDW code [15, 41] is the modified version of DW code. The MDW code weight can be any even number that is greater than two. The MDW can also be represented using the $K \times N$ matrix. The basic MDW code denoted by (9, 4, 1) is shown below:

$$H_m = \left| \begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \end{array} \right| \tag{4}$$

The same mapping technique as for the DW code is used to increase the number of users. An MDW code with a weight of four denoted by $(N, 4, 1)$ for any given code length N , which can be related to the number of users, K through:

$$N = 3K + \frac{8}{3} \left[\sin\left(\frac{K\pi}{3}\right) \right]^2 \tag{5}$$

3.3 EDW Code Construction

Hasoon et al. [29] state that the EDW code is the enhanced version of double weight DW code as reported in [15]. The DW code weight can be any even number that is greater than two while the enhanced double weight EDW code weight can be any odd number greater than one.

EDW code can be represented by using a $K \times N$ matrix. In EDW codes structures, the matrix K rows and N columns represent the number of users and the minimum code length respectively. In this chapter, the EDW code with the weight of three is used as an example. The basic EDW code denoted by (6, 3, 1) is shown below:

$$H_1 = \begin{vmatrix} 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{vmatrix} \quad (6)$$

From the basic matrix, a larger value of K can be achieved by using a mapping technique as shown below:

$$H_2 = \begin{bmatrix} 0 & H_1 \\ H_1 & 0 \end{bmatrix} \quad (7)$$

EDW codes have the subsequent properties: ideal maximum cross-correlation $\lambda_{\max} = 1$, EDW code weight, which can be any odd number greater than one, the weight pair structure maintained, the chip combination is maintained 1, 2, 1 for every consecutive pairs of codes, and the relation between the number of users K and code length N at weight of 3 is given by [29]:

$$N = 2K + \frac{4}{3} \left[\sin\left(\frac{K\pi}{3}\right) \right]^2 \left[\frac{8}{3} \left[\sin\left(\frac{(K+1)\pi}{3}\right) \right]^2 + \frac{4}{3} \left[\sin\left(\frac{(K+2)\pi}{3}\right) \right]^2 \right] \quad (8)$$

4 Experimental Simulation Result

A simple schematic block diagram consisting of 2 users is illustrated in Figs. 2 and 3 as an illustrative example (the study was carried out for 3 users). Each chip has a spectral width of 0.2 nm. The tests were carried out using Optisys, an established commercial software at the rates of 2.5 and 10 Gbps for 5–55 km.

The fibre used had the values of parameters taken from the data which are based on the G.652 Non Dispersion Shifted Fibre (NDSF) standard. This included the attenuation, group delay, group velocity dispersion, dispersion slope and effective index of refraction, which were all wavelength dependent. The non-linear effects such as the Four Wave Mixing and Self Phase Modulation (SPM) were also activated. At 1550 nm wavelength, the attenuation co-efficient was 0.25 dB/km, and the chromatic dispersion co-efficient was 18 ps/nm-km and the polarization mode dispersion (PMD)

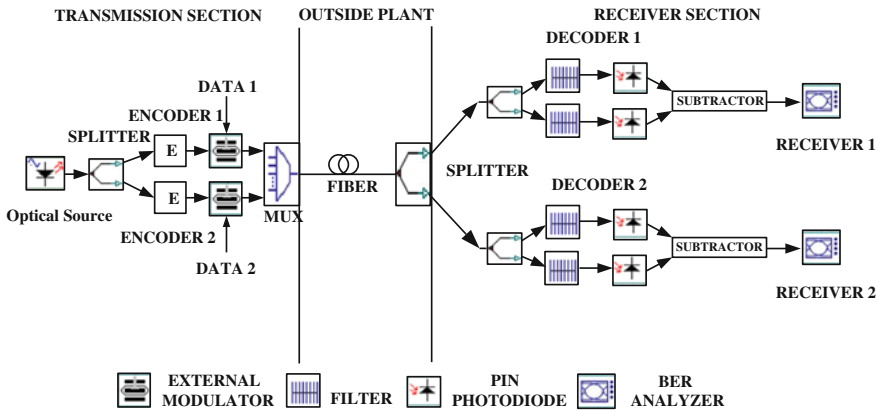


Fig. 2 Simulation setup for the OCDMA system with complementary technique

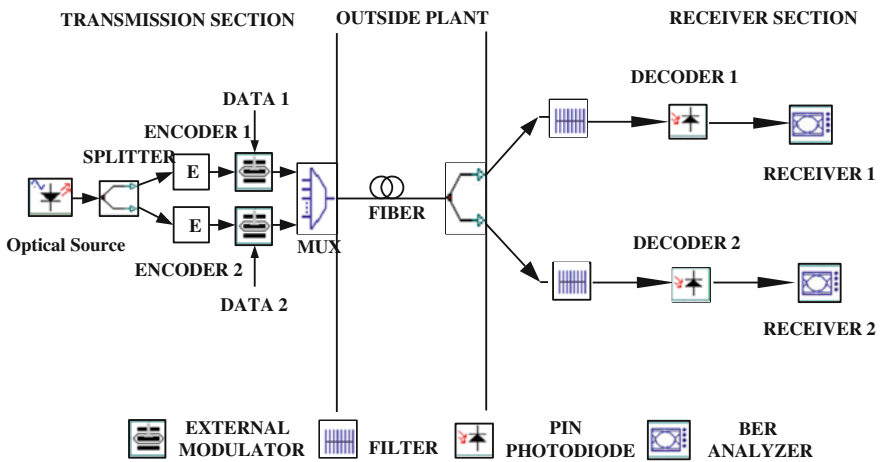


Fig. 3 Simulation setup for the OCDMA system with direct technique

co-efficient was $5 \text{ ps}/\sqrt{\text{km}}$. The transmit power used was 0 dBm out of the broadband source. The noises generated at the receivers were set to be random and totally uncorrelated. The dark current value was 5 nA and the thermal noise co-efficient was $1.8 \times 10^{-23} \text{ W/Hz}$ for each of the photo-detectors. The performance of the system was evaluated by referring to the bit error rate, output power and optical signal to the noise ratio (OSNR).

Figure 2 show that the incoming signal was split into two parts at the receiver side, one to the decoder that had an identical filter structure with the encoder and the other to the decoder that had the complementary filter structure. A subtractor was used to subtract the overlapping data from the desired one. In Fig. 3, no subtractors are needed at the receivers.

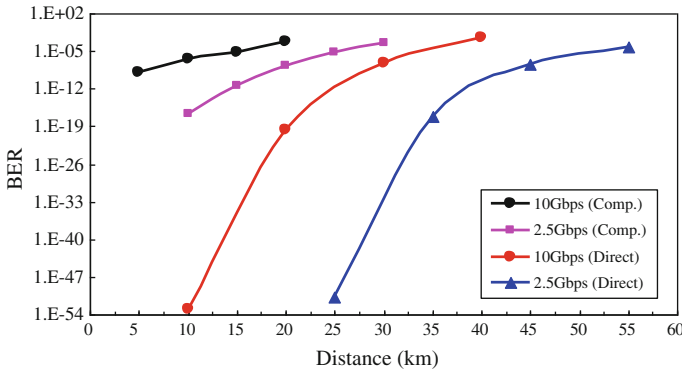


Fig. 4 BER versus distance for the OCDMA system using complementary and direct techniques at different transmission rates

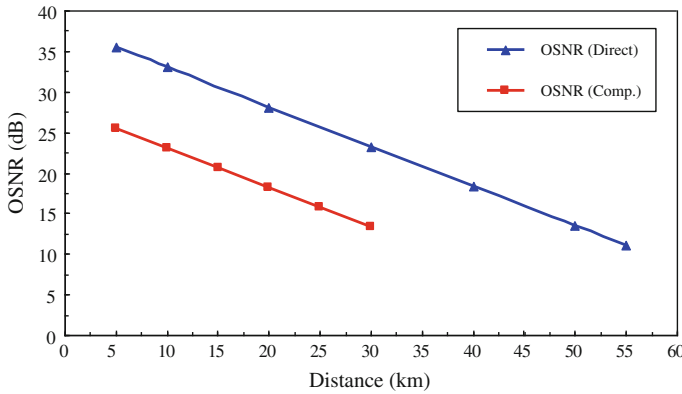


Fig. 5 Optical signal to noise ratio for the OCDMA system using complementary and direct techniques at different transmission distance for 2.5 Gbps transmission rates

Figure 4 shows that the use of the direct detection technique gives better BER system performance at the same data rate as compared to that of the complementary subtraction technique. For example, by using BER of 10^{-12} as the transmission quality cut-off, it was found that the system using complementary subtraction technique could perform well up to 12 km only at STM-16 rate as compared to the direct detection technique which still gives excellent performance at the distance of 38 km. The performance of the direct technique is evident at all rates with supportable distance double of that supported by the conventional technique [42].

Figure 5 shows the optical signal to noise ratio against the transmission distance for both detection techniques; the complementary and SDD. The OSNR for the complementary technique is lower than for the direct technique at 2.5 Gbps. However, it is worth to note that the result for the complementary technique is measured up to 30 km only, because the system cannot support longer distance at acceptable BER

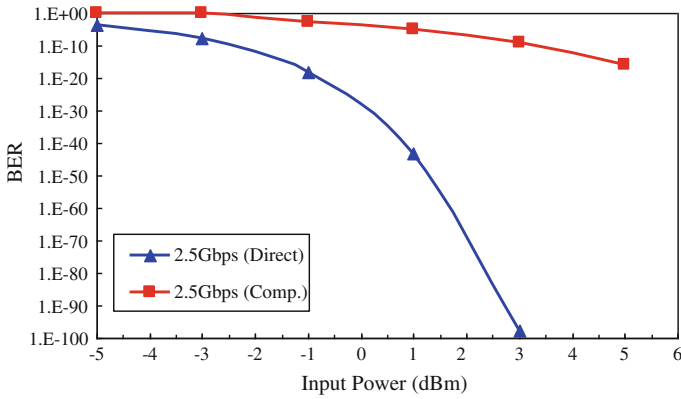


Fig. 6 BER versus input power for the different detection techniques

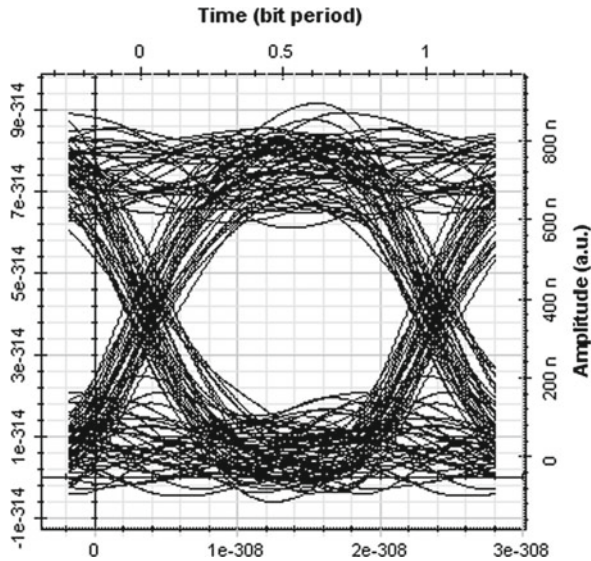


Fig. 7 Eye diagram of the direct technique at 10Gbps with BER of 2.36×10^{-10} at 27km transmission

performance. In addition, Fig. 5 shows a nearly linear reduction of the OSNR with distance for both direct and complementary techniques. For example, at a distance of 10 km, the OSNR for the direct technique is about 33.073 dB as compared to 22.98 for the complementary technique. This is about 10 dB higher for the direct technique. It should be noted that for the result of the direct technique, the power loss can be reduced so the distance can be extended, and the total power loss for the direct technique can be reduced because of the lesser number of filters used in the decoder.

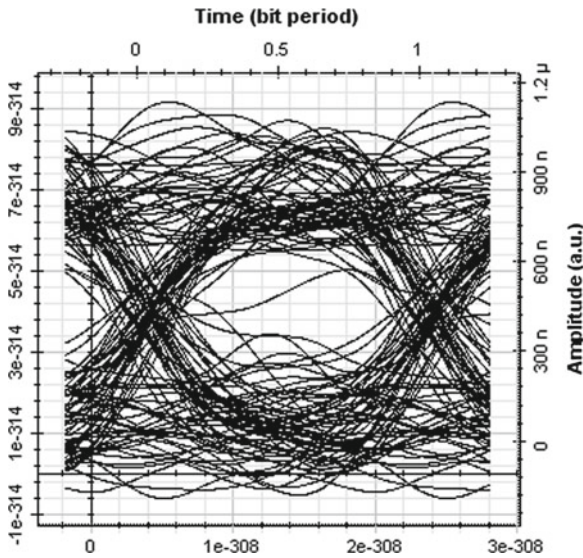


Fig. 8 Eye diagram of the complementary technique at 10 Gbps with BER of 4.47×10^{-4} at 5 km transmission

We also investigate the effect of the input signal power on the BER performance as depicted on Fig. 5. Figure 6 shows that increasing the input power from -5 to 3 dBm for both techniques improve the BER performance. For complementary technique there is small improvement in the performance of the system, while increasing the input power from -3 to 3 dBm. On the other hand, the performance of the system using the direct technique improves significantly. Therefore, the system with complementary technique gives small improvement in the performance as clearly shown.

The eye patterns shown in Figs. 7 and 8. clearly depict that the direct detection technique had a better performance with a larger eye opening. The BER for the direct and complementary techniques were 10^{-10} , and 10^{-4} , respectively at data rate of 10 Gbps.

5 Conclusion

Optical code division multiple access is a one of multiplexing technique used in the filed of optical communication systems. In this type of access, every user is assigned with a unique code sequence that serves as its address.

The EDW code is a double weight (DW) code family variation that has weight greater than one. This code possesses numerous advantages including the efficient and easy code construction, simple encoder/decoder design, and ideal maximum cross-correlation $\lambda = 1$ and higher SNR.

Various optical CDMA detection techniques were also studied in this chapter, such as the complementary techniques and direct technique based on the EDW code to improve the OCDMA system performance. At the same data rate, it has been shown through simulation that the BER performance of the OCDMA system using direct detection technique is better as compared to that of complementary subtraction technique. For example, by using BER of 10^{-12} as the transmission quality cut-off, it was found that the system using complementary subtraction technique could perform well, i.e., up to 12 km only at the data rate of 2.5 Gbps as compared to direct detection technique that could still give excellent performance at the distance of 38 km. In addition, at data rate of 10 Gbps the direct detection technique can support distance up to 30 km as compared to only 5 km using complementary detection technique for BER 10^{-9} .

References

1. Salehi JA (1989) Code division multiple access techniques in optical fiber network-Part I: fundamental principles. *IEEE Trans Commun* 37(8):824–833
2. Salehi JA, Brackett CA (1989) Code division multiple access techniques in optical fiber network-Part II: system performance analysis. *IEEE Trans Commun* 37(8):834–842
3. Pearce MB, Aazhang B (1994) Multiuser detection for optical code division multiple access systems. *IEEE Trans Commun* 42:1801–1810
4. Maric SV, Kostic ZI, Titlebaum EL (1993) A new family of optical code sequences for use in spread-spectrum fiber-optic local area networks. *IEEE Trans Commun* 41:1217–1221
5. Prucnal P, Santoro M, Ting F (1986) Spread spectrum fiber optic local area network using optical processing. *J Lightwave Technol* 4(5):547–554
6. Lam CF, Tong DTK, Wu MC (1998) Experimental demonstration of bipolar optical CDMA system using a balanced transmitter and complementary spectral encoding. *IEEE Photon Technol Lett* 10(10):1504–1506
7. Wei Z, Shalaby HM, Ghafouri-Shiraz H (2001) New code families for fiber-brag-grating-based spectral-amplitude-coding optical CDMA Systems. *IEEE Photon Technol Lett* 13(8):890–892
8. Wei Z, Ghafouri-Shiraz H, Shalaby HM (2001) Performance analysis of optical spectral-amplitude-coding CDMA systems using a super-fluorescent fiber source. *IEEE Photon Technol Lett* 13(8):887–889
9. Weng C-S, Wu J (2001) Optical orthogonal codes with non-ideal cross-correlation. *J Lightwave Technol* 19(12):1856–1863
10. Kwong WC, Yang G-C (2002) Design of multilength optical orthogonal codes for optical CDMA multimedia networks. *IEEE Trans Commun* 50:1258–1265
11. Wen J-H, Lin J-Y, Liu C-Y (2003) Modified prime-hop codes for optical CDMA systems. *IEEE Proc Commun* 150(5):404–411
12. Wei Z, Ghafouri-Shiraz H (2002) Proposal of a novel code for spectral amplitude-coding optical CDMA systems. *IEEE Photon Technol Lett* 14(3):414–416
13. Wei Z, Ghafouri-Shiraz H (2002) Unipolar codes with ideal in-phase cross-correlation for spectral amplitude-coding optical CDMA systems. *IEEE Trans Commun* 50(8):1209–1212
14. Wei Z, Ghafouri-Shiraz H (2002) Codes for spectral-amplitude-coding optical CDMA systems. *J Lightwave Technol* 20(8):1284–1291
15. Aljunid SA, Ismail M, Ramli AR, Borhanuddin MA, Abdullah MK (2004) A new family of optical code sequences for spectral-amplitude-coding optical CDMA systems. *IEEE Photon Technol Lett* 16(10):2383–2385

16. Hasoon FN, Aljunid SA, Abdullah MK, Shaari S (2007) New code structure for spectral amplitude coding in OCDMA system. *IEICE Electron Express* 4(23):738–744
17. Fadhil HA, Aljunid SA, Ahmad RB (2009) Performance of random diagonal code for OCDMA systems using new spectral direct detection technique. *J Opt Fiber Technol* 15(3):283–289
18. Lei X, Glesk I, Baby V, Prucnal PR (2004) Multiple access interference (MAI) noise reduction in A 2D optical CDMA system using ultrafast optical thresholding, lasers and electro-optics society 2004, LEOS 2004, the 17th annual meeting of the IEEE, vol 2, pp 591–592
19. Jen-Fa H, Chao-Chin Y (2002) Reductions of multiple-access interference in fiber-grating-based optical CDMA network. *IEEE Trans Commun* 50(10):1680–1687
20. Nguyen L, Aazhang B, Young JF (1995) All-optical CDMA with bipolar codes. *Electron Lett* 31:469–470
21. Smith EDJ, Blaikie RJ, Taylor DP (1998) Performance enhancement of spectral- amplitude-coding optical CDMA using pulse-position modulation. *IEEE Trans Commun* 46:1176–1185
22. Yim RMH, Bajcsy J, Chen LR (2003) A new family of 2-D wavelength-time codes for optical CDMA with differential detection. *IEEE Photon Technol Lett* 15:165–167
23. Djordjevic IB, Vasic B (2003) Novel combinatorial constructions of optical orthogonal codes for incoherent optical CDMA systems. *J Lightwave Technol* 21:1869–1875
24. Zaccarin D, Kavehrad M (1994) Performance evaluation of optical CDMA systems using non-coherent detection and bipolar codes. *J Lightwave Technol* 12:96–105
25. Hasoon FN, Abdullah MK, Aljunid SA, Shaari S (2007) Performance of OCDMA systems using complementary subtraction technique. *Opt Netw* 6:854–859
26. Hasoon FN, Aljunid SA, Abdullah MK, Shaari S (2008) Spectral amplitude coding OCDMA using and subtraction technique. *J Appl Opt* 47(9):1263–1268
27. Abdullah MK, Hasoon FN, Aljunid SA, Shaari S (2008) Performance of OCDMA systems with new detection schemes using enhanced double weight (EDW) code. *J Opt Commun* 281(18):4658–4662
28. Hassan YA, Ibrahim F, Naufal MS, Aljunid SA (2010) OCDMA system: new detection scheme and encoder-decoder structure based on fiber bragg gratings (FBGS) for vcc code. *Int J Comput Appl* 32(4):461–468
29. Hasoon FN, Aljunid SA, Abdullah MK, Shaari S (2007) Construction of a new code for spectral amplitude coding in optical code-division multiple-access systems. *Opt Eng J* 46(7):75004–75008
30. Perrier PA, Kwong WC, Prucnal PR (1991) Performance comparison of asynchronous and synchronous code-division multiple-access techniques for fiber-optic local area networks. *IEEE Trans Commun* 39(11):1625–1634
31. Walle H, Killat U (1995) Combinatorial BER analysis of synchronous optical CDMA with prime sequences. *IEEE Trans Commun* 43(12):2894–2895
32. Park E, Mendez AJ, Galiardi RM, Dale MR (1993) Fiber-optic digital video multiplexing using optical CDMA. *J Lightwave Technol* 11(1):20–26
33. Park E, Mendez AJ, Garmire EM (1992) Temporal/spatial optical CDMA network design, demonstration and comparison with temporal networks. *IEEE Photon Technol Lett* 4(10):1160–1162
34. Griffin RA, Sampson DD, Jackson DA (1992) Optical phase coding for code division multiple access networks. *IEEE Photon Technol Lett* 4(12):1401–1404
35. Griffin RA, Sampson DD, Jackson DA (1994) Photonic CDMA by coherent matched filtering using time-addressing coding in optical ladder networks. *J Lightwave Technol* 12(11):2001–2010
36. Griffin RA, Sampson DD, Jackson DA (1995) Coherence coding for photonic code-division multiple access networks. *J Lightwave Technol* 13(9):1826–1837
37. Yang GC, Kwong WC (1995) Performance analysis of optical CDMA with prime Codes. *Electron Lett* 10(7):569–570
38. Yang GC, Jaw J-Y (1994) Performance analysis and sequence designs of synchronous code-division multiple access systems with multimedia services. *IEE Proc Commun* 141(6):371–378

39. Meghavorian DM, Baghdasaryan HV (2001) Code-division multiple access: novel multiplexing strategy in optical fiber networks. In: Proceedings of 3rd international conference on transparent optical, nNetworks, pp 299–303
40. Aljunid AS, Maisara O, Hasnurollhaya S, Abdullah MK (2003) A new code structure for optical code division multiple access systems. The 3rd international conference on advances in strategic technologies (ICAST'03), knowledge-based technologies for sustainable development, vol 1, pp 553–558
41. Hasoon FN, Aljunid SA, Abdullah MK, Shaari S (2007) New code structure for spectral amplitude coding in OCDMA system. IEICE Electron Express 4(23):738–744
42. Hasoon FN, Al-Mansoori MH, Kazem HA, Ghazi Zahid AZ, Saini DK, Shaari S (2012) Performance of OCDMA systems with different detection schemes using enhanced double weight (EDW) code, lecture notes in engineering and computer science: proceedings of the world congress on engineering , WCE 2012, London, UK, vol 2198(1), pp 979–981

Multigate RADFET Dosimeter for Radioactive Environment Monitoring Applications

Fayçal Djéffal and Mohamed Meguellati

Abstract In this chapter, a new radiation sensitive FET (RADFET) dosimeter design (called the Dual-Dielectric Gate All Around DDGAA RADFET dosimeter) to improve the radiation sensitivity performance and its analytical analysis have been proposed, investigated and expected to improve the sensitivity behavior and fabrication process for RADFET dosimeter-based applications. Analytical models have been developed to predict and compare the performance of the proposed design and conventional (bulk) RADFET, where the comparison of device architectures shows that the proposed design exhibits a superior performance with respect to the conventional RADFET in term of fabrication process and sensitivity performances. The proposed design has linear radiation sensitivities of approximately $95.45 \mu\text{V}/\text{Gy}$ for wide irradiation dose range (from Dose = 50 Gy to Dose = 3000 Gy). Our results showed that the analytical analysis is in close agreement with the 2-D numerical simulation over a wide range of devices parameters. The proposed device and the Artificial Neural Networks (ANNs) have been used to study and show the impact of the proposed dosimeter on the environment monitoring and remote sensing applications. The obtained results make the DDGAA RADFET dosimeter a promising candidate for environment monitoring applications.

Keywords Artificial neural networks (ANNs) · Dosimeter · Environment monitoring · Genetic algorithm (GA) · Irradiation · RADFET · Remote sensing · Sensitivity · Traps

F. Djéffal (✉) · M. Meguellati
LEA, Department of Electronics, University of Batna, 05000 Batna, Algeria
e-mail: faycal.djeffal@univ-batna.dz; faycaldzdz@hotmail.com

M. Meguellati
e-mail: m_meguellati@yahoo.fr; meguellatidz@gmail.com

1 Introduction

Radiation sensitive MOSFETs (RADFETs) have been focus of interest both from applications and fundamental research point of views. In electronic industry these devices are considered as attractive alternatives for nuclear industry, space, radio-therapy and environment monitoring applications due to their reliability, low power consumption, non-destructive read-out of dosimetric information, high dose range, and compatibility to standard CMOS technology and on-chip signal processing [1–3].

The conventional (bulk) RADFET is a unique radiation dosimeter with a high scale dimension (the dimensions of sensor elements are $\approx 1 \text{ mm}^2$).

The pMOS dosimeter advantages, in comparison with other dosimetric systems, include immediate, non-destructive read out of dosimetric information, extremely small size of the sensor element, the ability to permanently store the absorbed dose, wide dose range, very low power consumption, compatibility with microprocessors, and competitive price (especially if cost of the read out system is taken into account). Figure 1 shows the using of pMOS dosimeter in radiation therapy [4]. The main bulk RADFET disadvantages are a need for calibration in different radiation fields (“energy response”), relatively low resolution (starting from about 1 rad) and nonreusability. In this context, the submicron multi-gate design may be considered as attractive alternative to overcome this disadvantage because of the high electrical performance and reliability provided by the multi-gate structure in comparison with single-gate one. However, as semiconductor devices are scaled into the deep sub-micron domain, short-channel effects (SCEs) begin to plague conventional planar CMOS-based devices. To avoid the electrical constraints and improve the sensitivity performance, a new design and enhancement of conventional (bulk) RADFET become important. In this context, the multigate design may provide an improved performance, in comparison to conventional design, for future RADFET-based applications.

The Gate All Around GAA MOSFETs have emerged as excellent devices to provide the electrostatic integrity needed to scale down transistors to minimal channel lengths, and allowing a continuous progress in digital and analog applications. In addition to a better electrostatics than the conventional bulk MOSFET, the uses of these devices have advantages relative to the electronic transport, mainly due to:

- (1) the reduced surface roughness scattering because the lower vertical electric field and
- (2) the reduction of the Coulomb scattering because the film is made of undoped/low-doped silicon [5–9].

Design and modeling guidelines of GAA MOSFETs have been discussed in previous work [6–9]. Employing this design for environment monitoring applications (irradiation measurement) becomes more beneficial if the device is made in vertical cylindrical recrystallized silicon due to highly flexible process integration options.

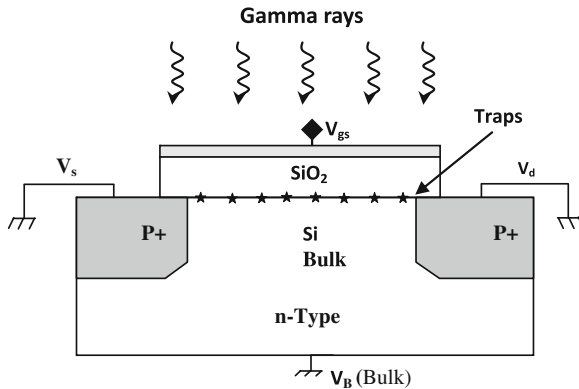


Fig. 1 Cross section of pMOS dosimeter (RADFET) with the defects created by radiation [4]

There have been several reports of MOSFETs fabricated in recrystallized silicon for high-density digital integrated circuits [9].

In this chapter, a new design of RADFET called the Dual-Dielectric Gate All Around (DDGAA) RADFET dosimeter, in which the manufacturing processes and sensitivity performances will be greatly improved, is proposed for deep submicron CMOS-based dosimeter applications. The (DDGAA) RADFET dosimeter design presented in this chapter is basically surrounded dual-dielectric layers (SiO_2 and Si_3N_4) with low p-channel (Si) doping concentration. The results showed that the analytical model is in agreement with the 2-D numerical simulation over a wide range of device parameters. The proposed structure has been analyzed and validated by the good sensitivity and electrical performance obtained in deep submicron regime in comparison with the conventional (bulk) design. In addition, we present the applicability of genetic algorithm optimization (GA) approach to optimize the radiation sensitivity of the DDGAA RADFET for integrated CMOS-based dosimeters.

Finally, the proposed dosimeter model was used to study and show the impact of the proposed design on the environment monitoring applications.

This chapter is organized as follows. In Sect. 2, we derive an analytical interface potential distribution including radiation-induced interface-traps. The threshold voltage shift model can then be determined based on the interface potential model. In Sect. 3, we investigate the performance of the proposed design. The conclusions will be drawn in Sect. 4.

2 Theory Development and Model Derivation

2.1 Interface Potential Analysis

Schematic cross-sectional view of the proposed (DDGAA) RADFET dosimeter is presented in Fig. 2. The insulator consists of a thermal oxide (SiO_2) grown on a

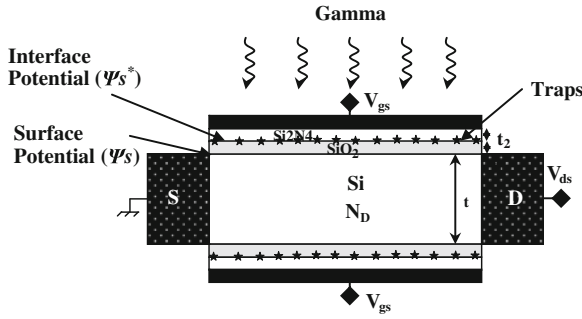


Fig. 2 Cross-sectional view of the proposed DDGAA RADFET design

(100) n on n⁺ epitaxial silicon substrate (channel), and a low pressure CVD silicon nitride layer (Si₃N₄) deposited on top of the oxide. ND/S represents the doping level of the drain/source region, respectively. The channel region is bounded by source and drain spacing at $x = 0$ and L , respectively, where L is the gate length. With a negatively applied gate bias, holes generated in the SiO₂ layer are transported and trapped at the SiO₂/Si₃N₄ interface producing a measurable threshold-voltage shift as it is shown in Fig. 2. The investigation reported in this work for gamma radiation sources can also be applied qualitatively to other radiation sources (protons, electrons, ...).

For deep submicron devices, the solution of 2D Poisson’s equation satisfying suitable boundary conditions is required to model the interface potential. Refer to Fig. 2, the 2D Poisson’s equation for the channel region is given by

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \psi(r, x) \right) + \frac{\partial^2}{\partial x^2} \psi(r, x) = \frac{q \cdot N_D}{\epsilon_{si}} \tag{1}$$

The boundary conditions for $\psi(x, r)$ are found by satisfying continuity of both the normal component of the electric displacement at the (Si/SiO₂) interfaces, and the potential at the source/drain sides.

Using the same parabolic potential profile in vertical direction [6] and applying the symmetry condition of $\partial\psi/\partial r = 0$ for $r = 0$, we obtained the following expressions of 2-D channel potential as

$$\psi(r, x) = \frac{C_{ox}}{\epsilon_{si} \cdot t_{si}} \left[V_g^* - \psi_s(x) \right] r^2 + \left(1 + \frac{C_{ox} t_{si}}{4\epsilon_{si}} \right) \psi_s(x) - \frac{C_{ox} t_{si}}{4\epsilon_{si}} V_g^* \tag{2}$$

where $\psi_s(x)$ represents the surface potential (the potential at the Si/SiO₂ interface), with C_{ox} represents the insulator capacitance ($C_{ox} = 2\pi\epsilon_1 L / \ln(1 + 2t_1/t_{si})$), t_{si} is the silicon thickness, the effective oxide and silicon nitride layer is defined as $t_{oxeff} = t_1 + t_2 \frac{\epsilon_1}{\epsilon_2}$ with t_1 is the thickness of the SiO₂ ($\epsilon_1 = \epsilon_{ox}$) layer and t_2 is the thickness of the Si₃N₄ layer (ϵ_2). V_{bi} is the junction voltage between the

source/drain and intrinsic silicon, $V_{bi} = (kT/q)\ln(N_{D/S}/n_i)$, n_i is the intrinsic silicon density, V_{ds} represents the drain-to-source voltage and k is the Boltzmann constant. V_g^* represents the effective voltage at the gate which is introduced to simplify notations and alleviate derivations for symmetric structure as $V_g^* = V_{gs} - V_{fb}$, with V_{fb} is the flat-band voltage.

Substituting (2) in (1), we obtain the differential equation that deals only with surface potential as

$$\frac{d^2\psi_s(x)}{dx^2} - \frac{1}{\lambda^2}\psi_s(x) = D_1 \quad (3)$$

$$\text{with } \lambda = \sqrt{\frac{\varepsilon_{si} \cdot t_{oxeff} \cdot t_{si}}{4 \cdot \varepsilon_{ox}}} \text{ and } D_1 = \frac{q \cdot N_D}{\varepsilon_{si}} - \frac{1}{\lambda^2} \cdot V_g^*$$

where λ represents the natural length of the analyzed (DDGAA) RADFET dosimeter. This parameter gives the scaling capability (downscaling ability) of the device. D_1 is a factor which represents the impact of the applied gate voltage and channel doping on the surface potential.

The boundary conditions in channel and oxide regions (Fig. 2) are given as,

$$\psi_s(x=0) = V_{bi} \quad (4a)$$

$$\psi_s(x=L) = V_{bi} + V_{ds} \quad (4b)$$

$$\varepsilon_2 \frac{V_{gs}^* - \psi_s^*}{t_2} = \varepsilon_1 \frac{\psi_s^* - \psi_s}{t_1} \quad (\text{continuity of the normal component of the electric displacement at the interface } SiO_2/Si_3N_4) \quad (4c)$$

where ψ_s^* represents the interface potential at SiO_2/Si_3N_4 interface which satisfies the continuity of the normal component of the electric displacement at the interface (Eq. (4c)).

Substituting (4c) in (3), we obtain the differential equation that deals only with interface potential as

$$\frac{d^2\psi_s^*(x)}{dx^2} - \frac{1}{\lambda^2}\psi_s^*(x) = D_2 \quad (5)$$

$$\text{with } D_2 = \alpha - \beta V_{gs}^* \text{ and } \alpha = \frac{qN_D\varepsilon_2t_1}{\varepsilon_{si}(\varepsilon_2t_1 + \varepsilon_1t_2)}, \beta = \frac{\varepsilon_2t_1}{\lambda^2(\varepsilon_2t_1 + \varepsilon_1t_2)}$$

This resolution of this Equation allows us the calculation of the interface potential without (before) irradiation.

In the case of RADFET under irradiation new term should be introduced in order to include the radiation-induced interface-traps effect [6, 10, 11]. So, the parameter D_2 can be written, in this case, as, $D_2 = \alpha - \beta V_{gs}^* - \frac{qN_f}{\varepsilon_2t_2}$, with N_f represents the irradiation induced localized interface charge density per square area. The second

term in this expression represents the impact of the irradiation induced localized interface charge density on the interface potential.

Using these boundary conditions (Eqs. 4a, 4b and 4c), the surface and interface potentials can be, respectively, expressed as

$$\psi_S(x) = -\lambda^2 D_2 + \frac{\phi_D \sinh\left(\frac{x}{\lambda}\right) - \phi_S \sinh\left(\frac{x-L}{\lambda}\right)}{\sinh\left(\frac{L}{\lambda}\right)} \quad (6)$$

With $\phi_D = V_{ds} + \lambda^2 D_2$ and $\phi_S = V_{bi} + \lambda^2 D_2$

$$\psi_S^*(x) = \frac{\varepsilon_1 t_2}{\varepsilon_2 t_1 + \varepsilon_1 t_2} V_{gs}^* + \frac{\varepsilon_2 t_1 x}{\varepsilon_2 t_1 + \varepsilon_1 t_2} \psi_S(x) \quad (7)$$

2.2 Threshold Voltage Shift Model

Schematic cross-sectional view of the proposed (DDGAA) RADFET The basic concept of RADFET dosimeter is to convert the threshold voltage shift, ΔV_{th} , induced by radiation, into absorbed radiation dose, where $\Delta V_{th} = V_{th} - V_{th0}$ with V_{th} and V_{th0} represent the threshold voltage after and before irradiation, respectively.

Based on the surface potential model given by Eq. (5), the threshold voltage can be derived using the condition of the minimum channel potential $\psi_{s\ min}|_{V_{gs}=V_{th}} = 2 \cdot \phi_B$, with $\psi_{s\ min} = \psi_s(x_{min})$, V_{th} is the threshold voltage value, and ϕ_B represents the bulk potential of silicon body given as $\phi_B = (K_B T/q) \cdot \ln(N_D/n_i)$. The location of the minimum surface potential can be obtained analytically by solving $\frac{d\psi_s(x)}{dx} = 0$ [1, 6].

The solution of the equation $\psi_{s\ min}|_{V_{gs}=V_{th}} = 2 \cdot \phi_B$ at low drain-source voltage for long channel lengths ($L \gg \lambda$) can be given as

$$V_{th} = \frac{\left(2A\phi_B + \lambda^2\alpha + \frac{qN_f}{\varepsilon_2 t_2}\right) \sinh\left(\frac{L}{\lambda}\right) + (V_{bi} - V_{ds}) \sinh\left(\frac{L}{2\lambda}\right)}{\left(\beta\lambda^2 - \frac{B}{A}\right) \sinh\left(\frac{L}{\lambda}\right) - 2\sinh\left(\frac{L}{2\lambda}\right)} \quad (8a)$$

with: $A = \frac{\varepsilon_1 t_2 - \varepsilon_2 t_1}{\varepsilon_1 t_2}$, $B = \frac{\varepsilon_2 t_1}{\varepsilon_1 t_2}$

$$V_{th0} = V_{th}|_{N_f=0} = \frac{\left(2A\phi_B + \lambda^2\alpha\right) \sinh\left(\frac{L}{\lambda}\right) + (V_{bi} - V_{ds}) \sinh\left(\frac{L}{2\lambda}\right)}{\left(\beta\lambda^2 - \frac{B}{A}\right) \sinh\left(\frac{L}{\lambda}\right) - 2\sinh\left(\frac{L}{2\lambda}\right)} \quad (8b)$$

From (8a) and (8b), the threshold voltage shift can be given as

$$\Delta V_{th} = \frac{\frac{qN_f}{\varepsilon_2 t_2} \sinh\left(\frac{L}{\lambda}\right)}{\left(\beta\lambda^2 - \frac{B}{A}\right) \sinh\left(\frac{L}{\lambda}\right) - 2\sinh\left(\frac{L}{2\lambda}\right)} \quad (8c)$$

3 Results and Discussion

The RADFET radiation sensitivity, S , given by [3, 10, 12]:

$$S = \frac{\Delta V_{th}}{D} \quad (9)$$

where D represents the absorbed radiation dose.

From an experimental study carried out by Jaksic et al. [3] an empirical relationship between the interface charge densities, N_f , and the absorbed radiation dose, D , for the silicon material at room temperature can be written as

$$N_f = d_{11}D + d_{12} \quad (10)$$

where $d_{11} = 1.6 \times 10^9 \text{ cm}^{-2}/\text{Gy}$ and $d_{12} = 5 \times 10^{10} \text{ cm}^{-2}$ are fitting parameters.

Substituting (10) in (8c), we obtain the RADFET sensitivity model that deals only with absorbed radiation dose as

$$S = \frac{\frac{q}{\epsilon_2 l_2} \sinh\left(\frac{L}{\lambda}\right) \left(d_{11} + \frac{d_{12}}{D}\right)}{\left(\beta \lambda^2 - \frac{B}{A}\right) \sinh\left(\frac{L}{\lambda}\right) - 2 \sinh\left(\frac{L}{2\lambda}\right)} \quad (11)$$

In Fig. 3, the variation of DDGAA RADFET sensitivity versus the absorbed radiation dose, D , has been compared with conventional (bulk) RADFET. For both designs, the output response of the RADFETs is linear with absorbed radiation dose. It is clearly shown that DDGAA RADFET has higher sensitivity, $S = 95.45 \mu\text{V}/\text{Gy}$, in comparison with conventional RADFET design, $S = 30.68 \mu\text{V}/\text{Gy}$. This means

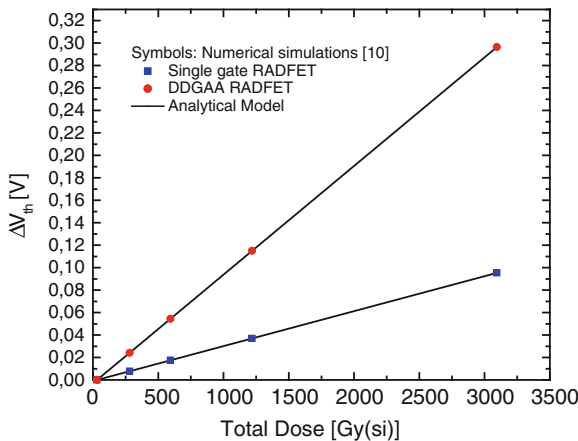


Fig. 3 Variation of threshold voltage shift in function of the absorbed radiation dose for the conventional and DDGAA RADFET designs

that DDGAA RADFET has better electrical and scaling performances in comparison with the conventional design. So, our design provides a high sensitivity, better electrical and technological performances in comparison with the conventional structure. These results make the proposed design as a promising candidate for CMOS-based dosimeters.

3.1 GA-Based Sensitivity Optimization

GA optimization has been defined as finding a vector of decision variables satisfying constraints to give acceptable values to objective function. It has recently been introduced to study the complex and nonlinear systems and has found useful applications in engineering fields. Due to the simple mechanism and high performance provided by GA for global optimization, GA can be applied to find the best design of DDGAA RADFET in order to improve the radiation sensitivity by satisfying of the following objective function:

– Maximization of the RADFET radiation sensitivity $S(X)$

where X represents the input normalized variables vector which is given as $X = (t_{si}, t_1, t_2, L)$.

For the purpose of GA-based optimization of the radiation sensitivity of DDGAA RADFET, routines and programs for GA computation were developed using MATLAB 7.2 and all simulations are carried out on a Pentium IV, 3 GHz, 1 GB RAM computer. For the implementation of the GA, tournament selection is employed which selects each parent by choosing individuals at random, and then choosing the best individual out of that set to be a parent. Scattered crossover creates a random binary vector. It then selects the genes where the vector is unity from the first parent, and the genes where the vector is zero from the second parent, and combines the genes to form the child. An optimization process was performed for 20 population size and maximum number of generations equal to 200, for which stabilization of the fitness function was obtained.

The steady decrease in objective function in each generation until it reaches a best possible value can be attributed to the selection procedure used namely Roulette wheel selection.

The radiation sensitivity values of the DDGAA RADFET with and without optimization are shown in Table 1. It is clearly shows that the radiation sensitivity, for

Table 1 DDGAA RADFET design parameters

Symbol	Optimized design	Design without optimization	Conventional design
L (nm)	100	100	100
tsi (nm)	50	20	20
t1 (nm)	5	5	5
t2 (nm)	15	5	–
S ($\mu\text{V}/\text{Gy}$)	162.22	95.45	30.68

optimized design ($162.22 \mu\text{V}/\text{Gy}$) is better than the both conventional RADFET ($S = 30.68 \mu\text{V}/\text{Gy}$) and DDGAA RADFET without optimization ($S = 95.45 \mu\text{V}/\text{Gy}$).

3.2 Radioactive Environment Sensing

In order to show the impact of the proposed design on the radioactive environment monitoring, we propose the study of a contaminated radioactive environment. This latter is considered a big challenge in the field of the environment monitoring. In this work, using simulated database (built from numerical data) of total dose radioactivity distribution in 2-D space and the Multi-Layer-Perception (MLP) tool, we will study a contaminated environment by gamma radiation.

Artificial neural network (ANN) based methods have been widely used for modeling various complex and nonlinear processes (classification, speech recognition, and signal processing). The model based on artificial neural network [13–15] assumes that input and output patterns of a given problem are related by a set of neurons organized in hidden layers. The layers in these networks are interconnected by communication links that are associated with weights that dictate the effect on the information passing through them. These weights are determined by the learning algorithm.

The output of node j in the hidden layer is given by

$$h_j = g \left(\sum_{i=14}^n w_{ij} \cdot x_i + b_j \right)$$

And the output of the network by

$$y = \sum_{i=14}^k w_{oi} \cdot h_i$$

where w_{ij} are the weights connecting the inputs to node j in the hidden layer, b_j is the bias to the node, and w_{oi} are the weights from the hidden to the output layer.

The activation function relates the output of a neuron to its input based on the neuron's input activity level. Some of the commonly used functions include: the threshold, piecewise linear, sigmoid, tangent hyperbolic, and the Gaussian function [14]. The learning process of the MLP network involves using the input–output data to determine the weights and biases. One of the most techniques used to obtain these parameters is the back-propagation algorithm [14–16]. In this method, the weights and biases are adjusted iteratively to achieve a minimum mean square error between the network output and the target value.

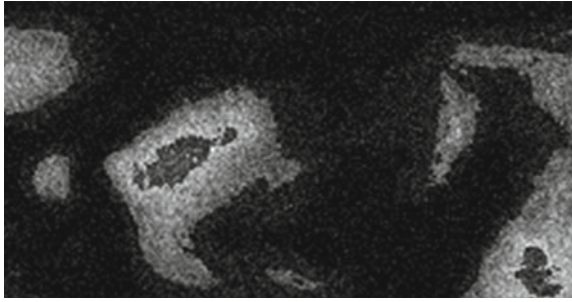


Fig. 4 The distorted image due to the transmission noise

The energy recorded by the sensor has to be transmitted, often in electronic form, to a receiving and processing station, where the data are processed into an image. Radiation that is not absorbed or scattered in the atmosphere can reach and interact with the Earth's surface. There are three forms of interaction that can take place when energy strikes, or is incident upon the surface [17].

In remote sensing, we are most interested in measuring the radiation reflected from targets. This reflection disgusting the image caption by the sensors (which are located at reception stations), we propose that interacting energy (noise) is a Gaussian noise.

In this work, the artificial neural network is used to denoising the image distorted by the transmission noise (Fig. 4). In this context, the database for MLP optimization consists of 49600 samples split into three categories: training, validation and test sets. The training and validation are used tune MLP configuration and the test is used to test the MLP configuration to denoise the different regions of the contenned environment. Test and training steps were run for a given MLP structure to obtain the optimal MLP configuration. The database is collected from several RADFETs, which have been located in different regions in the contenned environment. In order to validate the denoising proprieties of the optimized MLP, test set is compared to the MLP response.

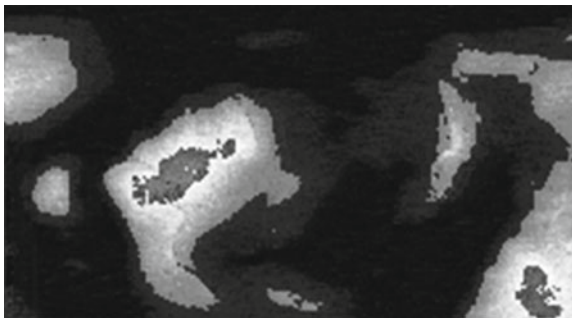


Fig. 5 The denoised image using MLP

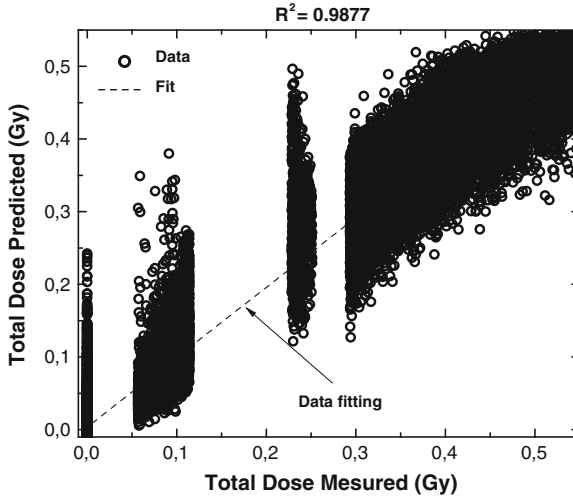


Fig. 6 Validation of the neural network result for test set

Figure 5 presents the space distribution of the gamma radiation in the investigated contaminated environment after the denoising process. It is shown that the different regions are clearly represented. This last observation shows the applicability and the efficiency provided by the MLP-based approach to study the radioactive environment.

Figure 6 shows that a good agreement between MLP and real results is found. Hence, the optimized structure can be used for the radioactive environment monitoring applications.

4 Conclusion and Future Work

In this chapter, we compared new sensor design, DDGAA RADFET, with conventional planar RADFET through 2-D analytical investigation. A two-dimensional analytical analysis comprising radiation-induced interface-traps effect, 2D surface and interface potentials, threshold voltage shift and sensitivity model for DDGAA RADFET has been developed. It has been found that incorporation of the gate all around design leads to an improvement threshold behavior while also enhancing the gate controllability, and thus provides better performance as compared to conventional planar RADFETs. The threshold voltage shift behavior of the proposed design was more effectively improved than those of the conventional planar RADFET. Also, we confirmed that DDGAA RADFET had advantages in CMOS scaling in comparison with planar RADFET. Moreover, the DDGAA RADFET has a linear sensitivity about $S = 95.45 \mu\text{V}/\text{Gy}$ in radiation dose ranging from $D = 0$ to $D = 3000 \text{ Gy}$. With continued progress towards fabricating RADFET-based

dosimeters, it is possible to fabricate DDGAA RADFET without much technological processes. Our analytical analysis provides the incentive for experimental exploration of the DDGAA RADFETs with around-gate and cylindrical-channel aspects. Application of the GA-based design approach to DDGAA RADFET has also been discussed. It can be concluded that proposed GA-based approach is efficient and gives the promising results. In order to show the impact of the proposed design on the radioactive environment monitoring, we developed a MLP-based approach to study a condemned radioactive environment. The proposed approach can be used for remote sensing applications, where the information about the condemned radioactive environment should be transmitted in electronic form to a receiving and processing station. It is to note that our work can be extended to implement the proposed design into Software tools in order to study the engineering systems under radiation conditions.

References

1. Djeflal F, Bendib T, Meguellati M, Arar D, Abdi MA (2012) New dual-dielectric gate all around (DDGAA) RADFET dosimeter design to improve the radiation sensitivity. In: Proceedings of the world congress on engineering 2012 Vol II WCE 2012, London, pp 917–921, 4–6 July 2012.
2. Kelleher A, Lane W, Adams L (1995) A design solution to increasing the sensitivity of pMOS dosimeters: the stacked RADFET approach. *IEEE Trans Nucl Sci* 42:48–51
3. Jaksic A, Ristic G, Pejovic M, Mohammadzadeh A, Sudre C, Lane W (2002) Gamma-ray irradiation and post-irradiation responses of high dose range RADFETs. *IEEE Trans Nucl Sci* 49:1356–1363
4. Holmes-Siedle A (1994) From space to therapy—the radiation sensitive silicon fet (RADFET). In: Proceedings of the technology transfer workshop, European Space Agency, ESA SP-364. ESA/ESTEC, Noordwijk, The Netherlands 1994:87–90
5. The international technology roadmap for semiconductors (ITRS) (2007) <http://public.itrs.net>
6. Ghoggali Z, Djeflal F, Lakhdar N (2010) Analytical analysis of nanoscale double- gate MOSFETs including the hot-carrier degradation effects. *Int J Electron* 97(2):119–127
7. Jiménez D, Iñiguez B, Suñé J, Marsal LF, Pallarès J, Roig J, Flores D (2004) Continuous analytic I-V model for surrounding-gate MOSFETs. *IEEE Electron Devices Lett* 25:571–573
8. Kaur H, Kabra S, Haldar S, Gupta RS (2007) An analytical drain current model for graded channel cylindrical/surrounding gate MOSFET. *Microelectron J* 38:352–359
9. Abdi MA, Djeflal F, Arar D, Hafiane ML (2008) Numerical analysis of double gate and gate all around MOSFETs with bulk trap states. *J Mater Sci Mater Electron* 19:S248–S253
10. Meguellati M, Djeflal F (2012) New dual-dielectric gate all around (DDGAA) RADFET dosimeter design to improve the radiation sensitivity. *Nucl Instrum Methods Phys Res A* 683:24–28
11. Djeflal F, Meguellati M, Benhaya A (2009) A two-dimensional analytical analysis of sub-threshold behavior to study the scaling capability of nanoscale graded channel gate stack DG MOSFETs. *Phys E* 41:1872–1877
12. Schwank JR, Roeske SB, Beutler DE, Moreno DJ, Shaneyfelt MR (1996) *IEEE Trans Nucl Sci* 43:2671–2678
13. Atlas User's manual, SILVACO TCAD, 200.
14. Guessasma S, Ghislain M, Christain C (2004) Modeling of the APS plasma spray process using artificial neural networks: basis requirements and an example. *Comput Mater Sci* 29:315

15. Djeflal F, Chahdi M, Benhaya A, Hafiane ML (2007) An approach based on neural computation to simulate the nanoscale CMOS circuits: application to the simulation of CMOS inverter. *Solid-State Electron* 51:48–56
16. Taylor JG (1996) *Neural networks and their applications*. Wiley, West Sussex
17. A Canada Centre for Remote Sensing, Remote Sensing Tutorial. http://www.uprm.edu/biology/profs/chinea/gis/g06/NRC1_1_1_5.pdf

Multi-Objective-Based Approach to Optimize the Analog Electrical Behavior of GSDG MOSFET: Application to Nanoscale Circuit Design

Toufik Bendib and Fayçal Djéffal

Abstract In this chapter, the small signal parameters behavior of Gate Stack Double Gate (GSDG) MOSFET are studied and optimized using multi-objective genetic algorithms (MOGAs) for nanoscale CMOS analog circuits' applications. The transconductance and the OFF-current are the small signal parameters which have been determined by the analytical explicit expressions in saturation and subthreshold regions. According to the analytical models, the objectives functions, which are the pre-requisite of genetic algorithms, are formulated to search the optimal small signal parameters in order to obtain the best electrical and dimensional transistor parameters to obtain and explore the better transistor performances for analog CMOS-based circuit applications. Thus, the encouraging obtained results may be of interest to practical applications. The optimized design is incorporated into circuit simulator to study and show the impact of our approach on the nanoscale CMOS-based circuits design. In this context, we proposed to study the electrical behavior of a ring oscillator circuit. In this study a great improvement of the oscillation frequency has been recorded in our case. The main advantages of the proposed approach are its simplicity of implementation and provide to designer optimal solutions that suites best analog application.

Keywords Analog application · CMOS · Double gate · Gate stack · Genetic algorithm · MOGA · Small signal · Submicron

T. Bendib (✉) · F. Djéffal
LEA, Department of Electronics, University of Batna, 05000 Batna, Algeria
e-mail: bendib05.t@gmail.com; toufikdzdz@gmail.com

F. Djéffal
e-mail: faycal.djeffal@univ-batna.dz; faycaldzdz@hotmail.com

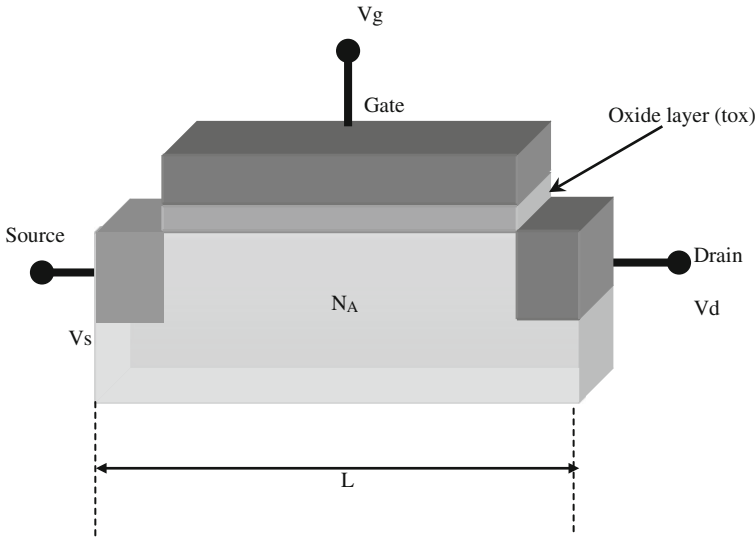


Fig. 1 Cross-sectional view of the conventional MOSFET

1 Introduction

Because the widely uses of the silicon based devices in many fields of physics, a lot of aspects related to the behavior of CMOS devices, and a global understanding of their effect structure and properties became increasingly important due to the reduction in chip sizes and to the increase of the operation speed [1]. The downscaling of device dimensions has been the primary factor leading to improvements in Integrated Circuits (CIs) performance and cost, which contributes to the rapid growth of the semiconductor industry.

As MOSFET gate length (Fig. 1) enters nanoscale field, small signal parameters and short channel effect such as off current, threshold voltage roll-off and drain-induced-barrier-lowering become increasingly significant, which limit the scaling capability of MOSFET design [2, 3]. Downscaling MOSFETs to their limits is a key challenge faced by the nanoelectronic industry. Therefore, a new designs and structures become necessary to overcome these challenges.

Double-Gate (DG) MOSFETs have become very attractive for scaling CMOS devices down to deep submicron sizes because of a number of advantages such excellent short-channel-effects immunity, as ideal subthreshold slope and unique mobility enhancement [4–6]. This structure utilizes a very thin body to eliminate sub-surface leakage paths between the source and drain. The use of an undoped body is desirable for immunity against dopant fluctuation effects which give rise to threshold-voltage roll-off, and also for reduced drain-to-body capacitance and higher carrier mobility which provide for improved circuit performance. The studied device presented in Fig. 2 is considered as symmetrical structure, with a double-layer gate

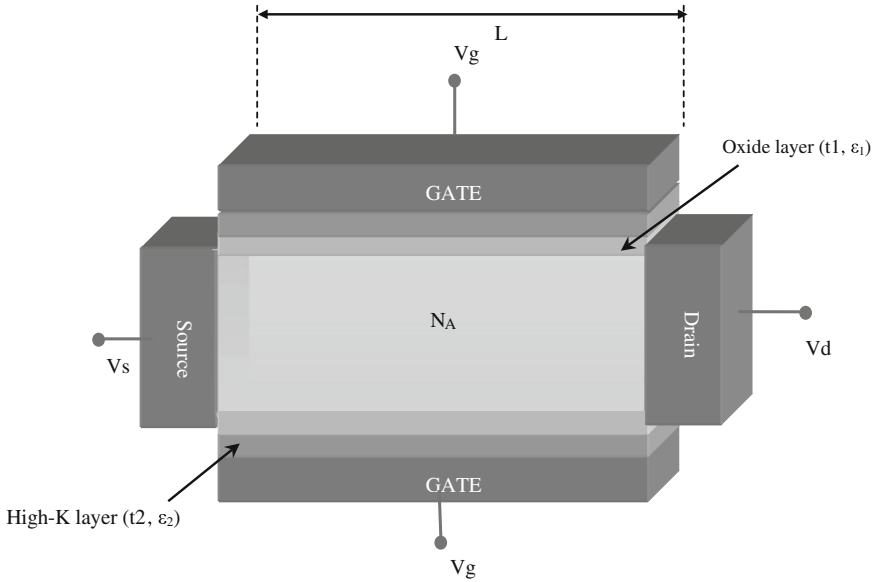


Fig. 2 Cross-sectional view of the proposed GSDG MOSFET design

stack, oxide and high- k layers, with no overlap with the source extensions. ND/S represents the doping level of the drain/source region respectively.

The small-signal and subthreshold parameters particularly the transconductance (gm) and the OFF-current are required basically for analog circuits design [5, 8]. These parameters can easily be derived from the device drain current models. To improve the device performances for analog circuits' applications, a new design approaches are required to enhance the reliability performances of the devices for analog applications (voltage amplifiers, Current conveyor,...). One preferable approach is the evolutionary-based model, which could provide practical solutions for a nanoscale CMOS circuits design. We called this the "intelligent simulator" approach [8].

Recently, novel structures for the DG MOSFETs with high performance and scalability are used for nanoscale analog circuits in order to improve the device immunity against SCEs and the small signal parameters. Therefore, new design approaches are required to enhance the reliability and electrical performances of the devices for nanoscale analog applications. Although a numerous of literatures have modeled and studied the subthreshold and saturation behaviours of the nanoscale DG MOSFET [5–14].

The key idea of these approaches is to find the best dimensions, electrical and biasing conditions of the transistor to facilitate the circuits design strategy. In order to facilitate a device design and improve electrical behavior for high-performances analog circuits, the proposed approach suggests two objective functions, which are the transconductance (which describes the transistor behavior in saturation regime)

and the OFF-current (which gives the power dissipation of the device), that simplifies the optimization procedure.

In this chapter, we present the applicability of multi-objective genetic algorithm optimization (MOGA) approach to optimize the small signal parameters of GSDG MOSFET for deep submicron CMOS analog applications. Design optimization, adopted in this work, is the process of finding the maximum/minimum of the device parameters called the objective functions and must also satisfy a certain set of specified requirements within constraints [3, 13, 14]. In the present work, we present an alternative approach based on MOGAs, where the designer can specify several objective functions simultaneously and the optimal results are presented as a global results. The main advantages of this approach are its simplicity of implementation and provide to designer optimal solutions that suites best analog application.

2 MOGAS Computation Methodology

Multi-objective optimization has been defined as finding a vector of decision variables satisfying constraints to give acceptable values to all objective functions [9, 10]. It has recently been introduced to study the complex and non linear systems and has found useful applications in engineering fields [9–11]. An ideal multi-objective optimization procedure constitutes of two steps. The first is to find some optimal solutions corresponding to multiple objectives considered in research space. The next step is to choose the most suitable solution by using higher level information. Due to the simple mechanism and high performance provided by MOGAs for multi-objective global optimization, MOGA can be applied to study the nanoscale GSDG MOSFETs.

In the present study, Pareto approach should be a suitable choice. Pareto approach searches non-dominant solutions called Pareto optimal solutions in the objective space. The number of Pareto optimal solutions is always not single [10, 14].

The objective in the design of optimal GSDG MOSFET for analog and digital CMOS-based devices is to find the better design of the transistor that satisfies the high working performances in subthreshold regime.

The first step of our approach consists of compact models of transconductance and OFF-current parameters for GSDG MOSFETs proposed in [15]. Using the SILVACO software [16], it was observed that the formulated parameters models can be used as objective functions, which are given as function of input design variables.

In the saturation region the drain current can be given as,

$$I_{ds} = \mu_{eff} C_{oxeff} \frac{W}{L} \left[(V_g - V_0)^2 - \frac{8rk^2T^2}{q^2} e^{q(V_g - V_0 - V_{ds})/kT} \right] \quad (1)$$

where μ_{eff} is the effective mobility, C_{oxeff} is the effective capacitance oxide, $r = \varepsilon_{si} t_{oxeff} / \varepsilon_{ox} t_{si}$ is a structural parameters, $t_{oxeff} = t_1 + \frac{\varepsilon_1}{\varepsilon_2} t_2$ is the effective

thickness oxide, t_1 is the thickness of the SiO_2 ($\epsilon_{ox} = \epsilon_1$) layer and t_2 is the thickness of the high-k layer (ϵ_2) V_0 is a weak function of silicon thickness and is close to the threshold voltage of DG MOSFETs. V_0 is given by: $V_0 = \frac{2kT}{q} \ln \left[\frac{2}{t_{si}} \sqrt{\frac{2\epsilon_{si}kT}{q^2 n_i}} \right]$

In the subthreshold region, the drain current can be given as,

$$I_{ds} = \mu_{eff} \frac{W}{L} kT n_i t_{si} e^{q(V_g)/kT} \left(1 - e^{-qV_{ds}/kT} \right) \quad (2)$$

where t_{si} is the silicon thickness

Note that the subthreshold current is proportional to the silicon thickness t_{si} , but independent of t_{oxeff} . In contrast, the current in saturation region, (1), is proportional to C_{ox} , but independent of silicon film thickness t_{si} . So, our optimized problem can be considered as deep non linear coupled objective functions.

According to the drain current expression in the saturation region (1) and the subthreshold region (2), the transconductance and the OFF-state current can be calculated by:

$$g_m(x) = \left. \frac{\partial I_{ds}}{\partial V_g} \right|_{V_{ds}} = \mu_{eff} C_{oxeff} \frac{W}{L} \left[2 \times (V_g - V_0) - \frac{8rkT}{q} e^{q(V_g - V_0 - V_{ds}/kT)} \right] \quad (3)$$

$$I_{OFF}(x) = I_{sub}|_{V_{gs}=0} = \mu_{eff} \frac{W}{L} kT n_i t_{si} \left(1 - e^{-qV_{ds}/kT} \right) \quad (4)$$

3 Results and Discussion

In what follows, first, we will consider the problem of electrical and geometrical synthesis to improve the electrical behavior of the transistor for both regimes of working, which are the subthreshold and saturation regimes. The obtained design can provide the best subthreshold and saturation parameters by satisfying of the following objective functions:

- Maximization of the transconductance function:
- Minimization of the OFF-current state:

where x represents the input variables vector which is given as, $x = (t_{si}, t_1, t_2, \epsilon_2, L, V_g, V_{ds})$.

Finally, the global optimization approach is adopted. This approach is a multi-objective function optimization where several objective functions are considered simultaneously.

Genetic algorithms have been shown to solve non-linear problems by exploring all regions of state space and exponentially exploiting promising areas through selection, crossover and mutation applied to individuals in populations [10].

The overall objective function is obtained by given weightage based on ‘Weighted sum approach method’ as follows:

$$F(x) = w_1 \cdot g_m(x) + w_2 \cdot I_{OFF}(x) \quad (5)$$

where w_1 and w_2 are weight functions. If high derived current, ONN-current, and low power dissipation transistor is the device produced by this process, then both transconductance and OFF-state current are equally important. Hence, w_1 and w_2 can be assigned equal values as 0.5. Lower value of OFF-state current is needed to design a transistor with low power dissipation and high transconductance values are needed to improve the transistor behavior in saturation regime. It is to note that the optimization of the transconductance and OFF-state current leads to increase of the ON-state current, and therefore an increasing of the ration I_{ON}/I_{OFF} can be obtained by our approach.

Given the clearly defined problem to be solved and a bit-string representation for candidate solutions, the adopted MOGA works as follows [10],

Start with a randomly generated population of ‘n’ chromosomes (candidate solutions to the problems):

- Calculate the overall objective function of each of the chromosome ‘x’ in the population.
- Create ‘n’ offspring from current population using the three MOGA operators namely selection, crossover and mutation.
- Replace the current population with the updated one.
- Repeat the above steps until the termination criteria are reached.

It is to note that at the end of each of evolutionary period, the non-dominated individuals (paretian solutions) are selected from dynamic population and added to elitist one (paretian population). The elitist population is then filtered to yield a non-dominated population. For our MOGA-based approach, the optimum solutions form what is called a ‘Pareto Front’. These solutions correspond to the non-dominated individuals which present the best solution of the objective functions simultaneously.

The MOGA parameters were varied and the associated optimization error was recorded. For this configuration, the fitness function, global objective function, was 28.877 and almost 99% of the submitted cases were learnt correctly. This resulted in 10,000 parameter set evaluations, and took about 20 s to complete using Windows XP with Pentium IV (1.5 GHz).

Figure 3 shows the variation of global overall fitness function, optimized transconductance and optimized OFF-state current with generation for a maximum iteration of 500.

The steady decrease in both transconductance and optimized OFF-state current of the best solution in each generation until it reaches a best possible value can be attributed to the selection procedure used namely tournament wheel selection. Final optimized (GSDG) MOSFET parameters are summarized in Table 1.

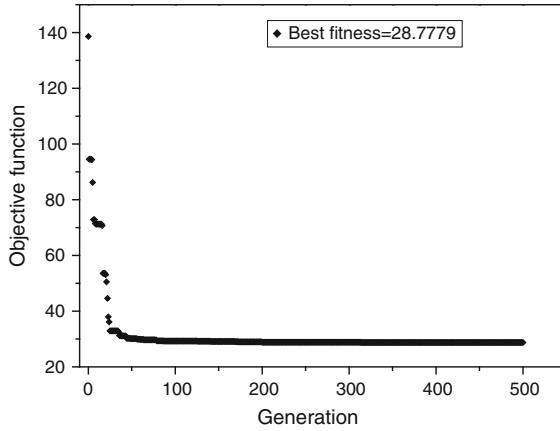


Fig. 3 Variation of normalized overall objective function

Table 1 Optimized GSDG MOSFET design parameters

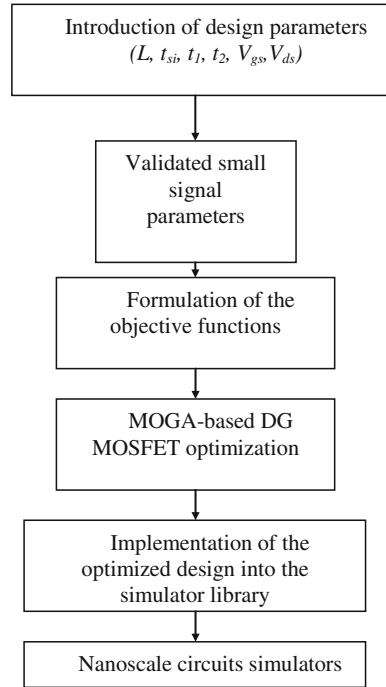
Optimized parameters			
Symbol	Quantity	Optimized design	Design without optimization
V_{ds}	Drain source	4.8512 V	3 V
V_g	Voltage	4.9999 V	2 V
	Gate voltage		
t_{si}	Silicon thickness	49.999 nm	40 nm
t_1	thickness of the SiO ₂	0.5046 nm	0.6 nm
t_2	thickness of the high-k layer	0.5010 nm	0.4 nm
L	Channel length	144.7187 nm	150 nm
ϵ_2	Permittivity of the high-k layer	39.9999	30
I_{OFF}	OFF-state current	9.0044×10^{-14} A/ μm	8.1264×10^{-12} A/ μm
gm	transconductance	1.7374×10^{-2} S/ μm	7.8873×10^{-3} S/ μm

3.1 Impact on Nanoscale Circuit Design

The numerical model of the current-voltage characteristics of a nanoscale ultra-thin body GSDG MOSFET developed, using the numerical simulator SILVACO [16], is explored to validate our proposed optimized device configuration. The developed MOGA-based approach can be used as interface between device compact modeling and circuit simulators, like SPICE, Cadence, and Anacad’s Eldo, in order to optimize the electrical circuit performances. A simplified overview is shown in Fig. 4.

In order to show the impact of our design methodology on the nanoscale circuit design, we propose to study the electrical behavior of a ring oscillator circuit, with and without transistor optimization. Ring oscillators are commonly studied due to their use in communications circuits such as PLLs and clock and data recovery circuits (CDRs).

Fig. 4 Our proposed approach for nanoscale circuit optimization



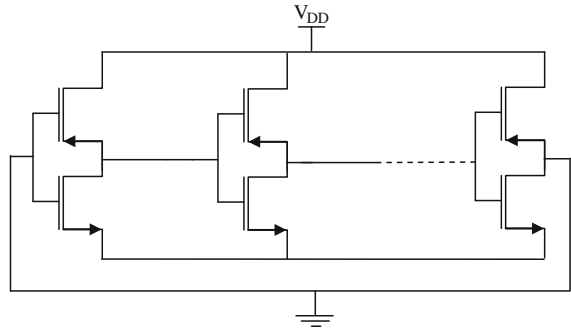
The ring oscillator consisted of an even number of inverters as shown in Fig. 5. It is comprised of a number of delay stages, with the output of the last stage fed back to the input of the first. The ring must provide a phase shift of 2π and have unity voltage gain at the oscillation frequency in order to obtain oscillation. In which, each delay stage must provide a phase shift of π/N , where N is the number of delay stages. The remaining π phase shift is provided by a DC inversion [17].

The most basic ring oscillator is simply a chain of single ended digital inverters, with the output of the last stage fed back to the input of the first stage. This circuit is shown in Fig. 5. Note that to provide the DC inversion, an important number of stages must be used.

There are numerous existing equations available to predict the oscillation frequency. The most common way to determine the frequency of oscillation of the ring is to assume each stage provides a delay of t_d [18]. The signal must go through each of the N delay stages once to provide the first π phase shift in a time of $N \times t_d$. Then, the signal must go through each stage a second time to obtain the remaining π phase shift, resulting in a total period of $2N \times t_d$. Therefore, the frequency of oscillation f is:

$$f = \frac{1}{\eta \times N \times t_d} \quad (6)$$

Fig. 5 GSDG MOSFET ring oscillator using single ended inverters



The oscillating frequency (f) mainly depends on the current flowing through the transistors of the inverter stages which in turn depends on the aspect ratio of the devices [19]. So, the oscillation frequency can be given by

$$f = \frac{I_{ds}}{\eta \times N \times C_{tot} \times V_{DD}} = \frac{g_m}{\eta \times N \times C_{tot}} \tag{7}$$

where I_{ds} is the current flowing through the transistors in the inverter stages, N is the number of stages, C_{tot} is the total capacitance, V_{DD} is the supply voltage and η represents a characteristic constant which is near about 0.75 [20].

The total capacitance is given by:

$$C_{tot} = C_{ox} \times L \times W \tag{8}$$

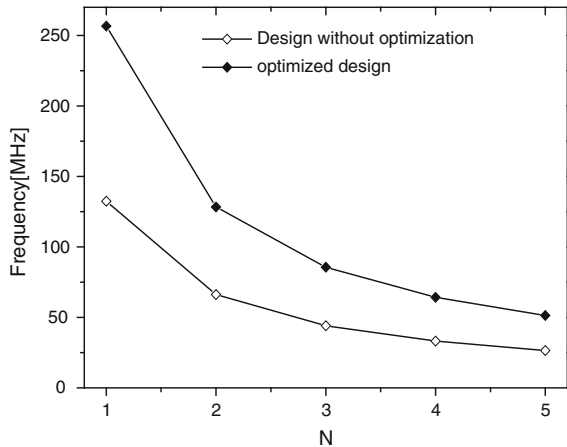
where C_{ox} is the oxide capacitance per unit area, L and W are the channel length and the width of the transistors, respectively.

The supply voltage (drain voltage) is set to give the optimized biasing conditions of the transistors. The frequency values of the ring oscillator are obtained using our GSDG MOSFET with and without device optimization. The frequency in $N = 1$, for optimized design, is about 256 MHz. However, the frequency degraded to about 132 MHz for device without optimization.

Figure 6 shows the calculated curves of the investigated GSDG MOSFET-based ring oscillator for different transistor configurations. It is clear shown that our optimized configuration provides better electrical and commutation circuit behaviors. Consequently they improve the performance (frequency) of the nano-CMOS ring oscillator. In practice, a high frequency can provide a high transition speed of nano-CMOS oscillator and better performance of analog operations.

From Fig. 6, it is observed that an improvement of the circuit performances, the ring frequency, can be obtained by introducing our MOGAs-based approach in the nanoscale circuits design. Hence, the proposed design can provide a new insight to improve the device electrical performance for the future nanoscale CMOS-based circuits.

Fig. 6 Calculated results for frequency versus N for different device configuration



4 Conclusion

In this work, MOGA-based approach is proposed to improve the subthreshold performances and saturation behavior of (GSDG) MOSFET for deep submicron CMOS digital applications. A global optimization problem was then formulated where all the parameters of the (GSDG) MOSFET were considered simultaneously, and the problem is presented as a multi-objective optimization one where the geometry and the electrical parameters were considered simultaneously. Such multi-objective optimization led to discovering of important relationships and useful optimal design principals in electrical behavior optimization of deep submicron devices both in the space of objective functions and device parameters. The proposed approach has successfully calculated the best possible transistor performance and the input design parameters that can yield those specific performances. The optimized design has been implemented into circuit simulator in order to study the electrical behavior of a ring oscillator, where a high improvement of the oscillation frequency has been recorded. Therefore, the proposed design can be considered as an alternative device for future CMOS-based circuits. Moreover, it can be concluded that the proposed MOGA-based approach is efficient and gives the promising results for circuits design and optimization problems.

References

1. Östling M, Malm BG, von Hartman M, Hallstedt J, Zhang Z, Hellstrom P, Zhang S (2007) Challenges for 10 nm MOSFET process integration. *J Telecommun Inf Technol* 2:25–32
2. The International Technology Roadmap for Semiconductors, 2007. <http://public.itrs.net>
3. Ghoggali Z, Djeflal F, Lakhdar N (2010) Analytical analysis of nanoscale double-gate MOSFETs including the hot-carrier degradation effects. *Int J Electron* 97(2):119–127

4. ITRS 2007. International Technology Roadmap for semiconductors. <<http://www.itrs.com>>
5. Djeflal F, Ghogkali Z, Dibi Z, Lakhdar N (2009) Analytical analysis of nanoscale multiple gate MOSFETs including effects of hot-carrier induced interface charges. *Microelectron Reliab* 49:377–381
6. Bendib T, Djeflal F (2011) Electrical performance optimization of nanoscale double-gate MOSFETs using multiobjective genetic algorithms. *IEEE Trans Electron Dev* 58:3743–3750
7. Reddy GV, Kumar MJ (2005) A new dual-material double-gate (DMDG) nanoscale SOI MOSFET-two dimensional analytical modeling and simulation. *IEEE Trans Nanotechnol* 4:260–268
8. Djeflal F, Abdi MA, Dibi Z, Chahdi M, Benhaya A (2008) A neural approach to study the scaling capability of the undoped double-gate and cylindrical gate all around MOSFETs. *Mater Sci Eng B* 147:239–244
9. Marseguerra M, Zio E, Cipollone M (2003) Designing optimal degradation tests via multi-objective genetic algorithms. *Reliab Eng Syst Saf* 79:87–94
10. Affi Z, EL-Kribi B, Romdhane L (2007) Advanced mechatronic design using a multi-objective genetic algorithm optimization of a motor-driven four-bar system. *Mechatronics* 17:489–500
11. Chang P-C, Hsieh J-C, Wang C-Y (2007) Adaptive multi-objective genetic algorithms for scheduling to drilling operation of printed circuit board industry. *Appl Soft Comput* 7:800–806
12. Murata T, Ishibuchi H, Tanaka H (1996) Multi-objective genetic algorithm and its application to flowshop scheduling. *Comput Ind Eng* 30:957–968
13. Djeflal F, Bendib T (2011) Multi-objective genetic algorithms based approach to optimize the electrical performances of the gatestackdoublegate (GSDG) MOSFET. *Microelectron J* 42:661–666
14. Bendib T, Djeflal F, Bentrchia T, Arar D, Lakhdar N (2012) Multi-objective genetic algorithms based approach to optimize the small signal parameters of gate stack double gate MOSFET. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, 4–6 July, London, UK*, pp 968–970
15. Taur Y, Liang X, Wang W, Lu H (2004) A continuous, analytic drain-current model for DG MOSFETs. *IEEE Electron Device Lett* 25:107–109
16. Atlas User's manual, SILVACO TCAD (2008)
17. Razavi B (1997) A 2-GHz 1.6-mW phase-locked loop. *IEEE J Solid-State Circuits* 32:730–735
18. Sun L, Kwasniewski TA (2001) A 1.25-GHz 0.35- μ m monolithic CMOS PLL based on a multiphase ring oscillator. *IEEE J Solid-State Circuits* 36:910–916
19. Baker RJ, Li HW, Boyce DE (2000) CMOS circuit design, layout and simulation. IEEE Press, New York
20. Lee TH (1998) The design of CMOS radiofrequency integrated circuits. Cambridge University Press, Cambridge

Performance Analysis of Series Hybrid Active Power Filter

M. A. Mulla, R. Chudamani and A. Chowdhury

Abstract In the area of active power filtering, the Series Hybrid Active Power Filter (SHAPF) has been taken into account increasingly. Existing methods used for controlling SHAPF are either based on detecting source current harmonics or load voltage harmonics. Generalised Instantaneous Power Theory (GIPT) gives simple and direct method of defining power quantities under sinusoidal and non-sinusoidal situations. In this paper the definition of GIPT is used to decompose voltage vector into different components, which represents different parts of the power quantity. The separated components of voltage vector are used to derive reference signal for the SHAPF. This paper presents a study on performance analysis of SHAPF where the method used for calculating reference is based on the GIPT. Steady state and transient performance of SHAPF used for compensating current type harmonic producing load and voltage type harmonics producing load are evaluated by the simulation study.

Keywords Active power filters · Geometric algebra · Harmonic compensation · Hybrid active power filter · Instantaneous power · Non-sinusoidal waveforms · Passive power filters · Power multi-vector · Power quality · Real-time control

M. A. Mulla (✉) · R. Chudamani · A. Chowdhury
Department of Electrical Engineering, S. V. National Institute of Technology,
Ichchhanath, Surat, Gujarat 395007, India
e-mail: mam@eed.svnit.ac.in

R. Chudamani
e-mail: rc@eed.svnit.ac.in

A. Chowdhury
e-mail: ac@eed.svnit.ac.in

1 Introduction

Load compensation in power engineering is the procedure used to obtain the supply currents that are sinusoidal and balanced. Active power line conditioners make it possible to obtain power-electronic solutions to power quality problems. In particular, balanced or unbalanced load compensation in non-sinusoidal supply situations is possible [1]. Active Power Filter (APF) has become the main research direction of load compensation as its filtering characteristic is not affected by system parameters. Various APF configurations and control strategies have been researched during the last decades. So far, a large number of shunt APF have been installed, but there are still some shortcomings like large capacity, high initial investment, harmonic circulation while working with the shunt Passive Power Filter (PPF) together, improving filtering characteristic when the PPF makes the system resonance etc. In order to reduce inverter capacity, the hybrid APF is becoming very popular in recent development. A variety of configurations and control strategies are proposed to reduce inverter capacity [2–4]. Series filters are normally preferred for eliminating voltage type harmonic but SHAPF configuration is capable of eliminating current type harmonics also.

To obtain efficient SHAPF performance, it is important to choose proper reference generation algorithm and an appropriate current or voltage control strategy. The publication of the instantaneous reactive power theory caused a great impact in reference generation. Many approaches have been published since then [5–12]. But all of these definitions are computational intensive and do not provides simple expression of instantaneous power quantity. In year 2004, Dai [13] introduced generalised instantaneous power theory, which gives a direct and simple expression for instantaneous power quantities. Chapter [14] proposed a simple method of decomposing voltage using a GIPT for series active power filter.

This chapter presents a performance analysis of reduced rating SHAPF using a control algorithm based on GIPT. It is proposed to decompose multiphase voltage vector into quantities that represent different components of power. Normally the instantaneous power has average component and oscillating component. Using vector algebra it is possible to obtain the voltage vectors corresponding to this average and oscillating powers. The separated voltage vector corresponding to average active power, oscillating active power, average inactive power and oscillating inactive power are used to calculate reference voltage for SHAPF. Steady state and transient performance of SHAPF is tested by simulation study, while it is used for eliminating current type harmonics as well as voltage type harmonics.

This chapter is organized as follows. First, a generalized definition of instantaneous active, inactive and apparent power is presented. Then, the decomposition of voltage vector which represents different components of power quantities is defined. The use of this decomposition for deriving the reference for SHAPF is discussed further. Finally, the simulation studies performed for evaluating performance analysis of SHAPF along with simulation results are discussed.

2 Definition of Generalized Instantaneous Power Theory

For a three-phase four-wire system, the instantaneous quantities of load voltage and currents are expressed as $\vec{v} = [v_a, v_b, v_c]^T$ and $\vec{i} = [i_a, i_b, i_c]^T$. The instantaneous power multi-vector which is defined as the geometric product of voltage and current vectors can be expressed as,

$$\begin{aligned}\vec{s}(t) &= v(t)i(t) = v(t) \bullet i(t) + v(t) \times i(t) \\ \vec{s}(t) &= p(t) + \vec{q}(t)\end{aligned}\quad (1)$$

Load instantaneous apparent power 's' is defined as $s = \|\vec{v}\| \|\vec{i}\|$, where $\|\vec{v}\| = \sqrt{v_a^2 + v_b^2 + v_c^2}$ and $\|\vec{i}\| = \sqrt{i_a^2 + i_b^2 + i_c^2}$.

Load instantaneous active power 'p' is defined as the inner product of voltage and current vectors.

$$p(t) = \vec{v}(t) \bullet \vec{i}(t) = v^T i = v_a i_a + v_b i_b + v_c i_c \quad (2)$$

Load instantaneous reactive power $\vec{q}(t)$ is defined as the outer product of voltage and current vector $\vec{q}(t) = \vec{v}(t) \times \vec{i}(t)$. The outer product is defined by means of the tensor product as $\vec{v}(t) \times \vec{i}(t) = \vec{i}(t) \otimes \vec{v}(t) - \vec{v}(t) \otimes \vec{i}(t)$.

$$\vec{q}(t) = \vec{v}(t) \times \vec{i}(t) = \begin{bmatrix} 0 & -q_{ab} & q_{ca} \\ q_{ab} & 0 & -q_{bc} \\ -q_{ca} & q_{bc} & 0 \end{bmatrix} \quad (3)$$

with each components defined as,

$$q_{ab} = v_a i_b - v_b i_a; \quad q_{bc} = v_b i_c - v_c i_b; \quad q_{ca} = v_c i_a - v_a i_c$$

$\vec{q}(t)$ is denoted as instantaneous inactive tensor and its norm is defined as instantaneous inactive power

$$\|\vec{q}\| = \sqrt{q_{ab}^2 + q_{bc}^2 + q_{ca}^2} \quad (4)$$

3 Decomposition of Voltage Vector

Using (1) representing voltage with two components, one in phase with current and another quadrature to the current, the apparent power $\vec{s}(t)$ can be written as $\vec{s}(t) = [\vec{v}_p(t) + \vec{v}_q(t)] \bullet \vec{i}(t) + [\vec{v}_p(t) + \vec{v}_q(t)] \times \vec{i}(t)$. After simplification we obtain

$$\vec{s}(t) = \vec{v}_p(t) \bullet \vec{i}(t) + \vec{v}_q(t) \times \vec{i}(t) = p(t) + \vec{q}(t) \quad (5)$$

Using the first part of (5), the component of load instantaneous voltage vector ‘ \vec{v}_p ’, corresponding to active power can be expressed as $\vec{v}_p(t) = \vec{i}^{-1} p(t) = \frac{\vec{i}}{\|\vec{i}\|^2} p(t)$.

$$\vec{v}_p = [v_{pa}, v_{pb}, v_{pc}]^T = \frac{\vec{i}}{\|\vec{i}\|^2} p(t) = \frac{p(t)}{\|\vec{i}\|^2} [i_a, i_b, i_c]^T \tag{6}$$

‘ \vec{v}_p ’ is denoted as instantaneous active voltage tensor and its norm is defined as instantaneous active voltage $\|\vec{v}_p\| = \sqrt{v_{pa}^2 + v_{pb}^2 + v_{pc}^2}$.

Using the second part of (5), the component of load instantaneous voltage vector ‘ \vec{v}_q ’, that represents inactive power, is expressed as $\vec{q}(t) = \vec{v}_q(t) \times \vec{i}(t)$. Multiplying both sides by current vector $\vec{i}(t)$ and using the simplification given in [14], following expression is obtained.

$$\vec{v}_q(t) = [v_{qa}, v_{qb}, v_{qc}] = \frac{\vec{i}(t) \times \vec{q}(t)}{\|\vec{i}\|^2} \tag{7}$$

Equation (7) expresses the load instantaneous voltage vector ‘ \vec{v}_q ’ which represents reactive power quantity. Using the procedure given in [14], the cross product is calculated as

$$\vec{v}_q(t) = \frac{1}{\|\vec{i}\|^2} \begin{bmatrix} 0 & q_{ab} & -q_{ca} \\ -q_{ab} & 0 & q_{bc} \\ q_{ca} & -q_{bc} & 0 \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{8}$$

‘ \vec{v}_q ’ is denoted as instantaneous inactive voltage tensor and its norm is defined as instantaneous inactive voltage $\|\vec{v}_q\| = \sqrt{v_{qa}^2 + v_{qb}^2 + v_{qc}^2}$.

It is clear that the quantities defined by ‘ \vec{v}_p ’ and ‘ \vec{v}_q ’ (6) and (8) are representing components of instantaneous voltages that corresponds to active power and inactive power drawn by the load. These components are directly associated with three phase instantaneous voltages and currents. It is also observed that these voltage components are separated without any form of artificial coordinate transformations.

4 Reference Generation for SHAPF

In series compensation, the compensator works as a variable voltage source and compensate the unwanted component of load voltage. Figure 1 show the SHAPF equivalent circuit, where load is represented as summation of different voltages corresponding to power components. The equivalence circuit of series active power filter is represented by variable voltage source ‘ v_c ’ connected in series with the load.

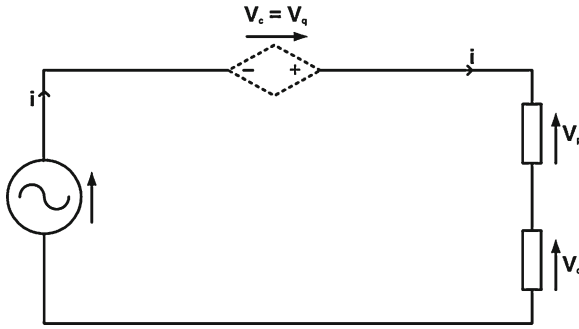


Fig. 1 Equivalent circuit of SHAPF system

The total instantaneous active power calculated using (2) is divided into average and oscillating active power $p(t) = \bar{p} + \tilde{p}$. The voltages corresponding to these components of active powers are expressed as,

$$\vec{v}_p = \frac{\vec{i}}{\|\vec{i}\|^2} [\bar{p} + \tilde{p}] = \vec{v}_{\bar{p}} + \vec{v}_{\tilde{p}} \quad (9)$$

where,

$\vec{v}_{\bar{p}}$ = Voltage corresponding to average active power.

$\vec{v}_{\tilde{p}}$ = Voltage corresponding to oscillating active power.

Similarly the total inactive power calculated using (3) is also divided into average and oscillating inactive power $\vec{q}(t) = \bar{q} + \tilde{q}$. The voltages corresponding to these two components of inactive powers are expressed as,

$$\vec{v}_q(t) = \vec{v}_{\bar{q}}(t) + \vec{v}_{\tilde{q}}(t) \quad (10)$$

where,

$\vec{v}_{\bar{q}}(t)$ = Voltage corresponding to average inactive power.

$\vec{v}_{\tilde{q}}(t)$ = Voltage corresponding to oscillating inactive power.

These components of voltage vectors, given by (9) and (10) are useful for generating reference voltage for SHAPF. For calculations of reference voltages for harmonic elimination, the required components of active and inactive power need to be compensated are $p_c(t) = \tilde{p}(t)$, $q_c(t) = \tilde{q}(t)$. The SHAPF reference is calculated as,

$$\vec{v}_c(t) = \vec{v}_{p_c}(t) + \vec{v}_{q_c}(t) = \vec{v}_{\tilde{p}}(t) + \vec{v}_{\tilde{q}}(t) \quad (11)$$

The voltage vector $\vec{v}_c(t)$ is a reference for injecting voltage in series in order to eliminate oscillating active and inactive component of power from the system.

5 ‘SHAPF’ Simulation Model

For evaluating performances of SHAPF, using the voltage reference calculated with GIPT, simulation study is performed in PSIM. Figure 2a shows the arrangement of power circuit configuration which is made up of non-linear load, Passive Power Filters (PPF) bank, Series Transformer and Inverter. The PPF bank is made up of 5th harmonic tuned filter, 7th harmonic tuned filter and a high-pass filter. Pulse width modulated voltage source inverter working at 100V dc link voltage along with 1:1 series transformer is used for series voltage injection.

Figure 2b shows the block diagram of control circuit used for reference voltage generation. Three phase currents, voltages are sensed from power circuit and reference voltage is calculated using Eqs. (6), (8) and (11). Table 1 shows the system parameter values with which the simulation study is done in PSIM software.

The performance of this simulation model is tested with two different types of harmonic producing loads: (1) current harmonic producing load and (2) voltage harmonic producing load. Diode rectifier with R load on dc side is modeled as current harmonic producing non-linear load while diode rectifier with R-C load on dc side is modeled as voltage harmonic producing non-linear load. For steady state performance analysis, the simulation model is run with one of this fixed load

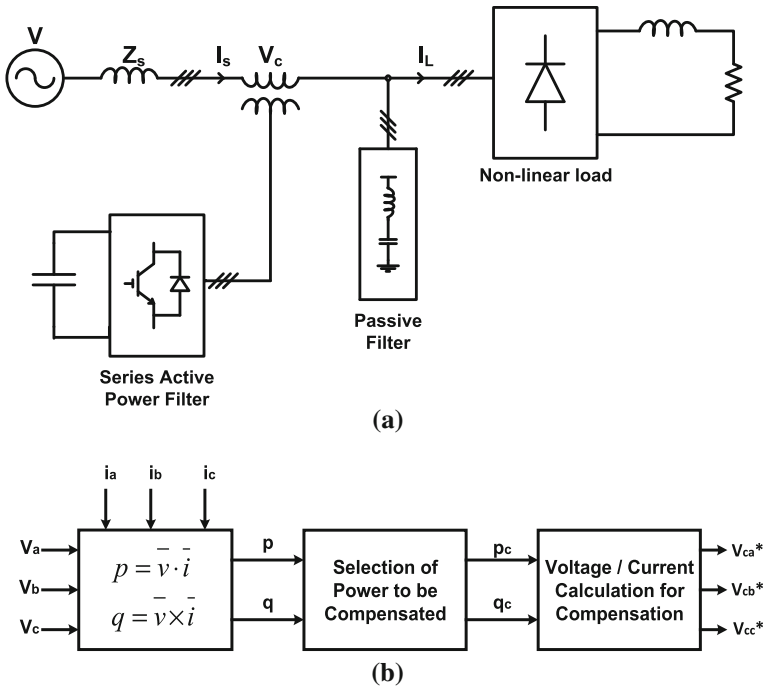


Fig. 2 SHAPF: a power circuit configuration and b control circuit

Table 1 Common system parameter of SHAPF simulation model

Sr. No.	Quantity	Value
1	Supply voltage	440 V, 50 Hz (line-line)
2	Source impedance	$R_s = 0.5 \text{ Ohm}$, $L_s = 0.1 \text{ mH}$
3	DC capacitor	1000 μF
4	DC link voltage	100 V
5	EMI filter	$L_f = 1.35 \text{ mH}$, $C_f = 50 \text{ }\mu\text{F}$
6	Tuned PPF (5th)	$L_5 = 12.32 \text{ mH}$, $C_5 = 32.88 \text{ }\mu\text{F}$
7	Tuned PPF (7th)	$L_7 = 6.29 \text{ mH}$, $C_7 = 32.88 \text{ }\mu\text{F}$
8	High pass PPF	$L = 2.36 \text{ mH}$, $C = 29.88 \text{ }\mu\text{F}$, $R = 17.75$
9	Series transformer	1:1
10	Load	Diode rectifier with $R = 10 \text{ Ohm}$ Diode rectifier with $R = 10 \text{ Ohm}$, $C = 2000 \text{ }\mu\text{F}$
11	Switching frequency	20 kHz

connected in the system and the performance is evaluated. For transient performance analysis another same value load is connected in the system after few cycles of operation of SHAPF and the performance of the system is evaluated.

6 Performance with Current Harmonic Producing Load

Before the SHAPF model is tested for compensating current harmonics produced by diode rectifier connected with the R load, the total harmonic distortion of load current is measured 25.35 % in phase A. The performance of the system is evaluated by observing source current without any filter, with PPF and with SHAPF connected in the system and are compared.

Figure 3a, b shows the waveform of load current and improved supply current respectively, when compensation is done by PPF bank. Since the PPF bank is made up of 5th harmonic tuned filter, 7th harmonic tuned filter and high pass filter, it is observed that 5th and 7th harmonics are removed and other harmonics are attenuated. The THD of supply current in phase A which is 25.35 % before compensation is reduced to 12.33 % after applying PPF and individual harmonic components are also reduced.

Figure 4a, b shows the waveforms of load current and supply current respectively, when compensation is done with SHAPF. It is observed that all three phases are drawing sinusoidal current from the source after compensation. The THD of supply current in phase A, which is 25.35 % before compensation is reduced to 3.01 % after applying SHAPF.

Table 2 shows the overall performance of the system. The THD of supply current in phase A is 25.35 % without compensation, which reduces to 12.33 % after applying PPF and which further reduces to 3.01 % after applying SHAPF.

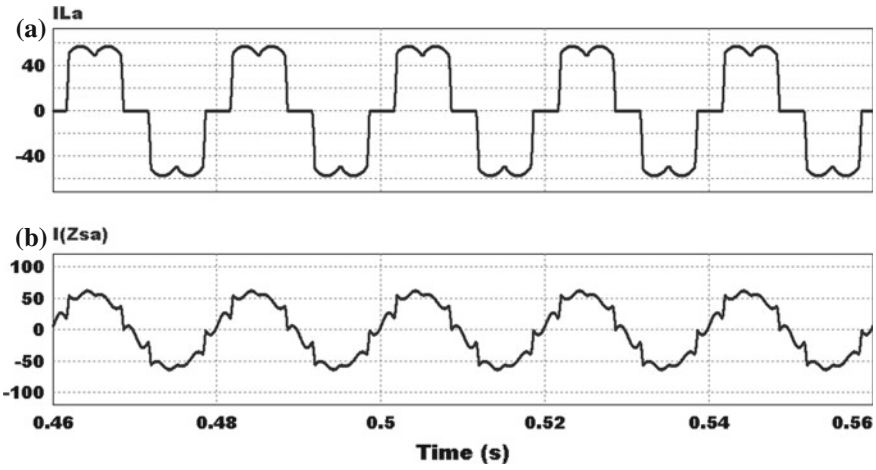


Fig. 3 Steady state performance while compensating current type harmonics with PPF: **a** load current of phase A, **b** source current of phase A

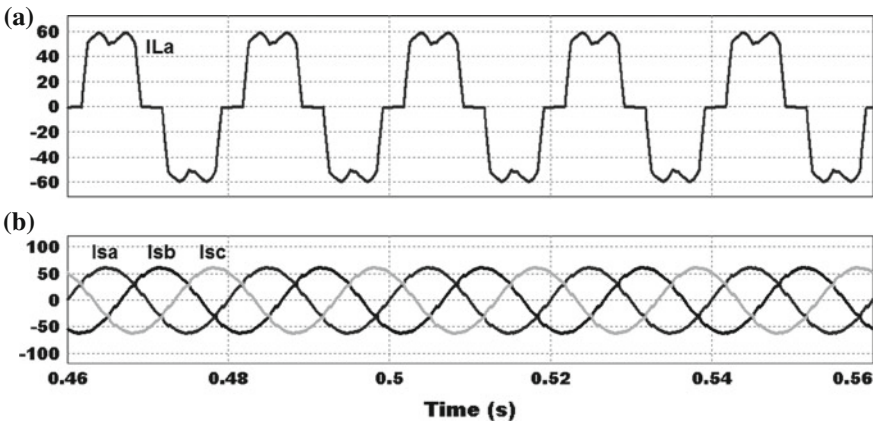


Fig. 4 Steady state performance while compensating current type harmonics with SHAPF: **a** load current of phase A, **b** source current of phases A, B, C

For evaluating transient performance of the SHAPF another diode rectifier with R load is connected in the system at time $t = 500$ ms. Figure 5a, b shows the waveforms of load current and supply current respectively, during this transient period. It is observed and measured from this waveform that the system is settled to new changed conditions in 0.31 ms which is almost instantaneous.

Table 2 Steady state performance while compensating current type harmonics

Sr. No.	Parameter	% THDi		
		Phase A	Phase B	Phase C
1	Load current	25.35	25.33	25.27
2	Source current without filter	25.35	25.33	25.27
3	Source current with PPF	12.33	12.33	12.33
4	Source current with SHAPF	3.01	3.01	3.02

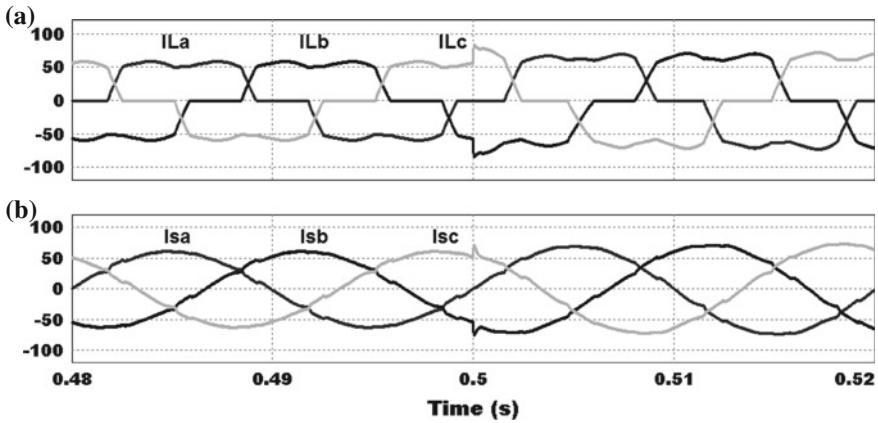


Fig. 5 Transient performance while compensating current type harmonics with SHAPF: **a** load current of phase A, B, C, **b** source current of phases A, B, C

7 Performance with Voltage Harmonic Producing Load

Before the SHAPF model is tested for compensating current harmonics produced by diode rectifier connected with the R-C load, the total harmonic distortion of load current is measured 78.42 % in phase A. The performance of the system is evaluated by observing source current without any filter, with PPF and with SHAPF connected in the system and are compared.

Figure 6a, b shows the waveform of load current and improved supply current respectively, when compensation is done by PPF bank. Since the PPF bank is made up of 5th harmonic tuned filter, 7th harmonic tuned filter and high pass filter, it is observed that 5th and 7th harmonics are removed and other harmonics are attenuated. The THD of supply current in phase A which is 78.42 % before compensation is reduced to 19.29 % after applying PPF and individual harmonic components are also reduced.

Figure 7a, b shows the waveforms of load current and supply current respectively, when compensation is done with SHAPF. It is observed that all three phases are drawing sinusoidal current from the source after compensation. The THD of supply current in phase A, which is 78.42 % before compensation is reduced to 3.21 % after applying SHAPF.

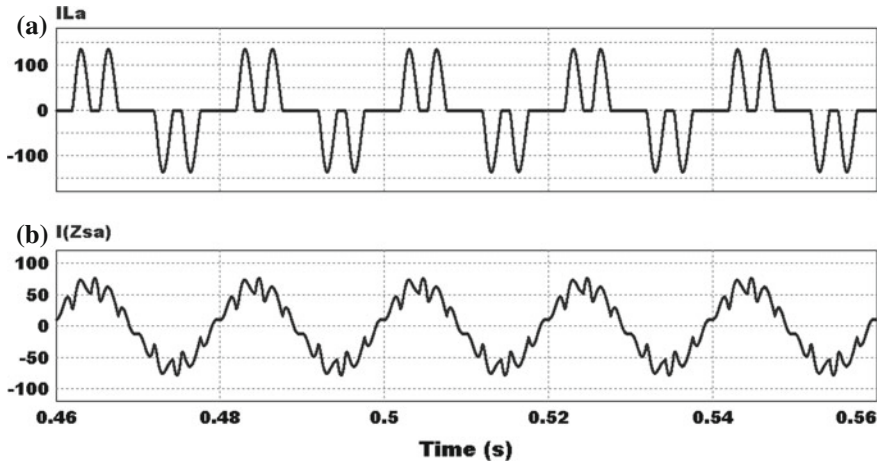


Fig. 6 Steady state performance while compensating voltage type harmonics with PPF: **a** load current of phase A, **b** source current of phase A

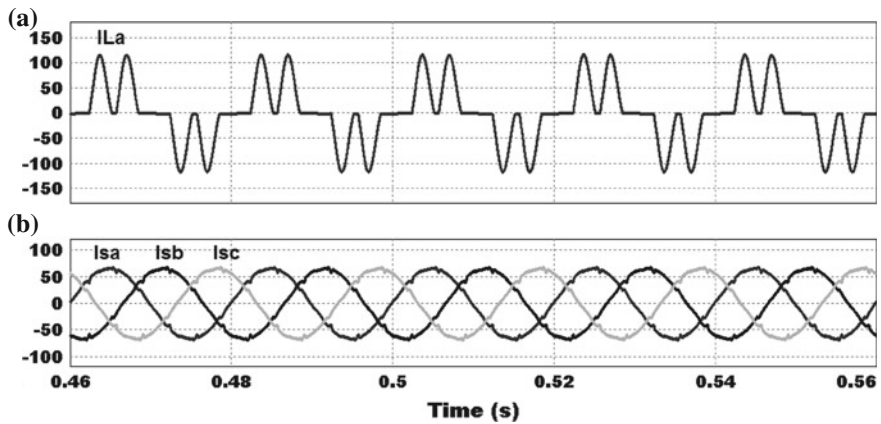


Fig. 7 Steady state performance while compensating voltage type harmonics with SHAPF: **a** load current of phase A, **b** source current of phases A, B, C

Table 3 shows the overall performance of the system. The THD of supply current in phase A is 78.42 % without compensation, which reduces to 19.29 % after applying PPF and which further reduces to 3.21 % after applying SHAPF.

For evaluating transient performance of the SHAPF another diode rectifier with R-C load is connected in the system at time $t = 500$ ms. Figure 8a, b shows the waveforms of load current and supply current respectively, during this transient period. It is observed and measured from this waveforms that the system is settled to new changed conditions in 1.94 ms which is less than one tenth of power cycle and almost instantaneous.

Table 3 Steady state performance while compensating voltage type harmonics

Sr. No.	Parameter	% THDi		
		Phase A	Phase B	Phase C
1	Load current	78.42	78.34	78.36
2	Source current without filter	78.42	78.34	78.36
3	Source current with PPF	19.29	19.28	19.28
4	Source current with SHAPF	3.21	3.17	3.21

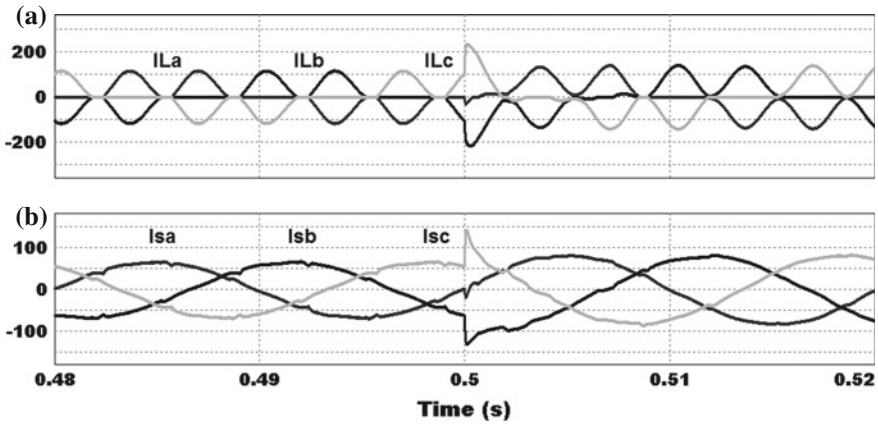


Fig. 8 Transient performance while compensating voltage type harmonics with SHAPF: **a** load current of phase A, B, C, **b** source current of phases A, B, C

These simulation results prove the filtering capabilities of SHAPF and proposed GIPT based voltage decomposition control scheme in compensating voltage type harmonic generating loads.

8 Conclusions

In this chapter, steady state and transient performance analysis of SHAPF, designed using control scheme based on GIPT, for eliminating current as well as voltage type harmonics is presented. Mathematical formulations of decomposing different power quantities in terms of voltage components are discussed. Application of this decomposition in generating reference signal for SHAPF is demonstrated. Following conclusions are derived from the simulation study:

- (i) During steady state performance of SHAPF while compensating current type harmonics, THD of supply current is reduced to 3.01 % from 25.35 %.

- (ii) During transient performance of SHAPF while compensating current type harmonics, the system is settled to the new changed conditions in 0.31 ms, which is almost instantaneous.
- (iii) During steady state performance of SHAPF while compensating voltage type harmonics, THD of supply current is reduced to 3.21 % from 78.42 %.
- (iv) During transient performance of SHAPF while compensating voltage type harmonics, the system is settled to new changed conditions in 1.94 ms, which is less than one tenth of power cycle and is almost instantaneous.

These simulation results prove the filtering capabilities of SHAPF and effectiveness of proposed GIPT based voltage decomposition control scheme.

References

1. Montano JC (2011) Reviewing concepts of instantaneous and average compensations in polyphase systems. *IEEE Trans Ind Electron* 58(1):213–220
2. Salmeron P, Litran SP (2010) A control strategy for hybrid power filter to compensate four-wires three-phase systems. *IEEE Trans Power Electron* 25(7):1923–1931
3. Tian J, Chen Q, Xie B (2012) Series hybrid active power filter based on controllable harmonic impedance. *IET J Power Electron* 5(1):142–148
4. Akagi H, Watanabe EH, Aredes M (2007) *Instantaneous power theory and applications to power conditioning*. IEEE Press, New Jersey
5. Akagi H, Kanazawa Y, Nabae A (1984) Instantaneous reactive power components comprising switching devices without energy storage components. *IEEE Trans Ind Appl IA-20*:625–631
6. Takahashi I (1988) Analysis of instantaneous current and power using space switching functions. In: *Proceeding conference Rec. IEEE power electronics specialists conference, 1988*, pp 42–49
7. Furuhashi T, Okuma S, Uchikawa Y (1990) A study on the theory of instantaneous reactive power. *IEEE Trans Ind Electron* 37(1):86–90
8. Willems JL (1992) A new interpretation of the Akagi-Nabae power components for nonsinusoidal three-phase situation. *IEEE Trans Instrum Meas* 41:523–527
9. Nabae A, Tanaka T (1996) New definition of instantaneous active-reactive current and power based on instantaneous space vectors on polar coordinates in three-phase circuits. *IEEE Trans Power Deliv* 11:1238–1243
10. Kim H, Blaabjerg F, Bak-Jensen B, Choi J (2002) Instantaneous power compensation in three-phase systems by using p-q-r theory. *IEEE Trans Power Electron* 17(5):701–710
11. Menti A, Zacharias T, Miliadis-Argitis J (2007) Geometric algebra: a powerful tool for representing power under nonsinusoidal conditions. *IEEE Trans Circuits Syst I* 54(3):601–609
12. Herrera RS, Salmeron P, Vazquez JR, Litran SP, Perez A (2009) Generalised instantaneous reactive power theory in poly-phase power systems. In: *13th European conference on power electronics and applications, 2009, EPE '09*, pp 1–10
13. Dai X, Liu G, Gretschek R (2004) Generalized theory of instantaneous reactive quantity for multiphase power system. *IEEE Trans Power Deliv* 19(3):965–972
14. Mulla MA, Chudamani R, Chowdhury A (2012) Series active power filter using generalised instantaneous power theory. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, vol II, WCE (2012)*. London, UK, pp 1002–1006, 4–6 July 2012

Protecting the End User Device in 4G Heterogeneous Networks

Hani Alquhayz, Ali Al-Bayatti and Amelia Platt

Abstract In recent years, there have been major developments in, and deployment of, diverse mobile technology. Security issues in mobile computing are now presenting significant challenges. Heterogeneous networks are the convergence of wired and wireless networks, other diverse end user devices and other communication technologies which provide very high speed connections. Major security challenges in 4G heterogeneous networks are inherent in current internet security threats and IP security vulnerabilities. In this paper, we propose a management system which is responsible for enforcing security policies and ensuring that security policies continued to be followed. The objective of this security management system is to prevent the mobile equipment from being abused or used as a malicious attack tool. The proposed security management system is consistent with the security specifications defined by ITU-T recommendation M.3400 TMN management functions. Finally, this paper will present a policy-based architecture for the security management system of 4G heterogeneous networks focusing on detection and prevention of malicious attacks.

Keywords 4G Networks · 4G Security · Heterogeneous networks · Heterogeneous security · Mobile security · Security management system

H. Alquhayz (✉)

Bede Island Building, The Gateway, De Montfort University,
Leicester LE1 9BH, UK
e-mail: hani373@gmail.com

A. Al-Bayatti

Bede Island Building, The Gateway, Demontfort University,
Room BI 2.30, Leicester LE1 9BH, England
e-mail: alihmohd@dmu.ac.uk

A. Platt

Bede Island Building, The Gateway, De Montfort University,
Room BI 2.12, Leicester LE1 9BH, England
e-mail: amp@dmu.ac.uk

1 Introduction

4G is the next generation of mobile networks. The International Telecommunications Union (ITU) defined the International Mobile Telecommunications-Advanced (IMT-Advanced) standard as the global standard for 4G wireless communications. As specified by the International Telecommunication Union's Recommendation (ITU-R), 4G provides very high speed connections such as 100 Mbps for outdoor environments and 1 Gbps for indoor environments. Also, it is recommended that a 4G heterogeneous network should have high capacity, low cost, low latency, good quality of service, and good coverage [1]. There are many candidates such as LTE Advanced and Wireless MAN Advanced which are trying to achieve these requirements, especially high speed, while, other candidates are trying to build a 4G heterogeneous network as a convergence between wired and wireless networks. There are some new architectures such as, for example, Y-Comm for this heterogeneous network which is a new architecture comprising of a fast core network and a slower peripheral network. The core network contains wired technologies such as optical networks and peripheral networks consisting of wireless technologies such as 3G [2]. 4G security vulnerabilities have been addressed well [3] and [4] and include current IP internet threats and other threats due to the open architecture and the diversity of end user devices' security levels. Some security solutions such as Y-Comm and Hockey have been presented for 4G heterogeneous mobile networks. However, they do not take into account the security of end user devices, which causes many security vulnerabilities and they do not achieve the security requirements of 4G systems.

Y-Comm uses a multi-layer security model to provide a security solution. However, there are researches which show that, because 4G is an IP-based and heterogeneous network, there are several security threats which could cause service interruption and hijack the data. These researches indicated that the current security threats, and also new threats, were inherent in 4G technology [3, 4]. The security requirements of 4G heterogeneous networks have been defined on two levels: firstly, these are on mobile equipment; and, secondly, on operator networks. Mobile equipment requirements include protecting the device's integrity, privacy and confidentiality, controlling access to data, and preventing the mobile equipment being stolen or compromised and the data being abused or used as an attack tool [5]. Existing research on security of 4G heterogeneous networks focused on the security such as authentication and authorization mainly, on the interface between the network and the operator. However, the protection of the mobile device from attacks and becoming an attack tool solves important security issues in the heterogeneous network.

Therefore, there is a demand to build security management systems for 4G heterogeneous networks. We are building a security management system which detects if the mobile device has been attacked and prevents using it as an attack tool by removing the user's access and severing the connection. We followed the ITU-T's definition of security management as detailed in standardization M.3400, where Security management is the combined function of groups of sets; these are Prevention, Detection, Containment, Recovery, and Security Administration.

The security policies can detect and prevent attacks on end user devices. Security management policy-based systems have a variety of mechanisms depending on the type and scale of the network. There is an increasing challenge when this is a large scale network which combines two different topologies. We propose a policy-based architecture which is independent from the network and can adapt to the changes to the network. Our proposal will achieve a security management system which maintains the security requirements of 4G heterogeneous mobile networks. The remainder of this paper is structured as follows. Section II presents background information about 4G heterogeneous network's security issues, ITU-T recommendations, and policy-based systems. Section III explains the approach and architecture of our security management system. Section IV provides an example of a security management system working in the case of an attack. Finally, Section V concludes this paper.

2 Background

2.1 Security of 4G Heterogeneous Networks

The open nature of 4G means that the infrastructure is accessed from many external connection points through peer operators, through the internet and via third party technologies. All these elements are at risk from providing holes in security and vulnerabilities. Also, different service providers share the core network infrastructure which means that one single provider being compromised affects the whole network infrastructure [4]. 4G network security concerns have been addressed well by many research groups. Y-Comm & Hocky, have worked on designing security architectures such as Hokey and Y-Comm for 4G networks. Y-Comm uses a multi-layer security model to provide a security solution. This model is applied together on peripheral and core frameworks. The four security layers work together through both frameworks. Security services in Y-Comm include Authentication; Authorization; Auditing; and other services related to protecting the entity of the network [2].

However, Aiash et al' research study was about the security challenges in 4G systems. Their research tried to address the security challenges by looking at the possibility of applying current security techniques on 4G networks. Their research indicated that the current security threats and, also, new threats were inherent to 4G technology. Their study used standard X.805 to investigate the possibility of applying the 3G's Authentication and Key Agreement (AKA) to a 4G communication framework. By applying X.805, they analyzed the AKA protocol in 4G networks. The result was that they found many threats to the network's security [3].

Yongsuk Park and Taejoon Park conducted another research study. They showed that, because 4G was an IP-based and heterogeneous network, there were a number of security threats which could cause service interruption and could hijack the data. They addressed, also, several outstanding open issues which required solutions.

In a traditional network security procedure, the network is secured by preventing threats from accessing network entities. However, this is inefficient with an open architecture network such as 4G because the attackers try to find security vulnerabilities in the operating system and in the network protocols or applications. From these vulnerabilities, they can create malware which abuses the network.

According to the new architecture, there are possible threats within a 4G network system. These threats are: IP address spoofing, User ID theft, Theft of Service (ToS), Denial of Service (DoS), and intrusion attacks.

Due to the open architecture and IP based environment, 4G heterogeneous networks receive new security threats and inherit threats from the internet. These threats were unseen in 3G because the network infrastructure was owned by the service providers and access was denied to other network equipment. Also, the diversity in end user devices and security levels leads to greater security threats [1]. The experience of internet protection which says that the protection should involve not only data but, also, entities led us to believe that the 4G should protect both the entities and infrastructure [4].

Also, in mobile communications, another security problem is when the end user device is disconnected from the network for reasons such as battery exhaustion. The transition from level of disconnection to connection presents an opportunity for the attacker to show himself as a mobile device or a mobile support station [6].

There is an increasing importance in protecting the end user device due to the increasing danger of root kits. These are malware which can modify for malicious reasons operating system code and data. McAfee stated that root kits had increased by 600% in the last few years [7]. Also, McAfee stated that most malware targeted Android operating systems [8].

In addition, new end user devices are sources of denial of service attacks, viruses, worms, and so on. Smart phones have become attractive targets for attackers and this make the social implications of the attacks more harmful.

2.2 ITU-T Recommendation

As specified by ITU-T, recommendation M.3400 belongs to the Telecommunications Management Network (TMN) recommendations. It provides the security management sets of specifications of TMN management function. It considers security management as a part of TMN Management which cannot be isolated from any telecommunication network [9]. Security Management includes four groups of function sets: these are Prevention, Detection, Containment and Recovery and Security Administration. In designing our system, we followed these specifications of security management. Specifications of security management contain many function sets but, as mentioned above, we considered which ones help us to achieve our security requirements. Some function sets, which met our requirements were, firstly, the Customer Security Alarm function set defined as “This set supports access to security alarm that indicates security attacks on their portion of network” [9]. We use

this function set to detect if the mobile equipment had been attacked and it help us afterwards to prevent its use as an attack tool. Secondly, there was the Investigation of Theft of Service function set defined as “This set supports investigation of customer and internal users whose usage patterns indicate possible fraud or theft of service” [9]. Also, we use it to lead us to the attacked mobile equipment. Thirdly, there was the Software Intrusion Audit function set defined as “This set supports checks for signs of software intrusion in the network” [9]. This helped, also, to detect if there was a violation in the network which leads us to an attack on the mobile equipment.

2.3 Policy Based System

A security policy is the rule which defines the functions to maintain security in a system. A procedure is the heuristic process to enforce the rule. Therefore, it is important, when we propose a security management system, to build a component which specifies the policies and another component which is responsible for ensuring that these policies have been followed. Another requirement for our system was to address the challenge of combining wired and wireless technologies.

The security management framework, used in a wired network, does not suit wireless networks because of the hosts’ dynamic topology and mobility. Consequently, the construction of a security management system for both wired and wireless technologies increased the complexity. There are many solutions on wireless networks which existing researches have followed. Research on a WLAN security management framework follows the concept of dividing the network into wireless policy zones and, therefore, enforcement and validation policies are easier and more efficient [10]. However, a WLAN security management framework requires expensive management and equipment with regard to the need for a local server for each wireless policy zone. Also, scalability was an important challenge which our system needed to address. Many researches on different networks achieved this feature such in [11]. They maintained scalability issues in wireless networks by presenting wireless network policy managers with local policy autonomy. However, they explained neither what type of security policies should be enforced nor the formal validation of such polices. Consequently, we could obtain some idea of autonomous policies by building an auto system administrator with regard to these network changes. An auto system administrator is very fast and cannot be managed easily by human capabilities.

3 Our Approach

This research’s goal was to design and implement a security management system which could detect an attack on a mobile in 4G heterogeneous networks and prevent this attacked mobile from being used an attack tool which could harm the network. In our system, we defined the attack as any malicious user who tried to access the

system’s configuration files such as the password file, the system log configuration file or the mail configuration file [12]. Our approach followed the M.3400 specifications of TMN management. As explained, the system requires detecting an attack on the mobile and preventing this attacked mobile from being used as an attack tool. There are function sets which achieve this requirement. These function sets are: Customer Security Alarm, Investigation of Theft of Service, Software Intrusion Audit, Exception Report Action, and Theft of Service Action.

As illustrated previously with regard to the security threats in a 4G heterogeneous network, we believe that there was a clear requirement for a security management system to address attacks on the end user device [5]. Based on these considerations, we present a policy-based concept of a security management system. Figure 1 shows the four main parts of the system’s architecture, which consists of: (1) Intelligent Agent, (2) Security Engine, (3) Security Policies Database, and (4) Security Administrator. We believe that the policies should be able to detect an attack and prevent any damage to the network. The system contains assurance functions to prevent this by removing the user’s access and severing the connection to the attacked mobile. Firstly, the Intelligent Agent collects information. Then, the Intelligent Agent obtains the policies from the Security Policies Database. Next, the Intelligent Agent analyses this information and sends the results to the Security Engine. Finally, the Security Engine finds that there is no attack and sends an instruction to execute the normal policy set and, when there is an attack, follows the appropriate procedures and stores a record in the database.

3.1 Intelligent Agent

The Intelligent Agent collects the information according to the Security Engine’s management policies. It collects different kinds of information such as the mobile device’s system information and files information. The system information contains hardware, operating system, etc.

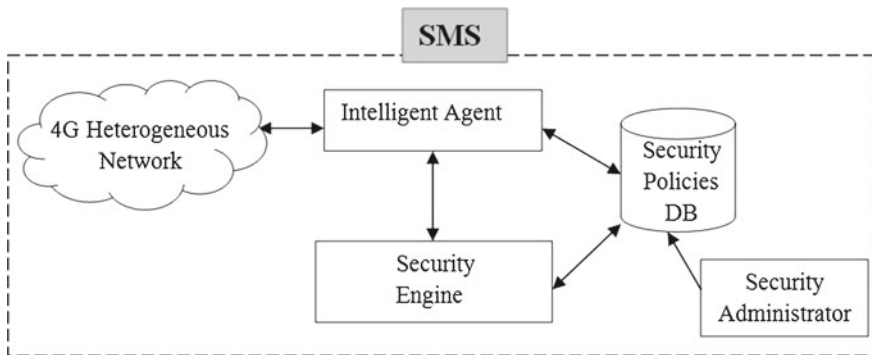


Fig. 1 Architecture of the security management system

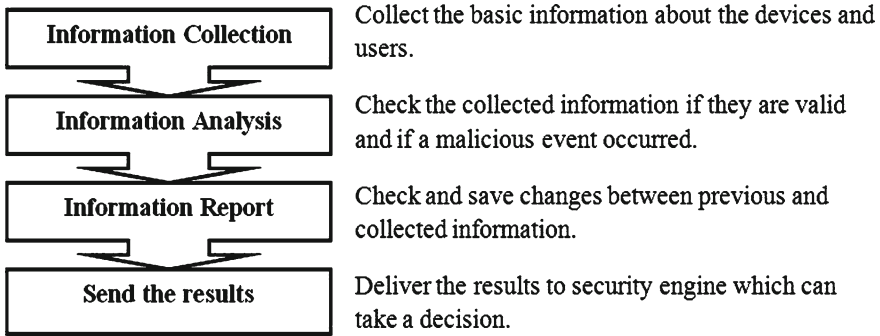


Fig. 2 Intelligent agent internal processes

As explained in Fig. 2 below, the Intelligent Agent follows four main steps.

3.2 Security Engine

The Security Engine obtains from the Intelligent Agent information about the events and saves this information in the Security Policies Database. When the security management system’s Security Engine receives event information that there is a violation, it should deny the network access to the mobile device.

The Security Engine makes the decision based on many factors such as the type of attack; the type of end user device; possible vulnerabilities in the same nodes; and previous records in the security database. The decision making process is heuristic and contains full details in order to generate a suitable security policy set [13]. Figure 3 shows the functional diagram of the security management system.

3.3 The Security Administrator

The Security Administrator updates policies and generates new settings for the network when the policies are breached. The Security Administrator’s roles are: (1) discovers the inconsistencies between the prescribed policies and current network status, and (2) confirms whether or not the policy rules, network domain and network entities are working together in consistent way.

If the Security Administrator detects a policy violation or inconsistency, it generates a new configuration setting and pushes a report to the security management system’s Security Engine which will prevent the mobile device from accessing the network.

The Security Administrator validates the policies by using network topology, the configuration state and previous records about the network.

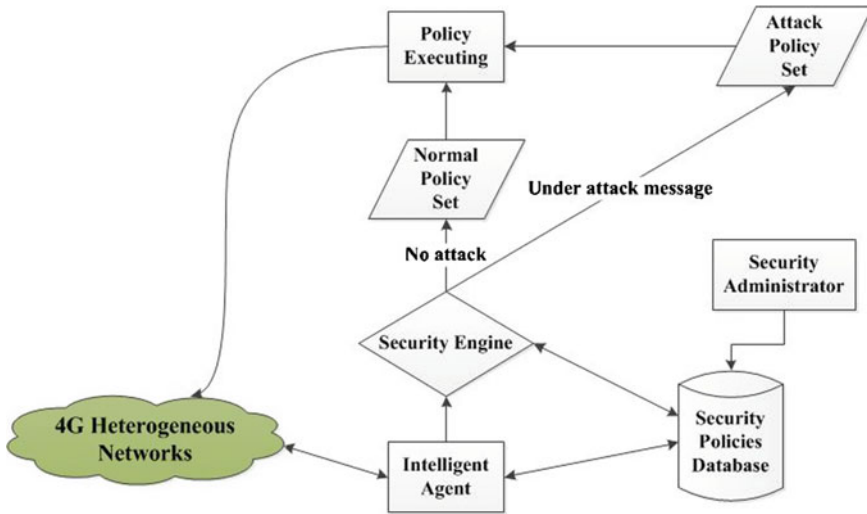


Fig. 3 SMS functional diagram

There are two main security policy levels which are: (1) normal level; and (2) danger level. When there is no attack, the security management system works on normal level with continuous monitoring. The security level moves to the danger level whenever an attack occurs. Also, the Security Engine should take a procedure from the Security Policies Database depending on its analysis and policies. When the security management system recovers from an attack, it moves the security policy level to the normal level and keeps a record in the database for future use [13].

3.4 Justification for Using an Automated Security Administrator

Due to the network and topology changes in wireless networks, human capabilities cannot match the rapid movements in network management movements.

The security policy management needs to be automated to prevent any malicious access to the network after any changes to it [1]. Also, because the limits of static policies are well known, there is a demand for dynamic policies. A dynamic policy is effective and responsive to the changes. Also, separating policy specification from policy management can offer robustness and automation of policies [11].

4 Example

We show an example of the security management system working in the case of an attack on a mobile device. The example assumes that there is a mobile device connected to the 4G heterogeneous network; this mobile device has an Android

operating system. This Android device contains system image located in /system/etc and this image contains Android configuration files. The device files can be accessed with READ and WRITE functions. In the environment of a heterogeneous network, the mobile device is connected via Bluetooth with other mobiles. The Android mobile device receives an attack from another device; the attacker is trying to access the configuration files and modify them [12]. The operating system provides events recognized by the Intelligent Agent; these events are triggered by changes in the configuration files. Therefore, the Intelligent Agent recognizes when a process is trying to access the configuration files. The Intelligent Agent knows that there is a malicious attack and sends a message to the Security Engine. This contains the Mobile's IP address, attack type and date and time of the attack. The Security Engine finds that there is danger from this mobile from being abused or used as an attack tool. As explained above, this is a clear security requirement. In this case of attack, the security management system should follow the TMN M.3400 Exception Report and Theft of Service actions function sets. These function sets state that the security breach should be limited by isolating the equipment to prevent the corruption from being propagated and removing the user's access. Therefore, the Security Engine makes a decision to isolate the mobile and remove the user's access to the network. The Security Engine is going to prevent the user's access to the network by contacting the service and application layer in Y-Comm security model. This layer is responsible for authenticating the users, so we can prevent the user's access and isolate the malicious device. Then, the Security Engine keeps a record in the database and the Security Administrator updates the policies for fast detection in case of the same attack happening in the future [13]. Figure 4 shows the sequence of the functions inside the security management system.

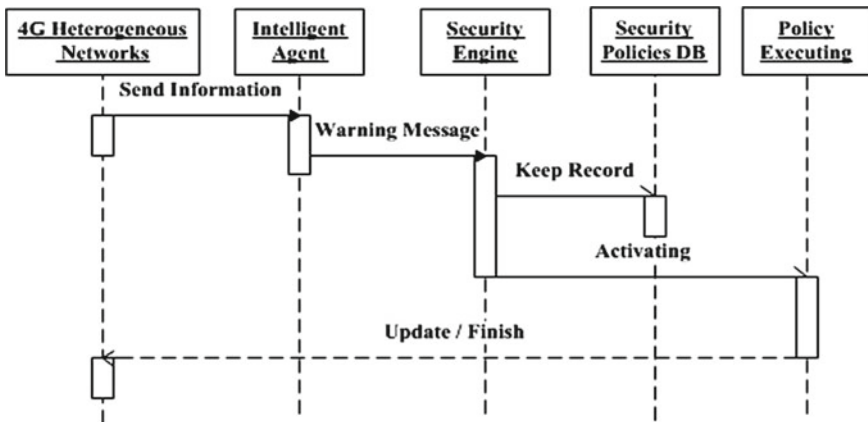


Fig. 4 Sequence of the functions in the proposed system

5 Conclusion

This research defined the architecture of a security management system. This security management system was policy-based and aimed to achieve the security requirements for protecting end user devices in 4G heterogeneous mobile networks. This system could be extended to achieve other requirements targeted at protecting the mobile entity from other attacks. We introduced the automatic security administrator which was reliable in addressing the challenges of fast changing networks. The system's processes are being implemented and the results will be published later.

References

1. Seddigh N, Nandy B, Makkar R, Beaumont JF (2010) Security advances and challenges in 4G wireless networks. In: Privacy security and trust (PST), 2010 eighth annual international conference, Ottawa, ON, 2010, pp 62–71
2. G. Mapp, M. Aiash, A. Lasebae, R. Phan (2011) Security models for heterogeneous networking. In: Proceedings of the 2010 international conference in security and cryptography (SECRYPT), Athens, Greece, pp 1–4
3. Aiash M, Mapp G, Lasebae A, Phan R (2010) Providing security, in 4G systems: unveiling the challenges. In: Sixth advanced international conference on telecommunications (AICT), 2010, Barcelona, pp 439–444
4. Park Y, Park T (2007) A survey of security threats on 4G networks. In: Globecom workshops 2007 IEEE Washington, DC, pp 1–7
5. Zheng Y, He D, Yu W, Tang X (2006) Trusted computing-based security architecture for 4G mobile networks. In: Sixth international conference on parallel and distributed computing, applications and technologies, 2005, PDCAT 2005, Sichuan, pp 251–255
6. Hardjono T, Seberry J (1995) Information security issues in mobile computing. In: Eleventh International Information Processing Conference-Security'95 Capetown. South Africa, pp 143–151
7. Bickford J, O'Hare R, Baliga Ar, Ganapathy V, Iftode L (2010) Rootkits on smart phones: attacks, implications and opportunities. In: Proceedings of the eleventh workshop on mobile computing systems & applications, pp 49–54
8. Greene T (2011) CSO. HYPERLINK <http://www.csoonline.com/article/694728/mcafee-android-is-sole-target-of-new-mobile-malware-in-q3>, <http://www.csoonline.com/article/694728/mcafee-android-is-sole-target-of-new-mobile-malware-in-q3>
9. ITU, ITU-T Recommendation M.3400, 2001
10. Bera P, Ghosh SK, Dasgupta P (2010) A spatio-temporal role-based access control model for wireless LAN security policy management. *Commun Comput Inform Sci* 54:76–88
11. Lapiotis G, Byungsuk Kim, S. Das, Anjum F (2006) A policy-based approach to wireless LAN security management. In: Security and privacy for emerging areas in communication networks, 2005. Workshop of the 1st international conference, pp 181–189
12. Vacca JR (2010) Network and system security. Elsevier, Oxford, UK
13. Alquhayz H, Al-Bayatti A, Platt A (2012) Security management system for 4G heterogeneous networks. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July, 2012, pp 1298–1302

Calibration Procedures for Indoor Location Using Fingerprinting

Pedro Mestre, Luis Reigoto, Luis Coutinho, Aldina Correia,
Joao Matias and Carlos Serodio

Abstract Fingerprinting is a location technique, based on the use of wireless networks, where data stored during the offline phase is compared with data collected by the mobile node during the online phase. When this location technique is used in a real-life scenario there is a high probability that the mobile node used throughout the offline phase is different from the mobile nodes that will be used during the online phase. This means that there might be very significant differences between the Received Signal Strength values acquired by the mobile node being located and the ones previously stored in the Fingerprinting Map. As a consequence, this difference between RSS values might contribute to increase the location estimation error. One possible solution to minimize these differences is to adapt the RSS values, acquired during the online phase, before sending them to the Location Estimation Algorithm. Also the internal parameters of the Location Estimation Algorithms, for example the weights of the Weighted k-Nearest Neighbour, might need to be tuned for every type of terminal. This paper focuses both approaches, using Direct Search optimization

P. Mestre (✉) · C. Serodio
CITAB-UTAD, Algoritmi-UM, Vila Real, Portugal
e-mail: pmestre@utad.pt

C. Serodio
e-mail: cserodio@utad.pt

L. Reigoto · L. Coutinho
UTAD, Vila Real, Portugal
e-mail: luisreigoto@gmail.com

L. Coutinho
e-mail: luis_coutinho_86@hotmail.com

A. Correia
CM-UTAD, ESTGF-IPP, Felgueiras, Portugal
e-mail: aldinacorreia@eu.ipp.pt

J. Matias
CM-UTAD, Vila Real, Portugal
e-mail: j_matias@utad.pt

methods to adapt the Received Signal Strength and to tune the Location Estimation Algorithm parameters. As a result it was possible to decrease the location estimation error originally obtained without any calibration procedure.

Keywords Direct search optimization methods · Fingerprinting · IEEE802.11 · Indoor location · LEA adaptation · RSS adaptation

1 Introduction

Fingerprinting location estimation is a wireless network based technique on which the values of a given property of the wireless signals, received by a mobile device, are compared with a set previously stored values, called the Fingerprinting Map (FM). This location estimation technique comprises two distinct phases: one on which data to generate the Fingerprinting Map is collected and stored in a database, called the offline phase; and a second phase, called the online phase, on which the estimation of the node location is made by comparing the collected data with the data stored in the Fingerprinting Map [1, 14].

Although any property of the wireless signal can be used, typically the Received Signal Strength (RSS) is used. During the offline phase, for each point of the spatial domain that will be mapped on the Fingerprinting Map, are collected RSS values from the wireless base stations used as references. In WiFi networks those references are the infrastructure Access Points (AP).

Mathematically, the power of the signal received by an wireless node from each Access Point can be expressed as in Eq. 1:

$$P_r = P_t + G_t + G_r - L_{sum} \quad (1)$$

where P_r is the received power (in dBm), P_t is the output power of the AP (in dBm), G_t and G_r the gains of the transmitting and the receiving antennas (in dB or dB_i) and L_{sum} is the sum of all losses in the path between the transmitter and the receiver (in dB).

It becomes obvious that if different types of mobile terminals are used then different values for the RSS are received. This is because different type of terminals might have different types of antennas that consequently might have different values of gain and directivity, therefore, under the same operating conditions they will have different values for the RSS. These differences can have a negative impact on the performance of Location Estimation Algorithms (LEA) because the mobile node to be located might not be the same that was used to collect data to build the Fingerprinting Map.

Collecting wireless network data to generate the FM using every type of mobile terminal that could be used during the online phase, and generate a Fingerprinting Map for each mobile terminal, is not a feasible solution to solve this problem because of the large number of existing mobile terminals that could potentially be used in

such applications. On the other hand, restricting the access to the location estimation application only to a subset of mobile terminals, might not be also the best solution because it would exclude many mobile terminals from the location system.

In this work are presented two techniques to overcome the differences between mobile terminals, [11], and minimize the impact of using different terminals during the offline and online phases. One of these techniques consists on adapting the RSS values and the second technique consists on adapting the LEA internal parameters.

The main objective of this work is to find a calibration procedure that enables the location estimation system to work correctly for a given mobile terminal, without the need to collect data in all (or too many) points of the scenario with it. The mobile terminal should be placed at some predefined calibration points in the scenario where the calibration would be executed.

2 RSS and LEA Adaptation

In this section are presented the two techniques proposed for the calibration procedure, and to minimize the impact of using different types of terminals.

In both cases, adaptation of the acquired signals and LEA adaptation, Direct Search Methods are used. These methods are useful when derivative-based methods cannot be used, [2], for example when the values of the objective function are experimentally collected data, which is the case.

2.1 RSS Adaptation

During the online phase, the acquired values of RSS must be sent to the LEA that estimates the current location of the mobile node based on these values.

This approach might not be suitable for the objectives of this work, therefore in the approach here presented, the raw values of the RSS acquired from the wireless network interface are not fed directly to the LEA.

Before sending the acquired RSS values to the LEA, they are adapted. This adaptation, as in [12, 13] consists in adding a calibration offset (c), Eq. 2, to the RSS values:

$$RSS_{LEA} = RSS_{acquired} + c \quad (2)$$

where RSS_{LEA} is the RSS value sent to the LEA, $RSS_{acquired}$ is the raw RSS value acquired from the wireless interface and c is the calibration offset.

To determine the optimal value for the calibration offset two strategies are presented below. One uses the average error between the values acquired by the wireless node and the values stored in the FM, and the other uses Direct Search Optimization Methods.

2.1.1 Using the Average Error

In this first method the average error between the expected values of RSS, which are stored in the FM, and the values acquired by the mobile node, as in Eq. 3, are used as the calibration parameter, c , of Eq. 2:

$$c = \frac{1}{K} \sum_{i=0}^{K-1} \left(\frac{1}{N} \sum_{j=0}^{N-1} (FM_{i,j} - RSS_{i,j}) \right) \tag{3}$$

where K is the number of points used in the calibration process, N is the number of Access Points detected at the calibration point i , $RSS_{i,j}$ is the RSS value of Access Point j at point i and $FM_{i,j}$ is the value of the RSS stored in the Fingerprinting Map for Access Point j at point i .

2.1.2 Using Direct Search Optimization Methods

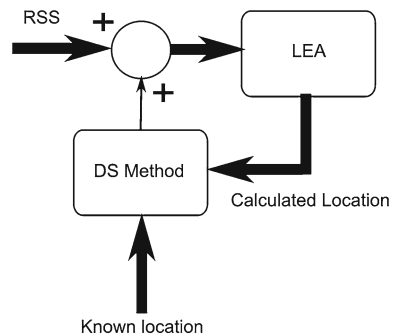
Although the above presented calibration procedure can boost the location estimation results, a second approach consisting on calculating the optimal calibration value using Direct Search Optimization Methods, is presented. These methods can be used in the optimization of both constrained and unconstrained optimization problems.

As depicted in Fig. 1, the inputs of the Direct Search Method are the location estimated by the LEA and the actual location of the mobile node. At each iteration, the Direct Search Method requests the LEA to recalculate the new value for the location estimation, as a function of the new calibration offset.

In this case we have an unconstrained optimization problem of the form:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{4}$$

Fig. 1 Using Direct Search Methods to calculate the calibration offset



where:

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the *objective function*.

The objective function, $f(x)$, for this particular problem is the average location error for the points of the scenario under test that were used during the calibration phase. The dimension of the problem is 1 because the only input value to the objective function is the value of the offset.

An API (Application Programming Interface) built using Java Technology, that implements the used optimization methods, developed by the authors in [8] was used. For the optimization of this particular problem it was used the Nelder and Mead algorithm, which was implemented in the API as in [4, 6, 7].

2.2 LEA Adaptation

While the above presented method consists in adapting the RSS values before sending them to the LEA, this approach is based on the tuning of the internal parameters of the LEA, using Direct Search Optimization Methods, adapting it to the mobile terminal characteristics.

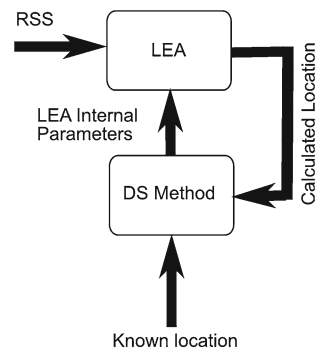
As depicted in Fig. 2, the inputs are the same as for the previous method (actual location and estimated location), however in this case the output of the optimization method are the internal parameters of the LEA.

In this case the values of the internal parameters of the algorithm might be subject to constraints, so we have a constrained problem of the form of 5:

$$\begin{aligned}
 & \min_{x \in \mathbb{R}^n} f(x) \\
 & s.t. \quad c_i(x) = 0, i \in \mathcal{E} \\
 & \quad \quad c_i(x) \leq 0, i \in \mathcal{I}
 \end{aligned}
 \tag{5}$$

where:

Fig. 2 Calibration of the LEA internal parameters using Direct Search Methods



- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the *objective function*;
- $c_i(x) = 0$, $i \in \mathcal{E}$, with $\mathcal{E} = \{1, 2, \dots, t\}$, define the problem *equality constraints*;
- $c_i(x) \leq 0$, $i \in \mathcal{I}$, with $\mathcal{I} = \{t + 1, t + 2, \dots, m\}$, represent the problem *inequality constraints*;
- $\Omega = \{x \in \mathbb{R}^n : c_i = 0, i \in \mathcal{E} \wedge c_i(x) \leq 0, i \in \mathcal{I}\}$ is the set of all feasible points, i.e., defines the *feasible region*.

As a case study, the optimization of the weights of the WkNN algorithm is presented and the following conditions were used:

- The objective function, f , is the average location error for the points of the testing scenario that were used during the calibration procedure;
- The input parameters of the objective function, x_i are the k weights of the k nearest neighbours;
- Each weight, W_i , must satisfy the following constraint: $W_i \geq 0$.

To solve the optimization problem it was used an implementation of the Penalty and Barrier Methods, [3], available on the above mentioned Java API. In Penalty and Barriers Methods the optimization problems are transformed into a sequence of unconstrained problems. These problems are then solved using the same algorithms that are used to solve unconstrained problems. This new sequence of unconstrained problems, that replaces the original problem, is defined by:

$$\Phi(x_k, r_k) : \min_{x_k \in \mathbb{R}^n} f(x_k) + r_k p(x) \quad (6)$$

where Φ is the new objective function, k is the iteration, p is a function that penalises (penalty) or refuses (barrier) points that violates the constraints and r_k is a positive parameter.

In this work the Penalty and Barrier Methods were used with the Hooke and Jeves algorithm, [5], as internal method. As penalty function it was used a Non-stationary Penalty function, [15].

3 Testing Scenario and Conditions

To test the feasibility of the above presented procedures, data was collected in the scenario depicted in Fig. 3, which are the headquarters of a local scouts group. These data were used to generate the Fingerprint Maps, calculate the calibration offset, tune the LEA and to make the location estimation tests.

Data was collected using two Android smartphones, from different manufactures and having different sizes, which will be referred as Smartphone 1 and Smartphone 2. In this scenario a total of 24 points were used to collect data with both mobile terminals and for each point, which are marked in Fig. 3, a total of 20 RSS samples was taken. Those samples were acquired using an application developed for Android, Fig. 4.

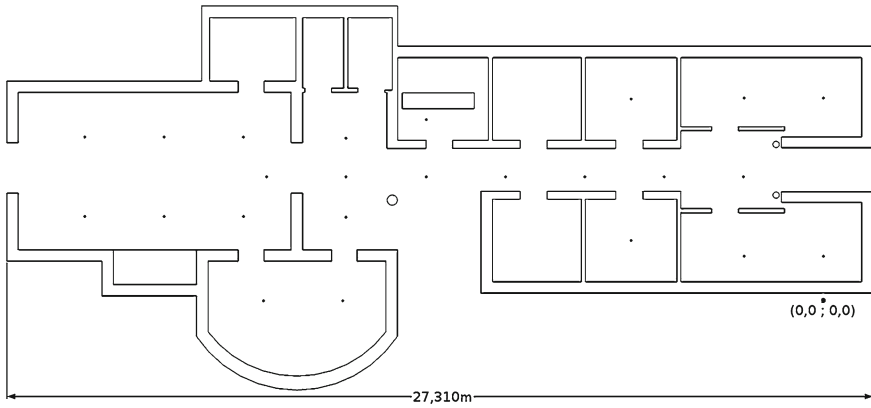


Fig. 3 Map of the building where the tests were made

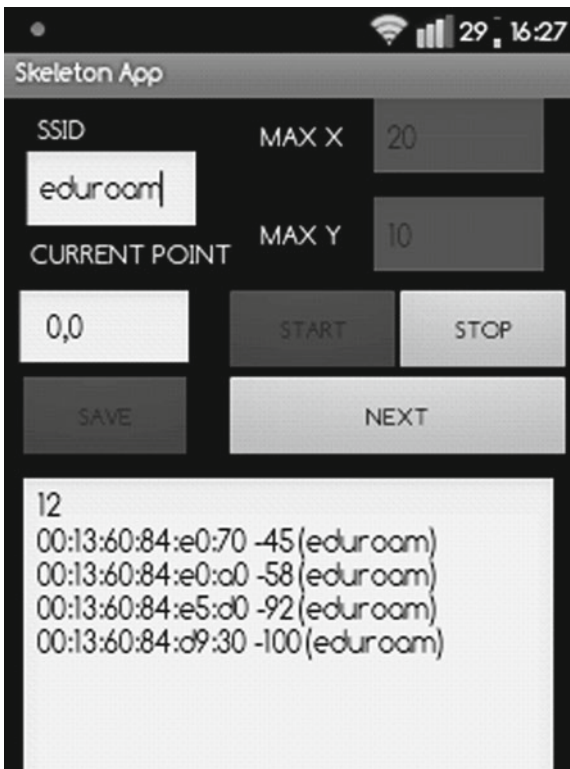


Fig. 4 Android application developed to acquire the data

Since data collected with this application are going to be used to test the several calibration procedures, this application does not make any location estimation. It only stores the acquired data in XML (eXtensible Markup Language) files for later processing. It is then possible to test offline all LEA and calibration procedures exactly with the same data.

Although Android smartphones were used in these experiments, any other type of smartphone and Operating System could be used, as long as it is possible to read the values of RSS and they can be expressed in *dBm*.

To access the feasibility of the above presented calibration procedures, three of the classic LEA were used: Nearest Neighbour (NN); k-Nearest Neighbour (kNN); Weighted k-Nearest Neighbour (WkNN). For these LEA it was considered that $k = 3$ for kNN, and the weights 0.7, 0.2, 0.1 for WkNN, as in [10]. There is an exception for the weights on those tests where the LEA parameters were calibrated for the mobile terminal.

Each of the calibration procedures, was made using different number of calibration points. The calibration procedures were made using three and one calibration points, located at the entrance hall of the building.

Although the use of all points, belonging to the scenario, in the calibration procedure is not a feasible solution in a real-life application, results obtained in such conditions are also presented. These results are presented as reference values for comparison purposes, and will be considered as the best results that could be achieved with the location system.

4 Numerical Results

In this section are presented the results obtained for both smartphones with the above presented LEA and calibration procedures.

4.1 Reference Values

Tables 1 and 2 present the values for the location precision (Prec.), standard deviation (StDev), maximum location error (MaxErr) and minimum location error (MinErr).

Table 1 Comparison of the used LEA, using both smartphones and the FM generated with data from Smartphone 1

	Smartphone 1			Smartphone 2		
	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	0.27	1.68	0.91	1.42	1.83	1.56
StDev (m)	1.01	0.87	0.72	1.70	1.09	1.15
MaxErr (m)	8.67	7.65	8.12	6.32	6.32	5.25
MinErr (m)	0.00	0.00	0.08	0.00	0.00	0.13

Table 2 Comparison of the used LEA, using both smartphones and the FM generated with data from Smartphone 2

	Smartphone 1			Smartphone 2		
	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	1.38	1.83	1.52	0.34	1.68	0.94
StDev (m)	1.58	0.94	1.02	1.00	0.90	0.65
MaxErr (m)	8.67	6.77	8.03	5.77	3.57	5.00
MinErr (m)	0.00	0.00	0.15	0.00	0.00	0.15

For the NN, kNN and WkNN algorithms, using both smartphones. For Table 1 the FM was generated using data acquired with Smartphone 1 and for Table 2 data acquired with Smartphone 2.

As already expected in beforehand there is a degradation of the LEA performance when the mobile node under test is not the same that was used during the offline phase, to build the Fingerprinting Map.

4.2 RSS Adaptation Using the Average Error

Numerical results of the tests using the average error as the calibration value, are presented in Tables 3 and 4. Table 3 presents the values for Smartphone 2 using the FM generated with Smartphone 1 and Table 4 presents the values for Smartphone 1 using the FM generated with Smartphone 2.

Comparing the values in Tables 1 and 3 (Smartphone 2) it can be observed that when all points are used in the calibration procedure there is a boost in the precision and standard deviation values for the NN and WkNN algorithms, and a better value for the maximum error in kNN. Similar behaviour can be observed for the value of precision and maximum error, when three points are used in the calibration procedure. When the calibration procedure is made using only one calibration point, worse results are obtained.

Table 3 Comparison of the used LEA using Smartphone 2, FM generated with Smartphone 1, and calibration values calculated by error average

	All Points			3 Points			1 Point		
	NN	kNN	WkNN	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	1.27	1.83	1.49	1.32	1.85	1.54	1.59	1.94	1.71
StDev(m)	1.66	1.10	1.13	1.71	1.11	1.17	1.96	1.26	1.42
MaxErr(m)	6.32	5.33	5.25	6.32	5.33	5.25	8.30	5.89	6.83
MinErr(m)	0.00	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.13

Table 4 Comparison of the used LEA using Smartphone 1, FM generated with Smartphone 2, and calibration values calculated by error average

	All Points			3 Points			1 Point		
	NN	kNN	WkNN	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	1.02	1.84	1.31	1.05	1.87	1.34	1.17	1.78	1.38
StDev(m)	1.55	1.05	1.08	1.68	1.05	1.16	1.50	0.92	0.98
MaxErr(m)	8.67	6.77	7.60	8.67	6.77	7.60	8.67	6.77	7.60
MinErr(m)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

For Smartphone 1 the absolute values are slightly better than for Smartphone 2, when comparing results from Tables 2 and 4. With this Smartphone also more precision values were better than the original ones, without calibration.

4.3 RSS Adaptation Using Direct Search Methods

Numerical results of the tests made with the optimization methods are presented in Tables 5 and 6. For Smartphone 2, all values of precision and standard deviation, using the calibration procedure with all points and three points, are better than the original ones. Even with a single calibration point, the precision and standard deviation values (NN and kNN) are better.

When comparing the results obtained with Smartphone 1, it can be concluded that all values for the precision are better than the original ones, except for one that had the same value. However the same is not true for the obtained standard deviation values. Nevertheless, those difference in the standard deviation values can be considered as small.

4.4 LEA Adaptation Using Direct Search Methods

This calibration procedure was used only with WkNN, to optimize the values of the weights. Table 7 presents the values of the weights obtained for Smartphone 1 (SP1) and Smartphone 2 (SP2). procedure.

Table 5 Comparison of the used LEA using Smartphone 2, FM generated with Smartphone 1, and calibration values calculated using direct search methods

	All Points			3 Points			1 Point		
	NN	kNN	WkNN	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	1.22	1.75	1.44	1.27	1.76	1.49	1.31	1.75	1.65
StDev(m)	1.57	1.01	1.02	1.66	1.00	1.13	1.59	1.01	1.35
MaxErr(m)	6.32	4.51	5.25	6.32	4.51	5.25	6.32	4.51	6.83
MinErr(m)	0.00	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.13

Table 6 Comparison of the used LEA using Smartphone 1, FM generated with Smartphone 2, and calibration values calculated using direct search methods

	All Points			3 Points			1 Point		
	NN	kNN	WkNN	NN	kNN	WkNN	NN	kNN	WkNN
Prec. (m)	0.97	1.75	1.28	1.06	1.78	1.34	1.06	1.83	1.35
StDev(m)	1.51	0.92	1.06	1.73	1.01	1.18	1.73	0.94	1.19
MaxErr(m)	8.67	6.77	7.60	8.67	6.77	7.60	8.67	6.77	7.60
MinErr(m)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7 Weights obtained with the Penalty and Barrier Methods for both smartphones

	All points	3 points	1 point
SP1	(1.00;0.00;0.00)	(0.31;0.29;0.40)	(0.39;0.55;0.15)
SP2	(1.00;0.00;0.00)	(0.52;0.24;0.24)	(1.00;0.00;0.00)

Table 8 Results obtained with Smartphone 2 and weight optimization of Weighted kNN using direct search methods

	All points	3 points	1 point
Prec. (m)	1.42	1.66	1.42
StDev(m)	1.70	1.01	1.70
MaxErr(m)	6.32	5.49	6.32
MinErr(m)	0.00	0.00	0.00

Table 9 Results obtained with Smartphone 1 and weight optimization of Weighted kNN using direct search methods

	All points	3 points	1 point
Prec. (m)	1.38	1.92	1.87
StDev(m)	1.58	0.98	0.91
MaxErr(m)	8.67	6.47	7.60
MinErr(m)	0.00	0.15	0.36

Data on this table, for all calibration points (both smartphones) and for one point (Smartphone 2), corroborates what the previous tables show: the best of the three LEA, considering the location error, for this case is NN. When three calibration points are used, there are no null weights, so for these testing conditions the WkNN algorithm is the best. Different values for the weights confirm that each type of terminal has its own optimal LEA values.

Using these weights a new set of tests was made using data of both smartphones and the Weighted kNN algorithms (Tables 8 and 9).

For the calibration procedure using all points better precision values (comparing to Tables 1 and 2) were obtained for both smartphones (using WkNN). Similar values

were also obtained for Smartphone 2 when one calibration point was used. This is a consequence of the weights being the same as for kNN.

However the results obtained with Smartphone 2 using three calibration points, and for Smartphone 1 using three and one calibration points, are worse than the original values. Also to be noticed that except for Smartphone 2 with all calibration points, none of the other values for the precision are better than those obtained for the previous calibration procedure (RSS adaptation).

Even though the use of direct search methods to optimize the LEA appears to have no positive impact in the performance of the location system, a new set of tests was made. While in the previous tests the LEA was calibrated using the raw RSS values acquired by the wireless network interface, in this new set of tests the two calibration procedures were used together, Fig. 5. This new calibration procedure consists in using two optimizations, Direct Search Methods are used to adapt the RSS values and the LEA parameters.

The weights obtained for these new set of tests are presented in Table 10, and for all tests it is confirmed that the best values for precision are obtained using the NN algorithm.

Tables 11 and 12 present the results obtained with both smartphones using these weights. The obtained values for the precision are better than the original values (Tables 1 and 2) and the values obtained using the RSS calibration (Tables 5 and 6). However the same is not true for the obtained values of standard deviation and the maximum location error which as expected had values very similar to the original values for NN.

Fig. 5 LEA and calibration offset optimization

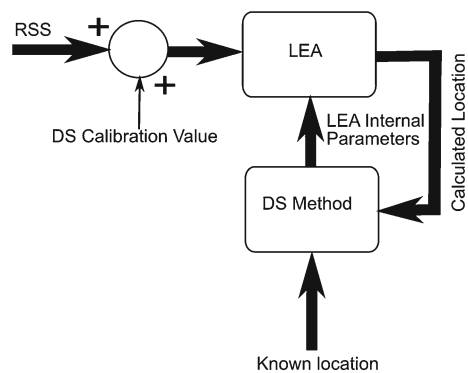


Table 10 Weights obtained with the Penalty and Barrier Methods for both smartphones

	All points	3 points	1 point
SP1	(1.00;0.00;0.00)	(1.00;0.00;0.00)	(1.00;0.00;0.00)
SP2	(1.00;0.00;0.00)	(1.00;0.00;0.00)	(1.00;0.00;0.00)

Table 11 Results obtained with Smartphone 2 and weight optimization of Weighted kNN using direct search methods

	All points	3 points	1 point
Prec. (m)	1.25	1.17	1.50
StDev(m)	1.57	1.66	1.90
MaxErr(m)	6.32	6.32	8.30
MinErr(m)	0.00	0.00	0.00

Table 12 Results obtained with Smartphone 1 and weight optimization of Weighted kNN using direct search methods

	All points	3 points	1 point
Prec. (m)	0.97	1.04	1.06
StDev(m)	1.51	1.72	1.73
MaxErr(m)	8.67	8.67	8.67
MinErr(m)	0.00	0.0	0.00

5 Conclusion and Future Work

In this chapter two techniques to minimize the location estimation error using wireless networks and fingerprinting were presented. These techniques have as objective to overcome the differences between mobile terminals, and are divided into two main categories: RSS adaptation and LEA adaptation.

For all tests, although the use of all points of the scenario in the calibration procedure had the best results, it is not feasible in a real life scenario. So, tests were also made using one and three calibration points located at the entrance hall of the building. These tests also had interesting results, especially those made using three calibration points. As it was expected, when more calibration points are used, the better are the results.

In the case of RSS adaptation, from the two proposed techniques, the one with better results was the use of Direct Search Methods. Considering only the tests with three and one calibration points, it was possible to reduce the location error by 10.56 % for Smartphone 1 and 23.91 % for Smartphone 2.

Using direct search methods to calibrate the values of RSS and tune the internal parameters of the LEA, it was possible to increase even further the precision of the location estimation. If used with WkNN, Direct Search Optimization Methods can be used to adapt the algorithm to the mobile terminal, and it can be used for automatic LEA selection (choosing between NN, kNN and WkNN) by adjusting the algorithm weights. In the presented tests it was confirmed that, for the test scenario and the used smartphones, NN was the algorithm with the best precision values.

For the tests made with three and one calibration points, the best values obtained when the two optimization procedures were used together are 25.00 % for Smartphone 2 and 31, 58 % for Smartphone 1 (using three calibration points).

Despite the fact that precision was used in the objective function of the optimization problems, other parameters such as the standard deviation, maximum error, minimum error or a combination of some of them could be used. This is an option to be explored in future developments of this work.

As future work, also the use of the proposed calibration procedures together with alternative techniques to build the Fingerprinting Map, such as the use of propagation models, [12], will be considered.

Furthermore, other types of LEA, such as the ones based on Fuzzy Logic, e.g. [9, 13] could benefit from the calibration procedures presented in this chapter.

References

1. Bahl P, Padmanabhan VN (2000) RADAR: an in-building RF-based user location and tracking system. In: Proceedings of nineteenth annual joint conference of the IEEE computer and communications societies, INFOCOM 2000, vol 2. IEEE 2, pp 775–784
2. Correia A, Matias J, Mestre P, Serodio C (2010) Derivative-free nonlinear optimization filter simplex. *Int J Appl Math Comp Sci (AMCS)* 4029(4):679–688
3. Correia A, Matias J, Mestre P, Serodio C (2010) Direct-search penalty/barrier methods. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering WCE 2010, U.K, London, pp 1729–1734, 30 June–2 July 2010
4. Dennis J, Woods D (1987) Optimization on microcomputers. the nelder-mead simplex algorithm. In: Wouk A (ed) *New computing environments: microcomputers in large-scale computing*, pp 116–122
5. Hooke R, Jeeves T (1961) Direct search solution of numerical and statistical problems. *J Assoc Comput Mach* 8(2):212–229
6. Kelley C (1999) *Iterative methods for optimization*. Number 18 in *frontiers in applied mathematics*. SIAM, Philadelphia, USA
7. Lagarias J, Reeds J, Wright M, Wright P (1998) Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM J Optim* 9(1):112–147
8. Mestre P, Matias J, Correia A, Serodio C (2010) Direct search optimization application programming interface with remote access. *IAENG Int J Appl Math* 40(4):251–261
9. Mestre P, Coutinho L, Reigoto L, Matias J, Correia A, Couto P, Serodio C (2011) Indoor location using fingerprinting and fuzzy logic. In: *Advances in intelligent and soft computing*, vol 107. Springer, Berlin, pp 363–374
10. Mestre P, Pinto H, Serodio C, Monteiro J, Couto C (2009) A multi-technology framework for LBS using fingerprinting. In: *Proceedings of industrial electronics, IECON '09 35th annual conference of IEEE*, pp 2693–2698
11. Mestre P, Reigoto L, Coutinho L, Correia A, Matias J (2012) RSS and LEA adaptation for indoor location using fingerprinting. In *Lecture notes in engineering and computer science: Proceedings of world congress on engineering WCE 2012*, 4–6 July 2012, London, U.K, pp 1334–1339
12. Mestre P, Serodio C, Coutinho L, Reigoto L, Matias J (2011) Hybrid technique for fingerprinting using IEEE802.11 wireless networks. In: *Proceedings of the International Conference on indoor positioning and indoor navigation (IPIN)*, pp 1–7
13. Serodio C, Coutinho L, Pinto H, Mestre P (2011) A comparison of multiple algorithms for fingerprinting using IEEE802.11. In *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering WCE 2011*, 6–8 July 2011, U.K, London, pp 1710–1715

14. Taheri A, Singh A, Agu E (2004) Location fingerprinting on infrastructure 802.11 wireless local area networks. In: LCN '04: Proceedings of the 29th annual IEEE international conference on local computer networks. IEEE Computer society, Washington DC, pp 676–683
15. Wang FY, Liu D (2006) Advances in computational intelligence: theory and applications (series in intelligent control and intelligent automation). World Scientific Publishing Co., Inc., River Edge

Handling the Congestion Control Problem of TCP/AQM Wireless Networks with PID Controllers

Teresa Alvarez and Diego Martínez

Abstract Internet users rely on the good capabilities of TCP/IP networks for dealing with congestion. The network delay and the number of users change constantly, which might lead to transmission problems. Usually, they are handled following approaches such as Drop Tail or Random Early Detection (RED) algorithms. During the last few years, automatic control techniques are providing practical solutions. Moreover, if there are wireless links, congestion control is more difficult to detect. This chapter presents a methodology to design a PID controller with a linear gain scheduling that allows a network with wireless links to deal with control congestion under a variety of configurations. The technique is compared with a standard PID and several tests are carried out using linear and non-linear environments.

Keywords AQM · Congestion control · Gain scheduling · PID · Robustness · TCP Westwood

1 Introduction

The importance of Internet cannot be denied. However, there are situations in which problems arise such as long delays in delivery, lost and dropped packets, oscillations or synchronization problems [1, 2]. Congestion is responsible for many of these problems. It is harder to detect if there are wireless links in the network. Thus, techniques to reduce congestion are of great interest. There are two basic approaches

T. Alvarez (✉)
Escuela de Ingenierías Industriales (Sede Mergelina),
Universidad de Valladolid, Valladolid, Spain
e-mail: tere@autom.uva.es

D. Martínez
Hewlett-Packard Española S.L., Parque Empresarial Madrid_las Rozas, Las Rozas, Spain
e-mail: diegomartinez29@hotmail.com

[2, 3]: *congestion control*, which is used after the network is overloaded, and *congestion avoidance*, which takes action before the problem appears.

Feedback control can help to solve congestion control. Congestion control is carried out at the router's transport layer. The Transmission Control Protocol (TCP) and the Active Queue Management (AQM) methodology are implemented at this layer. AQM techniques [4] must achieve an *efficient queue utilization* (minimizing queue overflows and underflows), *small queuing delay* (minimizing the time a data packet requires to be serviced by the routing queue) and *robustness* (maintaining closed-loop performance in changing network conditions).

AQM schemes enhance the performance of TCP, although they have disadvantages and they do not work perfectly in every traffic situation. Numerous algorithms have been proposed (see [3] for a revision). The most widely used AQM mathematical models were published in [4]. There are interesting approaches such as [5]. Since then, several control approaches have been tested: fuzzy, predictive control, robust control, etc. The automatic control techniques that have received the most attention are the Proportional Integral (PI) controller, followed by the Proportional Integral Derivative (PID, [6]). AQM techniques should be robust and give good results when the network is not operating in nominal situation, when the delays are changing, or the number of users or the intensity of the packet flow vary, i.e. the network should deliver the users' data on time and without errors regardless of the changes in the settings. The AQM congestion control algorithm should be robust, stable and perform in a changing environment. If PID is chosen as the AQM algorithm, there are many useful references [6–8] that work with PIDs and take into account the delay of the system. In [7], a graphical characterization of the stable region of a PID controller applied to the AQM router is given considering a network with fixed parameters.

Motivated by these issues, this paper presents how to derive PIDs with guaranteed stability and robustness and that perform well in a network with wireless links and changing traffic conditions. Some preliminary results can be found in [9]. The work presented in this chapter applies the generic method in [7] for finding stable PID controllers for time delay systems to a dynamic TCP Westwood/AQM router model. As novelty, a linear gain scheduling is included in the design (reducing the gain variation) to enhance the robustness and ensure adequate behaviour in extreme working scenarios. The metrics applied to study the controller performance are: the router queue size, the link utilization and the probability of packet losses.

First, the controller is obtained and tested with the linear system. Non-linear simulations in Matlab using the differential equation model show the goodness of the method. Finally, the network simulator ns-2, working with two routers that are connected in a dumbbell topology, is used for testing the method in an almost real environment. The three approaches show the advantages of the technique described in this chapter.

This chapter is organized as follows. Section 2 briefly describes TCP Westwood and presents a fluid flow model, in Sect. 3, the methodology is described. Results are shown in Sect. 4. Finally, some conclusions are presented.

2 Fluid Flow Model for TCP Westwood

TCP Westwood (TCPW) [10, 11] is a modification of TCP NewReno at the source side, intended for networks where losses are not only due to congestion, such as the wireless networks studied in this chapter. The protocol from the receiver point of view is the same in both TCPs, but there are some changes in the way the source calculates the available bandwidth. These modifications affect the dynamics of the system, which justifies the necessity of specific controllers for TCPW.

This section presents a fluid flow model derived from [10]. As in the original approach [12, 13], there is a single bottleneck. Now, all the TCP connections are assumed to follow the Westwood formulation.

2.1 Nonlinear Model

The nonlinear model can be described by two coupled nonlinear differential equations. The first one describes the TCP window control dynamic and Eq. (2) models the bottleneck queue length.

$$\dot{W} = \frac{1}{\frac{q(t)}{C} + T_p} - \frac{W(t) W(t - R_0)}{\frac{q(t-R_0)}{C} + T_p} p(t - R_0) + T_p \left(\frac{W(t - R_0)}{\frac{q(t-R_0)}{C} + T_p} \right)^2 p(t - R_0) \tag{1}$$

$$\dot{q} = \begin{cases} -C + \frac{N}{\frac{q(t)}{C} + T_p} W(t) & \text{if } q(t) > 0 \\ \max \left\{ 0, \frac{N}{\frac{q(t)}{C} + T_p} W(t) - C \right\} & \text{if } q(t) = 0 \end{cases} \tag{2}$$

where

W : average TCP window size (packets),

\dot{q} : average queue length (packets),

R : round-trip time = $q/C + T_p$ (secs),

C : link capacity (packets/sec),

T_p : propagation delay (secs),

$N_{TCP} = N$: load factor (number of Westwood TCP sessions) and

p : probability of packet mark.

The queue length and window size are positive, bounded quantities, i.e., $q \in [0, \bar{q}]$ and $W \in [0, \bar{W}]$, where \bar{q} and \bar{W} denote buffer capacity and maximum window size, respectively.

2.2 Linearized Model

Although an AQM router is a non-linear system, a linearized model is used to analyze certain required properties and design controllers. To linearize (1) and (2), it is assumed that the number of active TCP sessions and the link capacity are constant: $N_{TCP}(t) = N_{TCP} = N$ and $C(t) = C$.

As described in [4, 14], the dependence of the time delay argument $t - R$ on the queue length q is ignored and it is assumed to be fixed at $t - R_0$. This is acceptable in the situations under study (as it is a good approximation when the round-trip time is dominated by the propagation delay, [11]). Local linearization of (1) and (2) around the operating point (q_0, p_0, W_0) results in the following equations:

$$\left. \begin{aligned} \partial \dot{W}(t) &= -\frac{N}{R_0 T_{q_0} C} \left(\partial W(t) + \partial W(t - R_0) \left(1 - 2 \frac{T_p}{R_0} \right) \right) \\ &\quad - \frac{1}{R_0^2 C} (\partial q(t) - \partial q(t - R_0)) \frac{R_0 - 2T_p}{T_{q_0}} \\ &\quad - \frac{R_0^2 C}{N^2} \left(1 - \frac{T_p}{R_0} \right) \partial p(t - R_0) \\ \partial \dot{q}(t) &= \frac{N}{R_0} \partial W(t) - \frac{1}{R_0} \partial q(t) \end{aligned} \right\} \quad (3)$$

where $T_{q_0} = R_0 - T_p$ and $\partial \dot{W}(t) = W - W_0$ and $\partial p = p - p_0$, represent the perturbed variables. Taking (W, q) as the state and p as input, the operating point (q_0, p_0, W_0) is defined by $\dot{W} = 0$ and $\dot{q} = 0$, that is,

$$\dot{W} = 0 \Rightarrow W_0 = -\frac{R_0 (R_0 - T_p)^2}{N^3 (2R_0 - T_p)} p_0, \dot{q} = 0 \Rightarrow q_0 = N \cdot W_0 \quad (4)$$

Equation (3) can be further simplified by separating the low frequency ('nominal') behavior $P(s)$ of the window dynamic from the high frequency behavior $\Delta(s)$, which is considered parasitic. Taking (4) as the starting point and following steps similar to those in [4], we can obtain the feedback control system (Fig. 1) for TCPW/AQM ([13, 15]).

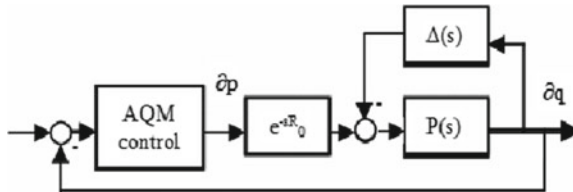


Fig. 1 Block diagram of AQM as a feedback control system

The action implemented by an AQM control law is to mark packets with a discard probability $p(t)$ as a function of the measured queue length $q(t)$. Thus, the larger the queue, the greater the discard probability becomes.

$$P(s) = \frac{\frac{-C^2}{N} \left(1 - \frac{T_p}{R_0}\right)}{s^2 + \frac{C \cdot R_0 + 2N}{C \cdot R_0^2} s + \left(\frac{N}{C \cdot R_0^3} \frac{2R_0 - T_p}{R_0 - T_p}\right)} \tag{5}$$

$$\Delta(s) = \frac{1}{\beta} \frac{1}{R_0 C} (1 - e^{-sR_0}) \quad \beta = \frac{C^2 R_0}{N^2} \left(1 - \frac{T_p}{R_0}\right) \tag{6}$$

3 Controller Description

This section presents the approach that has been followed to deal with the AQM congestion control problem in a network with wireless links under varying traffic. The round trip delay R_0 and the number of users N_{TCP} generally vary, whereas the link capacity C seldom changes. The proposed technique is straightforward, but works well in networks with a wide range of users and varying delays. The changes in traffic disturb the performance of the system. The controller (for example, a PID) is tuned to work well for a certain set of conditions, but the network parameters change and the performance decays. The method proposed here will improve the router’s operation no matter how many users or how much delay there is in the network.

The PID is tuned following a model-based approach, taking the linearized dynamic model of TCPW described in the previous section as the starting point. Figure 2 depicts the block diagram corresponding to the feedback control approach followed in the chapter: $P(s)$ is the transfer function obtained in Sect. 3. The delay e^{-sR_0} is explicitly considered in the design. There are two elements in the controller $C(s)$: $C_{PID}(s)$ (a PID controller) and K_1 (a variable gain: a simple gain scheduling that allows the controller to work satisfactorily in a broad range of situations).

The steps in the design are:

1. Choose the worst network scenario: the biggest R_0 and the smallest N_{TCP} .
2. Choose the best network scenario: the smallest R_0 and the biggest N_{TCP} .
3. Design a controller $C_{PID}(s)$ that is stable in these two situations and the scenarios in between.
4. Add a simple gain scheduling K_1 to improve the system’s response, based on the expected variations of the system’s gain.

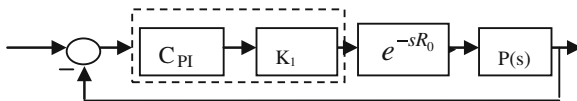


Fig. 2 Block diagram of the AQM feedback control system

We have chosen the PID controller (Eq. 7, see [6]) as the AQM congestion control algorithm, as it is the most common form of feedback control technique and relevant results can be found in the literature regarding its stability properties and tuning. The structure of the controller is:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt} \quad (7)$$

The PID should be tuned to be stable in the above mentioned situations. It is also required to have good responses in terms of settling time and overshoot. Thus, the tuning of the controller is done following the method presented in [7]: the system can be of any order and the roots can be real or complex. Using the provided Matlab toolbox, a three dimensional region is obtained. K_p , K_i and K_d chosen inside this region give a stable closed-loop response. The simple gain scheduling approach proposed in the chapter takes into consideration the big gain variation that the AQM dynamic model has:

$$\frac{-C^2}{N} \left(1 - \frac{T_p}{R_0} \right) \quad (8)$$

Moreover, if a fair comparison of all the network configurations were done, the open-loop gain of the systems should be the same, so the independent term of the denominator of the transfer function should be included:

$$\frac{N}{C \cdot R_0^3} \frac{2R_0 - T_p}{R_0 - T_p} \quad (9)$$

This term greatly affects the size of the stable region and the magnitude of the PID's tuning parameters (as will be shown in the next section). So we take out this term from $P(s)$ and calculate the stable region for:

$$P_{normalized}(s) = P_n(s) = \frac{\frac{N}{C \cdot R_0^3} \frac{2R_0 - T_p}{R_0 - T_p}}{s^2 + a \cdot s + \frac{N}{C \cdot R_0^3} \frac{2R_0 - T_p}{R_0 - T_p}} \quad (10)$$

And:

$$P(s) = \frac{-C^3 R_0^2 (R_0 - T_p)^2}{N^2 (2R_0 - T_p)} \cdot P_n(s) = P_{cte} \cdot P_n(s) \quad (11)$$

So, the gain scheduling term is given by (12):

$$K_1 = -\frac{2R_0 - T_p}{(R_0 - T_p)^2} \frac{N^2}{C^3 R_0^2} \quad (12)$$

It is important to choose the worst and best scenarios properly. So a good knowledge of the network is required. The network parameters are defined in Sect. 4. The stable regions are depicted and the PID is tuned. Then, linear and non-linear simulations show the goodness of the method.

4 Simulations

This section presents how to tune the controller and the tests that have been carried out. First, the controller is validated using the transfer function model. Then, the non-linear model defined by Eqs. (1) and (2) is used. Finally the controller is implemented in ns-2 (the non-linear network simulator) and tested. The results are rather promising in the three working environments.

4.1 Tuning the Controller

The basic network topology used as an example to test the controller is depicted in Fig. 3. It is a typical single bottleneck topology. The link capacity is kept constant in all experiments. The different scenarios (Table 1) come from changing the number of users (N) and the round trip time (R_0).

The situations deal with a broad range of conditions: few users and big delay, many users and small delay and in-between situations. In [9], a detailed discussion on how

Fig. 3 Dumbbell topology

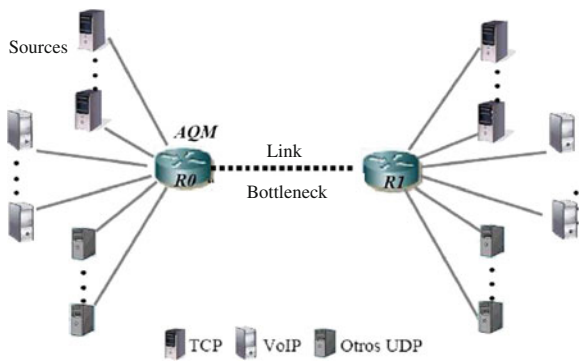


Table 1 Network configurations

	N	R_0
Best case	80	0.1
Extreme worst case	40	0.9
Working case	60	0.25
Worst case	45	0.45

the poles and zeros of the different scenarios are placed on the s-plane is studied. The open-loop step response for each of the studied settings of the original linear system ($P(s)$), and the normalized one ($P_n(s)$), have the same settling time [9], but the behavior once the loop is closed will be very different. The worst scenario gives the slowest open-loop response (settling time: 14.7 s) and the best scenario gives the fastest one (settling time: 0.74 s).

Let us consider the normalized transfer function approach. Using Hobenbichler’s technique [7], the 3-dimensional stable region for the PID is obtained. If the PID parameters (K_p , K_i and K_d) are chosen inside these regions, the closed loop response will be stable. Due to the normalization procedure, choosing these parameters is easier than for the standard PID. The same 3-dimensional representation can be obtained for the standard approach, where the scales are in the range of 10^{-3} – 10^{-4} . Moreover, the PID controller will give a worse performance, as is shown in the simulations. It is more difficult to see where all the regions overlap. A detailed discussion is given in [9].

4.2 Linear Simulations

The controller’s parameters, when the method proposed in the chapter is applied, are chosen as $K_p = 0.5$, $k_i = 0.3$ $k_d = 0.01$. Figure 4 shows the closed loop step response and settling time (black marker) of the linear systems for the four scenarios under study. The settling time is small, and smaller than 30 s in all the situations.

If the standard approach is followed, the controller parameters are $K_p = -4 \cdot 10^{-5}$, $k_i = -10^{-5}$ and $k_d = -2 \cdot 10^{-5}$. As shown in Fig. 5, the settling times are very different. The best case scenario is the slowest (settling time around 250 s) and the worst extreme is the fastest (10 s). These settling times are very different: there

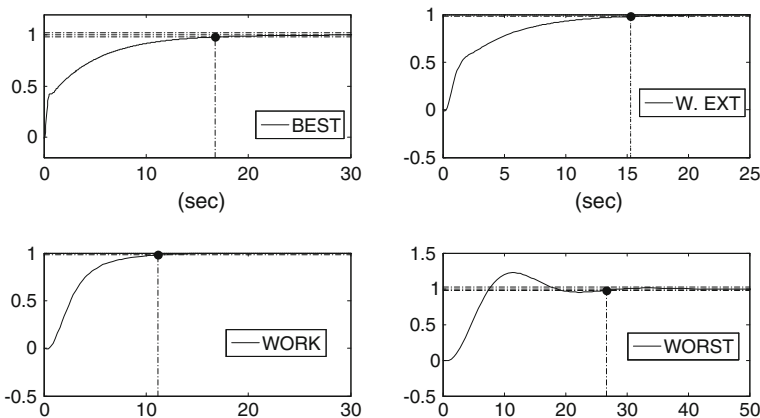


Fig. 4 Closed-loop step response with proposed method

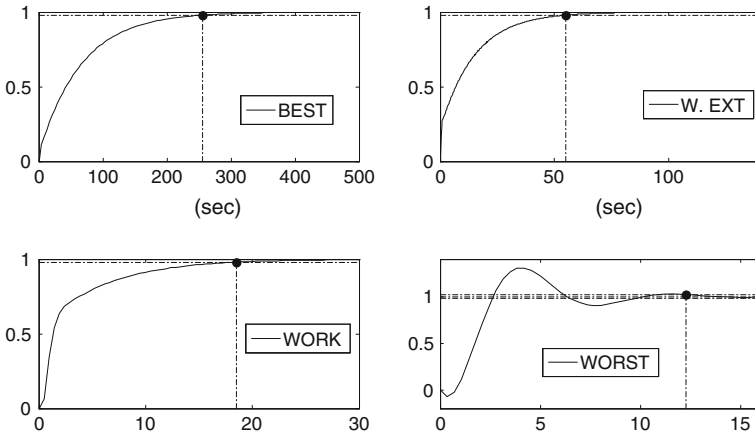


Fig. 5 Closed-loop step response with standard approach

are two orders of magnitude between them. It is not, therefore, feasible to have the same PID controller giving good closed-loop responses.

As can be seen, using the approach presented in the chapter (Fig. 4), the settling times have the same order of magnitude and the same PID can work well for all the scenarios.

4.3 Nonlinear Simulations in Simulink/Matlab

Taking the model described in Sect. 2 by Eqs. 1 and 2; the following experiments were carried out. In the first experiment, the reference is set at 230 packets (30 above the working point). Looking at the probability of marking a packet, there are not many differences between the two approaches (Fig. 6). However, the queue evolution

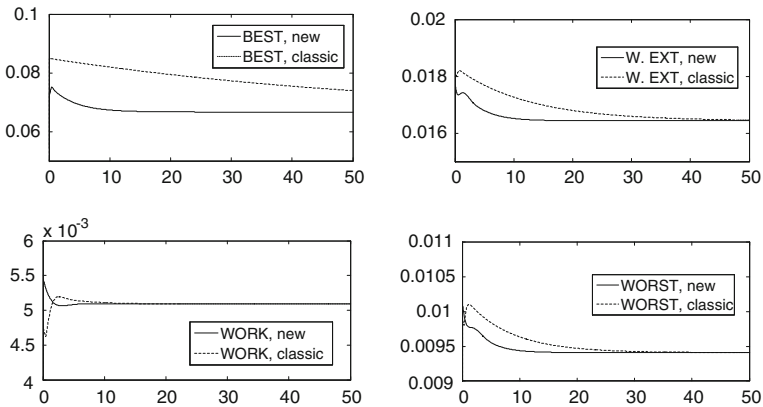


Fig. 6 Experiment 1, probability with new and standard methods

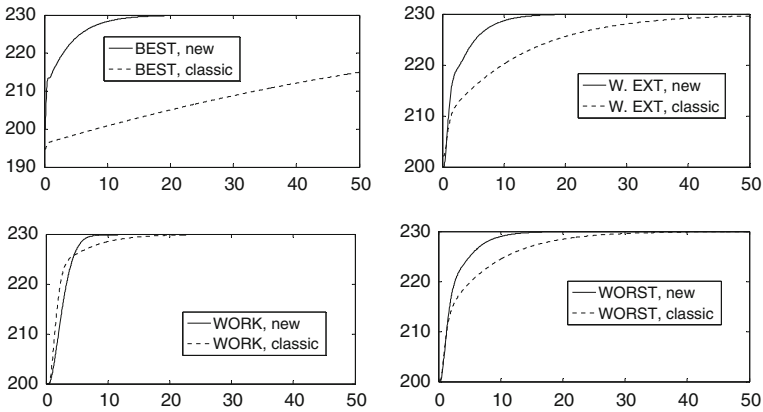


Fig. 7 Experiment 1, queue with new and standard approach

(Fig. 7) shows significant differences. It should be noted that the settling time when the proposed method is applied is smaller than when the standard approach is followed. As expected, the slowest response corresponds to the system with the biggest number of users.

In experiment 2, a variable reference was applied. First, the queue set-point (in packets) was increased by 30, then by 60 and finally decreased by 80 packets. All the plants performed well (Figs. 8, 9). The performance of the system follows the same patterns as in the first experiment. Figure 8 shows the evolution of the probability for the proposed and standard method.

The peaks are smaller than when using the standard approach (Fig. 8). Figure 9 shows the queue values with the modified and classic PID for each case under study.

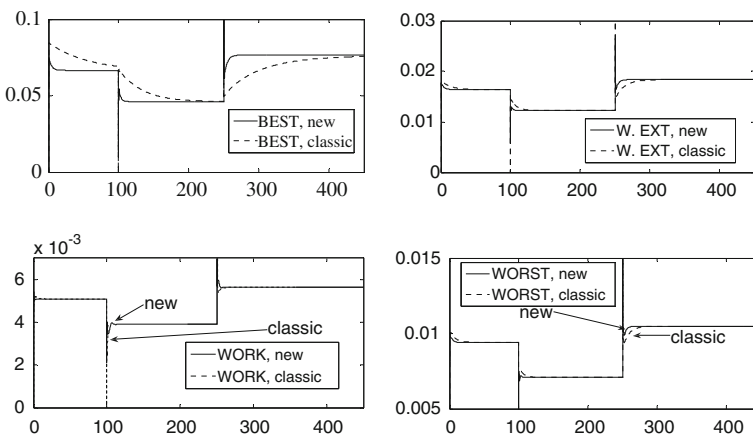


Fig. 8 Experiment 2, new and classic PID: Probability

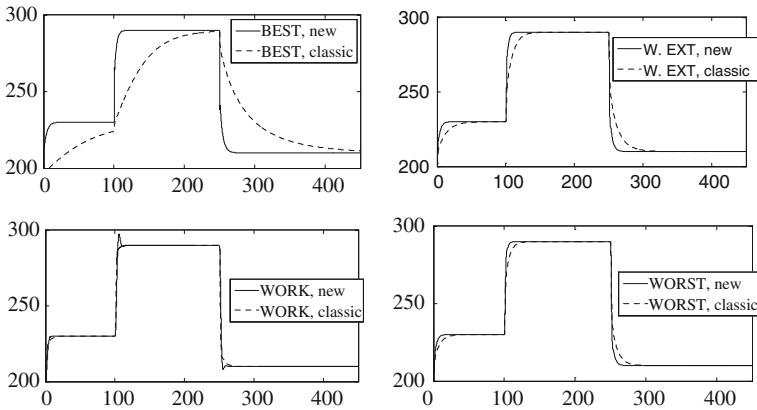


Fig. 9 Experiment 2, queue with new and classic approach

Clearly, the settling times are smaller with the proposed method than when the standard approach is followed (as happened in the first experiment). Again, the slowest response corresponds to the system with the biggest number of users.

Finally, let us consider an experiment where the number of users changes dynamically. The network default configuration is the working scenario described in Table 1. The set point does not change and is equal to the nominal value of the queue (200 packets).

The changes in the number of users are shown in Table 2 and the probability and queue size for the standard and proposed approaches in Figs. 10 and 11, respectively. Looking at Fig. 10, it can be concluded that the changes and the real value of the probability are smaller with the new method. Smaller probability values mean that the probability of discarding packets is smaller. As shown in Figs. 10 and 11, the standard approach cannot handle the sudden change in the number of users, whereas the new approach recovers very fast from the variations. The queue has almost no changes with the proposed PID, but the changes are huge with the standard approach.

The mean, standard deviation and variance for the queue size following the standard and new approaches are shown in Table 3. The reference is set at 200 packets and the mean value for the proposed methodology is 201.4 packets. This is a value very close to the reference, especially if it is compared with the mean of 210 packets of the standard approach.

Table 2 Changing N

Time N	
20	47
40	60
60	72
80	78
100	65

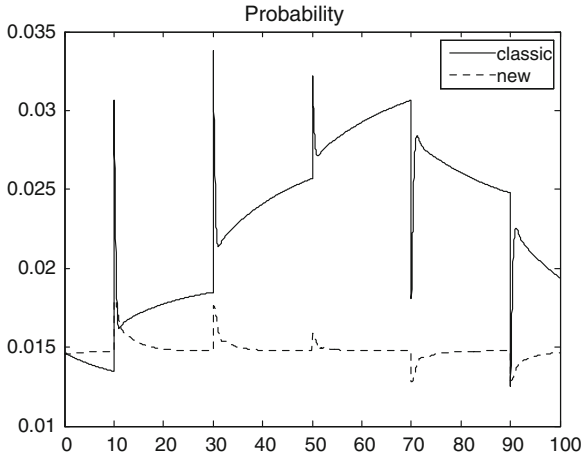


Fig. 10 Probability when the number of users changes

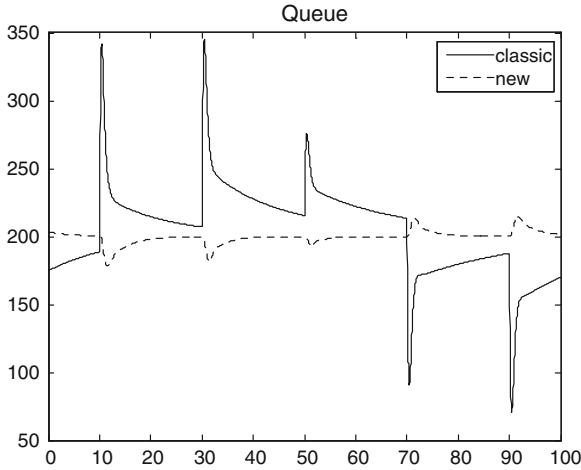


Fig. 11 Queue size when the number of users changes

Table 3 Mean and variance

	Mean	St. deviation	Variance
Standard	210.5	37.3	1393.4
New	201.4	31.9	1019.7

4.4 *Ns-2 Nonlinear Simulations*

Finally, the proposed technique was migrated to ns-2. This simulation tool provides a realistic environment for testing techniques. The same set of four situations (see Table 1) was considered. Again the new technique and the standard PID were

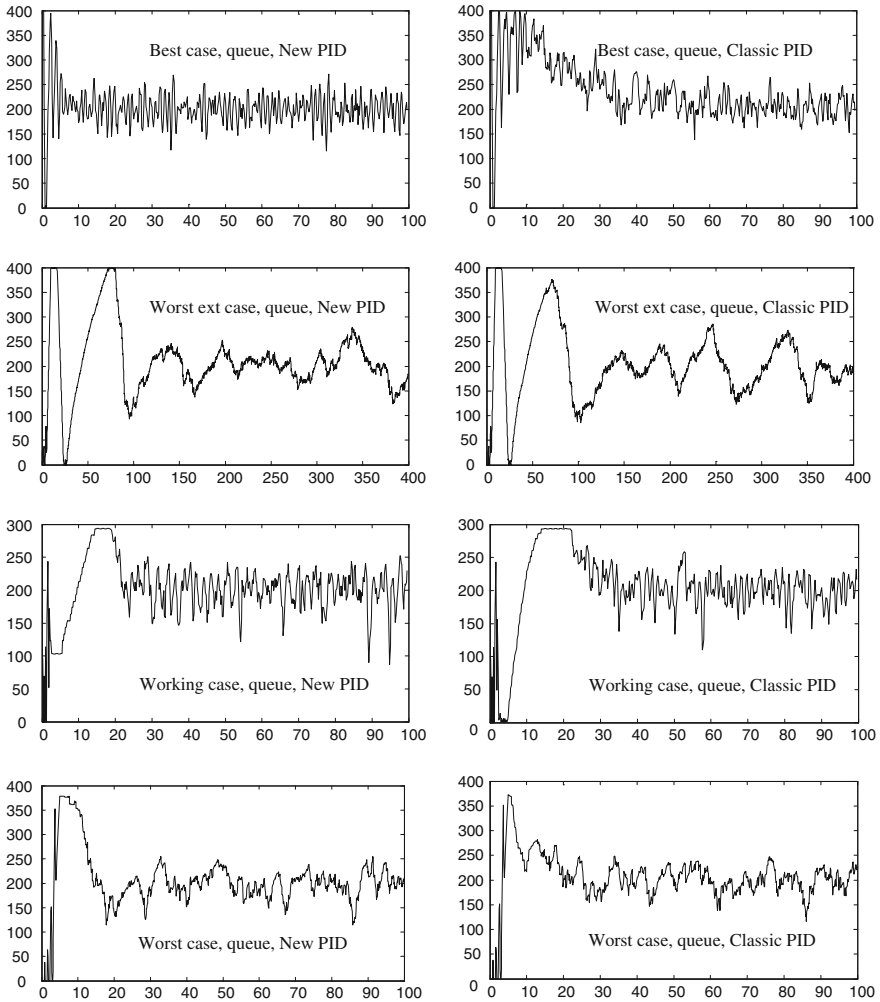


Fig. 12 Queue size in ns-2

Table 4 Queue mean and variance: ns-2 experiment

	Best		Wor Ext		Working		Worst	
	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Standard	219.1	31.6	200.5	56.32	207.9	29.5	199.8	22.42
New	199.8	25.7	206.3	56.84	201.6	26.8	197.8	25.31

compared. The tendency of the results support the conclusions obtained in previous sub-sections.

Figure 12 shows the queue evolution when a set point of 200 packets is considered. The queue follows the reference better when using the new PID (softer lines, Fig. 12, Table 4). Although the probability of marking the packets changes more abruptly with the new approach, this is not a big issue, because this is a software system opposed to physical systems such as deposits or valves.

5 Conclusion and Future Work

This chapter has presented a PID design with linear gain scheduling and normalized values that works very well under different network traffic conditions. The controller is tuned for the worst scenario and works properly in a wide range of situations.

As we are working with TCP Westwood, the analysis of the results is valid either in wired or wireless routers. Linear and nonlinear simulations confirm the technique's potential. This is especially convenient, as many design procedures are tested in restricted situations with no guarantee that the controller will behave similarly in other scenarios.

The technique presented in this chapter gives more uniform results than the standard approach. It does not matter if the scenario is close to the nominal situation or a worst case setting, the controller reacts adequately and the settling times are all of the same order. The standard approach cannot deal well with extreme situations. Results are encouraging and more work should be done.

Acknowledgments The authors would like to thank Prof. F. Tadeo for so many fruitful discussions.

References

1. Azuma T, Fujita T, Fujita M (2006) Congestion control for TCP/AQM networks using State Predictive Control. *Electr Eng Jpn* 156:1491–1496
2. Deng X, Yi S, Kesidis G, Das CR (2003) A control theoretic approach for designing adaptive AQM schemes. *GLOBECOM'03*, 5:2947–2951
3. Jacobson V (1988) Congestion avoidance and control. *ACM SIGCOMM'88*, Stanford
4. Ryu S, Rump C, Qiao C (2004) Advances in active queue management (AQM) based TCP congestion control. *Telecommun Syst* 25:317–351
5. Parthasarathy V, Anandakumar P, Rajamani V (2011) Design, simulation and FPGA implementation of a novel router for bulk flow TCP in optical IP networks. *IAENG Int J Comput Sci* 38(4):343–349
6. Aström KJ, Häggglund T (2006) *Advanced PID control*. ISA, North Carolina
7. Hohenbichler N (2009) All stabilizing PID controllers for time delay systems. *Automatica* 45:2678–2684
8. Silva G, Datta A, Bhattacharyya S (2005) *PID controllers for time-delay systems*. Birkhäuser

9. Alvarez T (2012) Design of PID controllers for TCP/AQM wireless networks, Lecture notes in engineering and computer science: proceedings of the world congress on engineering (2012) WCE 2012, 4–6 July 2012 U.K , London, pp 1273–1280
10. di Bernardo M, Grieco LA, Manfredi S, Mascolo S (2004) Design of robust AQM controllers for improved TCP Westwood congestion control. In: Proceedings of the 16th international symposium on mathematical theory of networks and systems (MTNS, 2004) Katholieke Universiteit Leuven, Belgium, July 2004
11. Wang R, Valla M, Sanadidi MY, Gerla M (2002) Adaptive bandwidth share estimation in TCP Westwood. In: Proceedings of IEE Globecom. Taipei, 2002
12. Hollot CV, Misra V, Towsley D, Gong W (2002) Analysis and design of controllers for AQM routers supporting TCP flows. *IEEE Trans Autom Control* 47:945–959
13. Hollot CV, Chait Y (2001) Nonlinear stability analysis for a class of TCP/AQM networks. In: Proceedings of the 40th IEEE conference on decision and control. Orlando, USA
14. Long GE, Fang B, Sun JS, Wang ZQ (2010) Novel graphical approach to analyze the stability of TCP/AQM networks. *Acta Automatica Sinica* 36:314–321
15. Alvarez T (2012) Designing and analysing controllers for AQM routers working with TCP Westwood protocol. (Internal Report, University of Valladolid. To be submitted)

On the Initial Network Topology Factor in Mobile Ad-Hoc Network

Ronit Nossenson and Adi Schwartz

Abstract The impact of the initial network topology on performance of routing algorithms is explored. Typically researchers use a randomly chosen network topology for performance evaluation of their protocols and algorithms. Here we show that the initial network topology can have a significant impact on algorithm performance and can lead to biased results, in particular, an initial topology that includes a major connectivity obstacle such as low connectivity level (e.g., a tree topology) or bridges. Although users move according to commonly implemented random mobility models, the effect of the initial topology can persist over time. To avoid biased results we recommend using multiple initial topologies instead of one, and/or running the simulation in an initialization phase until the effect of the initial topology fades.

Keywords MANET · MANET benchmark · Network topology · Performance evaluation · Routing algorithms · Traffic generation

1 Introduction

In wireless multi-hop ad-hoc networks each node not only acts as a possible source and sink of traffic, but also as a router enabling the forwarding of traffic between nodes in the network. Mobile routing protocols in such networks must be efficient and

R. Nossenson (✉)
Faculty of Computer Science, Jerusalem College of Technology (JCT),
Jerusalem, Israel
e-mail: nossenso@jct.ac.il

A. Schwartz
Faculty of Computer Science, The Academic College of Tel-Aviv,
Tel-Aviv, Israel
e-mail: adishh@gmail.com

must either keep their knowledge proactively up to date, or be highly reactive when routes are required. Over the past decade, hundreds of different routing protocols have been proposed for wireless ad-hoc networks.

The performance evaluation of routing algorithm is frequently carried out via a simulation on a randomly generated ad-hoc network topology. In addition, a mobility model is assumed to mimic the users' mobility. Researchers usually assume that using a randomly generated network topology with proper mobility is an objective environment to study the behavior of their protocols and algorithms. In this chapter we show that this common assumption is flawed and demonstrate that a connectivity obstacle in the initial network topology can bias performance results. This chapter is the extended version of our previous publication [1].

Previous works have studied the performance of routing protocols in ad-hoc networks; see for example [2–7] but none have investigated the impact of the initial network topology.

To understand the impact of the initial topology on performance we extended the Network Simulation 2 (NS2) [12] with a topology generator tool. The topology generator creates a network topology graph of nodes according to given prerequisites that makes it possible to create specific characteristic topologies such as a tree topology, a graph with a bridge edge, a well connected graph, etc.

Using the topology generation tool we generated three types of network topologies starting from a *tree* type having exactly one path between every two nodes, up to a *well connected* topology graph. We used three mobility scenarios: no mobility, low mobility and medium mobility. We tested the performance of four well known routing algorithms on these topologies and compared the results. The four algorithms were: Dynamic Source Routing (DSR) [8], Ad-hoc On-Demand Distance Vector (AODV) [9], Destination-Sequenced Distance Vector Routing (DSDV) [10], and Ad-Hoc On-demand Multipath Distance Vector (AOMDV) [11].

The first very striking finding was that a connectivity obstacle in the initial topology of the network resulted in lower performance results of all the algorithms we tested. Furthermore, the initial topology can bias the comparative results across algorithms. Furthermore, the impact of user mobility on algorithm performance varies as a function of the initial topology. In a well-connected initial topology, the algorithm performance decreases with increased user mobility whereas in an initial topology with connectivity obstacles performance can improve when user mobility increases.

The direct implication of these findings is that researchers must be aware of the fact that their randomly generated initial topology can bias performance results. To avoid this pitfall, we suggest using more than one random initial topology (that is, repeat the simulation several times with different initial topologies), and run the simulation in an initialization phase before starting to collect the statistics.

This chapter is organized as follows. In the next section we describe the four routing algorithms. Section 3 consists of the simulation description. The results are presented in Sect. 4. Finally, the implications are discussed in Sect. 5.

2 Routing Protocols

There are two main categories of protocols in ad hoc routing: proactive and reactive. Proactive routing protocols are based on ‘normal’ routing protocols used in wired networks, such as today’s Internet. Algorithms based on distance vectors or link states are commonplace. In proactive protocols, a local representation of routes at each node within the network is built before they are put into use. The routing tables are usually built up periodically through normal operation of the protocol by exchanging routing update packets. In normal operation, this has the advantage that the routes are already pre-computed and packet forwarding can take place as soon as a packet for a particular destination appears at a node. The drawback is that routes may be calculated and re-calculated (for example due to node mobility) when they are not actually required. This wastes bandwidth and, for mobile nodes, also wastes battery power because unnecessary routing updates are sent and received.

Reactive routing takes an alternative approach by building routes solely on demand. It can also cache route information according to some short time-out or staleness policy. Cached routes can be used as required, but if a route is not known then it has to be ‘discovered’. This has the advantage that routes are only evaluated when needed, although this approach adds latency to packet forwarding when routes are not already known.

2.1 *Dynamic Source Routing (DSR)*

The DSR routing algorithm is a reactive routing algorithm that uses source routing; in other words, the sender knows the complete route to the destination. These routes are stored in a route cache. The data packets carry the source route in the packet header. When a node in the ad hoc network attempts to send a data packet to a destination for which it does not already know the route, it uses a route discovery process to dynamically determine it. Route discovery works by flooding the network with route request (RREQ) packets. Each node receiving an RREQ rebroadcasts it, unless it is the destination or it has a route to the destination in its route cache. This node then replies to the RREQ with a route reply (RREP) packet that is routed back to the original source. If any link on a source route is broken, the source node is notified using a route error (RERR) packet and the route is removed from the cache [8].

2.2 *Ad-Hoc On-Demand Distance Vector (AODV)*

The AODV routing algorithm is a reactive routing algorithm. It discovers routes on an as-needed basis via a route discovery mechanism. It uses traditional routing tables, with one entry per destination. AODV relies on routing table entries to propagate an RREP back to the source and to route data packets to the destination. AODV

uses sequence numbers maintained at each destination to determine the freshness of routing information and to prevent routing loops. All routing packets carry these sequence numbers [9].

A routing table entry expires if not used recently. In addition, when a failure occurs, RERR packets in AODV are used to inform all sources using this link.

2.3 Destination Sequenced Distance Vector (DSDV)

The DSDV routing algorithm is a proactive routing algorithm. It is based on the Bellman-Ford routing algorithm with certain improvements. Every mobile station maintains a routing table that lists all available destinations, the number of hops to reach the destination and the sequence number assigned by the destination node. The sequence number is used to distinguish stale routes from new ones and thus avoid the formation of loops. The stations periodically transmit their routing tables to their immediate neighbors. A station also transmits its routing table if a significant change has occurred. Thus the update is both time-driven and event-driven. The routing table updates can be sent in two ways: full or incremental [10].

2.4 Ad-Hoc On-Demand Multipath Distance Vector (AODMV)

The AODMV is a reactive routing algorithm. It extends AODV to discover multiple link-disjoint paths between the source and the destination in every route discovery. It uses the routing information already available in the AODV protocol as much as possible. It makes use of AODV control packets with a few extra fields in the packet header such as advertised hop count and a route list which contains multiple paths [11].

3 Simulation Description

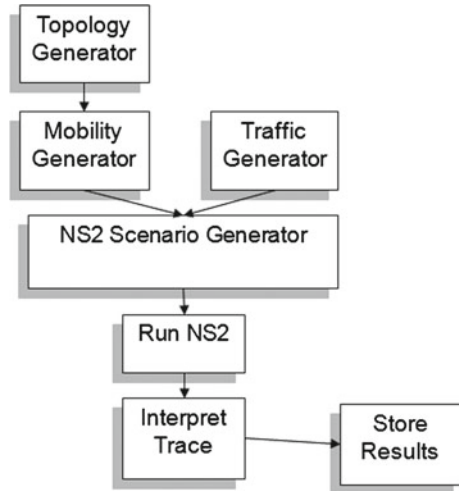
The simulation was performed using the NS2 simulator version 2.34 [12]. We compared the DSR [8], AODV [9], DSDV [10], and AODMV [11] routing algorithms under different initial topologies and mobility scenarios.

3.1 The Simulation Process

The simulation process consisted of the following steps as shown in Fig. 1.

1. The topology generator creates a network topology graph of nodes according to given prerequisites that create specific characteristic topologies (described in the next sub-section).

Fig. 1 The simulation process



2. The mobility generator uses the CANU mobility simulator [5]. It creates user profiles: static, slow walking and vehicular movement.
3. The traffic generator creates different traffic types. It creates VoIP traffic, using Constant Bit Rate, with small packets transmitted over UDP. Web-browsing traffic is created using an ON/OFF process with a Pareto distribution, with large packets transmitted over TCP. In addition, it creates video traffic using a smoother Pareto ON/OFF process with large packets transmitted over TCP.
4. The NS-2 scenario generator composites the mobility and traffic generated data into a specific NS-2 scenario template. The NS-2 scenario template runs the same scenario with the same initial network topology on different routing algorithm, leaving the routing algorithm as a parameter.
5. The trace interpreter reads the NS-2 generated trace file and extracts performance related information such as throughput, average drop rate and average packet delay.

3.2 Topology Characteristics

To study the impact of the initial network topology we extended the NS-2 to include a topology generator tool. The topology generator generates a random ad-hoc network topology based on the prerequisite list below. Each of the tested topologies was used as a starting point for the mobile nodes (before they started moving according to the mobility scenario). We created ten different instances of each topology type. All the generated topologies had the same number of nodes and were located in the same area size.

This random topology generator function can create the following three types of characterized topologies:

1. A tree
2. A bridged graph, having a link such that removing it decouples the graph into two connected graphs
3. A well connected graph having two disjoint paths from each node to each node

3.3 Traffic Generation

The simulator modeled three different traffic generator functions:

1. The VoIP traffic generator was a Constant Bit Rate application with a 100 kbps connection based on a UDP connection.
2. The Web browsing traffic generator was a Pareto application over TCP with a packet size of 1000 bytes, an average burst time (ONN period) of 100 ms, an average idle time (OFF period) of 700 ms, a rate of 500 kbps and a shape parameter (alpha) of 1.1.
3. The video traffic generator was also a Pareto application with a shape parameter of 2.2, an ON period average of 500 ms, an average idle time (OFF period) of 100 ms and a rate of 500 kbps.

We used a mixed traffic type scenario with 1/3 of the nodes generating VoIP traffic, 1/6 generating web browsing traffic and another 1/6 generating video-like traffic (the other nodes do not generate traffic).

3.4 The Mobility Generator

The mobility generator takes the topology area that was produced by the topology generator and generates a mobility scenario using the CANU mobility simulator [13]. There are three different mobility profiles for a given user:

1. A static user: without any movement
2. A random walking user with a speed of 1–5 kmh and a 5–50 s pause duration
3. A random driving user with a speed of 30–120 kmh and a 0–25 s pause duration

We considered three mobility scenarios in our simulation: (i) a no-mobility scenario, (ii) a low-mobility scenario in which 2/3 of the users are static and 1/3 have a random walking profile; and (iii) a medium-mobility scenario in which 1/3 of the users are static, 1/3 have a random walking profile and 1/3 have a random driving profile.

3.5 The Network Interface

The network interface was set to wireless PHY with a propagation distance of 100m and a radio propagation model of Two Ray Ground with an Omni antenna. The Queue is a priority queue that gives a higher priority to routing protocol packets, the queue size was set to 50 and the MAC was based on the 802_11 wireless LAN.

4 Results

The key simulation results are presented in this section. The impact of the initial network topology on the packet drop percentages is presented in Figs. 2, 3, 4, 5. As expected, initial topologies of weakly connected graphs such as the tree-graph and the bridge-graph increase the packet drop percentages in all algorithms. The DSR outperformed the other algorithms and maintained low packet drop percentages.

Strikingly, increasing the mobility level of the users had a differential impact on the drop percentages as a function of the initial network topology. For initial network topologies that are well connected, the increase in user mobility generally resulted in an increase in the drop percentage. For example in the DSDV algorithm (Fig. 3), increasing the mobility level from low to medium with a well connected initial topology resulted in drop percentages of 33 % instead of 14 %. By contrast, for initial network topologies with connectivity obstacles, the reverse was true: the increase in user mobility from low to medium mobility resulted in a decrease in the drop percentages. For example in the DSDV algorithm (Fig. 3), increasing the

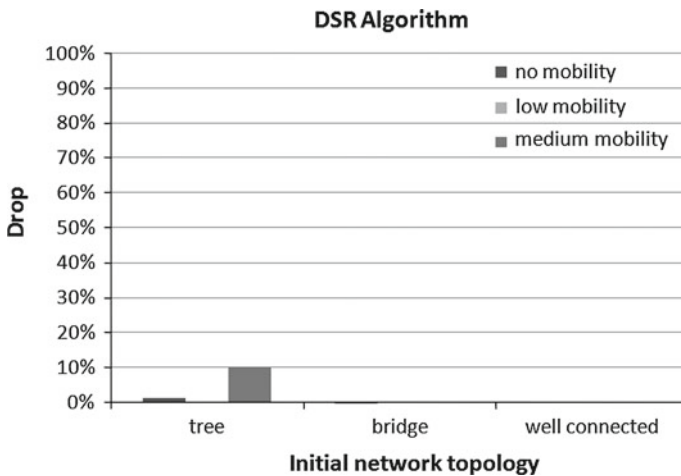


Fig. 2 DSR drops with different initial network topologies

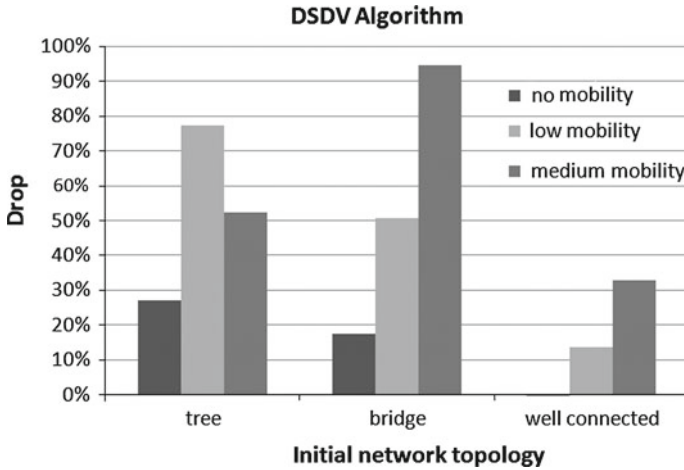


Fig. 3 DSDV drops with different initial network topologies

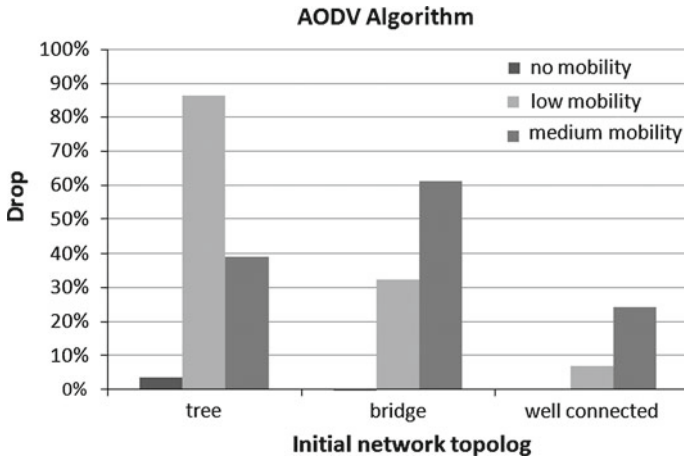


Fig. 4 AODV drops with different initial network topologies

mobility level from low to medium in the tree topology resulted in drop percentages of 52 % instead of 77 %.

Another interesting observation is that the initial network topology can change the comparative findings. For example using an initial network topology with a bridge edge in the low mobility scenario resulted in 33 % packet drop in the AODV algorithm which thus outperformed the DSDV algorithm which had a 51 % packet drop. But using an initial network topology of a tree (in the low mobility scenario) reversed these results. The DSDV algorithm had a 77 % packet drop, and thus outperformed the AODV algorithm which had a 86 % packet drop.

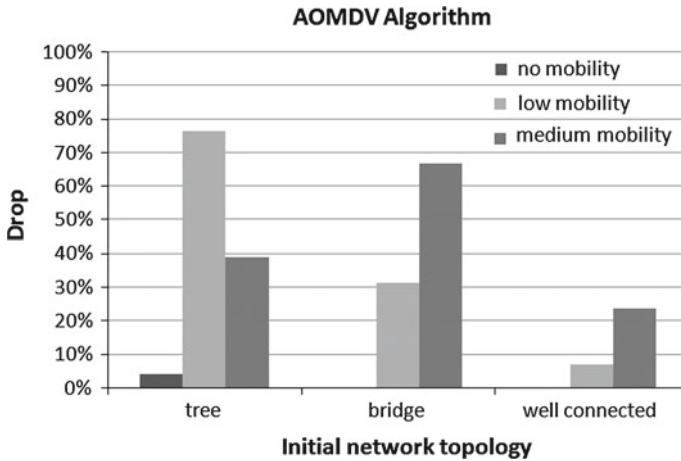


Fig. 5 AOMDV drops with different initial network topologies

5 Conclusion and Future Work

This chapter explored the impact of initial network topology on the performance evaluation of four well known routing protocols for mobile ad-hoc networks. The results show that the initial network topology has a substantial impact on performance. If the initial network has a connectivity obstacle, performance is much worse than when the initial network is well connected. Algorithms that outperform other algorithms with well connected initial network topologies might not do so in different initial network topologies with connectivity obstacles. Increasing the user mobility level usually results in a decline in performance of routing algorithms in simulations that start with a well connected network topology. However when the simulation starts with a weakly connected network topology, increasing in the user mobility level enhances performance. All these findings strongly suggest that the initial network topology is an important factor in an objective simulation environment. Researchers should be aware of its influence to avoid biased results.

Acknowledgments We thank Gabriel Scalosub for useful discussions.

References

1. Nossenson R, Schwartz A (2012) The impact of initial network topology on performance of routing algorithms in MANETs. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012, pp 1269–1272.
2. Abolhasan M, Hagelstein B, Wang JC-P (2009) Real-world performance of current proactive multi-hop mesh protocols. <http://ro.uow.edu.au/infopapers/736>

3. Murray D, Dixon M, Koziniec T (2010) An experimental comparison of routing protocols in multi hop ad hoc networks. In: Proceedings of telecommunication networks and applications conference (ATNAC), 2010 Australasian, 31 Oct 2010–3 Nov 2010, pp 159–164.
4. Shah S, Khandre A, Shirole M, Bhole G (2008) Performance evaluation of ad hoc routing protocols using NS2 simulation. In: Mobile and pervasive computing (CoMPC)
5. Timcenko V, Stojanovic M, Rakas SB (2009) MANET routing protocols vs. mobility models: performance analysis and comparison. In: Proceedings of the 9th WSEAS international conference on applied informatics and communications (AIC '09).
6. Johansson P, Larsson T, Hedman N (1999) Scenario-based performance analysis of routing protocols for mobile ad-hoc networks. In: Mobicom '99, Washington, USA.
7. Barakovi S, Kasapovi S, Barakovi J (2010) Comparison of MANET routing protocols in different traffic and mobility models. *Telfor J* 2(1):8–10
8. Johnson D, Maltz D (1996) Dynamic source routing in ad hoc wireless networks. In: Imielinski T, Korth H (eds) *Mobile computing*, Chap. 5. Kluwer Academic, Dordrecht, pp 153–181.
9. Perkins C, Belding-Royer E, Das S (2007) Ad hoc on-demand distance vector (AODV) routing. IETF mobile ad hoc networking working group.
10. Perkins C, Bhagwat P (1994) Highly dynamic destination-sequenced distance vector routing (DSDV) for mobile computers. In: Proceedings of ACM SIGCOMM'94.
11. Marina M, Das S (2001) On-demand multipath distance vector routing for ad hoc networks. In: Proceedings of 9th IEEE international conference on network protocols, pp 14–23.
12. The NS-2 network simulator–NS2 website: <http://www.isi.edu/nsnam/ns/>
13. CANU Mobility Simulator Environment Website: <http://canu.informatik.uni-stuttgart.de/mobisim/>

Internet Key Exchange Protocol Using ECC-Based Public Key Certificate

Sangram Ray and G. P. Biswas

Abstract Internet Key Exchange (IKE) protocol helps to exchange cryptographic techniques and keying materials as prior security association (SA) between two IP hosts. Similar to the several enhancements, the present paper proposes an efficient implementation of IKE using ECC-based public-key certificate that provides required security properties with much reduction in computation complexity and communication cost. The proposed method addresses both the Phase I and Phase II of IKE, where the main mode of the former instead of six, requires four rounds of message exchange. The formats specified in ISAKMP have been used for message exchanges in our implementation, thus the cookies of initiator-responder have been used to protect attacks like DoS, parallel session etc. The security analysis of the proposed method and comparison with other techniques are given and satisfactory performance is found.

Keywords Certificate authority (CA) · Elliptic Curve Cryptography (ECC) · Internet Key Exchange (IKE) protocol · Internet Security Association and Key Management Protocol (ISAKMP) · IP Security (IPSec) · Public Key Infrastructure (PKI) · Security Association (SA)

1 Introduction

The secure communication over the Internet has become increasingly important and since it has no self security, the IPSec protocol has been developed for the protection of the same. IPSec comprises two protocols called AH (Authentication

S. Ray (✉) · G. P. Biswas
Department of Computer Science and Engineering,
Indian School of Mines, Dhanbad 826004, India
e-mail: sangram.ism@gmail.com

G. P. Biswas
e-mail: gpbiswas@gmail.com

Header) and ESP (Encapsulated Security Payload) that provide security to the IP data packet and source authentication, data confidentiality and integrity, respectively with provision for protection against replay attack in both the security protocols. The Internet Key Exchange (IKE) protocol [1–13] designed based on Oakley [6] and ISAKMP [7] is usually used for security association (SA) to be used by two end-host machines for secure transmission of IP packet over Internet [4]. That is, the IKE protocol provides SA to the IPsec through negotiating of MAC, encryption algorithms to be used together with the required secret keys for providing security to the IP datagrams. In essence, the SA comprises two communication entities/security gateways for their mutual authentication, the generation of shared secret session keys, and the negotiation and exchange confidential parameters between them. The parameters exchanged are SA lifetime, sequence number counter, security parameter index (SPI), IPsec mode, different cryptographic techniques, key related materials etc. [12, 13].

The IKE protocol works in two phases: Phase I starts at the beginning to provide SA for the Phase II, which finally provides the SA to IPsec. In phase I, after SA-offered by initiator and negotiated by responder, a common secret key (SKEYID) is established between two entities, from which different required secret keys are derived. In Phase II, the SA for IPsec is established and the final keying materials are generated. In fact, phase I is implemented in two different modes namely Main Mode and Aggressive Mode with six and three messages exchange, respectively. The four different authentication methods for implementation of Phase I are defined in [2, 3] and they are namely pre-shared key, public key signature, public key encryption and revised public key encryption techniques. The phase II is normally implemented using a single mode called “quick mode”, which uses three messages for establishing the SA of IPsec protocol [2, 3].

In fact, each of the methods used in main-mode implementation directly or indirectly requires a public key certificate issued by a CA (certificate authority) and for this, a RSA based public key certificate is mainly used. But the overhead for maintaining and using it is much higher than an ECC based public key certificate with comparable security. This is because it involves scalar point multiplication and it is seen that an ECC based cryptosystem [14–16] with 160-bit key size provides equivalent security of the RSA cryptosystem with 1024-bit key size [14–16]. Thus designing IKE using ECC is more efficient and cost-effective.

The Public-Key Infrastructure (PKI-X.509) working group of Internet Engineering Task Force (IETF) similar to the X.509 RSA based public key certificate [17–19], provides standardized ECC based public key certificate as the PKI-X.509, which specified as PKIX. Subsequently, the Elliptic Curve Digital Signature Algorithm (ECDSA), Elliptic Curve Diffie-Hellman (ECDH) public keys and the generation of ECC based certificate with ECDSA signature of PKIX are proposed in [16, 19]. It may be noted that the PKIX is easily interoperable with PKI (X.509) and the Certificate Authority (CA) issues and certifies both the certificates. Thus, the efficient ECC based certificate scheme PKIX can be implemented on the existing PKI infrastructure, and the present paper without additional overheads uses it for the implementation of the both phases of IKE protocol.

In brief, this paper addresses the development of a secure and efficient implementation of IKE protocol based on PKIX using elliptic curve cryptosystem. We assume that the existing tree-type hierarchical model [17, 18] for CAs with PKI is capable for creating, storing, issuing and revoking any number of PKIX certificates. This model verifies the ECC based public keys of any entity in a chaining fashion from leaf nodes towards the root of the tree, where the root CA, which has self-signed, self-issued certificate, completes the verification processes. All intermediate CAs issue certificates to the entities to relief the burden of the root CA and participate in the chaining verification process as mentioned above.

The remaining parts of the paper are organized as follow: Sect. 2 briefly introduces the basics of the Elliptic Curve Cryptography (ECC) with its computational problems and ECC-based public key certificate. In Sect. 3, the ECC-based IKE protocol is proposed. The security analysis of proposed scheme against different related attacks and a comparison with existing RSA based schemes is given in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Elliptic Curve Cryptography

The elliptic curve cryptosystem was initially proposed by Koblitz [14] and then Miller [15] in 1985 to design public key cryptosystem and presently, it becomes an integral part of the modern cryptography. Let E/F_p , which denotes an elliptic curve E over a prime finite field F_p , can be defined as $y^2 = x^3 + ax + b$, where $a, b \in F_p$ and the discriminant

$$D = 4a^3 + 27b^2 \neq 0$$

The points on E/F_p together with an extra point O called the point at infinity used for additive identity form an additive group A as

$$A = \{(x, y) : x, y \in F_p, E(x, y) = 0\} \cup \{O\}$$

Let the order of A is n , which is very large and can be defined as $n \times G \text{ mod } q = O$, where G is the generator of A . The A be a cyclic additive group under the point addition “+” defined as follows: $P + O = P$, where $P \in A$. The scalar point multiplication over A can be defined as $tP = P + P + \dots + P$ (t times). If $P, Q \in A$, the addition $P + Q$ be a point $-R$ (whose inverse is R with only changing the sign of y coordinate value and lies on the curve) on the E/F_p such that all the points P, Q and $-R$ lie on the straight line. Note that if $P = Q$, it becomes a tangent at P or Q that intersects the curve at the point $-R$. The security strength of the ECC lies on the difficulty of solving the Elliptic Curve Discrete Logarithm Problem (ECDLP) [14, 15]. An overview of ECC is given below.

2.1 Computational Problems

Similar to the DH problem [20], some computational hard problems on ECC are defined below:

- **Elliptic curve discrete logarithm problem (ECDLP)**
Given $Q, R \in A$, find an integer $k \in F_p^*$ such that $R = k \cdot Q$.
- **Computational Diffie-Hellman assumption (CDHA)**
Given $P, xP, yP \in A$, it is hard to compute $xyP \in A$.
- **Decisional Diffie-Hellman problem (DDHP)**
Given $P, aP, bP, cP \in G$ for any $a, b, c \in F_p^*$, decide $cP = abP$ or not.

2.2 ECC Based Certificate

As stated earlier, the ECC based certificate has been standardized by IETE as PKIX-X.509, which is almost similar to X.509 with a main difference of using ECC based public key signed by the ECDSA. A simple ECC-based X.509 certificate format [16] to combine user’s identity and the ECC-based public key proposed by ITU is described in Fig. 1.

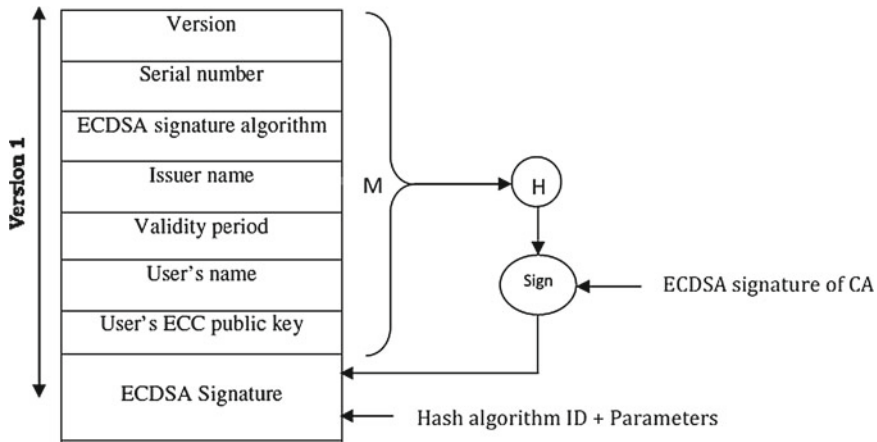


Fig. 1 ECC based X.509 certificate format

3 Proposed ECC Based IKE Protocol

In this section, the proposed two phase scheme requires for IKE protocol is discussed below, where the following notations are used:

- p, n Two large prime numbers;
- F_p A finite field;
- E An elliptic curve defined on F_p with prime order n ;
- G The group of elliptic curve points on E ;
- P A point on elliptic curve E with order n ;
- $H(_)$ One-way hash function (e.g. SHA1, MD5);
- I Initiator;
- R Responder;
- HDR ISAKMP-Header;
- SA_{PROP} Security association proposal of I ;
- SA_{SELEC} Security association selected by R ;
- ID_I Identity of initiator I ;
- ID_R Identity of responder R ;
- CA_I Public key certificate of initiator I ;
- CA_R Public key certificate of responder R ;
- (s_I, V_I) Private/public key pair of I , where $V_I = s_I \cdot P$;
- (s_R, V_R) Private/public key pair of R , where $V_R = s_R \cdot P$;

3.1 Proposed ECC Based Main Mode Protocol

The initiator I and the responder R initially collects ECC based public key certificate from CA and start to exchange four messages in phase I for SA negotiation and establishment of a secure key similar to SKEYID without requiring any pre-shared secret or others as used in the main mode of original IKE. The corresponding message-flow diagram is shown in Fig. 2. Note that each certificate contains a user's public key and the signature of the CA over the hash value of the public key, user-ID, issue-date, issuer-name etc using ECDSA algorithm.

The messages 1–3 as shown are the parameters for negotiation as well as the mutual authentication of I and R through ECC-based public key certificate, where each message is preceded by HDR , the standard ISAKMP message format that contains the information required by the protocol to maintain state, process payloads, and possibly to prevent *replay attack* and *denial-of-service attack*. The ISAKMP-Header [7] contains the following fields:

Initiator's Cookie (8 bytes); Responder's Cookie (8 bytes); Next Payload (1 byte); Major Version (4 bits); Minor Version (4 bits); Exchange Type (1 bytes); Flags (1 byte); Message ID (4 bytes); Length (4 bytes).

The cookie [7] of initiator (C_I) and responder (C_R) are separately formed by using the following information as subfields:

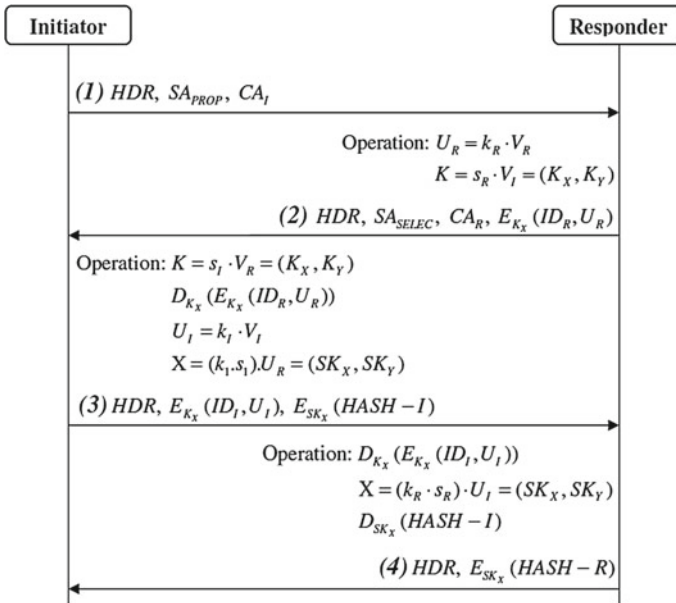


Fig. 2 Proposed phase I of IKE protocol

- Hash value of the IP address, port number, and protocols;
- A secret random number known to the I (or R); and finally
- A timestamp.

Note that *I* and *R* in messages 1 and 2 only send their respective cookies, however, in subsequent message headers include both the cookies $\langle C_I, C_R \rangle$. The SA_{PROP} , a list of cryptographic proposals, is sent by the initiator to the responder for negotiation and SA_{SELEC} , the cryptographic protocols, is selected by the responder from the list sent by the initiator. If necessary, the responder can reject the entire list sent by *I* and sends back an error message in reply. The proposed protocol comprises four steps as given below.

Step 1: *Initiator* → *Responder*: HDR, SA_{PROP}, CA_I

Initiator generates cookie C_I and sends SA proposal SA_{PROP} and ECC-based public key certificate CA_I to responder.

Step 2: *Responder* → *Initiator*: $HDR, SA_{SELEC}, CA_R, E_{K_X}(ID_R, U_R)$

Responder selects the cryptographic proposal SA_{SELEC} from SA_{PROP} , generates his cookie C_R and calculates the following parameters:

- (i) A random number k_R from $[1, n - 1]$,
- (ii) A secret key K_X for symmetric encryption using $K = s_R \cdot V_I = s_R \cdot s_I \cdot P = (K_X, K_Y)$,

- (iii) A private value $U_R = k_R \cdot V_R = (k_R \cdot s_R) \cdot P$, which is then encrypted with his identity using the symmetric key K_X as obtained in (ii).

The responder now sends SA_{SELEC} , ECC certificate and encrypted $E_{K_X}(ID_R, U_R)$ to the initiator.

Step 3: *Initiator* \rightarrow *Responder*: $HDR, E_{K_X}(ID_I, U_I), E_{SK_X}(HASH-I)$
Initiator similarly computes the following:

- (i) The decryption key K_X by using

$$K = s_I \cdot V_R = s_I \cdot s_R \cdot P = (K_X, K_Y),$$

- (ii) U_R by decrypting $E_{K_X}(ID_R, U_R)$,
 (iii) A random number k_I from $[1, n - 1]$,
 (iv) A private value $U_I = (k_I \cdot s_I) \cdot P$, which is then encrypted using K_X as obtained in (i).

Now, the initiator calculates a point X (say) on EC as shown below, the x-coordinate of which (SK_X) is taken as the secret encryption key (SK) negotiated in phase I, i.e., $SK = SK_X$.

$$X = (k_I \cdot s_I) \cdot U_R = k_I \cdot s_I \cdot k_R \cdot s_R \cdot P = (SK_X, SK_Y)$$

The initiator also generates the message $HASH-I$ as given below, and sends to the responder after encrypting with the encryption key SK_X ,

$$HASH-I = H(U_I \parallel U_R \parallel C_I \parallel C_R \parallel SA_I \parallel ID_I)$$

Step 4: *Responder* \rightarrow *Initiator*: $HDR, E_{SK_X}(HASH-R)$

After receiving, responder decrypts it using K_X , compares the received ID_I with ID_I stored in initiator's certificate and the cookies. If it passes, the responder then calculates the encryption key $SK = SK_X$ as

$$X = (k_R \cdot s_R) \cdot U_I = k_R \cdot s_R \cdot k_I \cdot s_I \cdot P = (SK_X, SK_Y)$$

The $HASH-I$ value send by initiator is obtained by decrypting $E_{SK_X}(HASH-I)$ using the encryption key SK_X and compared with its own calculated $HASH-I$. If match is found, initiator becomes authenticated to responder; otherwise it terminates the execution. Finally, responder generates the message $HASH-R$ as given below, and sends it to initiator after encrypting with the encryption key SK_X

$$HASH-R = H(U_I \parallel U_R \parallel C_I \parallel C_R \parallel SA_I \parallel ID_R)$$

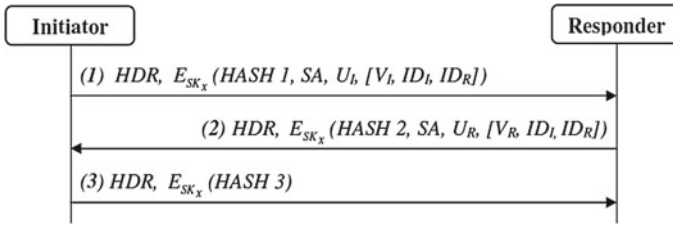


Fig. 3 Proposed phase II of IKE protocol

On receiving, initiator decrypts it and compares with his own computed *HASH-R*. If it passes, initiator is assured that responder is authenticated and the valid common secret encryption key is negotiated. If not, initiator terminates the execution and closes SA.

It can be seen that the proposed main mode protocol uses four random numbers in establishing the common secret *SK* and thus, according to [21], it can be said that the method is well secured as a shared secret with only one random number is assumed to be compromised. The security analysis of the proposed protocol against different attacks is given in Sect. 4.

3.2 Proposed ECC Based Quick Mode Protocol

After completion of phase I, *quick mode* of phase II uses the secret key (*SK_X*) negotiated in phase I for establishing final SA to be used in IPsec. Note that either of the initiator or responder can initiate phase II.

The detail of the proposed ECC-based phase II protocol is given in Fig. 3 along with three steps of message-exchange, which is actually the same round of messages exchanged in the quick mode of the original IKE protocol. However, the following points may be noted:

- V_I* and *V_R*: New public keys of initiator and responder, which are optional, and exchanged if PFS [Perfect Forward Security] is desired.
- ID_I* and *ID_R*: Identities of initiator and responder used optionally, which however may be included to describe the SA-payload being established in phase I.

Step 1: *Initiator* → *Responder*: HDR, *E_{SK_X}*(HASH 1, SA, *U_I*, [*V_I*, *ID_I*, *ID_R*])
 Initiator generates a hash value as

$$HASH\ 1 = hmac(SK_Y, MsgID \parallel SA \parallel U_I)$$

where *hmac* is a keyed hash function.

It is used to authenticate the message generated in Phase I, where SK_Y is the y-coordinate of the point X securely calculated in Phase I. U_I is a new private value to be used if new public value V_I in phase II is used. Now initiator encrypts $HASH\ 1$, SA , U_I and optional parameters if needed using the key SK_X of phase I (x-coordinate of point X) and sends to the responder.

Step 2: *Responder* → *Initiator*: $HDR, E_{SK_X}(HASH\ 2, SA, U_R, [V_R, ID_I, ID_R])$
 Responder decrypts the received encrypted message using SK_X of phase I and calculates its own $HASH\ 1$ value using received U_I and $MsgID$, and compares it with the received $HASH\ 1$. If they match, only then the responder calculates a hash digest as shown below; otherwise phase II is terminated.

$$HASH\ 2 = hmac(SK_Y, MsgID \parallel SA \parallel U_R)$$

The hash value $HASH\ 2$ is used to authenticate the messages as shown, where U_R is a new private value of responder for phase II. Now it encrypts $HASH\ 2$, SA , U_R and optional parameters if needed using the encryption key SK_X of phase I, and sends it to initiator.

Step 3: *Initiator* → *Responder*: $HDR, E_{SK_X}(HASH\ 3)$
 After receiving, initiator decrypts it using SK_X and verifies for the authenticity by calculating its own $HASH\ 2$ using received U_R and $MsgID$, and comparing it with the received value of $HASH\ 2$. If they match, only then initiator calculates a new hash digest as

$$HASH\ 3 = hmac(SK_Y, MsgID \parallel SA \parallel U_I \parallel U_R).$$

It is then encrypted using SK_X and finally sends the same to the responder for final authentication for phase II operation. Note that $HASH\ 3$ includes both U_I and U_R to prove each other about the real, live and active participant in the exchange to protect DoS attack.

Since multiple Quick Mode exchanges may occur simultaneously, there must be some way to identify and determine the phase II messages for each SA, and for this, a unique message ID is assigned to each Quick Mode exchange using ISAKMP header. In addition to this, initiator needs a liveness proof that the responder is on-line and has actually processed his initial Quick Mode message. To do this, the responder includes the initiator’s private value and the message ID of the exchanged message in the authenticating hash payload that not only provides message integrity and source authentication to the initiator, but also provides a liveness proof. After completion of phase II, a key material (KM) is generated as follows:

$$KM = hmac(SK_X, protocol|SPI|U_I|U_R) \quad \text{--- without PFS}$$

$$KM = prf(SK_X, SK_Y|protocol|SPI|U_I|U_R) \quad \text{--- with PFS}$$

where, SK_Y is new y-coordinate of the common secret key generated during phase II for providing Perfect Forward Security (PFS). Since a SPI (Security Parameter Index) is associated with each SA, which depends on the direction from initiator to response; thus two sets of SAs are required for IPsec implementation and they are—one for *inbound traffic* handling and other for *outbound traffic* handling, and the KM generated in one set SA becomes unidirectional and two unidirectional KMs are established.

4 Performance Analysis

An in-depth security analysis of the proposed two phased IKE protocol is given in this section and for this, a number of related attacks are considered, each of which as shown below is seen to be protected.

- **Man-in-the-middle attack:**

The proposed IKE protocol is free from man-in-the-middle attack since entity's ECC-based public key certificate is used in phase I for authentication. Since the messages 2 and 3 as shown in Fig. 2 contain the identity of responder and initiator, and encrypted using ECDH symmetric key K_X , they are authenticated and the attack is protected. For example, after receiving message 3, the responder does the following:

Decrypts $E_{K_X}(ID_I, U_I)$ using the symmetric key K_X ; Gets ID_I ; Retrieves ID_I from CA_I ; Compares retrieved $ID_I =$ received ID_I ? If fails, man-in-the-middle attack is detected and responder terminates the process.

- **Replay attack**

In our proposed protocol, the ISAKMP-Header is used for exchange of each and every message, which contains the initiator/responder's cookie (C_I/C_R) generated by hashing a unique identifier of the peer (such as IP address, port number, and protocol), a secret random number known to the party and a *timestamp*. Now when initiator (or responder) sends a message in phase I/phase II, it estimates and includes present time in the cookie and the responder (or initiator) only accepts messages if *timestamp* having reasonable tolerance. Thus if an attacker wants to replay any previous message, it can be easily detected by comparing with the timestamp of a previous cookie. Thus the proposed protocol prevents the replay attack.

- **Denial-of-service (DoS) attack**

Since the ISAKMP-Header HDR precedes every payload exchanged, and since the cookies of initiator (C_I) and responder (C_R) are included, the *denial-of-service attack* is prevented. This is because if an attacker acts as an initiator using a bogus IP address, he does not receive the reply message and thus, he is not capable to return the same cookie to the sender. Thus the denial-of-service attack is not possible in our proposed IKE protocol.

- **Impersonation attack**

If an attacker makes an effort to impersonate the initiator/responder to exchange a session key, then it is impossible for the attacker to figure out U_R, U_I from messages 2 and 3 of phase I since these are encrypted by a symmetric secret key K_X , known to the initiator and the responder. Then a wrong reply by the attacker directs the termination of the process. Thus our proposed protocol prevents the impersonation attack.

- **Perfect forward/backward secrecy**

The Perfect forward/backward secrecy is the property that the disclosure of the initiator/responder's private key (or any session key) does not compromise the secret key negotiated from earlier/latter runs. In our proposed phase I protocol, the initiator/responder's private key is used for authentication purpose whereas the secret key negotiation is done by the initiator/responder's secret random number (k_I/k_R). Now if the initiator/responder's private key is known to an attacker, then he is able to

- (i) Compute $K = s_I \cdot V_R = s_I \cdot s_R \cdot P = s_R \cdot V_I = (K_X, K_Y)$,
- (ii) Decrypt $E_{K_X}(ID_I, U_I)$ and $E_{K_X}(ID_R, U_R)$ using K_X
- (iii) Get U_R and U_I from the messages 2 and 3,

Even then he cannot derive the session key SK_X from

$$X = k_I \cdot s_I \cdot k_R \cdot s_R \cdot P = (SK_X, SK_Y).$$

This is because the attacker tries to compute the session key SK_X from the pair $(U_I, U_R) = (k_I \cdot s_I \cdot P, k_R \cdot s_R \cdot P)$ directly, which is impossible due to difficulties of Computational Diffie-Hellman Assumption (CDHA). Also if any session key is leaked, the attacker cannot derive any other session keys or the current one. Hence, the proposed phase I protocol is protected.

In phase II, if an attacker is able to determine the SK_X from an IKE SA (phase I), he would easily determine all the keys of a subsequent IPsec SA. Also a new shared secret key SK_Y , which is generated from the optional parameters exchanged during phase II and used in generating the key material KM , is deleted as soon as exchange is finished. Thus, our proposed phase II protocol supports the PFS.

- **Known key security**

The proposed phase I protocol results in a unique shared session key after completion of each negotiation. The compromise of one secret key (K_X) in one negotiation is never compromised with the shared session key (SK_X) agreed on any other negotiation.

- **Identification privacy**

The proposed protocol does not disclose the identity of I and R ; since it is encrypted using ECDH shared session key and CA signed public key certificate is used. In message 2 and 3 of our proposed phase I protocol (Fig. 2), the identity of R (or I) is only verified by I (or R) respectively.

- **Key control**

In our proposed protocol, initiator/responder generates their public-private key pair themselves and no pre-shared secret is used to calculate new shared session key, thus the key control of the initiator/responder is supported.

- **Explicit key confirmation**

The explicit key confirmation means that before using the key to encrypt confidential data, one communication party has to confirm that the other party has actually computed the correct shared session key. In message 3 and 4 of Fig. 2, and all three messages of Fig. 3, the responder/initiator makes a message digest and sends it to each other for verification. This supports the explicit key confirmation in our proposed protocol.

Thus, the proposed ECC-based IKE protocol not only supports less computation- and communication-cost, but also protects all relevant attacks, and in this regard, a theorem is given below.

Theorem 1 Proposed ECC based IKE protocol is efficient and secure.

Proof: The following points regarding the processing cost and the security aspects of the proposed scheme may be noted:

- *For public challenges*, the existing RSA based IKE protocols use Diffie-Hellman key exchange protocol [20] in which the required public challenges generated by using expensive modular exponential operation are $V_I = g^{s_I} \bmod n$ and $V_R = g^{s_R} \bmod n$, where the size of the modulus n should be at least 1024 bits length for its security. On the other hand, the public challenges in ECC $V_I = s_I \cdot P$ and $V_R = s_R \cdot P$ require 160 bits only for comparable security, which shows cost reduction in ECC based technique.
- *For encryption/decryption*, the existing protocols apply RSA-based public key encryption/decryption technique which is due to modular exponentiation operation, much slower than the scalar point multiplication used in ECC. This is because the modular exponentiation is used over 1024-bit discrete logarithm problem (DLP) in RSA, whereas the ECC requires point multiplication using 160-bit ECDLP. The processing speed in ECC is further enhanced by incorporating symmetric encryption rather than the RSA-based public key encryption as used in existing techniques. Thus, the proposed protocol reduces computation cost over the RSA-based techniques.
- *In terms of operation*, the existing phase I of RSA based IKE protocols require six messages exchange, implicit key confirmation, long 1024-bit key size, longer negotiation time, whereas in ECC, it has four messages exchange, explicit key confirmation, short 160-bit key size, shorter negotiation time. Therefore, the proposed protocol have low communication cost, faster processing speed and low network traffic.
- *In terms of security*, some relevant cryptographic attacks of IKE like man-in-the-middle attack, replay attack, denial-of-service attack, impersonation attack etc. are discussed. It has been shown that the proposed scheme prevents these attacks. Hence, the theorem is verified (Table 1).

Table 1 Comparison of existing RSA based schemes [1–11] with the proposed scheme

Parameters	Existing RSA based schemes [1–11]	Proposed scheme
Cryptosystem used	RSA	ECC
Encryption based on	Public key	Symmetric key
Key exchange based on	DH exchange	ECDH exchange
Key length to provide same level of security	1024 bits	160 bits
Processing speed	Low	High
Efficiency	Low	High
No. of message exchanges in phase I	6	4
Message payloads, computation cost	High	Low

5 Conclusion

An IKE implementation based on ECC is proposed in this paper, where the main mode of phase I uses four messages exchange and the quick mode of phase II involves three messages. Due to use of ECC, it achieves many advantages over RSA-based schemes like less computation cost, high processing speed, low network traffic and comparable security even using small secret key-size and thus suitable for efficient implementation. The security analysis of the proposed scheme and comparison mainly with other RSA-based techniques are given, which show improved performance.

References

1. Zhou J (2000) Further analysis of the Internet key exchange protocol. *Comput Commun* 23:1606–1612
2. Forouzan BA (2007) *Cryptography and network security*. Special Indian edition 2007, TMH, pp 563–588
3. Zhu J-m, Ma J-f (2004) An internet key exchange protocol based on public key infrastructure. *J Shanghai Uni (English Ed)*. Article ID: 1007-6417(2004)01-0051-06
4. Kaufman C (2004) The internet key exchange (IKEv2) protocol. IETF draft-ietf-ipsec-ikev2-17, Sept 2004
5. Haddad H, Berenjkoub M, Gazor S (2004) A proposed protocol for internet key exchange (IKE). *Electrical and computer engineering, Canadian conference*, May 2004
6. Orman H (1998) The OAKLEY key determination protocol, RFC 2412
7. Maughan D et al (1998) Internet security association and key management protocol (ISAKMP), RFC 2408
8. Su M-Y, Chang J-F (2007) An efficient and secured internet key exchange protocol design. *Proceedings of the fifth annual conference on communication networks and services research (CNSR'07)*, pp 184–192
9. Fereidooni H, Taheri H, Mahramian M (2009) A new authentication and key exchange protocol for insecure networks. In: *Proceedings of the fifth international conference on wireless communication, networking and mobile computing (WiCom'09)*, pp 1–4

10. Nagalakshmi V, Rameshbabu I (July 2007) A protocol for internet key exchange (IKE) using public encryption key and public signature key. *Int J Comput Sci Netw Secur* 7(7):342–346
11. Nagalakshmi V, Rameshbabu I, Avadhani PS (2011) Modified protocols for internet key exchange using public encryption key and signature keys. In: *Proceedings of the 8th international conference on information technology: new generations 2011*, pp 376–381
12. Ray S, Nandan R, Biswas GP (2012) ECC based IKE protocol design for internet applications, *Procedia Technology*, Elsevier: *Proceedings of 2nd international conference on computer, communication, control and information technology (2012) C3IT 2012*, Hooghly, WB, India, 25–26 Feb 2012, pp 522–529
13. Ray S, Biswas GP (2012) Establishment of ECC-based initial secrecy usable for IKE implementation. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012*, London, UK, 4–6 July 2012, pp 530–535
14. Koblitz N (1987) Elliptic curve cryptosystem. *J Math Comput* 48(177):203–209
15. Miller V (1985) Use of elliptic curves in cryptography. In: Williams HC (ed) *Advances in cryptology-CRYPTO 85*, LNCS 218. Springer, Berlin, pp 417–426
16. Dang Q, Santesson S, Moriarty K, Brown D, Polk T (2010) Internet X.509 public key infrastructure: additional algorithms and identifiers for DSA and ECDSA, RFC 5758, Jan 2010
17. Weise J (2001) Public key infrastructure overview, Sun PSSM global security practice. Sun Blue Prints™ Online, Aug 2001
18. National Institute of Standards and Technology (2001) Introduction to public key technology and the federal PKI infrastructure. National Institute of Standards and Technology, 26 Feb 2001
19. Schaad J, Kaliski B, Housley R (2005) Additional algorithms and identifiers for RSA cryptography for use in the internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile, RFC 4055, June 2005
20. Diffie W, Hellman ME (1976) New directions in cryptography. *IEEE Trans Inf Theory* 22(6):644–654
21. Biswas GP (2011) Establishment of authenticated secret session keys using digital signature standard. *Inf Secur J: A Glob Prosp* 20(1):09–16

Intrusion Alert Correlation Framework: An Innovative Approach

Huwaida Tagelsir Elshoush and Izzeldin Mohamed Osman

Abstract Alert correlation analyzes the alerts from one or more collaborative intrusion detection systems (IDSs) to produce a concise overview of security-related activity on a network. The process consists of multiple components, each responsible for a different aspect of the overall correlation goal. The sequence order of the correlation components affects the process performance. The total time needed for the whole process depends on the number of processed alerts in each component. An innovative alert correlation framework is introduced based on a model that reduces the number of processed alerts as early as possible by discarding the irrelevant and false alerts in the first phases. A new component, *shushing the alerts*, is added to deal with the unrelated alerts. A modified algorithm for fusing the alerts is presented. The intruders' intention is grouped into attack scenarios and thus used to detect future attacks. DARPA 2000 ID scenario specific datasets is used to evaluate the alert correlator model. The experimental results show that the correlation model is effective in achieving alert reduction and abstraction. The performance is improved after the attention is focused on correlating higher severity alerts.

Keywords Alert correlation · Alert correlation datasets · Alert reduction · Collaborative intrusion detection systems · False alarm rate · Intrusion detection

H. T. Elshoush (✉)

Department of Computer Science, Faculty of Mathematical Sciences,
University of Khartoum, Khartoum, Sudan
e-mail: htelshoush@uofk.edu

I. M. Osman

Sudan University of Science and Technology, Khartoum, Sudan
e-mail: izzeldin@acm.org

1 Introduction

Alert correlation is a process that contains multiple components with the purpose of analyzing alerts and providing high-level insight view on the security state of the network surveillance [2, 21, 22]. Correlation aims to relate a group of alerts to build a big picture of the attacks, hence can be used to trace an attack to its source.

The core of this process consists of components that implement specific function, which operate on different spatial and temporal properties [17].

The correlation components are effective in achieving alert reduction and abstraction. Research shows that the effectiveness of each component depends heavily on the nature of the data set analyzed [17]. Moreover, the performance of the correlation process is significantly influenced by the topology of the network, the characteristics of the attack, and the available meta-data [16].

Since alerts can refer to different kinds of attacks at different levels of granularity, the correlation process cannot treat all alerts equally. Instead, it is necessary to have a set of components that focus on different aspects of the overall correlation task. Some components, see Fig. 1, e.g. those at the initial and second units, implement general functionality applicable to all alerts, independent of their type. Other components (e.g. in the third unit) are responsible for performing specific correlation tasks that cannot be generalized for arbitrary alerts, but for certain class of alerts.

Thus, one cannot, in general, determine a ranking among components with respect to their effectiveness. Each component can contribute to the overall analysis. Therefore, the most complete set of components should be used [17].

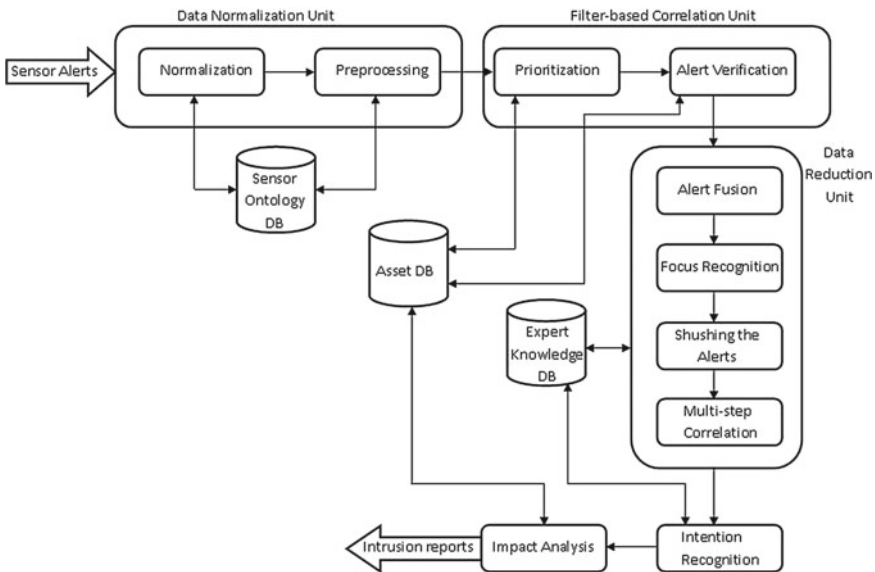


Fig. 1 The improved correlation model

An innovative framework focuses on reordering the correlation components such that redundant, irrelevant and false alerts are reduced as early as possible thus reducing the number of processed alerts to enhance the performance. The unrelated alerts that are not correlated are dealt with in a separate component, *shushing the alerts*. Hence, the overall effectiveness of the correlation process is improved.

2 Overview of Improved Alert Correlation Framework

Our architecture, see Fig. 1, is composed of ten components: *normalization*, *preprocessing*, *prioritization*, *alert verification*, *alert fusion*, *focus recognition*, *shushing the alerts*, *multi-step correlation*, *intention recognition*, and *impact analysis* [7].

In the *normalization* component, alerts that are generated by multiple IDSs are collected and stored in a database before they are modeled and converted into a standard format called Intrusion Detection Message Exchange Format (IDMEF). Then data *preprocessing* is required in order to clean the data, do feature extraction and selection, and finally deal with any incomplete or missing data.

The *filter-based correlation* unit either assigns a priority to each alert or identifies irrelevant alerts. Thus, alerts are ranked based on their severity level in order to distinguish between the high and low risks alerts depending on information in the asset DB. In the *alert verification* component, alerts are checked to find out the verifiable alerts, false positives and unverifiable alerts.

Redundant alerts are fused based on similarity functions [6] in the *alert fusion* component in the *data reduction unit*. In the *focus recognition* component, alerts are aggregated then classified using feature similarity. Unrelated and false alerts tend to be random and will not correlate, hence uncorrelated alerts are removed by *shushing the alerts* component. Lastly, *multi-step correlation*, is expected to achieve substantial improvement in the abstraction level and data reduction [18]. In this component, priori information of the network topology, known scenarios, etc are provided by the expert knowledge DB; hence high level patterns are specified.

In the *intention recognition* component, relevant behavior is grouped into attack scenarios to extract attack strategy and plan recognition.

In the final component, *impact analysis*, the asset DB is consulted to determine all services that are dependent on a specific target. The heartbeat monitor checks whether all dependent services are still operational. If any service is failed, this information can be added to the alert as a likely consequence of the attack [17].

The asset DB stores information about installed network services, dependencies among services, and their importance to the overall operation of a network installation. So the DB does not represent an absolute measure of the importance of any asset, but rather reflect the subjective view of a security administrator. It is updated if there is any new information from the impact analysis or prioritization components.

The *knowledge DB* is a complete repository including all the necessary information about attacks, vulnerabilities, and the topological and configuration information about the protected networks and hosts.

3 Evaluation of Alert Correlators

3.1 Evaluating Alert Correlators

The effectiveness of IDS sensors are evaluated using the detection rate and the false positive rate. These values cannot be easily calculated for an alert correlator as [9]:

- the false positive is not clearly defined for a correlator, as it may receive a false positive (which is an IDS sensors output) and therefore draws a wrong assumption that an attack took place. Thus, validation of the input alerts is necessary.
- if there exists an attack creating a single low-level alert, no correlation will be performed and thus this may be considered a missed attack and hence results in a reduction in the detection rate.

According to [17], the evaluation of the correctness of the correlation process has to be performed manually as:

- there exist very few shared datasets for correlation evaluation, and
- generating alerts from raw data might be risky and may bias the correlation evaluation process,
- no truth files are associated with datasets, which makes it difficult to know if the correlated alerts are representative of a meaningful grouping of detected attacks.

Correlation tools are evaluated “as a whole”, without an assessment of the effectiveness of each component of the analysis process. As a result, it is not clear if and how the different parts of the correlation process contribute to the overall goals of correlation [17]. Recent research proved that the correlation components are effective in achieving alert reduction and abstraction.

3.2 Datasets

The available datasets, e.g. the DARPA-sponsored Cyber Panel Correlation Technology Validation (CTV) effort 2002 and the Defcon 9 dataset, have problems in evaluating the effectiveness of the components of the correlation process [9].

- To assess correlation systems, the sensor alerts output of the IDSs is needed, therefore these datasets are sui to evaluate IDSs and not correlation tools. Thus, IDS alerts have to be generated by running specific sensors on the collected event streams. As a result, the alert stream associated with a dataset may be different if different sensors are utilized or if the configuration is different. This, in turn, makes it harder to compare the correlation systems’ results.

Hence, the effectiveness of correlation is not only highly dependent on the nature of the datasets alone, but also on the different sensors used and the different configuration, if IDSs datasets are used.

- Alert verification should be done in real time, and this is hindered by the offline nature of these datasets.
- The lack of a mission model and its relationship to the network assets hinders the impact analysis.
- To determine the actual impact of the attacks, a health monitoring is needed but not usually provided.
- All presented datasets are not real world traffic.
- Finally, also these datasets have no internet anomalous traffic as none were recorded on a network connected to the internet.

3.3 Correlation Evaluation Truth Files

The correlation evaluation truth file should be the high-level plan behind the attacks in the test data, where the plan should describe for example how each low-level alert is related to other alerts [9]. Some problems are:

- Representing this information is difficult, as there is no good definition of a high-level plan and also no standardized way to express such a plan exists.
- It is difficult to compare the correlator output to the plan contained in the truth file, as two widely different high-level views of an attack might both be correct.

3.4 Factors Affecting the Alert Reduction Rate

When testing correlation systems, it is important to calculate reduction rates for different datasets. Furthermore, it is useful to calculate the reduction rate during each correlation step in addition to the total reduction rate as [9].

- experiments showed that the achieved alert reduction rate (RR) is highly dependent on the features of the dataset being processed, and
- a particular dataset that experience a high reduction rate during one correlation step might achieve poor reduction rates during other steps.

4 Implementation and Experimental Results

4.1 Experiments on DARPA 2000 Datasets

DARPA 2000 [10] is a well-known IDS evaluation dataset created by the MIT Lincoln Laboratory. It consists of two multistage attack scenarios, namely LLDOS 1.0 and LLDOS 2.0.2. Both attack scenarios contain a series of attacks in which an attacker

probes, breaks-in, installs the components necessary to launch a Distributed Denial of Service (DDoS) attack against an off-site server, with different stealth levels. LLDOS 2.0.2 is a bit more sophisticated than LLDOS 1.0.

In both scenarios, the attacker tries to use the vulnerability of Sadmind RPC service and launches buffer overflow attacks against the vulnerable hosts. The attacker installs the mstream distributed DOS software after he breaks into the hosts successfully. Finally, the attacker launches DDOS attacks from the victims. The differences between these two scenarios lay in two aspects: First, the attacker uses IPSweep and Sadmind Ping to find out the vulnerable hosts in LLDOS 1.0 while DNS HInfo is used in LLDOS 2.0.2. Second, the attacker attacks each host individually in LLDOS 1.0, while in LLDOS 2.0.2, the attacker breaks into one host first and then fans out from it [3, 13].

Each scenario includes network traffic collected from both the demilitarized zone (DMZ) and the inside part of the evaluation network. We performed eight experiments, four on the improved model and four on the comprehensive approach model.

4.2 Analysis and Performance Evaluation of the Improved Correlation Components

In this section, each component's function of the innovative alert correlation framework is explained in details, together with the implementation of the model based on the improved framework. In the implementation, Microsoft SQL Server 2005 was used as the relational database to store the alert datasets, the intermediate data, and the analysis results of each component as well as the correlated alerts. Programs written in C#, Microsoft Visual Studio 2010, were created to implement the correlation components' functionalities. The alert log files generated by RealSecure IDS of the DARPA simulation network is used [11].

Data Normalization Unit

The data normalization unit contains two components, namely the *normalization* and the *preprocessing* components.

- **Normalization**

The interaction of collaborators from distinct IDSs poses various requirements. Dependent on the level of collaboration, these include a common language, protocol or even a complete framework. Hence, in a distributed environment with heterogeneous IDSs, specifying a common alert format is fundamental for providing interpretability among IDSs. Moreover, high level analysis such as alert correlation also requires that the alerts that are processed to be in a generic format. There exists a variety of exchange formats; prominent examples include IDMEF and IODEF. Therefore, in this component, all attributes of each sensor alert will be translated

into a common data format, IDMEF in particular [5, 10]. The IDMEF data model is implemented using a Document Type Definition (DTD) to describe Extensible Markup Language (XML) documents. IDMEF is also an object-oriented representation and a Unified Modeling Language (UML) model. If only one type of sensor was used, then that sensor’s format could be used as the standard format. DARPA alerts were normalized in IDMEF standard format.

• **Preprocessing**

The goal of this component is to supply, as accurately as possible, missing alert attributes that are important for later correlation components. Reducing the dimensionality of the data improves the correlation process considerably and detection accuracy is improved as a result of removing irrelevant and redundant features. Hence, this component handles null values, missing and incomplete data [4]. Feature selection is used to reduce the alert attributes, as it is revealed from recent research [1, 19, 20] that it improves detection accuracy and performance. The type information is useful because it allows one to group attacks with a similar impact together.

In both scenarios, there are 45 features, of which only 7 features were extracted, namely EventID, timesec, SrcIPAddress, DestPort, DestIPAddress, OrigEventName, and SrcPort. The date attribute was represented in date/time format, and I converted it to time in seconds (represented as timesec). 5 alerts, representing incomplete data, were removed in all datasets, except for the inside segment of scenario 1.0., see Table 1.

Filter-Based Correlation Unit

The primary goal is to reduce the number of alerts to be correlated by eliminating false, irrelevant and low risk alerts. False alerts need to be handled at an early stage as they will have negative impact on the correlation result, and moreover the number of processed alerts will be greatly reduced.

• **Prioritization**

In this component, depending on the information contained in the asset DB, high and low risks alerts are identified. It takes into account the security policy and the security requirements of the site where the correlation system is deployed. The low risks that do not have significant effect on the protected system are discarded. By alert ranking, the data fed into the remaining components is reduced as only the high and medium risk alerts are considered in the later components. Thus

Table 1 Impact of preprocessing component on LLDOS scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	891	922	430	494
Output alerts	886	922	425	489

Table 2 Impact of prioritization component on LLDOS scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	886	922	425	489
Output alerts	188	167	54	71
Reduction rate (%)	78.78	81.89	87.29	85.48

as acknowledged in [15], when the number of processed alerts is reduced, the performance is improved as the total time needed for the whole process depends on the number of processed alerts in each component. The ranking/priority of alerts of LLDOS scenarios from [14] is used. It is a very naive alert prioritization policy just to give a rough idea of how much reduction this step could achieve, but in most cases a much more complex policy is needed. Thus low risk alerts are discarded, and only the medium and high risk alerts are sent to the next component. Table 2 shows the implementation results.

• Alert Verification

The verification component issue a non-relevant positive attack. Thus, it distinguishes between successful and failed intrusion attempts.

The verification of an attack can be accomplished by extending ID rules with an expected “outcome” of the attack that describes the visible and verifiable traces that a certain attack leaves at a host or on the network [16, 17].

Verification can be performed using both *passive* and *active* techniques, and each approach requires different support from the ID infrastructure [9, 16, 17].

Passive verification mechanisms can be done by storing the actual security status of the network in the knowledge base. *Active alert verification* is a technique designed to reduce the false positive rate of IDSs by actively probing for a vulnerability associated with detected attacks. If the vulnerability corresponding to a detected attack is found to exist in the host or network against which the attack was directed, the alert is generated, invoking any logging and response functions as normal. If, however, the vulnerability is determined not to exist, the alert is considered a false positive and is suppressed. The alert verification process requires that the protected assets be available for real-time verification of the actual exposure of the system and/or that a detailed model of the installed network services be available. Unfortunately, this information is not available for the data set analyzed and there is no sufficient information found about the asset DB, so this component could not be implemented.

Data Reduction Unit

Similar alerts are fused and thus data is reduced by eliminating data redundancies, and irrelevant, false and unreal alarms using alert correlation. False alerts are usually less likely to be correlated using alert correlation.

• **Alert Fusion**

It combines a series of alerts that refer to attacks launched by one attacker against a single target. This component removes duplicates created by the independent detection of the same attack by different sensors, and also correlates alerts that are caused by an attacker who tests different exploits against a certain program or that runs the same exploit multiple times to guess correct values for certain parameters (e.g., offsets and memory addresses for buffer overflow) [9, 16, 17].

The alert fusion component keeps a sliding timewindow of alerts. The alerts within the timewindow are stored in a time-ordered queue. When a new alert arrives, it is compared to the alerts in the queue, starting with the alert with the earliest timestamp.

A *fusion* match is found if all overlapping attributes are equal and the new alert is produced by a different sensor. The timestamp of the meta-alert is assigned the earlier of the sub-alerts times. The later timestamp simply indicates a delayed detection by the other sensor. On the other hand, attack *threads* are constructed by merging alerts with equivalent source and target attributes that occur in a certain temporal proximity but the alerts need not be produced by different sensors. The timestamp of the meta-alert is assigned the earlier of the two start-times and the later of the two end-times.

The value of the time window should be a good trade-off between a small value, which would cause several attack threads to go undetected, and a larger value, which would slow down the system by requiring the component to keep a large number of alerts in the queue.

Algorithm 1 is the alert fusion component method [7].

Algorithm 1 Alert Fusion

Parameter *window-size, fuse-window, thread-window*

Global *alert-queue, fuse, thread*

```

fuse(alert)
fuse ← false
thread ← false
al ← get a:alert with lowest start-time from alert-queue where
if alert.analyzer ∩ a.analyzer is empty and all overlapping attributes except start-time, end-time, analyzer,
alertid are equal then
    fuse
    window-size = fuse-window
else
    if alert.victimhosts = a.victimhosts and alert.attackerhosts = a.attackerhosts then
        thread
        window-size = thread-window
    end if
end if
    
```

There were two sensors in DARPA data sets, but all the alerts generated by one of the sensors contained null and incomplete values and thus were removed by the

```

if  $al \neq null$  then
  replace  $al$  in alert-queue with fuse-merge(alert, al)
else
  add alert to alert-queue
  remove all  $a:alert$  from alert-queue where
     $a.start-time < (alert.start-time - window-size)$ 
  pass removed alerts to next correlation component
end if

fuse-merge(alert1, alert2)
 $r \leftarrow$  new alert
 $r.alertid \leftarrow$  get unique-id()
 $r.start-time \leftarrow$  min(alert1.start-time, alert2.start-time)
 $r.reference \leftarrow$  (alert1.alertid  $\cup$  alert2.alertid)
if fuse then
   $r.end-time \leftarrow$  min(alert1.end-time, alert2.end-time)
  for each attr:attribute except start-time, end-time, reference, alertid do
     $r.attr \leftarrow$  alert1.attr  $\cup$  alert2.attr
  end for
  fuse  $\leftarrow$  false
else
  if thread then
     $r.end-time \leftarrow$  max(alert1.end-time, alert2.end-time)
     $r.analyzer =$  alert1.analyzer  $\cup$  alert2.analyzer
    thread  $\leftarrow$  false
  end if
end if
if alert1.name = alert2.name then
   $r.name \leftarrow$  alert1.name
else
   $r.name \leftarrow$  "Attack Thread"
end if
for each attr:attribute except start-time, end-time, reference, analyzer, alertid do
  if alert1.attr = alert2.attr then
     $r.attr \leftarrow$  alert1.attr
  else
     $r.attr \leftarrow$  null
  end if
end for
return r

```

Table 3 Impact of alert fusion component on LLDOS scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	188	167	54	71
Output alerts	92	110	34	45
Reduction rate (%)	51.06	34.13	37.04	36.62

preprocessing component. Thus, there were no fusion in the data set used as all traffic injected into this component were seen by one sensor, but there were thread reconstruction. Table 3 shows the results of the implementation.

- **Focus Recognition**

Table 4 Impact of focus recognition component(one-to-many)

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	92	110	34	45
Output alerts	42	57	23	27
Reduction rate (%)	56.52	48.18	32.35	40

This component has the task of identifying hosts that are either the source or the target of a substantial number of attacks. This is used to identify denial-of-service (DoS) attacks or port scanning attempts. It aggregates the alerts associated with single hosts attacking multiple victims (called a one-to-many scenario) and single victims that are targeted by multiple attackers (called a many-to-one scenario).

The one-to-many scenario has two tunable parameters: the size of the timeout, which is used for the initial window size, and the minimum number of alerts for a meta-alert to be generated. On the other hand, the many-to-one scenario has three tunable parameters: the first two are the same as for the one-to-many scenario. The third parameter is the number of meta-alerts required before a many-to-one alert is labeled as a denial-of-service attack [9, 16, 17].

Specifically, a many-to-one attack is classified as a DDoS attack when the total number of attacks against a victim exceeds a user-defined limit. A one-to-many alert is classified as a horizontal scan when a single port is scanned on all victim machines, or as a horizontal multiscan when more than one port is scanned [17].

In the carried out experiments, the minimum number of alerts in a meta-alert was two. We first applied one-to-many focus recognition on DARPA datasets, then followed by many-to-one focus recognition. Some horizontal scan and multi-scan attacks were observed. Tables 4 and 5 show the reduction rates. DMZ in scenario 2.0.2 shows a great RR as is expected being a multistage attack scenario.

- **Shushing the alerts**

As shown in [13], alert correlation can be used to differentiate between false and true alerts. False alerts and unreal alarms tend to be more random than actual alerts, and are less likely to be correlated. Thus, based on this founding, we intentionally removed the alerts that are not correlated in the alert fusion and focus recognition, resulting in Table 6, which shows great reduction rates.

- **Multi-step Correlation**

The goal of this component is to identify high-level attack patterns that are composed of several individual attacks. The high-level patterns are usually specified using some form of expert knowledge [6, 9, 12, 17].

Table 5 Impact of focus recognition component(many-to-one)

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	42	57	23	27
Output alerts	31	28	5	24
Reduction rate (%)	26.19	50.88	78.26	11.11

Table 6 Impact of shushing the alerts component on LLDOS scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	31	28	5	24
Output alerts	6	7	3	5
Reduction rate (%)	80.65	75	40	79.17

Specifically, it may also associate network-based alerts with host-based alerts that are related to the same attack, called *attack session reconstruction*. It automatically links a network-based alert to a host-based alert when the victim process listens to the destination port mentioned in the network-based alert. This requires either real-time access to the systems being protected or very detailed auditing information in order to map network traffic to host activity. Identifying relations between these two types of alerts is difficult because the information that is present in the alerts differs significantly [17].

While multistep attack analysis may not generate the same level of alert reduction achieved by other components, it often provides a substantial improvement in the abstraction level of the produced meta-alerts. Newly detected attack strategies are fed back to the expert knowledge DB to keep it up-to-date.

Relying on the information in [3], attack patterns are identified, and used to implement this component resulting in Table 7.

Intention Recognition

Intention or plan recognition is the process of inferring the goals of an intruder by observing his/her actions [8]. It deduces strategies and objectives of attackers based on attack scenarios that are output by correlation systems. Failed attacks can be useful to know so to be avoided in the future. Using alert correlation, the intruders' relevant behavior can be grouped into attack scenarios, and later on, their attack strategy or plan can be extracted and fed back to update the expert knowledge DB.

Inadequate information of attack strategies or plans of intruders in the data set used hindered the implementation of this component.

Table 7 Impact of multi-step correlation component on LLDOS scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	6	7	3	5
Output alerts	6	5	2	4
Reduction rate (%)	0	28.57	33.33	20

Table 8 Total alert reduction for the improved model

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	891	922	430	494
Output alerts	6	5	2	4
Reduction rate (%)	99.33	99.46	99.53	99.19

Impact Analysis

This component contextualizes the alerts with respect to a specific target network. It combines the alerts from the previous correlation components with data from an asset DB and a number of heartbeat monitors to determine the impact of the detected attacks on the operation of the monitored network and on the assets that are targeted by the attacker. Thus, it requires a precise modeling of the relationships among assets in a protected network and constant health monitoring of those assets. Hence, insufficient information of asset DB of LLDOS scenarios deters the implementation.

4.3 Summary of Experimental Results

Hereby, we summarize the results of our experiments.

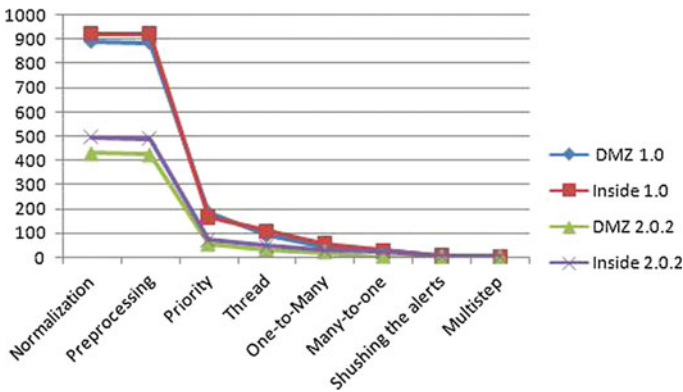


Fig. 2 Effect of Improved Correlation Model on LLDOS Scenarios 1.0 and 2.0.2.

Table 9 No. of processed alerts using improved model for scenario 1.0

	Prepr.	Prio.	Fus.	1:M	M:1	Shush.	Multi	total
DMZ	886	188	92	42	31	6	6	1251
Inside	922	167	110	57	28	7	5	1296

Table 10 No. of processed alerts using comprehensive approach for scenario 1.0

	Prepr.	Fus.	1:M	M:1	Multi.	Prio.	total
DMZ	886	619	208	175	118	13	2019
Inside	922	622	193	151	107	16	2011

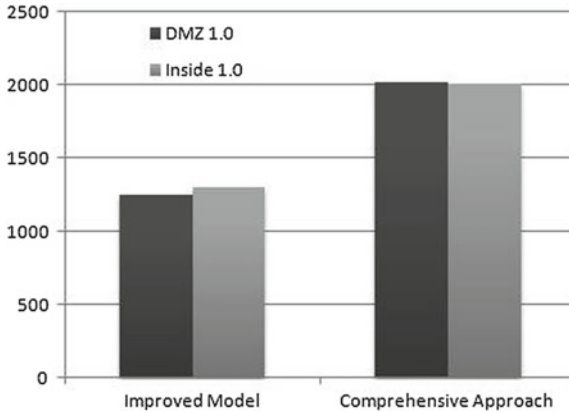


Fig. 3 Comparison of processing time of improved correlation model and comprehensive approach on LLDOS scenario 1.0

The Total Alert Reduction of the Improved Model

Table 8 shows the total alert reduction for each dataset. Figure 2 shows the effect of the improved correlation model on LLDOS 1.0 and 2.0.2 scenarios. There is a substantial drop in the number of alerts in the priority component for all datasets. This reduces the number of processed alerts considerably and thus improves the correlation process performance.

Comparison of the Performance of the Improved Model with the Comprehensive Approach on LLDOS Scenario 1.0

Tables 9 and 10 show the number of processed alerts in each component for scenario 1.0 for the improved model compared to the Comprehensive approach discussed in [17]. Since the processing time is proportional to the number of processed alerts, hence Fig. 3 shows that the improved model gives better results.

Comparison of the Performance of the Improved Model with the Comprehensive Approach on LLDOS Scenario 2.0.2

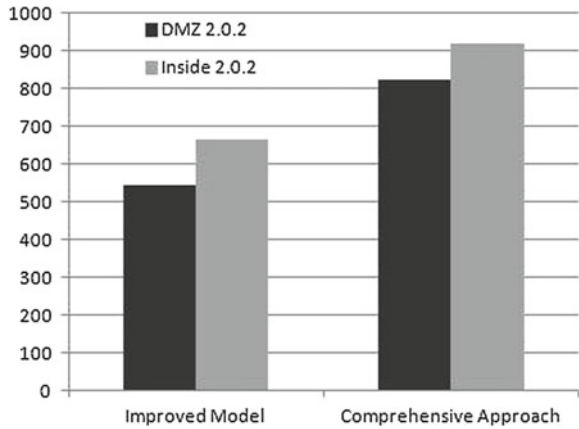
Table 11 No. of processed alerts using improved model for scenario 2.0.2

	Prepr.	Prio.	Fus.	1:M	M:1	Shush.	Multi	total
DMZ	425	54	34	23	5	3	2	546
Inside	489	71	45	27	24	5	4	665

Table 12 No. of processed alerts using comprehensive approach for scenario 2.0.2

	Prepr.	Fus.	1:M	M:1	Multi.	Prio.	total
DMZ	425	241	63	46	44	5	824
Inside	489	276	71	44	33	7	920

Fig. 4 Comparison of processing time of improved correlation model and comprehensive approach on LLDOS scenario 2.0.2



Tables 11 and 12 show the number of processed alerts in each component for scenario 2.0.2 for the improved model and the Comprehensive approach [17] respectively. The graph of Fig. 4 assured the affirmation of the better performance of the innovative improved model.

References

1. Amiri F, Yousefi MMR, Lucas C, Shakery A (2011) Improved feature selection for intrusion detection system. *J Netw Comput Appl*
2. Bye R, Camtepe SA, Albayrak S (2010) Collaborative intrusion detection framework: characteristics, adversarial opportunities and countermeasures
3. Cui Y (2002) A toolkit for intrusion alerts correlation based on prerequisites and consequences of attacks, MSc thesis
4. Davis JJ, Clark AJ (2011) Data preprocessing for anomaly-based network intrusion detection: a review. *J Comput Secur* 30:353–375
5. Debar H, Curry D, Feinstein B (2007) The intrusion detection message exchange format (IDMEF). <http://www.ietf.org/rfc/rfc4765.txt>
6. Elshoush HT, Osman IM (2011) Alert correlation in collaborative intelligent intrusion detection systems—a survey. *J Appl Soft Comput* 11(7):4349–4365
7. Elshoush HT, Osman IM (2012) An improved framework for intrusion alert correlation. Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE, U.K, London, pp 518–523, 4–6 July 2012
8. Ghorbani AA, Lu W, Tavallaei M (2010) Network intrusion detection and prevention: concepts and techniques. Springer, Heidelberg

9. Kruegel C, Valeur F, Vigna G (2005) *Intrusion detection and correlation—challenges and solutions*. Springer, Boston
10. MIT Lincoln laboratory (2000) DARPA intrusion detection scenario specific datasets. <http://www.ll.mit.edu/index.html>
11. Ning P (2007) TIAA: a toolkit for intrusion alert analysis. <http://discovery.csc.ncsu.edu/software/correlator/>
12. Ning P, Cui Y, Reeves DS (2002) Analyzing intensive intrusion alerts via correlation. In: *Proceedings of the 5th international symposium on recent advances in intrusion detection (RAID 2002)*, LNCS 2516, Zurich, Switzerland, pp 74–94, Oct 2002
13. Ning P, Cui Y, Reeves DS (2002) Constructing attack scenarios through correlation of intrusion alerts. In: *Proceedings of the 9th ACM conference on computer and communications security*, Washington, DC, pp 245–254, Nov 2002
14. Siraj MM, Maarof MA, Hashim SZM (2009) Intelligent alert clustering model for network intrusion analysis. *Int J Advanc Soft Comput Appl* 1(1), ICSRS Publication, ISSN 2074–8523
15. Taha AE, Ghaffar AI, Bahaa Eldin AM, Mahdi HMK (2010) Agent based correlation model For intrusion detection alerts. IEEE Computer Society, London
16. Valeur F (2006) *Real-time ID alert correlation*, PhD thesis. Barbara, USA
17. Valeur F, Vigna G, Kruegel C, Kemmerer R (2004) A comprehensive approach to intrusion detection alert correlation. *IEEE Trans Dependable Secure Comput* 1(3):146–169
18. Yusof R, Selamat SR, Sahib S (2008) Intrusion alert correlation technique analysis for heterogeneous log. *Int J Comput Sci Netw Secur (IJCSNS)* 8(9):132–138
19. Zainal A, Maarof MA, Shamsuddin SM (2007) Features selection using rough-PSO in anomaly intrusion detection
20. Zainal A, Maarof MA, Shamsuddin SM (2006) Feature selection using rough set in intrusion detection
21. Zhou CV, Leckie C, Karunasekera S (2009) Decentralized multidimensional alert correlation for collaborative intrusion detection. *J Netw Comput Appl* 32:1106–1123
22. Zhou CV, Leckie C, Karunasekera S (June 2009) A survey of coordinated attacks and collaborative intrusion detection. Elsevier Ltd, *Computer Security*

An Interactive Shadow Removing Tool: A Granular Computing Approach

Abhijeet Vijay Nandedkar

Abstract This work proposes a tool to remove shadow from colour images with the help of user interaction. Shadow detection and removal is an interesting and a difficult image enhancement problem. In this work, a novel granule based approach for colour image enhancement is proposed. The proposed method constructs a shadow classifier using a Granular Reflex Fuzzy Min-Max Neural Network (GrRFMN). Classification and clustering techniques based on granular data are up-coming and finding importance in various fields including computer vision. GrRFMN capability to process granules of data is exploited here to tackle the problem of shadows. In this work, granule of data represents a group of pixels in the form of a hyperbox. During the training phase, GrRFMN learns shadow and non-shadow regions through an interaction with the user. A trained GrRFMN is then used to compute fuzzy memberships of image granules in the region of interest to shadow and non-shadow regions. A post processing of image based on the fuzzy memberships is then carried out to remove the shadow. As GrRFMN is trainable on-line in a single pass through data, the proposed method is fast enough to interact with the user.

Keywords Granular computing · Granular neural network · Shadow detection and removal · Reflex fuzzy min-max neural network · Compensatory neurons

A.V. Nandedkar (✉)

Department of Electronics and Telecommunication Engineering,
S.G.G.S. Institute of Engineering and Technology, Vishnupuri, Nanded,
Maharashtra 431606, India
e-mail: avnandedkar@yahoo.com

1 Introduction

Human vision system is very immune to shadows. We do not find any difficulty in recognizing, tracking objects even with shadows. But in the case of computer vision, shadows create problems and reduce the reliability of the system. In addition, shadows are also responsible to degrade the image quality. Therefore, shadow removal is an important pre-processing step for computer vision and image enhancement algorithms [1]. It is noted that standard approaches, software, and evaluation datasets exist for a wide range of important vision tasks, from edge detection to face recognition; there has been comparatively little work on shadows during last four decades [2].

It may be noted that the proposed methods in the literature can be grouped into two categories, (i) automatic shadow detection, and (ii) shadow detection with user interaction. The automatic shadow detection mainly takes the help of multiple images for identifying shadows i.e. prior knowledge about shadow is necessary. Such approaches that use multiple images [3], time-lapse image sequences [4, 5] are available in the literature. An explicit focus on the shadows cast by objects onto the ground plane can be found in [2]. This approach consists of three stages. First, it computes edges in an image, then it extracts features around the edges, and finally these features are used with a trained decision tree classifier to detect whether an edge is shadow or not. Training of this classifier required in total 170 selected images from LabelMe [6], Flickr, and the dataset introduced in [7], with the only conditions being that the ground must be visible, and there must be shadows. It may be noted that training of such classifier in the above method is a crucial and a difficult task.

The other approaches such as [1, 8, 9, 13] attempt is made to remove shadow from given single image. To acquire knowledge about the shadow in the image, a user interaction is needed and is found helpful to refine the output in the intended manner. Wu and Tang's method [8, 9] remove shadows when given user-specified shadow and non-shadow regions. It adopts a continuous optimization method that requires lot of iterations to converge. As a result, it is not straightforward to use their method in an interactive and incremental manner [1]. The approach in [1] solves this problem by formulating the problem in an MRF framework. It is observed that this method requires more number of strokes for a complicated image.

The present work proposes an application of granular data processing to detect and remove shadows. It is observed that granulation of information is an inherent and omnipresent activity of human beings carried out with intent of better understanding of the problem [10]. In fact, information granulation supports conversion of clouds of numeric data into more tangible information granules. The concept of information granulation within the frame work of fuzzy set theory was formalized by Zadeh in his pioneering work [11]. He pointed out that humans mostly employ words in computing and reasoning; and information granulation is a part of human cognition [11]. This work is a small step towards implementation of this concept for shadow detection and removal. To implement the concept of data granulation, this work uses Granular Reflex Fuzzy Min-Max Neural Network (GrRFMN) proposed by Nandedkar and

Biswas in [12]. The major advantage of the work presented here is that it incorporates user’s knowledge for shadow removal and works at an interactive speed.

The rest paper is organized as follows. Section 2 briefly discusses GrRFMN architecture and its learning. Section 3 elaborates the proposed method. Section 4 presents experimental results and conclusions.

2 GRRFMN for Shadow Classification

In the proposed technique, the first step is to learn shadow and non-shadow regions from the given image. User interacts with the system to point-out shadow and non-shadow region. The classifier required in the proposed system must be fast enough to work at an interactive speed with the user. Looking towards the need of the problem at hand, GrRFMN (Fig. 1) is used as a shadow classifier. It is capable to learn data online, in a single pass through data. GrRFMN learns different classes by aggregating hyperbox fuzzy sets [12].

A hyperbox is a simple geometrical object which can be defined by stating its min and max points (a 3D example is shown in Fig. 1). The architecture of GrRFMN is divided into three sections, (i) Classifying neurons (CNs), (ii) Overlap compensation

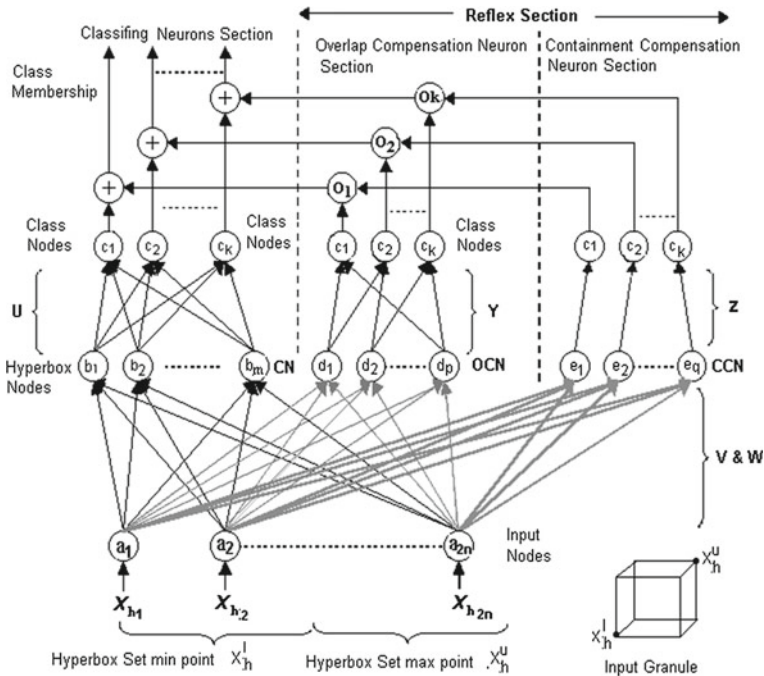


Fig. 1 GrRFMN architecture

neurons (OCNs) and (iii) Containment compensation neurons (CCNs). CNs deal with class regions i.e. shadow and non-shadow regions in this case. OCNs and CCNs tackle overlap regions amongst these classes. The working of OCNs and CCNs is inspired from reflex mechanism of human nervous system, which gets activated in hazardous conditions only. OCNs and CCNs are active if the input data belongs to overlap region of classes.

An n-dimensional input granule is represented in the form of, $X_h = [X_h^u, X_h^l]$ where $X_h^l, X_h^u = (x_{h1}, x_{h2}, \dots, x_{hn})$ are min and max point vectors of the input granule respectively. A point data is a special case with $X_h^l = X_h^u$. Appending min and max point vectors, the input is connected to the nodes $x_{h1} - x_{h2n}$. Here it is assumed that all input features are scaled in the range [0–1]. The training algorithm for GrRFMN and other details regarding activation functions are detailed in [12]. The output of this network is fuzzy membership of input vector to different classes.

3 Interactive Shadow Removal System

The task of shadow detection and removal from given colour image is an important issue in the field of computer vision as well as in commercial photography. The current approach uses granular computing to solve the problem of shadow detection and removal. This method utilizes capability of GrRFMN to acquire knowledge through granules of data. The proposed Interactive Shadow Removal (ISR) system is shown in Fig. 2.

The proposed ISR system can be divided into seven steps:

- (i) Interact with user to acquire information about shadow, non-shadow regions,
- (ii) Extract features of shadow and non-shadow regions,
- (iii) Train GrRFMN using extracted data,
- (iv) Interact with user to acquire Region of Interest (ROI) from which shadow is to be removed.

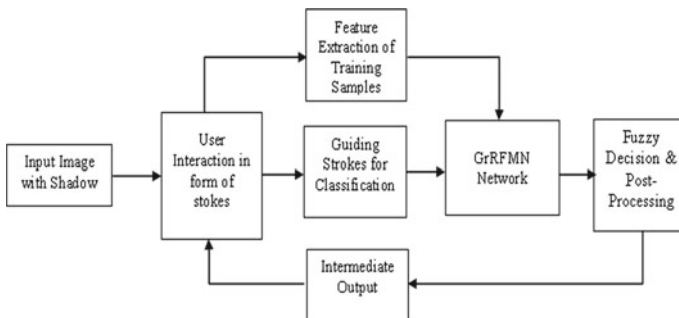


Fig. 2 Proposed SDR system

- (v) Find out fuzzy membership of pixels in the ROI to shadow and non-shadow regions,
- (vi) Use extracted properties of non-shadow region to correct pixels in shadow region, and
- (vii) Present output to user and if required obtain strokes (interaction) from user to remove shadow from the desired area.

The details of each step are as follows:

- (i) Interact with user to acquire information about shadow and non-shadow regions: In this step a user interact with the system in form of strokes and to point out shadow and non-shadow regions. The system uses RGB colour space to represent a colour image.
- (ii) Feature extraction of shadow and non-shadow regions: Image samples of shadow and non-shadow region provided by the user in the previous step are sub-divided into grids of size $(k \times k)$ and granules are produced to represent each grid in the form of hyperboxes which is represented by simply stating its min and max vertices. Thus, a hyperbox is computed by finding min-max values of the pixels in that grid, i.e.

$$\begin{aligned} V &= [R_{\min}, G_{\min}, B_{\min}], \\ W &= [R_{\max}, G_{\max}, B_{\max}] \end{aligned} \quad (1)$$

Along with this information, a mean value of the grid for the three planes is also computed and added to the min-max vector as, i.e.

$$\begin{aligned} V &= [R_{\min}, G_{\min}, B_{\min}, R_{\text{mean}}, G_{\text{mean}}, B_{\text{mean}}], \\ W &= [R_{\max}, G_{\max}, B_{\max}, R_{\text{mean}}, G_{\text{mean}}, B_{\text{mean}}] \end{aligned} \quad (2)$$

- (iii) Training of GrRFMN: Granules created in the previous steps are used to train GrRFMN. The network learns two classes i.e. shadow and non-shadow regions. GrRFMN architecture consists of 12 input nodes and 2 output nodes. The hidden layer grows during learning as per the requirement.
- (iv) Interact with user to acquire Region of Interest (ROI) from which shadow is to be removed through strokes.
- (v) Fuzzy membership of pixels in the ROI to shadow and non-shadow regions: In this step, fuzzy membership of pixels in the ROI is obtained for the two classes i.e. shadow and non-shadow. Around each pixel a neighborhood of size $(n \times n)$ is considered and granules are formed using Eq. 2. These granules are fed to GrRFMN to compute membership of pixels to shadow (μ_S) and non-shadow classes (μ_{NS}).
- (vi) Extract properties of non-shadow region to correct pixels in shadow region: To improve the pixels in shadow region, a correction factor is computed based on the difference between mean RGB values of non-shadow and shadow regions as,

$$\begin{aligned}
\Delta R &= R_{NS_mean} - R_{S_mean} \\
\Delta G &= G_{NS_mean} - G_{S_mean} \\
\Delta B &= B_{NS_mean} - B_{S_mean}
\end{aligned}
\tag{3}$$

where NS—non-Shadow, S-Shadow

Each pixel $P(R,G,B)$ in ROI is corrected using a fuzzy rule:

$$\text{Correction IF } (\mu_{PS} > 0) \text{ and } (\mu_{PNS} < 0.7) \text{ and } (I_{\text{mean_NS}} > I_p); \tag{4}$$

where $I_{\text{mean_NS}}$ is mean intensity of non-shadow region and I_p is the intensity of the pixel. They are calculated as:

$$\begin{aligned}
I_{\text{mean_NS}} &= (R_{NS_mean} + G_{NS_mean} + B_{NS_mean}) / 3; \\
I_p &= (R_p + G_p + B_p) / 3
\end{aligned}
\tag{5}$$

The correction to a pixel is done as follows:

$$\begin{aligned}
R_{\text{correct}} &= (R + \Delta R), \\
G_{\text{correct}} &= (G + \Delta G), \\
B_{\text{correct}} &= (B + \Delta B)
\end{aligned}
\tag{6}$$

- (vii) Once correction of ROI is done, the enhanced image is presented to the user. User can give a new ROI for shadow removal. The above steps (4–7) are repeated till the user interacts with the system for corrections. There are situations where shadow may fall on regions belonging to different properties. In such cases, it is required to repeat above steps (1–7) for each region with the help of user interaction.

The following section demonstrates detailed procedure and results for the proposed system.

4 Experimental Results

Here aim is to:

- (1) Demonstrate the system working, and
 - (2) Test the proposed system on some real life images.
- (1) **System working:** Consider an image such as Fig. 3a is presented to the system. User interacts with the figure to point out the shadow and non-shadow regions required to train GrRFMN. Figure 3b, c depicts samples of shadow and non-shadow to train GrRFMN acquired from user. After this step, the shadow and non-shadow regions are granulized as per Eq. 2 for a (3×3) grid size. These granules are used to train GrRFMN. The expansion coefficient (θ) of GrRFMN is 0.15. After completion of the training, pixels of ROI (Fig. 3d) are enhanced

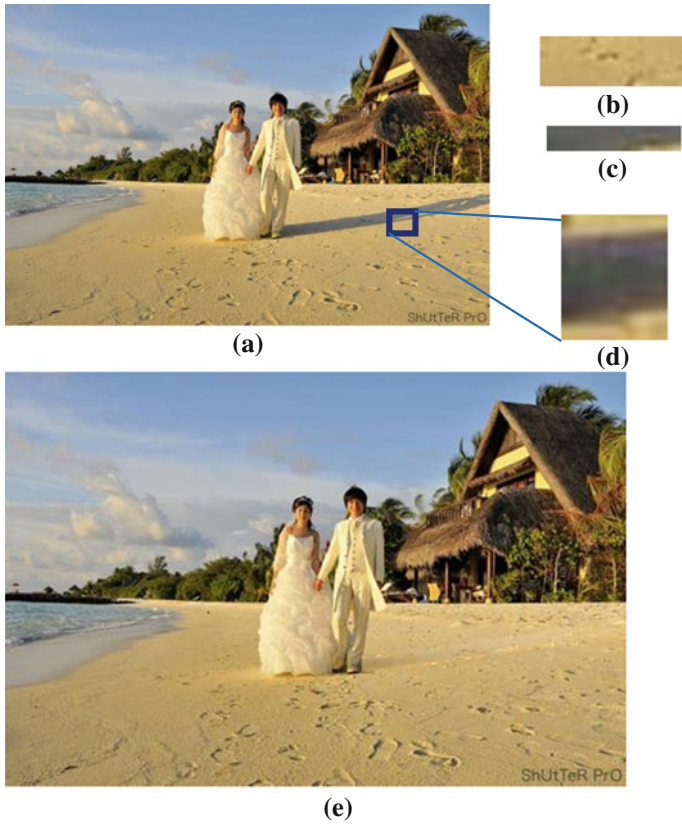


Fig. 3 Shadow removal using proposed system

using a (3×3) neighborhood as per steps 5 and 6. The user further interacts and may give next ROI for shadow detection and removal. In this way, user interacts till complete removal of shadow as shown in Fig. 3e is obtained.

- (2) **Results on other images:** To test the capability system performance was tested on few other real life images are shown in Fig. 4a–g.

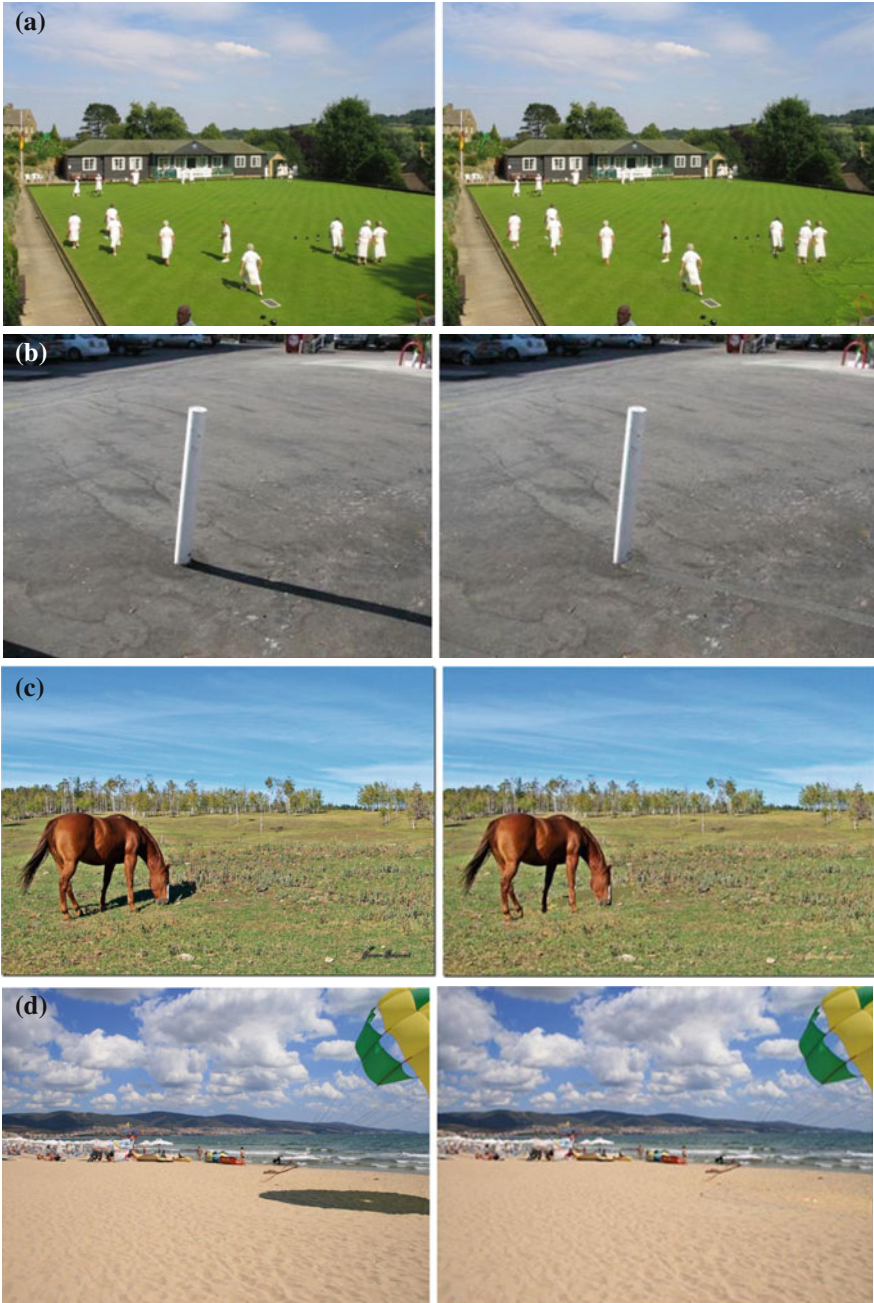


Fig. 4 Results on other images

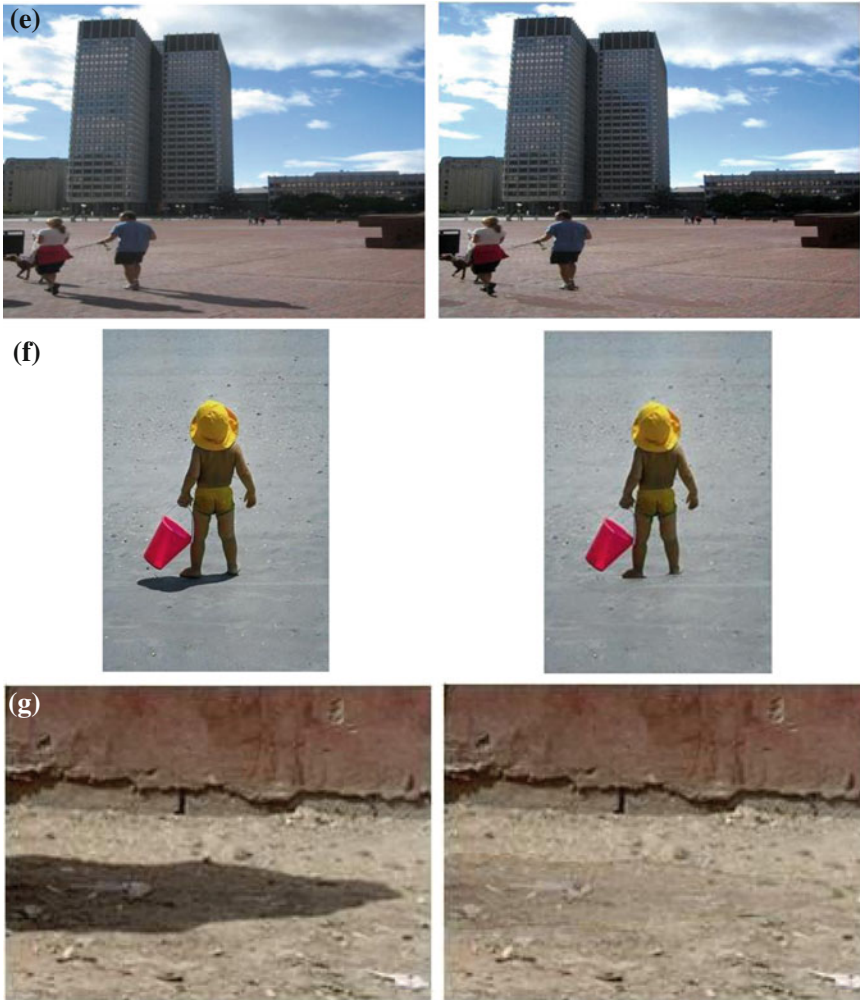


Fig. 4 Continued.

5 Conclusion

An interactive tool for shadow detection and removal is proposed. Granular computing is a powerful tool and is found suitable in shadow detection and removal problem. It is observed that due to on-line learning capability of GrRFMN, the system operates at an interactive speed. In future this method may be extended for unsupervised mode of operation.

References

1. Daisuke M, Yasuyuki M, Katsushi I (2010) Interactive removal of shadows from a single image using hierarchical graph cut. *IPSN Transactions on computer vision and applications*, vol 2, pp 235–252
2. Lalonde JF, Alexei A. Efros and Narasimhan SG (2010) Detecting ground shadows in outdoor consumer photographs. *Proceedings of European conference on computer vision, ECCV, Part II*. Heraklion, Crete, Greece, pp. 322–335, 5–11 Sept 2010
3. Finlayson GD, Fredembach C, Drew MS (2007) Detecting illumination in images. *IEEE international conference on computer vision*, Rio de Janeiro, pp 1–8, 14–21 Oct 2007
4. Weiss Y (2001) Deriving intrinsic images from image sequences. *IEEE international conference on computer vision*, vol 2. Vancouver, BC, pp 68–75
5. Huerta I, Holte M, Moeslund T, Gonzalez J (2009) Detection and removal of chromatic moving shadows in surveillance scenarios. *IEEE international conference on computer vision*, Kyoto, pp 1499–1506, 29 Sept–2 Oct 2009
6. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe, a database and web-based tool for image annotation. *Int J Comput Vision* 77:157–173
7. Zhu J, Samuel KGG, Masood SZ, Tappen MF (2010) Learning to recognize shadows in monochromatic natural images. *IEEE conference on computer vision and pattern recognition*, San Francisco, CA, pp 223–230, 13–18 June 2010
8. Wu TP, Tang CK (2005) A Bayesian approach for shadow extraction from a single image. *Proceedings of international conference on computer vision (ICCV)*. Beijing, China, vol 1, pp 480–487, 17–21 Oct 2005
9. Wu TP, Tang CK, Brown MS, Shum HY (2007) Natural shadow matting. *ACM Trans Graph* 26(2), Article 8
10. Pedrycz W (2001) Granular computing: an introduction. In: *Proceedings of joint IFSA world congress and 20th NAFIPS*, vol 3. Vancouver, BC, pp 1349–1354
11. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst* 90:111–127
12. Nandedkar AV, Biswas PK (2009) A reflex fuzzy min max neural network for granular data classification. *IEEE Trans Neural Netw* 20(7):1117–1134
13. Nandedkar AV (2012) An interactive shadow detection and removal tool using granular reflex fuzzy min-max neural network. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE*, vol II. London, UK, pp 1162–1165, 4–6 July 2012

Resolution Enhancement for Digital Off-Axis Hologram Reconstruction

Nazeer Muhammad and Dai-Gyoung Kim

Abstract A new method of digital off-axis hologram reconstruction based on the Fresnel transform is proposed. A combination of composite filtering, Abbe's limitation, and digital lens formulae has been used with an appropriate handling of Fresnel impulse response propagator. A clear image of microscopic object is efficiently reconstructed from hologram using a plane wave with involvement of electric field along bi-cubic interpolation in the final reconstruction step. In particular, the proposed method automatically suppresses the zero order term and virtual image. The image can be reconstructed with large size using interpolation scheme with the Haar wavelet. The proposed method facilitates the transverse high resolution of microscopic image, which has better applicability than other approaches. Moreover, the advantages of this method are its simplicity and convenience in data processing.

Keywords Bi-cubic interpolation · Digital holography · Fresnel transforms · Lens formula · Microscopy · Wavelet transforms

1 Introduction

The idea of recording and reconstructing holograms digitally has been proposed in early studies by Goodman [1]. Digital holography is a novel imaging technique, which uses a charge-coupled device (CCD) camera for hologram recording and a numerical method for hologram reconstruction. With advances in computer performance and electronic image acquisition devices, digital holography has become more

N. Muhammad · D.-G. Kim (✉)

Division of Applied Mathematics, the ERICA campus, Hanyang University,
Ansan 426-791, South Korea
e-mail: nazeer@hanyang.ac.kr

D.-G. Kim

e-mail: dgkim@hanyang.ac.kr

attractive for many applications. However, off-axis holographic configuration suffers from zero order and twin images which are partially superimposed to the real image, so that a noisy image is developed [2, 3]. This is mainly due to the spatial resolution of Digital Holographic method (DH) which is limited by the digital recording device and the field of view (FOV) of DH system.

In this chapter, we modified the reconstruction algorithm of efficient Fresnel impulse response propagation for digital off-axis hologram (FIRP-DOAH) reconstruction [4]. The proposed method reconstructs the hologram image at larger distance and improves the resolution of the FOV.

The chapter is organized as follows. Section 2 describes the theoretical expression of an efficient FIRP-DOAH reconstruction method with the complementary conditions. Section 3 provides the quality measurement of RHI. The proposed algorithm to enlarge the FOV at longer distance is experimented in Sect. 4, followed by conclusion in Sect. 5.

2 Theoretical Expression

Among the different techniques of digital recording of microscopic holograms; the most common Mach-Zehnder sketch is shown in Fig. 1. It is based on the same foundations as optical holography [1].

Figure 1 shows the experimental setup used in digital thin lens formulae in validation of proposed method. A collimated and expanded plane wave is divided into two beams by the beam splitter BS-1. One of the beams serves as reference wave u_r , and the other is object wave u_1 . A microscopic (MO) collects the object wave u_1 transmitted by the specimen with the help of beam splitters BS-2 where magnified image of the specimen using mirror (M) at a distance z behind the CCD. Two identical polarizer P-1 and P-2 are placed into the path of reference beam and object wave: along with two turning mirror $M-1$ and $M-2$ respectively. As explained in [1], an object wave u_1 emerging directly from the magnified image of the specimen and not from the specimen itself. In order to improve the sampling capacity of the CCD, a thin-lens can be optionally introduced through a plane reference wave. The initial

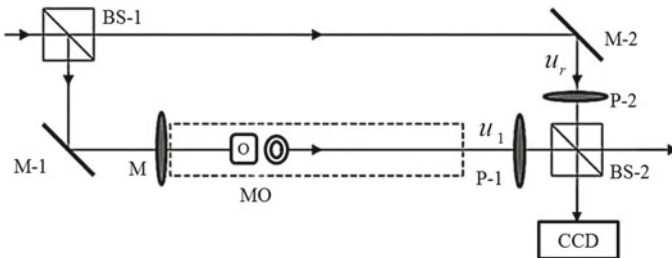


Fig. 1 Setup used in the validation of proposed method

distance for propagating object and reference wave towards CCD is considered to be the same in FIRP-DOAH technique. At the exit of the interferometer, the interference between the object wave and the reference wave creates the hologram intensity on the CCD plane.

Through the adjustment of the following complementary parameters in the proposed method in the first stages, it is possible to enlarge the size of the reconstructed image, at a larger distance in the second stage.

A. Fresnel-Kirchhoff Approximation

Under the condition of approximation of the Fresnel-Kirchhoff integral [1] the construction of a Fresnel hologram can be attained by letting the object wave to CCD plane with in the limit of Fresnel number using Eq. (1)

$$u_2(\xi, \eta) = \frac{jA}{\lambda z} \exp \left[-j \frac{\pi}{\lambda z} (\xi^2 + \eta^2) \right] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_1(x, y) \times \exp \left[-j \frac{\pi}{\lambda z} (\xi^2 + \eta^2) \right] \times \exp \left[-j \frac{2\pi}{\lambda z} (x\xi + y\eta) \right] dx dy \quad (1)$$

where A is an amplitude constant and λ is the wavelength of wave propagation. And u_2 is the diffracted object in complex wave. Furthermore, human eyes cannot normally observe complex wave, so the interference of object wave with plane reference wave in the terms of light intensity is recorded on CCD plane.

B. Plane Reference Wave

A plane reference wave u_r and the propagated object u_2 is made to interfere at (ξ, η) plane as shown in following equation

$$u_r = A \exp [ik (\xi \cos \alpha + \eta \cos \beta)] \quad (2)$$

where the k is the wave-number and A is the amplitude of constant values [5], which gives rise to a zero order in the FIRP-DOAH reconstruction. To minimize the effect of zero order, we use the composite filter on hologram recorded intensity [6]. The plane reference wave can be of any geometry as long as the sampling theorem is fulfilled. Moreover, the plane reference waves are electromagnetic fields of real functions in the spatial domain. Consequently, the Eq. (3) shows that the electric field does not only provide the part of real object hologram in large, but also reduces the computational time. Hence, we use a very small angle in an off-axis scheme (less than 0.0000001°) along vertical direction due to key involvement of electric field vector [5, 7]. Plane reference wave travels along the z-axis $\theta (\gamma = 0) = 0$, such that the sum of angles α and β should be 90° . By knowing the recording distances of an object and a plane reference wave, the hologram recording is nothing less than the digitized version of the wave fields that impinge on the CCD surface. Hence, reconstruction becomes more accurate for larger distances when the CCD size is larger than the size of the aperture of an object. Sampling interval (SI) increases in CCD plane and decreases the resolution intensity, consequently. In our proposed method, the limitations regarding the increase in CCD size has been adjusted according to Abbe's theory.

C. *Abbe's Limitation for Image Formation*

Object and plane reference waves approaching at the CCD at a specific angle on the base of optical system converge at points consistent to diffraction peaks [1–5]. Though, relation between corresponding lengths and distance z can be expressed as in Eq. (3):

$$\text{Magnification} = \frac{L_1}{L}, \quad \text{and} \quad N.A. = \sin \theta \approx \frac{L}{z} \tag{3}$$

The optical resolution of reconstructed image is firm by the numerical aperture (N.A) of the object wave confined in the CCD plane. The length of distance z is required to focus a reconstructed hologram image (RHI).

D. *Digital Thin Lens Formulae*

For determination of reconstruction distance in our proposed FIRP-DOAH method, we supposed the CCD plane as a digital thin lens. The choice of z_1 in practice is supposed by the required focusing of the real image with appropriate handling of digital thin lens formulae [7]

where, L is the side length of an object, L_1 is the side length of a CCD plane and L_2 is the side length of the RHI plane, respectively. Whereas, L'_1 is the side length of CCD plane and L'_2 is the side length of the RHI plane, respectively.

Figure 2 shows the magnification criteria for reconstruction of the RHI. This setup is best suited for highly specular micro-size objects [4]. The spatial resolution of CCD plane is expressed with total distance, $z_2 = z_1 + z$ from object to image plane. The relation of these factors with spatial resolution are investigated and presented in [1–5, 7, 8]. Consequently, the spatial resolution in the reconstructed plane can be amended by increasing the CCD aperture size. CCD size is increased relative to object for attaining the desired reconstruction of hologram image at larger distance and can be calculated as in Eq. (4). However, this method may lead to under-sampling.

$$z_1 = z \left(\frac{L_1}{L - L_1} \right) \tag{4}$$

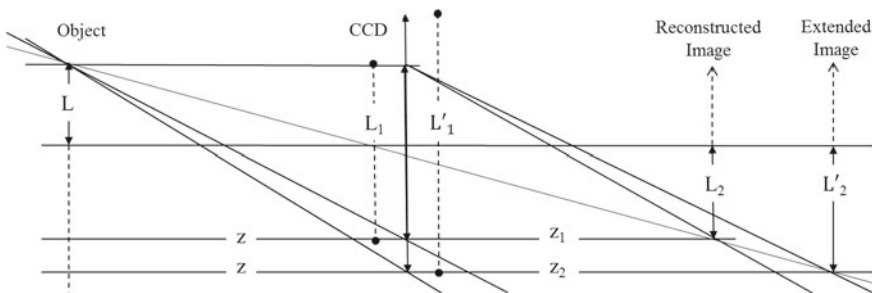


Fig. 2 Digital thin lens formulae for the reconstruction distance

E. Nyquist Shannon Theorem

In general practice the optical hologram is sampled into a digital form, on the base of the diffraction phenomena with the help of scanner in laboratory [9]. Nyquist rate is well known as the lower limit for image sampling, which avoids aliasing. However, to obtain RHI, only a finite number of samples are taken at the Image plane, whereas according to the Nyquist theorem, whereas frequency must be twice the image bandwidth in order to avoid the loss of information. The Nyquist Shannon theorem must be followed for getting an appropriate level of up-sampling. Thus, the composite filtering on the recorded intensity of hologram plane (CCD plane) is performed for enhancing detail and impulse response function of the ideal band-limited interpolation as shown by

$$h_s = \text{sinc} \left(\frac{L\xi}{\lambda z} \right) \quad (5)$$

A CCD plane is nothing less than the digitized version of the wave fields (the object wave, and the reference wave) that impinge on the CCD surface, which is determined by

$$I_{(x,y)} = |u_2|^2 + |u_r|^2 + u_2^* u_r + u_2 u_r^* \quad (6)$$

Two of the last terms of the above equation are directly proportional to the intensity distributions $I_{(x,y)}$. However, only one of these terms, either the real term or the virtual term, is generally subject to our interest. In our proposed method, performing of high pass filtering on hologram recorded intensity by means of composite filtering has suppress the virtual term and enhance the real term. Its application is helpful to understand what's going on behind or depth in an object details and ultimately reduce the zero order effect at the RHI. In this way eliminating the zero-order diffraction is much more convenient and faster than with other methods [2–4]. The RHI analysis starts with the calculation of the input and output impedances and the impulse response function. The hologram recorded intensity is further processed for composite filtering [6] because of the requirement for better matching of the original object resolution level to the RHI resolution level. This is a type of point operation based on a simple approach. Point operations are one to one mapping between object and recorded intensity of hologram where each pixel is independently adjusted and the average value of the surrounding pixel is utilized for anti-aliasing effect. A composite filter being used for suppressing the zero order term and enhancing the information detail of recorded intensity of hologram is given by

$$I_h(x, y) = T [I(x, y)] \rightarrow I_h(x, y) = \begin{pmatrix} 1 - I_{x,x} & 1 - I_{x,y} \\ 1 - I_{y,x} & 1 - I_{y,y} \end{pmatrix} \quad (7)$$

where $I_{h(x,y)}$ denotes the row and column index of hologram recorded intensity matrix after the application of composite filtering. The general formula of composite filtering is applied on hologram recorded intensity is given by:

$$I_h(x, y) = |u_2|_h^2 + |u_r|_h^2 + [u_2^*u_r]_h + [u_2u_r^*]_h \tag{8}$$

After treatment of composite filtering, we get the recorded intensity with improved data rate and proceeded for impinging the same plane reference wave again for reconstruction of hologram image.

F. FIRP-DOAH Reconstruction Model

Henceforward, the plane reference wave from Eq. (3) becomes the reconstruction wave and the reconstructed matrix becomes: $u_3 = u_r I_{rh}$. For instance with classical holography, the reconstructed wave front contains three different terms: a real image, a virtual image and a zero order of diffraction. These terms can be observed separately as an application of the FIRP-DOAH geometry. The spatial resolution of FIRP-DOAH system is limited by pixel averaging: finite CCD plane size, SI and object magnitude. All these are investigated in different studies and presented in [1–6] can be illustrated in Fig. 3 and in Eq. (10)

where, L is the side length of an object, L_1 is the side length of CCD plane and L_2 is the side length of the RHI plane, respectively. In above Fig. 3, the (x, y) , (ξ, η) and (x', y') denote the coordinates in the object, the CCD plane and RHI plane, respectively. Whereas, u'_3 is the extended FOV of u_3 matrix which is reconstructed using FIRP-DOAH method at longer distance as shown in Fig. 3, [4], and mathematically expressed by:

$$u_3 = u_r |u_2|_h^2 + u_r |u_r|_h^2 + u_r [u_2^*u_r]_h + u_r [u_2u_r^*]_h \tag{9}$$

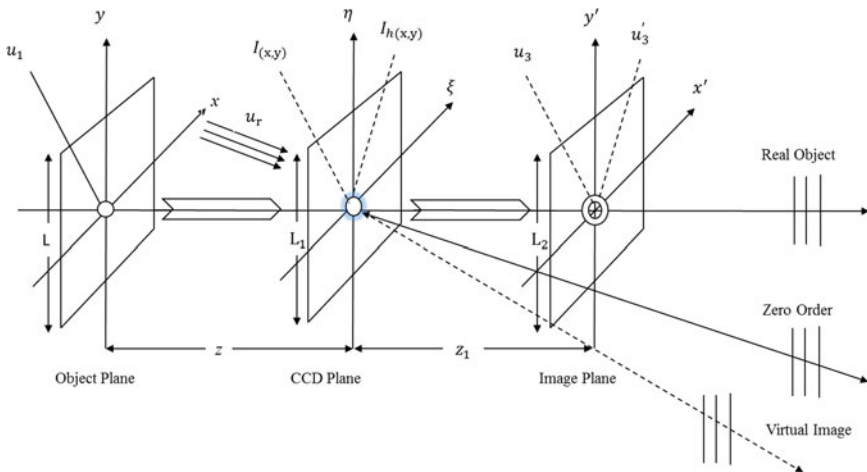


Fig. 3 FIRP-DOAH recording and reconstruction process

We noticed three parts in a basic synopsis of the proposed algorithm as given in Eq. (9). The spatial frequencies corresponding to the object image are located in the center of it, and the other two parts are the spatial frequencies corresponding to the zero order. Hence, the real image is shifted away at a larger distance due to the special adjustment of the plane reference wave and the reconstruction distance z_1 . The real part of normalized hologram is obtained in the constraint of Shannon’s Nyquist theorem which is discussed in Sect. 2 E.

G. Holographic Interpolation

The interpolation technique [10] is implemented for fast adjustment of pixel size in the reconstructed image at the rate of the SI of an object. The SI of an object and a CCD plane are given by:

$$dx = \frac{L}{M}, \quad \text{and} \quad d\xi = \frac{\lambda z}{M dx} \tag{10}$$

In the FIRP-DOAH proposed method, the size of CCD is large compare to the object plane. For this reason, the SI is increased at the CCD plane. If the CCD array has different SI in ξ and η directions, or if the extent of the object is different in x and y directions, then there is an optimal manipulation of the pixel array since both the zero orders and twin images are reconstructed simultaneously. As a consequence, we observed the enlargement of object size accordingly. If the SI in RHI is expanded into a larger area, as shown in following Fig. 4, the image gets blurred and the perceived sharpness decreases in the RHI plane. The result RHI on the base of Fresnel transform method is always affected by aliasing. However, the aliasing can be reduced by decreasing the SI or by increasing the number of pixels. In this study, we have developed an algorithm that allows us to arbitrarily change in SI of RHI of large size with standard resolution level (resolution level of an object) [9]. It shows by Nyquist criterion, that SI of recording at the CCD plane with coordinates $(d\xi, d\eta)$, needs a padding operation to get the increased number of pixels by employing Eq. (10).

Therefore, by decreasing the SI of CCD plane equal to SI of an object plane, the desired resolution of RHI plane can also be achieved. We verify the spatial resolution with the Abbe’s theorem explained in Eq. (3). After applying interpolation technique on RHI, we verified our result for Nyquist theorem in order to obtain the standard number of pixels in RHI plane using

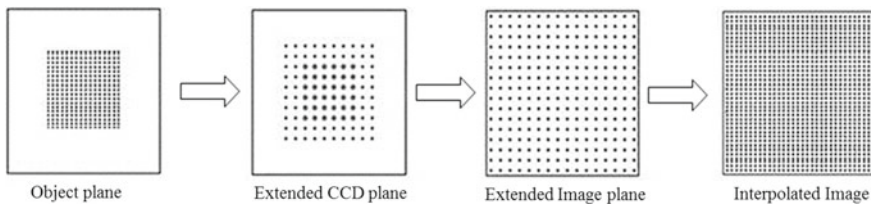


Fig. 4 Hologram recording and reconstruction Process

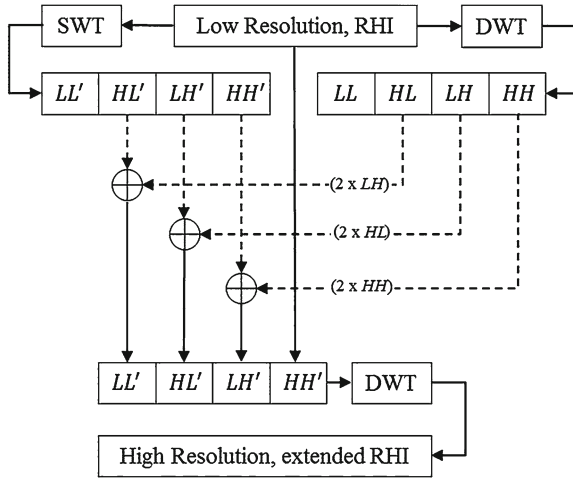


Fig. 5 FIRP-DOAH interpolation process

$$M_1 = \frac{\lambda z}{d\xi dx} \rightarrow M_2 = M_1 \left(\frac{z_2}{z_1} \right) = \frac{\lambda z}{d\xi dx} \left(\frac{z_2}{z_1} \right) \tag{11}$$

where, M_1 is the number of pixels in CCD plane, and M_2 is the number of pixels in the RHI plane, respectively. From the Eq. (11) it is clear that the resolution, and consequently the reconstruction pixel area of the RHI plane is depended on the wavelength, distance, number of the pixels $N \times M$ of an object, and $N_1 \times M_1$ is the size of CCD array. The size of RHI which has a reduced spatial lateral resolution for higher reconstruction distances can be reconstructed seamlessly using an appropriate interpolation technique according to Abbe’s theorem by following relation

$$L_2 = M_2 \left(\frac{L}{M} \right) \tag{12}$$

Thus, an image which is spread over a larger area by using FIRP-DOAH is interpolated and its perceived sharpness arises by filling the empty space around the pixels. This implies that higher number of pixels are required for RHI plane, to get the anticipated $SI (dx', dy')$ in the extended RHI plane.

$$dx' = \left(\frac{L_2}{M_2} \right) \rightarrow dx' = dx \tag{13}$$

Hence, images on prime-resolution shown in Eq. (13), required the standard number of pixels (number of pixels of an object).

H. Interpolation Techniquen

One level DWT is performed with the Haar as a wavelet function to decompose the RHI. The obtained high frequency subbands are resized upto the size of original RHI using the bicubic interpolation technique. Furthermore, to achieve the standard resolution level of extended RHI, an inherently redundant stationary wavelet scheme (SWT) of one level is separately performed on RHI [10]. The set of high frequency subbands obtained by SWT factorization are added respectively with the set of high frequency subbands (resized high frequency subbands) of DWT decomposed subbands which are obtained separately on employing RHI. Thus, the set of modified high frequency subbands is obtained. These modified high frequency subbands are carrying the sharp edges and able to reconstruct the standard size high resolution image (extended RHI). In proposed method, we do not use the low frequency subband of RHI, because it is the low resolution of original RHI. Instead of this, the original RHI is used as the low frequency subband for inverse transformation of DWT where the extended RHI is achieved (twice the size of original input image) with standard resolution (resolution of an input object image) according to following flow chart:

I. Discussion

It is important that, even if the size of RHI can be improved by zero padding on the digital hologram, the corresponding FOV remains unaffected. In fact, it depends on the real intensity of hologram system by using the generalized plane reference wave Eq. (2) and composite filter Eq. (11) with appropriate reconstruction distance Eq. (4). The reason behind the increasing number of pixels to RHI is to get large size of RHI. The reconstruction process shows that it is possible to control image parameters like focus distance, image size, and image resolution [4]. From Eq. (13), it is clear that the RHI is enlarged or contracted according to the reconstruction distance and that the size of the SI depends on the lateral number of the object standard pixels.

3 Image Quality Assessment

It is not possible to measure the PSNR of an input object image with respect to extended size of RHI (RHI is doubled the size of an input image). Thus for relative comparison, the original object size is extended equally to the size of extended RHI. Then the resized input image u'_1 and the extended RHI u'_3 : both of the same size $M' \times N'$ are evaluated on the relative scale of peak signal to noise ratio (PSNR) and the structural similarity index measure (SSIM) [11].

A PSNR technique has been used to assess the image quality reconstructed by the proposed FIRP-DOAH method. It is most easily defined through mean-squared error (MSE) shown in Eq. (14):

$$MSE = \frac{1}{M} \times \frac{1}{N} \sum_{k=1}^M \sum_{l=1}^N \|u'_1(x, y) - u'_3(x', y')\|^2 \quad (14)$$

where k, l, m and n are integers with ranges to maximum value of N (height of image) and M (width of image). The bigger the value of PSNR: better the image quality which is analyzed by

$$PSNR = 10 \times \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{15}$$

The SSIM comparison for resized u'_1 and the extended RHI u'_3 is defined as

$$SSIM(x, y) = l(x, y) c(x, y) s(x, y) \tag{16}$$

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{17}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{18}$$

$$s(x, y) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{19}$$

where $l(x, y)$ is the function of luminance comparison to measure the images closeness on the base of mean luminance values, i.e. μ_x and μ_y whereas maximum value of $l(x, y)$ is equal to 1, if and only if $\mu_x = \mu_y$. The second term $c(x, y)$ is used in Eq. (16) to measure the contrast on the base of standard deviation σ_x and σ_y . The maximal value of contrast term as given in Eq. (18) is achieved at $\sigma_x = \sigma_y$. Third term in Eq. (16) is the $s(x, y)$ which measure the structure comparison between the u'_1 (extended input object size) and the u'_3 (extended RHI), respectively. the σ_{xy} used in 19 shows the covariance which is useful to analyze the correlation between two images. In the above Eqs. (17–19) the positive constants C_1, C_2 and C_3 are used, respectively, to eliminate the null denominator. The quality value of SSIM is varied in the positive index of [0, 1]. A value of 1 shows that $u'_1 = u'_3$, and value of 0 shows no correlation between the two images.

4 Simulation Criteria

To prove the veracity of the proposed method, simulations in Matlab were done. The simulated object is a black-and-white film, as shown in Fig. 7. The holograms recorded at a distance of 35 cm on a CCD camera with 256×256 pixels were illuminated with a coherent light of the 633 nm wavelength [7, 9]. CCD is assumed to be a thin lens. From the Eqs.(10) and (13), it is observed that the SI increases with the reconstruction distance so that the RHI, in terms of resolution, is reduced

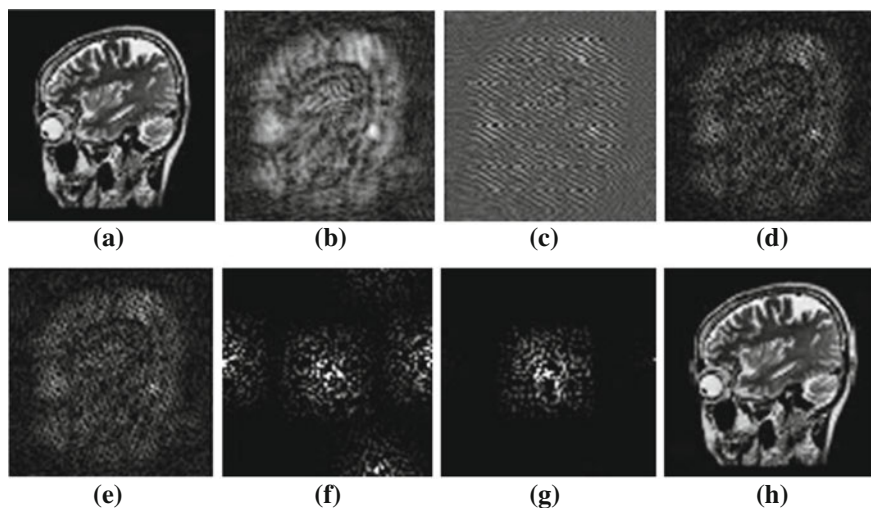


Fig. 6 Proposed simulation stages for test target object brain cells. **a** Original object. **b** Propagated image. **c** Hologram intensity. **d** Filtering intensity. **e** Hologram reconstruction. **f** FFT of reconstructed hologram. **g** FFT of propagated hologram. **h** Hologram reconstructed image

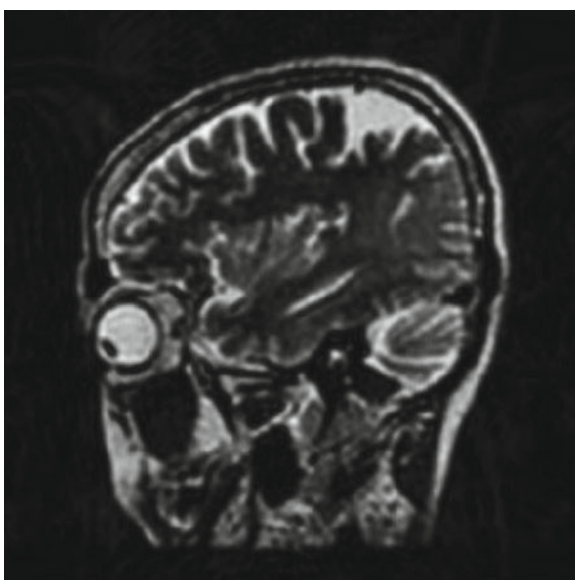


Fig. 7 Extension of RHI by using wavelet interpolation approach of Fig. 5h along with good value of PSNR (dB) = 55.70 and the high index of SSIM = 0.98/[0,1]

for a longer distance. For this reason, we would increase the number of pixels to 512×512 in a RHI on employing bi-cubic interpolation technique to achieve large DOAH image. Therefore, well-focused RHI is obtained at larger distance of 91.26 cm.

In such scenarios, the proposed FIRP-DOAH method has significant advantages over conventional augmenting holographic techniques because it offers finer FOV at a larger distance. The results are shown in Figs. 6 and 7, respectively.

The SI of RHI is same as the SI of the original object. In order to improve the performance of the extracted algorithms, wavelet transform is used with the Haar as a wavelet function to produce high resolution of large size image and efficient in terms of a reduced computational time.

5 Conclusion and Future Work

A simple method has been developed for enlarging the size of the reconstructed hologram image at a larger distance in digital holography. Our proposed method has the advantage of sweeping out the *dc*-term from the reconstructed hologram. The results of the simulation are also found to be efficient in terms of computation time achieving half that of previous approaches. The application of this method not only provides the enlargement of the image size but also produces better resolution and higher accuracy even at a larger distance.

Acknowledgments This work was by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2011-0026245).

References

1. Goodman JW (2004) Introduction to fourier optics, 3rd edn. Roberts & Company Publishers, Englewood
2. Schnars U, Werner PO (2002) Digital recording and numerical reconstruction of holograms. Meas Sci Technol 13(9):R85
3. Chen GL, Lin CY, Kuo MK, Chang CC (2008) Numerical reconstruction and twin-image suppression using an off-axis Fresnel digital hologram. Appl Phys B 90:527–532
4. Voelz DG (2011) Computational fourier optics: a MATLAB tutorial. SPIE, Bellingham
5. Poon T-C (2007) Optical scanning holography with MATLAB. Springer, New York
6. Ferraro P, Coppola G, Alfieri D, De Nicola S, Finizio A, Pierattini G (2004) Controlling images parameters in the reconstruction process of digital holograms. IEEE J Sel Topics Quantum Electron 10(4):829–839
7. Muhammad N, Kim D-G (2012) A simple approach for large size digital off-axis hologram reconstruction. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012. London, UK, 2012: pp 1183–1188, 4–6 July 2012
8. Gonzalez RC, Woods RE (2002) Digital image processing, 2nd edn. Prentice Hall, Upper Saddle River
9. Lancaster D, A review of some image pixel interpolation algorithms. <http://www.tinaja.com/gurgrm01.as>

10. Demirel H, Anbarjafari G (2011) Discrete wavelet transform-based satellite image resolution enhancement. *IEEE T Geosci Remote Sens* 49(6–1):1997–2004
11. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612

A Novel Two-Scan Connected-Component Labeling Algorithm

Lifeng He, Yuyan Chao, Yun Yang, Sihui Li, Xiao Zhao and Kenji Suzuki

Abstract This chapter proposes a novel two-scan labeling algorithm. In the first scan, all conventional two-scan labeling algorithms process image lines one by one, assigning each foreground pixel a provisional label, finding and resolving label equivalences between provisional labels. In comparison, our proposed method first scans image lines every four lines, assigns provisional labels to foreground pixels among each three lines, and finds and resolves label equivalences among those provisional labels. Then, it processes the leaving lines from top to bottom one by one, and for each line, assigns provisional labels to foreground pixels on the line, and finds and resolves label equivalences among the provisional labels and those assigned to the

L. He (✉) · Y. Yang · S. Li · X. Zhao

College of Electrical and Information Engineering, Shaanxi University of Science and Technology, Weiyang-Qu, Daxue-Lu, Xi'an, Shaanxi 710021, China
e-mail: helifeng@sust.edu.cn, helifeng@ist.aichi-pu.ac.jp

Y. Yang

e-mail: yangyun@sust.edu.cn

S. Li

e-mail: lisihui@sust.edu.cn

X. Zhao

e-mail: zhaox@sust.edu.cn

L. He

Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute, Aichi 480-1198, Japan

Y. Chao

Graduate School of Environment Management, Nagoya Sangyo University, Owariasahi, Aichi 488-8711, Japan
e-mail: chao@nagoya-su.ac.jp

K. Suzuki

Department of Radiology, Division of the Biological Sciences,
The University of Chicago, Chicago, IL 60637, USA
e-mail: suzuki@uchicago.edu

foreground pixels on the lines immediately above and below the current line. With our method, the average number of times for checking pixels for processing a foreground pixel will decrease; thus, the efficiency of labeling can be improved. Experimental results demonstrated that our method was more efficient than conventional label-equivalence-based labeling algorithms.

Keywords Computer vision · Connected component · Fast algorithm · Label equivalence · Labeling · Pattern recognition

1 Introduction

Labeling of connected components in a binary image is one of the most fundamental operations in pattern analysis, pattern recognition, computer (robot) vision, and machine intelligence [6, 20]. Especially in real-time applications such as traffic-jam detection, automated surveillance, and target tracking, faster labeling algorithms are always desirable.

Many algorithms have been proposed for addressing this issue, because the improvement of the efficiency of labeling is critical in many applications. For ordinary computer architectures and 2D images, there are mainly two types of labeling algorithms:

- (1) Raster-scan algorithms. These algorithms process an image in the raster-scan way. There are multi-scan algorithms [7, 9, 25] and two-scan algorithms [8, 10–15, 18, 19, 22].
- (2) Label propagation algorithms. These algorithms access an image in an irregular way. There are run-based algorithms [2, 24] and contour-tracing algorithms [3, 17].

According to experimental results on various types of images, the algorithm proposed in Ref. [12], which is an improvement on the two-scan algorithm proposed in Ref. [10], is the most efficient one, and has been used for various applications [1, 4, 5]. For convenience, we denote this algorithm as *MECTSL* (Most Efficient Conventional Two-Scan Labeling) *algorithm*.

Two-scan labeling algorithms complete labeling in two scans by four processes: (1) provisional label assignment, i.e., assigning a provisional label to each foreground pixel; (2) equivalent label finding and recording, i.e., finding all provisional labels assigned to foreground pixels that are connected and using some data structures to record them as equivalent labels; (3) label-equivalence resolving, i.e., finding a representative label for all equivalent provisional labels; (4) label replacement, i.e., replacing each provisional label by its representative label. The first process is completed in the first scan, the second and the third processes are completed during the first scan and/or between the first scan and the second scan, and the fourth process is completed in the second scan.

The *MECTSL* algorithm uses equivalent label sets and a representative label table to record equivalent labels and resolve the label equivalences. For convenience, an

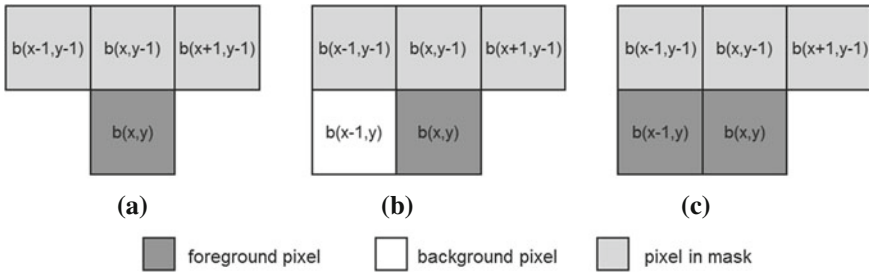


Fig. 1 Mask for the eight-connected connectivity

equivalent label set with the representative label u is denoted as $S(u)$, and that the representative label of a provisional label l is t is denoted as $T[l] = t$.

Moreover, this algorithm processes foreground pixels following background pixels and those foreground pixels following foreground pixels in a different way, which can reduce the number of times for checking neighboring pixels and, thus, enhance the efficiency.

In the first scan, this algorithm uses the mask shown in Fig. 1a, which consists of three scanned (processed) neighbor of the current foreground pixels, to assign provisional labels to foreground pixels, and to record and resolve label equivalences. At any moment in the first scan, all equivalent provisional labels are combined in an equivalent label set with the same representative label.

For the case where the current foreground pixel follows a background pixel (Fig. 1b), if there is no label (foreground pixel) in the mask, this means that the current foreground pixel does not connect with any scanned foreground pixel, and the current foreground pixel belongs to a new connected component up to now. The algorithm assigns a new provisional label m to the current foreground pixel, which is initialized to 1, and establishes the equivalent label set $S(m) = \{m\}$; it sets the representative label table as $T[m] = m$, and $m = m + 1$ for later processing.

In both cases, if there are provisional labels belonging to different equivalent label sets, all provisional labels in those sets are equivalent labels, and they need to be combined. Suppose that u and v are equivalent labels that belong to $S(T[u])$ and $S(T[v])$, respectively. If $T[u] = T[v]$, the two equivalent label sets are the same, thus, nothing needs to be done. Otherwise, without loss of generality, suppose that $T[u] < T[v]$, i.e., $T[u]$ is the smallest label in the two equivalent label sets, then the combination of the two equivalent label sets can be completed by the following operations:

$$S(T[u]) = S(T[u]) \cup S(T[v]);$$

$$(\forall s \in S(T[v]))(T[s] = T[u]).$$

As soon as the first scan is finished, all equivalent labels of each connected component have been combined into an equivalent label set with a unique representative label. In the second scan, by replacement of each provisional label with its

representative label, all foreground pixels of each connected component will be assigned a unique label.

As introduced above, in the first scan, all two-scan labeling algorithms process an image line by line. For each line, they resolve the connectivities of the foreground pixels on the line and their foreground neighboring pixels on the line immediately above the current line. In fact, the connectivities of the foreground pixels on the current line and their foreground neighboring pixels on the line immediately below the current line can also be resolved easily. Comparing to resolve these connectivities later when scanning the next line, it would be more efficient to do that at the same time processing the current line.

This paper presents a new two-scan labeling algorithm, which is an extension of Ref. [16]. In our algorithm, the label equivalences are recorded and resolved in exactly the same way introduced above. Our proposed first-scan method consists of two parts: In the first part, it scans image lines every four lines. For each scan line, it assigns to the foreground pixels in the line and its two neighboring lines provisional labels and resolves the label equivalences among them. In the second part, it scans the unprocessed lines in the first step one by one. For each line, it assigns to the foreground pixels in the line provisional labels and resolves the label equivalences among them and those in its two neighbor lines. Experimental results showed that the efficiency of our first-scan method is superior to that of the first scan of the MECTSL algorithm.

2 Outline of Our Proposed First-Scan Method

For an $N \times M$ binary image, we use $b(x, y)$ to denote the pixel value at (x, y) in the image, where $1 \leq x \leq N$, $1 \leq y \leq M$, and $v(x, y)$ for the value of $b(x, y)$. For convenience, we suppose that the value of foreground pixels is 1 and that of background pixels is 0. All pixels in the edge of an image are considered to be background pixels. Moreover, similar to the MECTSL algorithm, we process pixels following a foreground pixel and those following a background pixel in different ways.

Our first-scan method consists of two parts: Scan 1-A and Scan 1-B. In Scan 1-A, from line 3, it scans image lines every four other lines, i.e., in the order of line 3, line 7, line 11, . . . (the black lines in Fig. 2). For each current line being processed, it assigns to the foreground pixels in the line and its neighbor lines (the gray lines in Fig. 2) provisional labels and resolves the label equivalences among them. By Scan 1-A, all foreground pixels in each area consisting of black and gray lines in Fig. 2 will be assigned provisional labels, and the label equivalences among them will be resolved.

Then, Scan 1-B scans the lines unprocessed in the Scan 1-A (the white lines in Fig. 2) in the order of line 5, line 9, line 13, . . . For each current line, it assigns to the foreground pixels in the line provisional labels and resolves the label equivalences among them and those in their neighboring lines.

Fig. 2 Scanning method of the first scan in our proposed method

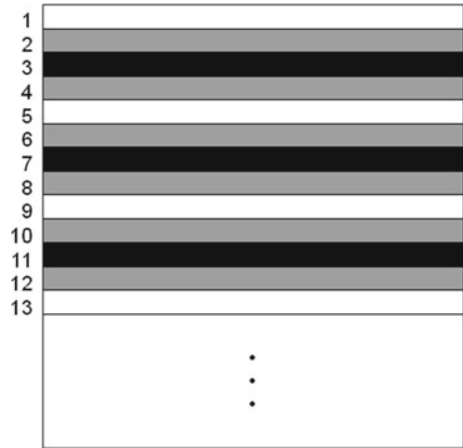
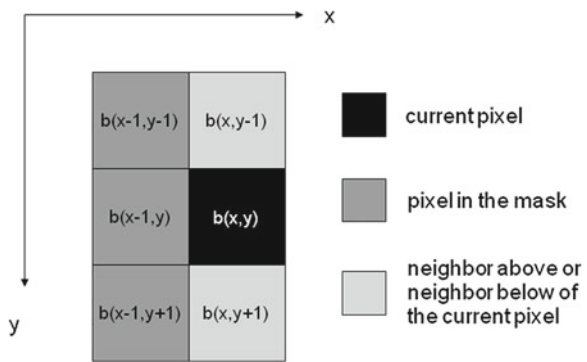


Fig. 3 Mask used in Scan1-A



When Scan 1-B is finished, all foreground pixels in the given image will be assigned provisional labels, and the label equivalences among them will be resolved, i.e., all equivalent labels will be combined into an equivalent label set with a unique representative label.

In the second scan, similar to other two-scan labeling algorithms, by replacing the provisional label of each foreground pixel with its representative label, we can complete the whole labeling process.

2.1 Scan 1-A

In Scan 1-A, our method uses the mask shown in Fig. 3 to assign to the current pixel $b(x, y)$, its neighbor above, $b(x, y - 1)$ and its neighbor below, $b(x, y + 1)$ provisional labels, and to resolve the label equivalences in the mask. The following four cases should be considered.

Case 1: The current pixel $b(x, y)$ is a background pixel that follows another background pixel (Fig. 4b, c, e, f). The neighbor above, $b(x, y - 1)$, and the neighbor below, $b(x, y + 1)$, of the current pixel can be processed as follows:

```

if  $b(x, y-1)$  is a foreground pixel
  if  $b(x-1, y-1)$  is a foreground pixel,
    assigning  $b(x, y-1)$  the label of  $b(x-1, y-1)$ ;
  else assigning  $b(x, y-1)$  a new label.
end of if
end of if
if  $b(x, y+1)$  is a foreground pixel
  if  $b(x-1, y+1)$  is a foreground pixel,
    assigning  $b(x, y+1)$  the label of  $b(x-1, y+1)$ ;
  else assigning  $b(x, y+1)$  a new label.
end of if
end of if

```

The pseudo codes of the above process, denoted as $process1(x, y)$, can be given as follows, where m means the current provisional label and is initialized as 1:

```

if  $v(x, y-1) > 0$ 
  if  $v(x-1, y-1) > 0$ ,  $v(x, y-1) = v(x-1, y-1)$ ;
  else  $v(x, y-1) = m$ ,  $m = m + 1$ ;
end of if
end of if
if  $v(x, y+1) > 0$ 
  if  $v(x-1, y+1) > 0$ ,  $v(x, y+1) = v(x-1, y+1)$ ;
  else  $v(x, y+1) = m$ ,  $m = m + 1$ .
end of if
end of if

```

Case 2: The current pixel is a background pixel following a foreground pixel (Fig. 4a, d).

The neighbor above, $b(x, y - 1)$, and the neighbor below, $b(x, y + 1)$, of the current pixel can be processed by the following pseudo codes, denoted as $process2(x, y)$:

```

if  $v(x, y-1) > 0$ ,  $v(x, y-1) = v(x-1, y)$ ;
end of if
if  $v(x, y+1) > 0$ ,  $v(x, y+1) = v(x-1, y)$ ;
end of if

```

Case 3: The current pixel is a foreground pixel following a background pixel (Fig. 5b–e).

The current foreground pixel and its above neighbor $b(x, y - 1)$ and below neighbor $b(x, y + 1)$ can be processed by the following pseudo codes, denoted as $process3(x, y)$, where $resolve(a, b)$ denotes to resolve the label equivalence between labels a and b :

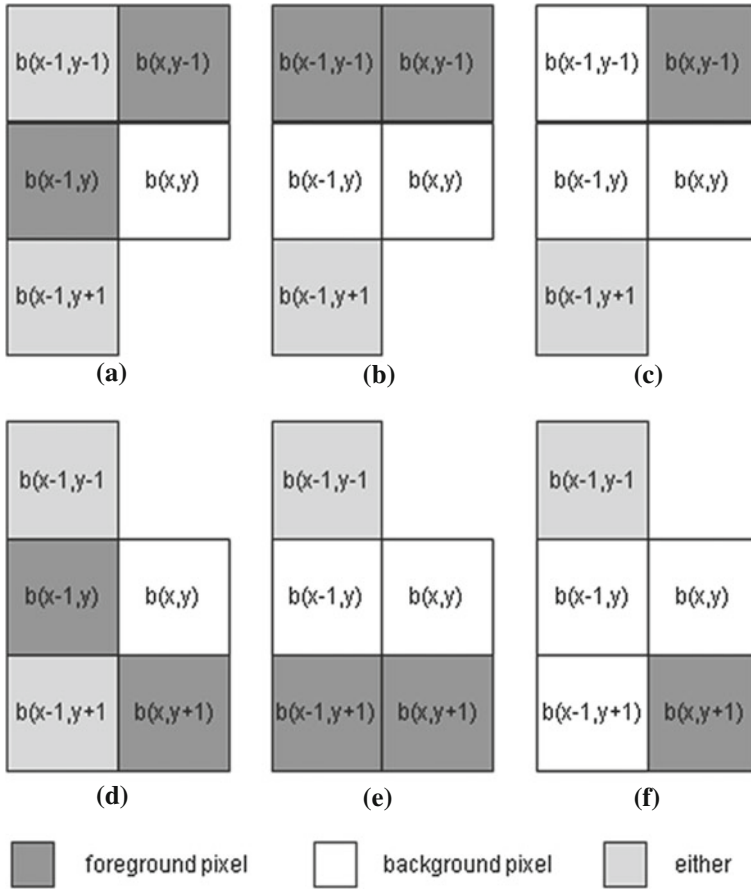


Fig. 4 Cases in Scan1-A where the current pixel is a background pixel: **a–c** for processing pixel $b(x, y - 1)$ when it is a foreground pixel; **d–f** for processing pixel $b(x, y + 1)$ when it is a foreground pixel

```

if  $v(x-1, y-1) > 0$ 
     $v(x, y) = v(x-1, y-1)$ ;
    if  $v(x-1, y+1) > 0$ ,  $resolve(v(x, y), v(x-1, y+1))$ ;
    end of if
else if  $v(x-1, y+1) > 0$ ,  $v(x, y) = v(x-1, y+1)$ ;
else  $v(x, y) = m$ ,  $m = m + 1$ ;
end of if
    
```

Case 4: The current pixel is a foreground pixel following a foreground pixel (Fig. 5a).

The current foreground pixel and its neighbor above, $b(x, y - 1)$, and the neighbor below, $b(x, y + 1)$, can be processed by the following pseudo codes, denoted as $process4(x, y)$:

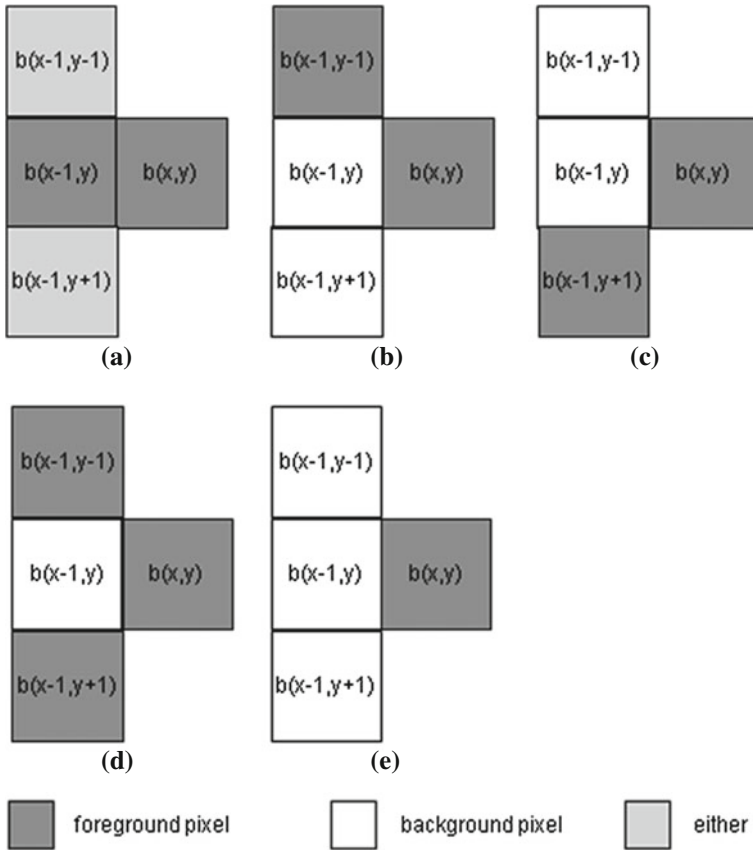


Fig. 5 Cases in Scan1-A where the current pixel is a foreground pixel

```

v(x, y)= v(x-1, y);
if v(x, y-1)>0, v(x, y-1)= v(x, y);
end of if
if v(x, y+1)>0, v(x, y+1)= v(x, y);
end of if
    
```

For each scanned line in Scan 1-A, from the second pixel to the pixel before the last pixel (the first pixel and the last pixel are edge pixels of the given image, thus, background pixels), it processes pixels one by one recursively as follows: While the current pixel is a background pixel, it calls *process1(x, y)* for processing; for the first foreground pixel after background pixels, it calls *process3(x, y)* for processing; then, for the following foreground pixels, it calls *process4(x, y)* for processing; Lastly, for the first background pixel after foreground pixels, it calls *process4(x, y)* for processing.

The pseudo codes of Scan 1-A are shown as follow, notice that, for any foreground pixel, its x -coordinate cannot be larger than $N - 1$.

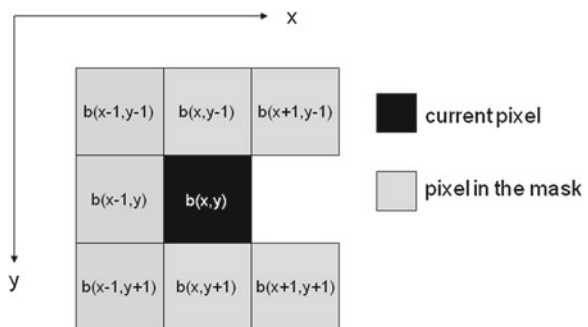
```

m=1;
for (y=3; y<M; y=y+4)
  for (x=2; x<N; x++)
    if (v(x, y)<1)
      process1(x, y);
    else
      process3(x, y);
      x++;
      while (v(x,y)>0)
        process4(x, y);
        x++;
      end while
      process2(x, y);
    end if
  end for
end for
    
```

2.2 Scan 1-B

After Scan 1-A, from line 5, in the order of line 5, line 9, line 13, . . . , Scan 1-B scans the lines unprocessed in Scan 1-A one by one. It does nothing for background pixels. For each foreground pixel, it uses the mask shown in Fig. 6 to assign to the pixel a provisional label and resolves the label equivalences in the mask. Because the foreground pixels that are connected each other in the mask, such a combination is called a *connected part*, belonging to the same connected component, and their provisional labels are equivalent labels and belong to the same equivalent label set; thus, a connected part can be considered as a single foreground pixel. For this reason, checking the pixels in the mask in the order of the largest number of the neighbors of a pixel first will reduce the number of times for checking pixels in the mask; thus, it leads to efficient processing [15].

Fig. 6 Mask used for Scan1-B



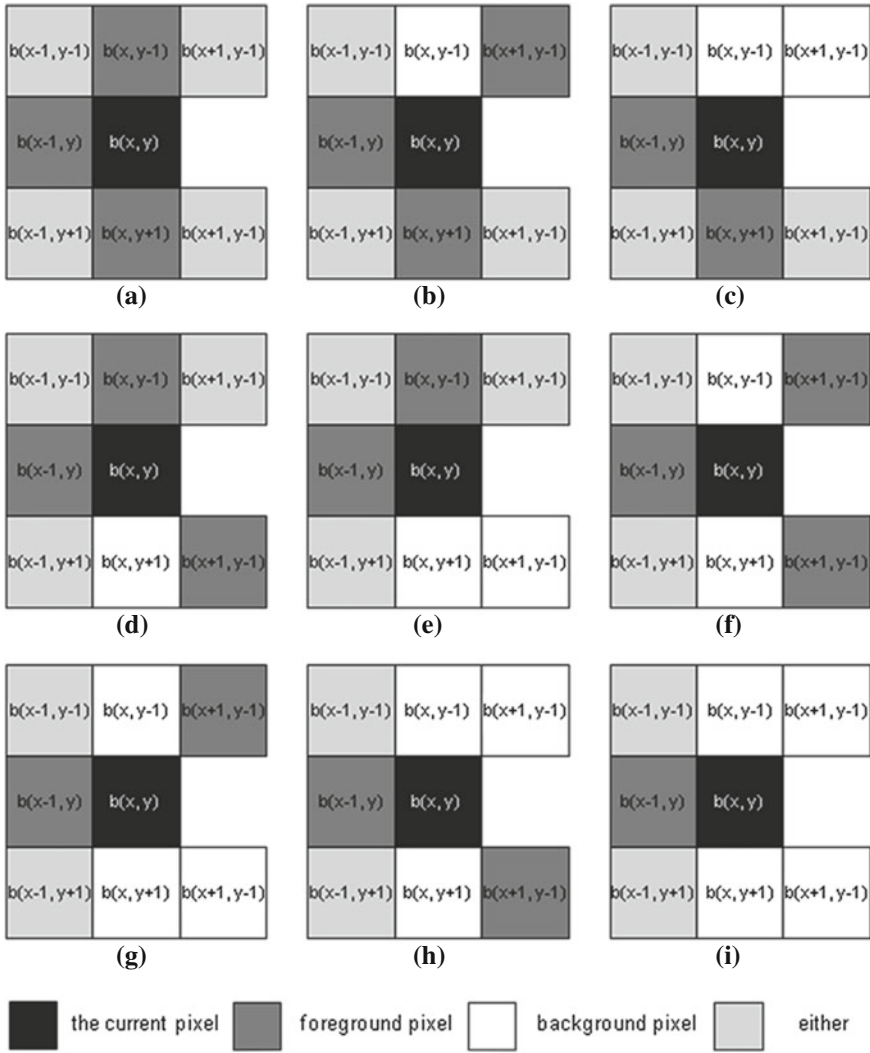


Fig. 7 Cases in Scan 1-B where the current pixel is a foreground one following a foreground pixel

In the case where the current foreground pixel follows another foreground pixel, there are nine subcases shown in Fig. 7. We process the current pixel $b(x, y)$ as follows:

- (1) because $b(x - 1, y)$ is a foreground pixel, we assign to the current foreground pixel the same provisional label of $b(x - 1, y)$;
- (2) because the number of connected parts does not depend on whether $b(x - 1, y - 1)$ and/or $b(x - 1, y + 1)$ is a background pixel or a foreground pixel, we do not need to check either of them;

(3) according to the number of neighbors of each pixel in the mask, the order for checking the pixels except for $b(x-1, y)$, $b(x-1, y-1)$ and $b(x-1, y+1)$ is $b(x, y-1)$, $b(x, y+1) \rightarrow b(x+1, y-1) \rightarrow b(x+1, y+1)$. If $b(x, y-1)$ is a foreground pixel (e.g., Fig. 7a), $b(x, y-1)$ and $b(x-1, y)$ belong to the same connected part, we need to do nothing for the pixel. Moreover, in this subcase, whether $b(x+1, y-1)$ is a foreground pixel or not does not change the number of connected parts in the mask, we also need to do nothing about this pixel. On the other hand, if $b(x, y-1)$ is a background pixel and $b(x+1, y-1)$ is a foreground pixel (e.g., Fig. 7b), we need to resolve the label equivalence of $v(x-1, y)$ and $v(x+1, y-1)$. Then $b(x, y+1)$ and $b(x+1, y+1)$ can be processed in a similar way. For convenience, we denote the above process to *process5*.

On the other hand, in the case where the current foreground pixel $b(x, y)$ follows a background pixel, there are 21 subcases, as shown in Fig. 8a–u. Except for $b(x-1, y)$, according to the number of neighbors of each pixel in the mask, the order for checking pixels is $b(x, y-1)$, $b(x, y+1) \rightarrow b(x-1, y-1)$, $b(x-1, y+1) \rightarrow b(x+1, y-1)$, $b(x+1, y+1)$. For each pixel, similarly in *process5(x, y)*, if there are more than one connected part in the mask (e.g., Fig. 8a–d), we need to resolve the label equivalences among them, and assign to the current foreground pixel its representative label. On the other hand, if there is only one connected part in the mask (e.g., Fig. 8e), we only need to assign to the current foreground pixel the representative label. Lastly, if there is no foreground pixel in the mask (Fig. 8u), we only need to assign to the current pixel a new provisional label. For convenience, we call the process for this case *process6*.

In conclusion, Scan 1-B scans a line from the second pixel, (1) if the current pixel is a background pixel, nothing is done; (2) for the current foreground pixel following a background pixel, it calls *process6*; and (3) for the following foreground pixels, it calls *process5*. The above processing is recursively called until the end of the line being processed.

The pseudo codes of Scan 1-B are shown as follows:

```

for (y=5; y<M; y=y+4)
  for (x=1; x<N; x++)
    if (v(x, y)>0)
      process6;
      x++;
      while (v(x, y)>0)
        process5;
        x++;
      end while
    end if
  end for
end for

```

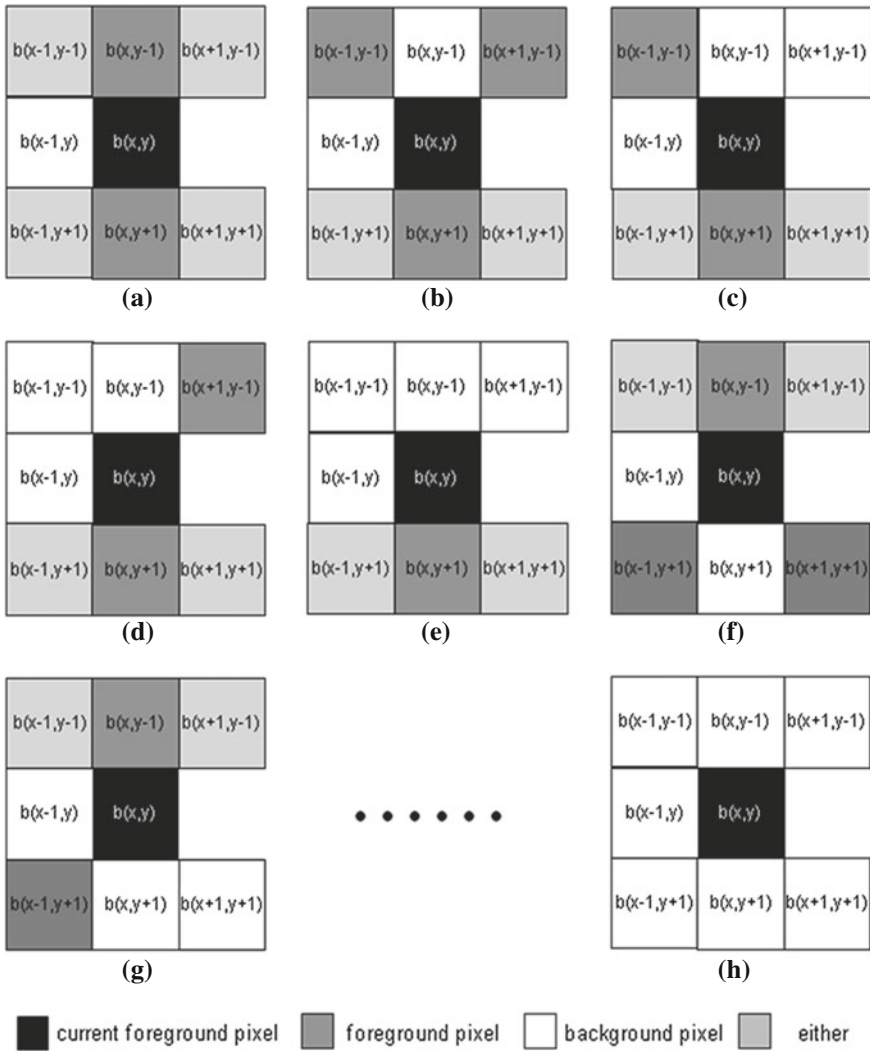
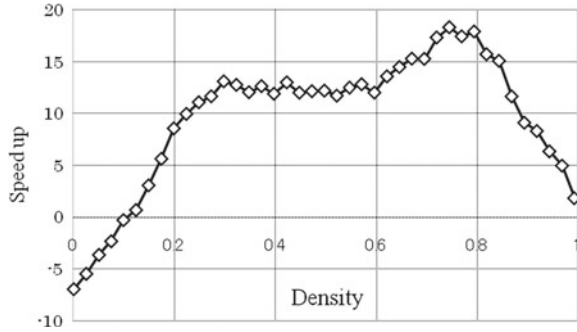


Fig. 8 Cases in Scan 1-B where the current pixel is a foreground one following a background pixel

2.3 The Second Scan

As soon as the first scan is finished, all provisional labels assigned to each connected component have been combined in an equivalent label set with a unique representative label. During the second scan, similar to all conventional two-scan labeling algorithms, by replacing each provisional label with its representative label, we can complete labeling.

Fig. 9 Speed-up of our method compared to the MECTSL method on the 512×512 noise images



3 Comparative Evaluation

We implemented the MECTSL algorithm and our algorithm with the C language on a PC-based workstation (Intel Pentium D 3.0 GHz + 3.0 GHz CPUs, 2 GB Memory, Mandriva Linux OS). Because our method is a new first-scan method (as we described above, the second scan of our method is exactly the same with the MECTSL method), we will compare the performances of the two methods only on the first scan. All data in this section were obtained by averaging of the execution time for 10,000 runs with a single core.

Images used for testing included of four types: noise images, natural images, texture images, and medical images.

Noise images consist of forty one 512×512 -sized noise images were generated by thresholding of the images containing uniform random noise with 41 different threshold values from 0 to 1000 in steps of 25.

On the other hand, 50 natural images, including landscape, aerial, fingerprint, portrait, still-life, snapshot, and text images, obtained from the Standard Image Database (SIDBA) developed by the University of Tokyo¹ and the image database of the University of Southern California², were used for realistic testing of labeling algorithms. In addition, seven texture images, which were downloaded from the Columbia-Utrecht Reflectance and Texture Database³, and 25 medical images obtained from a medical image database of The University of Chicago were used for testing. All of these images were 512×512 pixels in size, and they were transformed into binary images by means of Otsu’s threshold selection method [20].

Figure 9 shows the speed-up of our method compared to the MECTSL method on the 512×512 noise images, where the vertical axis is defined as $(t_1 - t_2)/t_1$, where t_1 is the execution time of the MECTSL algorithm and t_2 is that of the proposed method.

¹ <http://sampl.ece.ohio-state.edu/data/stills/sidba/index.htm>

² <http://sipi.usc.edu/database/>

³ <http://www1.cs.columbia.edu/CAVE/software/curet/index.php>

Table 1 Comparison of various execution times [*msec*] for natural images, medical images, and textural images

Image type		The MECTSL algorithm	Our proposed algorithm
Natural	Max.	1.83	1.57
	Mean	0.96	0.89
	Min.	0.44	0.39
Medical	Max.	0.98	0.95
	Mean	0.73	0.70
	Min.	0.59	0.29
Textural	Max.	1.48	1.38
	Mean	1.09	0.95
	Min.	0.79	0.53

The experimental results on the natural images, the medical images, and the textural images are shown in Table 1.

4 Concluding Remarks

In this chapter, we presented a new method for the first scan of label-equivalence-based two-scan labeling algorithms. In our proposed method, the first scan consists of two subscans: Scan 1-A and Scan 1-B. In Scan 1-A, we process image lines every four lines. For each current line being scanned, we assign to the foreground pixels in the line and its neighboring lines provisional labels and resolve the label equivalences among them. In Scan 1-B, we scan the lines that were unprocessed in Scan 1-A one by one. For each current line, we assign to the foreground pixels in the line provisional labels and resolve the label equivalences among them and those in its neighboring lines processed in Scan 1-A. By our method, the number of times for checking pixels for assigning provisional labels and processing label equivalences is decreased; thus, the efficiency of labeling is improved. Our experimental results demonstrated that our method was more efficient than the first scan of conventional label-equivalence-based two-scan labeling algorithms.

Acknowledgments This work was supported in part by the Ministry of Education, Science, Sports and Culture, Japan, Grant-in-Aid for Scientific Research (C), 23500222, 2011.

References

1. Alexey A, Tomas K, Florentin W, Babette D (2011) Real-time image segmentation on a GPU. Facing Multicore Chall Lect Notes Comput Sci 6310:131–142
2. Ballard DH (1982) Computer vision. Prentice-Hall, Englewood

3. Chang F, Chen CJ, Lu CJ (2004) A linear-time component-labeling algorithm using contour tracing technique. *Comput Vis Image Underst* 93:206–220
4. Christopher W, Nicholas Graham TC, Pape JA (2010) Seeing through the fog: an algorithm for fast and accurate touch detection in optical tabletop surfaces. In: *ACM international conference on interactive tabletops and surfaces (ITS '10)*. ACM, New York, USA, pp 73–82 (2010)
5. Dellen B, Erdal EA, Wrgtter F (2009) Segment tracking via a spatiotemporal linking process including feedback stabilization in an n-D lattice model. *Sensors* 9(11):9355–9379
6. Gonzalez RC, Woods RE (1992) *Digital image processing*. Addison Wesley, Reading
7. Haralick RM (1981) Some neighborhood operations. In: *Real time/parallel computing image analysis*. Plenum Press, New York, pp 11–35
8. Haralick RM, Shapiro LG (1992) *Computer and robot vision I*. Addison-Wesley, Reading, pp 28–48
9. Hashizume A, Suzuki R, Yokouchi H et al (1990) An algorithm of automated RBC classification and its evaluation. *Bio Med Eng* 28(1):25–32
10. He L, Chao Y, Suzuki K (2007) A linear-time two-scan labeling algorithm. In: *2007 IEEE international conference on image processing (ICIP)*. San Antonio, Texas, USA, pp V-241-V-244
11. He L, Chao Y, Suzuki K (2008) A run-based two-scan labeling algorithm. *IEEE Trans Image Process* 17(5):749–756
12. He L, Chao Y, Suzuki K, Wu K (2009) Fast connected-component labeling. *Pattern Recognit* 42:1977–1987
13. He L, Chao Y, Suzuki K (2010) An efficient first-scan method for label-equivalence-based labeling algorithms. *Pattern Recognit Lett* 31:28–35
14. He L, Chao Y, Suzuki K (2011) A run-based one-and-a-half-scan connected-component labeling algorithm. *Int J Pattern Recognit Artif Intell* 24(4):557–579
15. He L, Chao Y, Suzuki K (2011) Two efficient label-equivalence-based connected-component labeling algorithms for three-dimensional binary images. *IEEE Trans Image Process* 20(8):2122–2134
16. He L, Chao Y, Suzuki K (2012) A new two-scan algorithm for labeling connected components in binary images. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE, 4–6 July 2012 U.K , London*, pp 1141–1146
17. Hu Q, Qian G, Nowinski WL (2005) Fast connected-component labeling in three-dimensional binary images based on iterative recursion. *Comput Vis Image Underst* 99:414–434
18. Lumia R, Shapiro L, Zungia O (1983) A new connected components algorithm for virtual memory computers. *Comput Vis Graph Image Process* 22(2):287–300
19. Naoi S (1995) High-speed labeling method using adaptive variable window size for character shape feature. *IEEE Asian Conf Comput Vis* 1:408–411
20. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9:62–66
21. Ronsen C, Denjiver PA (1984) *Connected components in binary images: the detection problem*. Research Studies Press, New York
22. Rosenfeld A, Pfalts JL (1966) Sequential operations in digital picture processing. *J ACM* 13(4):471–494
23. Samet H (1984) The quadtree and related hierarchical data structures. *Comput Surv* 16(2):187–260
24. Shima Y, Murakami T, Koga M, Yashiro H, Fujisawa H (1990) A high-speed algorithm for propagation-type labeling based on block sorting of runs in binary images. In: *Proceedings of 10th international conference pattern recognition*, pp 655–658
25. Suzuki K, Horiba I, Sugie N (2003) Linear-time connected-component labeling based on sequential local operations. *Comput Vis Image Underst* 89:1–23
26. Udupa JK, Ajjanagadde VG (1990) Boundary and object labeling in three-dimensional images. *Comput Vis Graph Image Process* 51(3):355–369

Approaches to Bayesian Network Model Construction

Ifeyinwa E. Achumba, Djamel Azzi, Ifeanyi Ezebili and Sebastian Bersch

Abstract Bayesian Network (BN) has sound mathematical basis, enables reasoning under uncertainty, and facilitates the update of beliefs, given new evidence. It also enables the visual representation of a model. These make BN suitable for solving uncertainty problems. This chapter details BN model construction approaches and presents our experiences with selecting the optimal construction approach.

Keywords Bayesian networks · BN model construction · BN model parameterization · Parameter learning · Performance index · Structure learning

1 Introduction

A Bayesian Network (BN) model consists of two component parts: *network structure* (the qualitative part of a BN)—a set of variables (nodes) and a set of directed edges interconnecting the nodes without creating directed loops, such that the nodes, together with the edges, form a Directed Acyclic Graph (DAG); and *parameters* (the

I. E. Achumba (✉) · D. Azzi · S. Bersch
Faculty of Technology, School of Engineering, University of Portsmouth,
Anglesea Building, Anglesea Road, Portsmouth, PO1 3DJ, UK
e-mail: Ifeyinwa.Achumba@port.ac.uk

D. Azzi
e-mail: Djamel.Azzi@port.ac.uk

S. Bersch
e-mail: Sebastian.Bersch@port.ac.uk

I. Ezebili
Department of Electrical and Electronic Engineering, School of Engineering
and Engineering Technology, Federal University of Technology, Owerri,
Owerri PMB 1526, Imo State, Nigeria
e-mail: ifezebili@yahoo.com

quantitative part of a BN) which are value entries in the Conditional Probability Tables (CPTs) and the Prior Probability Tables (PPTs) associated with the nodes in the BN. Each node represents a random variable, and each directed edge represents a relationship between the two nodes it links; thereby creating a parent \rightarrow child relationship between the nodes. Often, semantics are used to give the edges and their directions particular meanings reflecting the relationship modeled by the edges. The commonest semantic is causality, which implies an asymmetric cause-effect relationship. Causality is not however, a requirement for Bayesian networks [1]. The edges in a BN can represent other relationships, such as containment, ownership, part, requirement, or any relationship that has meaning within the context of the domain being modeled. For example, Collins et al. [2] defined the edges in their BN model as representing the relationship of “skill \rightarrow sub-skill” where the parent node represents a skill and the child node represents a sub-skill (component) of the parent skill. A child node can have from zero to a finite number of parents. Every child node is characterized by a CPT that models the combined impact of its parents, while each leaf node (node without any parents) is characterized by a PPT describing the prior knowledge about the variable. There is an exponential relationship between the size of the CPT of a child node, the number of its parent nodes, and the number of possible states of its parents. For example, if a two-valued node, A , has k number of parents and each parent has two possible states, then the CPT for node A will contain, 2^k entries such that, $k = 2$, $k = 8$, and $k = 16$ will result in 4, 256, and 65536 entries respectively. If each of the parents has three possible values, 3^k entries are required such that for $k = 2$, $k = 8$, and $k = 16$ the number of entries will be 9, 6561, and 43046721, respectively. Generally, the number of entries in the CPT of a child node, A , that has k number of parents, where each parent has n possible states, is n^k .

Each discrete variable in a BN can take a finite set of values (states) and can assume only one of its states at any one time. Each state is associated with a probability value, such that the sum of the probability values of all the states adds up to 1. BN is subject to the standard axioms of probability theory. The basic axioms of probability theory, Bayes theorem, and the concept of Joint Probability Distribution (JPD) underpin a BN. The work presented in this chapter addresses the different possible approaches to Bayesian network model construction and is based on the paper by Achumba et al. [3].

2 Model Construction Approaches

BN model construction involves three ordered tasks: identification of the variables and their possible values (states), identification of the relationships between the variables, and parameterization. These tasks can be accomplished through three different approaches. One, all three tasks could be accomplished “manually”, which entails committed involvement of domain experts. This so called manual approach is referred to, in this context, as total expert (*totalExpert*) approach. Two, all three tasks could be accomplished by “learning”, which involves the acquisition of relevant domain

data and appropriate BN software tools. This approach is referred to, in this context, as total data (*totalData*) approach. Three, all three tasks could be accomplished by a combination of “manual” and “learning” approaches, which requires the involvement of domain experts and use of domain data. This is referred to, in this context, as semi data (*semidata*) approach.

BN model constructors, ab initio, relied on domain experts to define both the structure and parameters of a model. There are no formal foundations for *totalExpert* BN model construction. The key player in this approach is the domain expert(s) from whom most of the knowledge required for the construction of the model is elicited. Domain expert refers to a person having or possessing high-level knowledge or skills in a particular subject area. Figure 1 highlights the different stages and key components of the *totalExpert* construction approach.

The issues often highlighted about this approach are that:

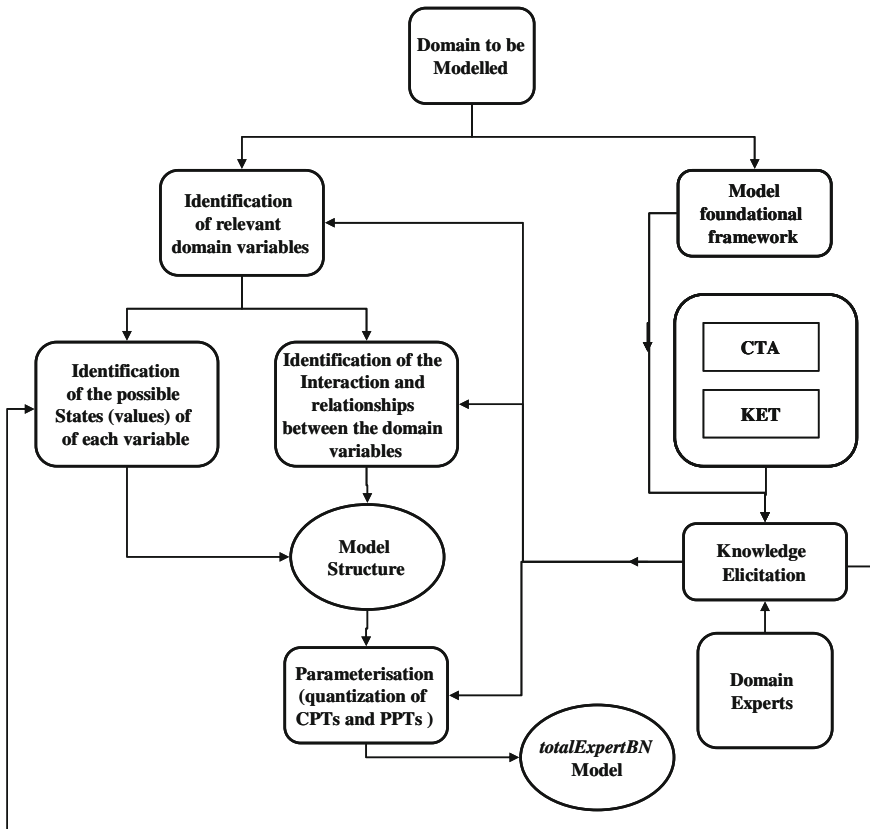


Fig. 1 Total expert (*totalExpert*) approach. KEY CTA—Cognitive Task Analysis; KET—Knowledge Elicitation Tools

- Expert knowledge is subject to bias. This issue is addressed through the involvement of more than one domain expert, and the knowledge elicitation process often goes through several stages of review. Also, the elicited model is usually subjected to sensitivity analysis which affords opportunities for the identification and minimization of bias, though existing knowledge elicitation were designed to minimize bias. Moreover, BN models are not overly sensitive to inaccuracies in their parameters.
- Knowledge elicitation can be a relatively time consuming and difficult process. Processes, methods, and tools for easing the elicitation process exist.
- Experts rarely agree. Experts' opinion disagreement is generally acknowledged. Methods for resolving expert opinion conflicts and obtainment of composite/consensus opinion are addressed in literature.

The total expert approach offers a number of benefits including that:

- Model variables, their states, and relationships are fully appreciated and the reasoning/rationale behind the model can be clearly articulated.
- Model creation is often based on the consensus or average of information and opinions of more than one domain expert thereby enabling the capture of uncommon or rare scenarios and knowledge.
- The technicalities of the domain represented by the model can be verified/discussed in details at each stage of the construction cycle.
- The process codifies knowledge so that it is available in the future for other projects and systems thereby promoting reliability in assessment of a family of systems that change within a changing usage environment.

The *totalData* approach entails the learning of both the model structure and parameters from existing domain data. This approach, represented graphically in Fig. 2, is based on the representation of a BN model as a variable, $B = \{G, \theta\}$, where G is the network structure with nodes corresponding to a set of random variables, $X = (X_1, \dots, X_m)$, while θ represents the set of parameters for the network. B is seen as encoding the Joint Probability Distribution (JPD), $p(X_1, \dots, X_m) = \prod_{i=1}^m p(x_i | pa(x_i))$, where $pa(x_i)$ represents the parent set of node x_i . The *totalData* approach entails learning both the structure, G , and parameters, θ , from domain data.

A dataset, D , is a table consisting of records of observations for a set of variables, such that, $D = [d_1, d_2, \dots, d_N]$, where $N =$ total number of records in D and $d_l = \{x_1[l], x_2[l], \dots, x_m[l]\} \in D$, $l = 1$ to N , represents a record of observation for all the variables, X (record of observation—a single set of findings or instantiation of all the variables, X). The variables in D form the nodes of the learnt model. A dataset can be complete or incomplete because incompleteness (missingness) is sometimes a feature of datasets. A complete dataset contains No Missing Values (NMV) (no missing instantiation or finding) for any of the variables, implying full observability. An incomplete dataset contains missing values for some variables in some or in all the cases in the dataset, implying partial observability or presence of latent (hidden) variables, respectively.

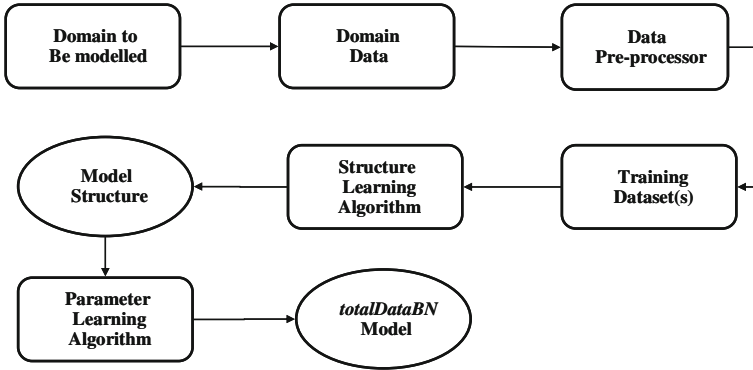


Fig. 2 Total data (*totalData*) approach

In order to *learn the model structure*, given the dataset, D , the structure learning algorithm works to find the most probable model structure, G_i , from among the set of all possible model structures (the search space) with respect to the domain variables represented in D . G_i is taken to be the model structure that most likely generated the dataset, D , which best describes the conditional independences suggested in D [4]. Structure learning algorithms are either based on Conditional Independence (CI) tests or Search and Score (SaS) technique. The CI approach uses *constraint-based* algorithms to find the structure whose implied independence constraints “match” those found in the data by performing CI tests on tuples of variables using statistical tests or information theoretic measures [5]. The SaS approach consists of three components: the search space, the scoring function, and the search engine. The search space consists of the set of all possible BN structures, G , given the domain variables. The score metric takes the dataset and a possible structure and returns a score reflecting the goodness-of-fit of the data to the structure [6]. Essentially, the dataset D , the scoring function, and the search space constitute the inputs to the search algorithm while the output is a network that maximizes the score, $P(D|G_i)$, and the probability of the most probable structure, G_i , given the dataset, D [7]. The issues often associated with structure learning include that:

- Structure learning is NP-hard. Research has aimed reduce the complexity of structure learning, without exact and exhaustive solution. Consequently, heuristic algorithms are often employed, which may yield an acceptable solution, but is not certain to arrive at an optimal solution.
- The search space grows super-exponentially with the number of variables, n , in the dataset, D . For n variables, the cardinality of the search space is given by the recursive function [8]: $f(n) = \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} f(n-k)$, where $f(1) = 1$.

In *parameter learning*, the structure, G , is known (already learnt or manually constructed) and the problem is the estimation of the parameters, $\theta = \{\theta_i\}_{i=1,\dots,m}$, from D given G , where θ_i is the set of parameters for node X_i . θ is the complete set of parameters that can best explain the set of observations in D [9]. The process could involve learning single or multiple parameters. *Single Parameter Learning* implies that the variable, X_i , has only two possible mutually exclusive states, x_i and \bar{x}_i , and $P(X_i)$ is defined by: $p(X_i = x_i) = \theta_i$ and $p(X_i = \bar{x}_i) = 1 - \theta_i$. Let r be the number of states of the variable, X_i . *Multinomial Parameter Learning* implies that X_i is a multinomial variable with $r > 2$ states, x_{i1}, \dots, x_{ir} , such that X_i has the set of parameters, $\theta_i = (\theta_{i1}, \dots, \theta_{ir})$, respectively, where $\sum_{k=1}^r \theta_{ik} = 1$.

The *semiData* approach, which scenario is depicted graphically in Fig. 3, entails the use of domain data and the involvement of domain expert(s).

Figure 3 highlights two different possible paths, (a) and (b), of the approach. Path (a) indicates the structure is created manually with assistance of domain experts, while the parameters are learnt from data. Path (b) indicates that structure is learnt from existing domain data, while the parameters are elicited from domain experts. The structure and parameter learning processes described earlier apply.

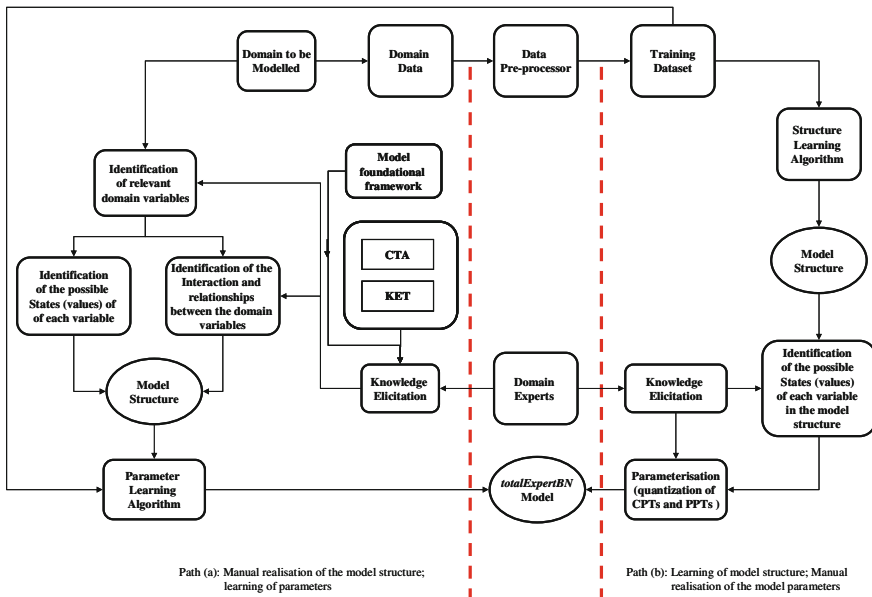


Fig. 3 Semi data (*semiData*) approach

3 Experience with Selecting the Optimal Construction Approach

The BN model construction approaches were empirically investigated in order to determine the optimal approach for the construction of a BN model, with respect to the application, undergraduate electronic engineering students' laboratory work performance assessment. Optimal, in this context, refers to the model which is best in terms of the performance index based on a set of adopted performance metrics. Several BN models were constructed using the three different approaches: *totalExpert*, *totalData*, and *semiData*.

The *totalExpert* approach was used to first construct the *totalExpertBN* model with the committed participation of three domain experts. To define the structure of the model, the following issues, highlighted by Fenton et al. [10], were taken into consideration:

- The more the number of states a node has, if the number of its parents becomes greater than two, it becomes more difficult to elicit parameters exhaustively. If the number of states increases up to seven and the number of parents more than two, exhaustive elicitation becomes infeasible.
- An BN need not model the same type of relationship throughout the network. Different fragments can model different relationships.

Also, a base framework for the model was first derived as shown in Fig. 4. The framework is based on Psychology of Learning, which is focused on understanding how the learning process (often depicted as learning theories and models) works and the effect of learning on behaviour. Learning theories and models are ideas about how learning may happen (conceptualization of the learning process), and are meant to be applied in the instructional process in order to facilitate learning by instruction and assessment, which drives learning. The framework is consistent with the definition of assessment, by [11] and [12], as a generic term for a set of processes that measure the outcomes of learning, in terms of knowledge acquired, understanding developed and abilities/skills gained. In order to build on the framework, Cognitive Task Analysis (CTA) technique was used to breakdown each of the core variables of the framework (knowledge, understanding, abilities/skills) into their constituent constructs. That is, the CTA process was used to facilitate knowledge elicitation. The *totalExpertBN* is served as the reference model.

Sample domain datasets and BN software tools were used for the construction of total data (*totalData*) models. There are two possible sources of sample data: domain historical data and/or empirically generated data. Where data from such sources are not available, which was the case in this context, the alternative is the use of simulated sample data. Often, researchers needing to undertake empirical investigations, with respect to structure and/or parameter learning, create frameworks that would facilitate the generation of the required sample data from the Joint Probability Distribution (JPD) represented by an existing reference model. Along this line, the bare structure (structure minus parameters) of the *totalExpertBN* was used to generate two sample

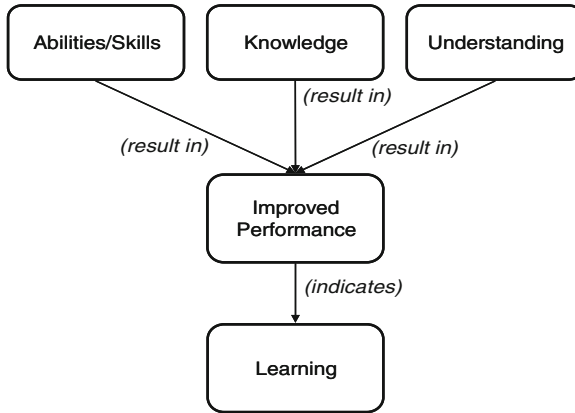


Fig. 4 Base framework for the *totalExpert* model

datasets, one training dataset (TDSMP dataset), and one test dataset, *TD*. A second training dataset (TDSPP dataset) was generated with the complete *totalExpertBN* model (structure plus parameters). This facilitated the construction of two different *totalDataBN* models: *totalDataBNSMP* and *totalDataBNSPP*, using the BN software tool, Genie. Genie supports both structure and parameter learning. Construction of *semiData* models was based on both paths (a) and (b) of Fig. 3. Two instances of the *totalExpertBN* model structure was used to construct two *semiData* models based on path (a) by learning their parameters using the training datasets, TDSPP and TDSMP.

3.1 Evaluation and Performance Comparison

The constructed BN models were evaluated using test dataset. The test procedure consisted of entering findings at selected evidence nodes of a model, and querying one or more target nodes. The process divides the nodes of the model into two sets: evidence and target nodes. Any node can belong to any one of the two sets, for the purposes of the test. It is often preferable to choose as target, the node that in the real context would be target of inference. The values in each record of the test dataset are split into two sets: values for the chosen evidence nodes, and values for the chosen target nodes. The values for the evidence nodes are entered as findings into the network, the network is updated and inference made at the target nodes. This process is repeated for each record in the test dataset. For each network update, the probability distribution of the target node is recorded and its prediction determined. That is, after each network update, the state with higher belief value (the most likely or maximum likelihood state), based on a cut-off threshold probability, is taken to be the prediction for the target node. For example, for a 50% cut-off threshold probability, of the two states of the target node, the state which belief level is higher than 50% is taken to

be the prediction. The predictions are then compared with the observations (the set of values for the target node in the test dataset taken as the actual “observations”), for each record of the test dataset. If a prediction corresponds to the observation, it is recorded as a success otherwise it is recorded as a failure. The statistics are then collected and used to assess the performance of the model, generating values for the various performance metrics and index that constitute the optimality criteria.

The evaluation process yields a set of performance scores for each model. The scores are then used to derive a Performance Index (PI) for the model. The score metrics include, *Briers score*, *error rate*, and *logarithmic (logloss) score* [5, 13–15]. *Sensitivity* is also often used as a model performance measure [5, 15]. These metrics were used together in this investigation in order to facilitate the drawing of more robust conclusions. The different metrics, though not complementary, evaluate performance from different perspectives, thereby collectively giving a more robust picture of the performance of a model. Error rate informs on the percentage failure rate of a model, the Brier score gives a measure of the accuracy of the probability estimates made by the model, and sensitivity informs on the percentage success rate of the model. The logloss score is similar to Brier score, however, the logloss score is *local* in that it only depends upon the probability assigned to the particular state and not on any of the probabilities assigned to the other states. Sensitivity (also referred to as the recall rate) is a statistical measure of model performance. It measures the proportion (in percentage) of actual values (observations) which are correctly predicted. A sensitivity of 100% means that the model correctly predicted all actual observations for the target variables (100% actual or true positives).

The error rate, based on the maximum likelihood state of the target node [16], is a way to analyze model predictions by dividing the number of predictive errors by the number of test cases in the test dataset. It gives the percentage failure rate. It identifies the percentage of the cases in a test dataset for which the network predicted a wrong value for the query node. For example, an error rate of 24% implies that in 24% of the cases for which the test dataset contains a value for the target node, the predictions did not match the observed values. The logarithmic (logloss) score was suggested by Good [17] and is defined as follows: let X denote a discrete random variable, with m (mutually exclusive) possible states, $(x_1, x_2, \dots, x_i \dots, x_m)$, which is to be observed for a sequence of cases, $i = 1, \dots, N$. Let $p(x_i)$ denote the estimated probability (referred to as the predicted value for the purposes of the test) for the i th state. Suppose the j th state is actually observed, then the particular observation is associated with a logloss score for the j th state given by Cowell [4] (1) as: $\ell_j = \log \frac{1}{p(x_j)} = -\log p(x_j)$. Then, by accumulating the scores for the N cases,

a total penalty for the N observations is obtained by: $\ell = \sum_{j=1}^N \ell_j$, and the average

logloss score for the N cases is: $\ell_{avg} = \frac{1}{N} \sum_{j=1}^N \ell_j = \frac{1}{N} \sum_{j=1}^N -\log p(x_j)$. The logloss value lies in the range $[0, \infty]$, where smaller (lower) values of the score imply better model performance. The Brier Score (b), also referred to as Quadratic Loss (QL) or Mean Squared Error of Prediction (MSEP), measures the accuracy of a set of

probability assessments. The Brier score function as used in BN model performance comparison is given as [15]:

$$b = \frac{1}{N} \left[\sum_{i=1}^N \left((1 - 2 \times p(y = c|x_i)) + \sum_{j=1}^k p(y = j|x_i)^2 \right) \right] \text{ where } p(y = c|x_i)$$

is the probability predicted for the actual (observed) state, c , of the target variable, y (the state of y in the particular record of the test dataset), given the evidence variables, x_i ; $p(y = j|x_i)$ is the probability predicted for the j th state of y , given the evidence variables; k is the number of states of the target variable, y ; N is the number of records in the test dataset. The QL is a measure of the average quadratic loss that occurred on each instance in the test dataset. It is averaged over all the records in the test dataset and not only accounts for the probability assigned to the actual (observed) state, but also the probabilities assigned to the other possible states of y . The value of Brier score lies in the range $[0, 1]$, with $b = 0$ indicating higher prediction accuracy, thus better performance.

The Performance Index (PI) of a model is derived based on the function, $\psi = [(100 - e) + (1 - b) * 100 + s + (1 - l) * 100]_{normalized}$. The function, ψ , takes the values of the performance metrics, error rate (e), brier score (b), logloss (l), and sensitivity (s) of a model, as inputs, and yields an index of performance for the model, for a test instance. The function assumes equal importance for all the metrics. The index was used as the main model performance measure, in this context.

3.2 Structure Comparison Measures

Measures of structural differences are often used to compare the structural differences between models, which may not take into consideration edge orientations. Orientation is important where the edges model causal relationships, else, edge orientation can be deemphasized [18]. A causal model is a ‘‘Bayesian network with added property that the parents of each node are its direct causes’’, and this implies an asymmetric relationship between parent and child nodes, such that in the case of edge reversal the resulting network will not be equivalent in terms of representational ability [19]. In learnt non-causal model structures, it is possible to ignore the direction of the reversible edges but not those of the compelled edges. The reversible edges are the edges that occur in the opposite direction in some other DAG that is equivalent (in terms of representational ability) to the current DAG [20]. ‘‘If two DAGs encode the same conditional independencies, they are called Markov equivalent. The set of all DAGs can be partitioned into Markov equivalence classes. Graphs within the same class can have the direction of some of their arcs reversed without changing any of the CI relationships. Each class can be represented by a PDAG (partially directed acyclic graph) called an essential graph or pattern. This specifies which edges must be oriented in a certain direction (compelled edges), and which are reversible. When learning graph structure from observational data, the best one can hope to do is to identify the model up to Markov equivalence’’.

In this context, ignoring the reversible edges, the link statistics of an induced model structure are categorized as:

- *correct positive (cp)*—a link is learnt between two nodes where a link exists between the same two nodes in the reference model (correct link)
- *false positive (fp)*—a link is learnt between two nodes where a link does not exist between the same two nodes in the reference model (extra link)
- *correct negative (cn)*—no link is learnt between two nodes where a link does not exist between the same two nodes in the reference model (correct nolink)
- *false negative (fn)*—no link was learnt between two nodes where a link exists between the same two nodes in the reference model (missing link).

3.3 Outcome

A total of five different models were actually constructed, where the intention had been to construct seven. The reason for this short fall will later become evident. The constructed models are: the *totalExpertBN*, two *totalDataBN*, and three *semi-DataBN*. The sizes of the training and tests datasets were 137024 and 7000 samples, respectively. Two different training datasets (TDSMP and TDSPP—both the same size) and one test dataset, *TD*, were used. The TDSPP dataset was generated with the complete *totalExpertBN* model (structure plus parameters), while the TDSMP dataset was generated with only the model structure (structure minus parameters) of the *totalExpertBN*, with the assumption of observability. It was also assumed that the training data samples are representative of the larger set of baselines samples. The test dataset was used for evaluating the models. The results of the empirical investigation are hereby presented with respect to the models' performance metric values and performance indices, shown in Table 1 and Fig. 5, respectively.

The results of the investigation highlighted some interesting themes. First, the performances of the semi data-centred models, *semiDataBNSPPa* and *semiDataBNSPPb*, constructed with the TDSPP training dataset were comparable to the performance of the reference model, *totalExpertBN*. So also was the performance of the

Table 1 Performance evaluation results

Approach	Model	Error rate	Logloss	Brier score	Sensitivity
Total expert centered	<i>ExpertBN</i>	29.54	0.5454	0.374	74.65
Total data centered	<i>totalDataBNSPP</i>	29.46	0.5456	0.3741	66.69
	<i>totalDataBNSMP</i>	*	*	*	*
Semi data centered	<i>semiDataBNSMPa</i>	55.06	0.693	0.4998	46.09
	<i>semiDataBNSMPb</i>	*	*	*	*
	<i>semiDataBNSMPa</i>	29.33	0.5444	0.3735	71.03
	<i>semiDataBNSMPb</i>	29.54	0.5454	0.374	74.65

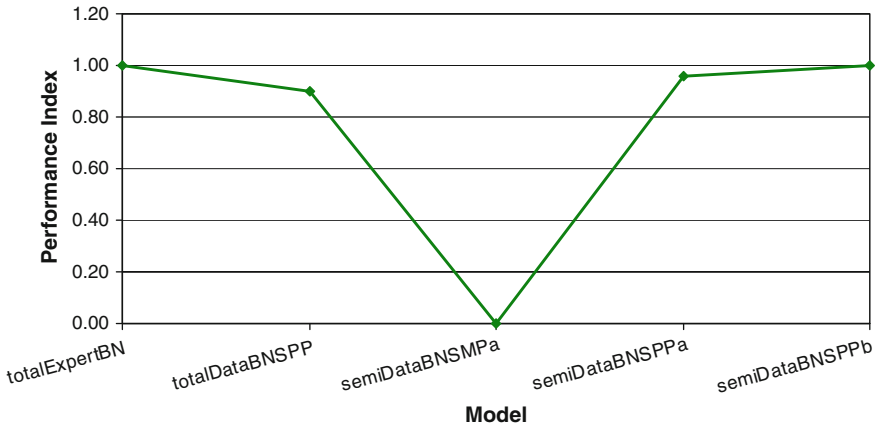


Fig. 5 Performance indices of the models

total data-centred model, *totalDataBNSPPb*, constructed with the *TDSPP* training dataset. Second, the performance of the semi data-centred model, *semiDataBNSMPa*, constructed with the *TDSMP* training dataset was relatively poor compared to the performance of the reference model, taking 0.50 index as the threshold between comparable and poor performance. Its learnt CPT entries were more or less inconclusive. Third, in all cases where the training dataset, *TDSMP*, was to be used for model construction involving structure learning (as in the cases of constructing the *totalDataBNSMP* and *semiDataBNSMPb*), no structures were learnt. That is, the structure learning algorithm failed to discover any relationship between the variables in the training dataset. This is why there are no performance metric value entries for the models in Table 1. Hence, only five models were constructed, instead of the intended seven. Finally, the *totalExpertBN* and the *semiDataBNSPPb* had the highest Performance Index (PI) of 1, followed by the *semiDataBNSPPa* with a PI of 0.96, and then *totalDataBNSPP*, with a PI of 0.90. The performances of the three data related models, *semiDataBNSPPb*, *semiDataBNSPPa*, and *totalDataBNSPP* are significant. However, the results indicate that it may not have been possible to construct the three models without first constructing the complete reference model (structure + parameters) (the *totalExpertBN* model) with the assistance of domain experts. It was observed, as shown by the results, that a complete reference model (that is knowledge of the relationship between the domain variables and their Conditional Probability Distributions (CPDs)) is a requirement for simulating sample datasets for structure and/or parameter that will yield meaningful and comparable models.

4 Conclusion and Future Work

The optimal approach for the construction of a BN model for the performance assessment of undergraduate electronic engineering students' laboratory work in a virtual laboratory environment has been investigated. The exercise has yielded significant results and highlighted an aspect requiring further investigation. The results provide reassurance that the procedure followed in the derivation of the assessment model was fit for purpose. In addition, the results provide additional insight for researchers and show that the data- and semi-centred BN model construction approaches depend on the availability of appropriate sample training datasets. The source of the training dataset could impact on the outcome of the model construction. This highlights the need for further investigation of the effect of historical domain training data samples not generated with knowledge of the relationship between the domain variables and their CPDs on structure and parameter learning.

Furthermore, from our experience, the data-centred BN model construction approach depends on the availability of appropriate software tools and sample training datasets. Model construction may be limited by the software tools available. Commercial software tools may be inaccessible, in which case freeware tools, which may have limited capabilities, are used. Also, there seems to be no standardized data and network file formats for BN software tools. Different tools support different data and network file formats. For example, some software tools may support only numeric data files, some string data files, while some may require the inclusion of record occurrence frequencies. This may be counterproductive for the data-centred BN construction approach.

Acknowledgments This work was supported in part by the Schlumberger Foundation under its Faculty For The Future (FFTF) scholarship programme aimed at encouraging women, in Science, Engineering and Technology (STE), in their pursuit for academic excellence.

References

1. Jenson FV (2001) Bayesian networks and decision graphs. Springer, New York
2. Collins V, Greer JE, Huang SX (1996) Adaptive assessment using granularity hierarchies and bayesian nets. *Lect Notes Comput Sci* 1086:569–577
3. Achumba IE, Azzi D, Ezebili I, Bersch SD (2012) On selecting the optimal Bayesian network model construction approach. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE 2012, 4–6 July 2012 U.K, London*, pp 690–695
4. Cowell RG (1999) Parameter learning from incomplete data for Bayesian networks, 1999. <http://www.staff.city.ac.uk/~rgc/webpages/aistats99.pdf>
5. Oteniya L () Bayesian belief networks for dementia diagnosis and other applications: a comparison of hand-crafting and construction using a novel data driven technique. Unpublished Ph.D. thesis, Department of Computing Science, University of Stirling, Stirling, FK9 4LA, Scotland
6. Chickering DM, Geiger D, Heckerman D (1995) Learning Bayesian networks: search methods and experimental results, 1995. <http://research.microsoft.com/en-us/um/people/dmax/publications/aistats95.pdf>

7. Forster M, Sober E (2010) AIC scores as evidence—a Bayesian interpretation, 2010. <http://philosophy.wisc.edu/sober/forster%20and%20sober%20AIC%20Scores%20as%20Evidence%20jan%2028%202010.pdf>
8. Robinson RW (1973) Counting labelled acyclic digraphs. In: Harary F (ed) *New directions in the theory of graphs*. Academic Press, New York, pp 239–273
9. Heinrich G (2010) Parameter estimation for text analysis. Technical note version 2.4, vsonix GmbH and University of Leipzig, August, 2008. <http://www.arbylon.net/publications/text-est.pdf>. Accessed 9 July 2010
10. Fenton N, Neil M, Caballero JG (2006) Using ranked nodes to model qualitative judgements in Bayesian networks. *IEEE Trans Knowl Data Eng* 19(10):1420–1432
11. QAA (Quality Assurance Agency), (2006). Code of practice Section 6: assessment of students. <http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/section6/default.asp>. Accessed 3 Feb 2008
12. Harvey L (2004) “Analytic quality glossary”, Quality Research International, 2004. <http://www.qualityresearchinternational.com/glossary/>
13. Morgan M, Henrion M (1990) *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, London
14. Pennock D (2006) Evaluating probabilistic predictions. <http://blog.oddhead.com/2006/12/26/evaluating-probabilistic-predictions/>. Accessed Dec 2006
15. Doshi P, Greenwald L, Clarke J (2002) Towards effective structure learning for large Bayesian networks. AAAI technical report WS-02-14, 2002. <http://www.aaai.org/Papers/Workshops/2002/WS-02-14/WS02-14-003.pdf>
16. NSC (Norsys Software Corp), Netica-J Reference Manual (Version 3.25), 2008. http://www.norsys.com/netica-j/docs/NeticaJ_Man.pdf
17. Good IJ (1952) Rational decisions. *J Royal Stat Soc* 14:107–114. In: Roulston M, *The logarithmic scoring rule a.k.a. “ignorance”*. <http://www.cawcr.gov.au/bmrc/wefor/staff/eee/verif/Ignorance.html>
18. de Jongh M, Druzdzal MJ (2009) A comparison of structural distance measures for causal Bayesian network models. *Recent advances in intelligent information systems, 2009*, pp 443–456. <http://iis.ipipan.waw.pl/2009/proceedings/iis09-43.pdf>. Accessed 07 Sept 2010
19. Pearl J, Russell S (2010) Bayesian networks, November 2000. <http://www.cs.berkeley.edu/~russell/papers/hbtt-n-bn.ps>. Accessed 12 Feb 2010
20. MR (Microsoft Research) (2011) WinMine Toolkit Tutorial. <http://research.microsoft.com/en-us/um/people/dmax/WinMine/Tutorial/Tutorial.html>. Accessed 02 Feb 2011

Fertilization Operator for Multi-Modal Dynamic Optimization

Khalid Jebari, Abdelaziz Bouroumi and Aziz Ettouhami

Abstract Solving Multi-modal Dynamic Optimization problems (MDO) has been a challenge for genetic algorithms (GAs). In this kind of optimization, an algorithm requires not only to find the multiple optimal solutions but also to locate a changing optimum dynamically. To enhance the performance of GAs in MDO, this paper proposes a New Genetic Operator NGO. The NGO is built on three components. First, a novel Genetic Algorithm with Dynamic Niche Sharing (GADNS) which permits to encourage the speciation. Second, an unsupervised fuzzy clustering that tracks multiple optima and enhances GADNS. Third, Spacial Separation (SS) which induces the stable sub-populations and allows local competition. In addition, NGO maintains diversity by a new genetic operators. To control the selection pressure, a new tournament selection is presented. Moving Peaks benchmark is applied to test the performance of NGO. The ability of the NGO to track multiple optima is demonstrated by a new diversity measure.

Keywords Dynamic niche sharing · Dynamic optimization · Evolutionary computation · Fuzzy clustering · Genetic algorithms · Unsupervised learning

K. Jebari (✉) · A. Ettouhami
Conception and Systems laboratory, Faculty of Sciences, University (UM5A),
Rabat, Morocco
e-mail: khalid.jebari@gmail.com

A. Ettouhami
e-mail: touhami@fsr.ac.ma

A. Bouroumi
Faculty of sciences Ben Msik,
Hassan II Mohammadia-Casablanca University, Casablanca, Morocco
e-mail: a.bouroumi@gmail.com

1 Introduction

Two challenges must be considered in many MDO problems. The first is their dynamic character and the second is their multi-modal aspect. This implies that the optimization algorithms must find and track the optimal solutions in an environment which continuously changes over time.

In literature many research based on Genetic Algorithms (GAs) are used to solve MDO. But, when solving MDO, standard GAs face a big problem which is the convergence. Once GAs converge, they are unable to adapt to the new environment when change occurs and they progressively lose diversity through generations.

There are several approaches based on GAs [3] to solve the problem of diversity such as: increasing mutation rate [6], using the immigrants approaches [5, 10, 14], enhancing GAs with memory [14, 17] and using subpopulations approaches [4].

In this paper, a new technique to maintain diversity and to track multiple optima is proposed. To achieve this goal, the NGO architecture is built on three components [11]. The first component is a GA with dynamic niche sharing (GADNS). The second one is an unsupervised fuzzy clustering algorithm. The third component implements the spatial separation (SS).

The first component is used to preserve genetic diversity, to encourage speciation and to form niches. However, GADNS suffers from some limitations. Therefore, a new genetic operator is proposed which is called fertilization. The principle of this operator is to add new individuals in the current population. These new individuals are the prototypes (V_i) of clusters given by the unsupervised fuzzy clustering algorithm. After, we apply mutation and restricted crossover for only those prototypes. This new subpopulation is added to the current population with elitism replacement. In addition, an Unsupervised Fuzzy Tournament Selection (UFTS) [12], is employed to dynamically adjust the selection pressure. The second component which is the unsupervised fuzzy clustering method is utilized in order to identify clusters that correspond to niches. Furthermore, the number of clusters (C) and the characteristics of each detected cluster (C_i) are automatically provided by the second component without requiring any prior knowledge on the distribution of population. The characteristics (C , the prototypes V_i , the cluster radius r_i) are exploited to overcome the limitations of dynamic niche sharing. The third component is based on Spatial Separation (SS) which involves the formation of real niches, and which promotes local competition in the same niches.

2 Clustering Algorithm

In the initialization phase of GAs, individuals are usually created at random manner. This random creation implies an absence of information on the distribution of individuals and a possible presence of overlapping clusters. Fuzzy clustering offers the advantage of efficiently dealing with overlapping clusters, and does not require

prior knowledge on individuals distribution. Based on an inter-points similarity measure, the clustering algorithm used consists of two main steps. The first step is an Unsupervised Fuzzy Learning (UFL) [2] procedure that explores the individuals of population $P(t)$ for discovering the clusters. In addition to the number of clusters (C), UFL provides the initial prototypes (V_i). The second step is an optimization procedure that applies Fuzzy C-Means (FCM) [1] algorithm in order to optimize the C learned prototypes.

UFL starts with generating the first cluster around the first encountered individual. The other individuals are then sequentially explored. A new cluster is automatically created whenever the current individual presents a small similarity, i.e., less than a certain prefixed threshold S_{min} , to the already existing prototypes.

The similarity measure controls how the clusters are formed. In our study, the similarity measure is based on Euclidean distance. To measure the similarity between any pair of p -dimensional vectors (individuals) I_i and I_j ($I_i, I_j \in \mathfrak{R}^p$), the following expression is used:

$$S(i, j) = \frac{d(I_i, I_j)}{\sqrt{p}} \quad (1)$$

where $d(I_i, I_j)$ is the Euclidean distance measure, calculated on the basis of the normalized values of I_i and I_j .

Generally, in the absence of any prior information of the good value for S_{min} , the algorithm is automatically iterated for different values of S_{min} , where $S_{min} \in [\xi_{min}, \xi_{max}]$:

$$\xi_{min} = \min_{i \neq k} (S(I_i, I_k)) \quad (2)$$

$$\xi_{max} = \max_{i \neq k} (S(I_i, I_k)) \quad (3)$$

However, this clustering algorithm is sensitive to the choice of the similarity threshold value S_{min} , in our clustering algorithm S_{min} is varied within the range ξ_{min}, ξ_{max} with a step equal to 0.01 (step = 0.01). Consequently, the different values of S_{min} may lead to different results. So, a criterion validity is used in the third phase. This validation (VAL) has used the normalized partition entropy [1] which is defined as follows:

$$h(U) = -\frac{1}{\log(C)} \frac{1}{N} \sum_{i=0}^N \sum_{j=0}^C [u_{ij} \log(u_{ij})] \quad (4)$$

where $U = [\mu_{i,j}]$ represents matrix of membership degrees; C : Number of clusters; N : Population Size.

The best solution U^* is the one that minimizes $h()$, in this case the number of clusters C^* given by UFL() is the optimum found and also the prototype V^* given

by FCM(). The proposed clustering algorithm UFL-FCM-VAL is described in the Algorithm 1.

```

Data: Individuals:  $(I_1, I_2, \dots, I_N)$ 
Result: Prototypes  $(V_1^*, V_2^*, \dots, V_C^*)$ ;  $C^*$ : Number of Clusters
Initialization: ;
 $S_{min} \leftarrow \xi_{min}; h_{min} \leftarrow 1; C^* \leftarrow 2;$ 
while  $S_{min} < \xi_{max}$  do
    Apply UFL();
    Apply FCM();
    if  $(h_{min} < h())$  then
         $h_{min} \leftarrow h();$ 
         $C^* \leftarrow C;$ 
         $V^* \leftarrow V;$ 
         $U^* \leftarrow U;$ 
    end
     $S_{min} \leftarrow S_{min} + step;$ 
end
    
```

Algorithm 1: Proposed Clustering Algorithm UFL-FCM-VAL()

The C prototypes V_1, V_2, \dots, V_C represent the tracking optima and also the individuals for fertilization operator.

3 GAs with Dynamic Niche Sharing

Dynamic niche sharing [15] defines a fixed number of dynamic niches with radii and centers determined by a population sorted in decreasing fitness order. For the individuals that are not in a niche, regular fixed sharing is used. The shared fitness value f_{ds} for an individual within a dynamic niche is as follows:

$$f_{ds,i} = \frac{f_i}{m_{ds,i}} \tag{5}$$

where f_i Fitness value, the dynamic niche count $m_{ds,i}$ is calculated by the following:

$$m_{ds,i} = \begin{cases} n_j & \text{if individual } i \text{ is within dynamic niche } j \\ m_i & \text{otherwise.} \end{cases} \tag{6}$$

where n_j is the cardinal of j th dynamic niche;

$$m_i = \sum_{j=1}^N sh(d(i, j)) \tag{7}$$

where N denotes the population size and $d(i, j)$ is a distance measure between the individuals i and j . The sharing function (sh) measures the similarity between two individuals

$$sh(d(i, j)) = \begin{cases} 1 - \left(\frac{d(i, j)}{\sigma_s}\right)^\alpha & \text{if } d(i, j) < \sigma_s \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

where σ_s denotes a threshold of dissimilarity and α is a constant which regulates the shape of the sharing function.

The primary problem of GADNS is the use of fixed sharing outside the dynamic niches [9]. Note also, the number q and the niche radii are often estimated.

Furthermore, making σ_s the same for all individuals means that the peaks are considered as nearly equidistant in the domain [16].

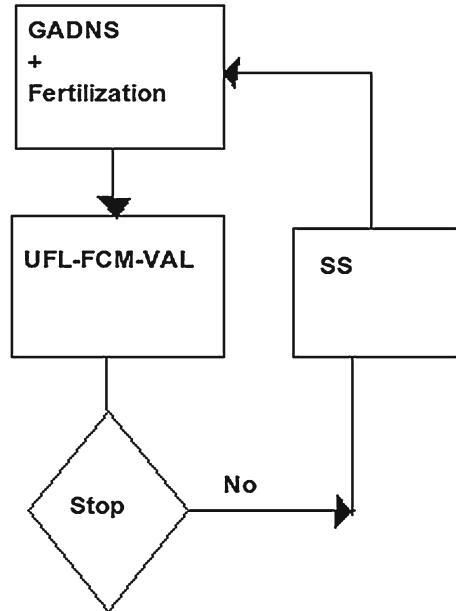
To overcome these limitations, our approach uses a fuzzy clustering technique in order to determine automatically the number of niches q (q = number of clusters given by UFL-FCM-VAL). Moreover, the radius of each niche is continuously updated.

4 Proposed Approach

The aim is to determine different optima of a dynamic multi-modal optimization using GADNS, the number of peaks q and characteristics corresponding niches (center, radius, cardinal, etc.). The idea is to apply UFL-FCM-VAL at population of solutions produced by GADNS, in order to detect the presence of homogeneous classes. If the entropy of the fuzzy c -partition obtained [2] is higher than (10^{-3}), these solutions will be again evolved by GADNS and classified by UFL-FCM-VAL. The principle of our iterative technique (Fig. 1) is based on a three-component. The first component (GADNS) is a GA, which combines dynamic niche sharing, and mating restriction to maintain diversity and to encourage speciation. The second component (UFL-FCM-VAL) is based on an unsupervised fuzzy clustering algorithm, which performs the partition of the individuals given by the first component into a set of C clusters so that each of them corresponds to a niche. The last component (SS) implements the principle of spatial separation to generate sub-populations from the resulting cluster characteristics (center, radius). Hence, individuals undergo a cyclic process throughout the three components of the system. The system is based on the following:

1. each cluster represents a niche;
2. prototypes represent the expected optima and individuals of fertilization operator;
3. the number of clusters (C) is computed by UFL-FCM-VAL;
4. the subpopulations and their appropriated subspaces are generated using the characteristics (center and radius) of each identified cluster;
5. in order to identify a non-detected niche in the previous cycle, GADNS is used again.

Fig. 1 NGO structure



4.1 GADNS Component

The real encoding scheme is used in the NGO. The initialization of the population is generated by a random process, the population size is $N = 100$. For the crossover operator, we have used a Simulated Binary Crossover SBX [7], with crossover rate = 0.8. For the mutation, we have considered the polynomial mutation [8], the mutation rate = 0.2. For the selection operator, we have used the Unsupervised Fuzzy Tournament Selection.

4.1.1 Unsupervised Fuzzy Tournament Selection

Unsupervised Fuzzy Tournament Selection, is based on standard tournament selection. But UFTS guards in each iteration the best individual. The tournament size (k) is the number of clusters (C) given by UFL-FCM-VAL. This tournament size controls better the selection pressure [12].

4.1.2 Fertilization Operator

After calculating and updating the prototypes by UFL-FCM-VAL, the fertilization operator allows to insert new individuals in the population. The first individuals in

this sub population are the prototypes $F(t)$, we evolved the individuals by respecting the following steps:

- apply crossover; Let I_f, I_m are two parents, the children are:
 - $C_1 = \alpha * I_f + (1 - \alpha) * I_m$
 - $C_2 = \alpha * I_m + (1 - \alpha) * I_f$; where $\alpha \in [0, 1]$
 - the result is the sub population $\hat{F}(t)$;
- apply Gaussian mutation (the result is the sub population $\ddot{F}(t)$);

For inserting new individuals given by the operator of fertilization, two strategies can be investigated. The first strategy is to insert a portion of the new individuals in the population (rate of fertilization). While the second, adopted in this study and called nearest replacement, is that each individual in $\{F(t) \cup \hat{F}(t) \cup \ddot{F}(t)\}$ replaces the nearest individual in the population $P(t)$ if it has a higher fitness.

```

Let  $I_N(t) \in \text{niche } k$  ;
Let  $I_S \in \text{niche } k$ ;
 $1 \leq k \leq C$ ;
foreach Individual  $I$  in niche  $k$  do
    | if  $d(I_N(t), I) \leq d(I_N(t), I_S)$  then
    | | //  $d(I_N(t), I_S)$  is the Euclidean distance measure;
    | |  $I_S \leftarrow I$ 
    | end
end
if  $f(I_N(t)) \leq f(I_S)$  then
    |  $I_N$  replaces  $(I_S)$ 
end
    
```

Algorithm 2: Nearest Replacement

4.2 UFL-FCM-VAL Component

Once UFL-FCM-VAL() is applied, a defuzzification procedure is performed in order to affect definitely each individual to its natural class for which it presents the maximum membership degree. This results in a final hard C-partition with C cluster centers $V_1; V_2; \dots; V_c$, which permit to track the expected optima.

4.3 SS Component

This component permit to affect each individuals to it final class, for which it presents the maximum membership degree. It can be done either from the matrix U^* , by assigning definitely each individual to the class for which he shows the higher degree of membership, or from the matrix V^* by applying the algorithm nearest prototype [2, 13].

Table 1 Setting for the MPB

Parameter	Setting
Basic function	NO
Correlation coefficient λ	0
Dimension	5
No. of evaluations between changes	5000
Number of peaks (p)	5
Peak heights (H)	[30, 70]
Peak shape	Cone
Peak widths (W)	[1, 12]
Shift severity (s)	1

5 Experimental Study

The experiments were based on the Moving Peaks Benchmark (MPB) proposed by [3]. The MPB problem has been widely used as dynamic benchmark problems in the literature. The optima in MPB problem can be varied following three aspects, location, height, and width of peaks.

The parameters concerning the MPB are presented in Table 1.

Our first experiment is to investigate the NGO mechanism of diversity, analyze the fertilization operator. In the experiments, SGA parameters are as follows: simulated binary crossover [7], with crossover rate = 0.8. For the mutation, we considered the polynomial mutation [8], with mutation rate = 0.2. We have used a Tournament Selection with size = 4. The population size N is equal to 100 For NGO and SGA on a MDO, 30 independent runs were executed with the same set of random seeds. The mean population diversity of a NGO on a MDO at generation over 30 independent runs were executed with the same set of random seeds and it is calculated according to the formula:

$$D\bar{iv}(t) = \frac{1}{30} \sum_{k=1}^{30} \left(\frac{1}{\ln(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n d_{i,j}(k,t) \right) \quad (9)$$

$D_{i,j}(k,t)$ Euclidean distance between the i th and j th individuals at generation t of the k th run. The diversity comparison for SGA, GADNS, NGO with the first strategy and NGO with the second strategy on MDO is shown in Fig. 2. In this experiment, we analyzed the two fertilization strategies. The first, which involves inserting a portion (fertilization rate) of new individuals, given by the operator of fertilization, in the population. For this strategy three fertilization rates are invested namely rate belongs to the set {20, 30, 40}. While the second, chosen by our study, is the nearest replacement. This strategy allows new individuals provided by the operator of fertilization to replace the nearest individual in the population $P(t)$ if it has a higher fitness. It can be seen that NGO does maintain the highest diversity level in the population while SGA maintains the lowest diversity level. This interesting

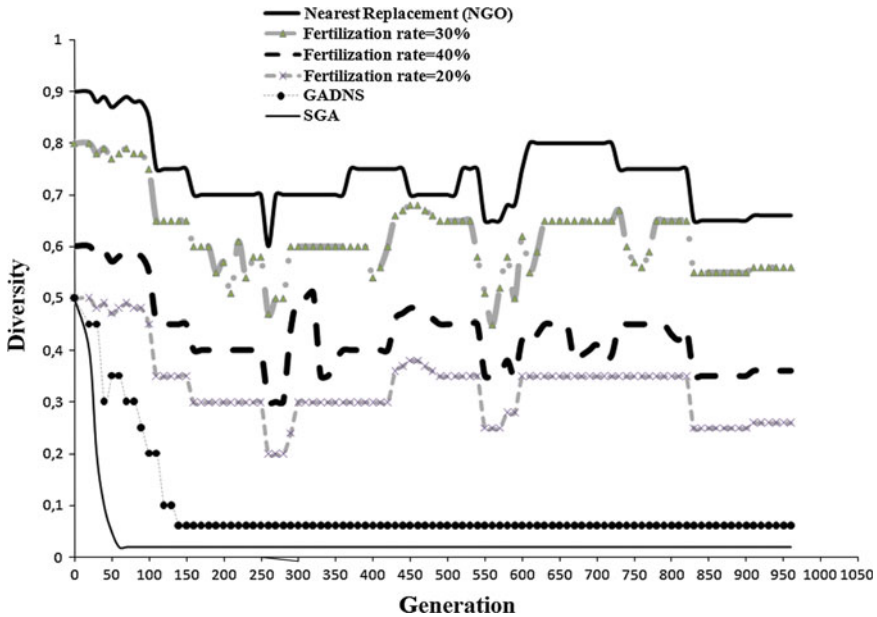


Fig. 2 Diversity comparison

result shows that approaches that aim at maintaining a high diversity level in the population in dynamic environments with fertilization operator.

To further demonstrate the diversity of the population maintained by our approach, we used the Shannon entropy normalized as diversity measure:

$$H^* = -\frac{1}{\log(C)} \sum_{i=1}^C p_i \log(p_i) \tag{10}$$

where p_i is often the proportion of individuals belonging to the i th cluster. When there is only one cluster in the population, Shannon entropy exactly equals zero. By cons, the abundances of the clusters in the population implies, the larger value of the Shannon entropy ($\simeq 1$).

Figure 3 provides results related to Shannon entropy measure. Figures 2 and 3 show that, by varying the rate of fertilization, the performance of the algorithm improves when the rate of fertilization increases. However, when this rate is greater than 0.3 (rate > 0.3), the algorithm performance starts to decrease. While, by using the nearest replacement NGO is very efficient. This result validates the benefit of introducing Fertilization into GADNS.

In the second experiment we used the off-line error as quality measure. The off-line performance is defined as the average of the best solutions at each time step:

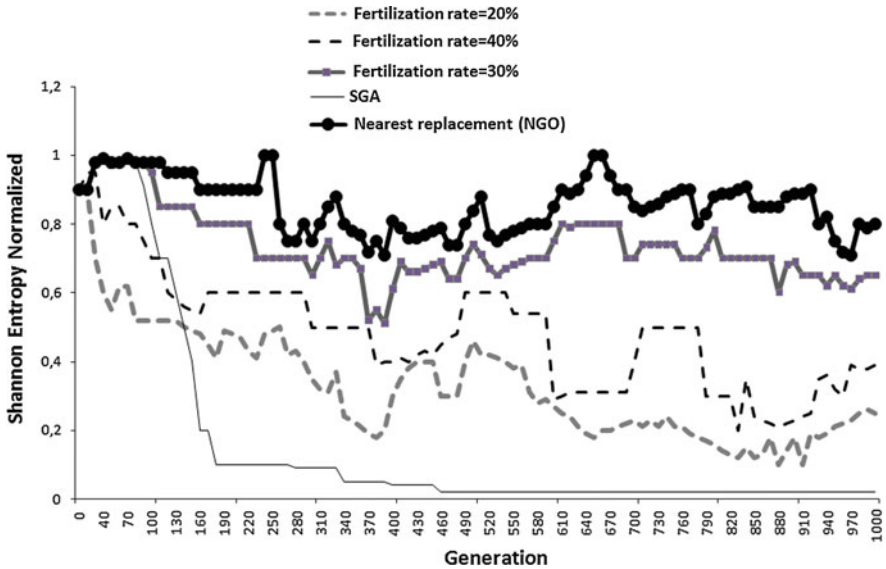


Fig. 3 Diversity comparison by Shannon entropy measure

$$AOE = \frac{1}{T} \sum_{t=1}^T (O_t - B_t) \tag{11}$$

where B_t is the best solution obtained by an algorithm just before the environmental change, O_t is the optimum value of the environment at time t , AOE is the average of all differences between O_t and B_t over the environmental changes. NGO is compared with SGA, Random Immigrants Genetic Algorithms (RIGA) [10] with the total number of immigrants $N_i = 30$, and also with Hyper-mutation Genetic Algorithm (HMGA) [6]. For this second experiment, SGA, HMGA, RIGA parameters are as follows: simulated binary crossover, with crossover rate = 0.85. For the mutation, we considered the Gaussian mutation, with mutation rate = 0.15. We have used a Tournament Selection with size = 4. The population size is $N = 100$. The followed results are based on the average over 50 independent runs with different random seeds.

1. Effect of Varying the Shift Severity:

This set of experiments compare the performance of NGO with SGA, HMGA and RIGA on the MPB problems with different settings of the shift length s . The experimental results regarding the off-line error is shown in Table 2, it can be seen that the results achieved by NGO are much better than the results of SGA, HMGA and RIGA algorithms on the MPB problems with different shift severities. As we know, the peaks are more difficult to track with the increasing of the shift length. Naturally, the performance of SGA, HMGA and RIGA degrade

Table 2 Off-line error for varying shift severity

s	NGO	SGA	HMGA	RIGA
0	0.75 ± 0.01	2.18 ± 0.06	1.78 ± 0.06	1.18 ± 0.07
1	1.05 ± 0.02	2.78 ± 0.06	2.09 ± 0.06	1.76 ± 0.06
2	1.39 ± 0.02	3.02 ± 0.07	2.93 ± 0.06	2.41 ± 0.06
3	1.57 ± 0.03	3.47 ± 0.07	3.57 ± 0.06	3.02 ± 0.06
4	1.73 ± 0.02	4.06 ± 0.08	3.89 ± 0.06	3.54 ± 0.06
5	1.88 ± 0.02	4.25 ± 0.1	4.07 ± 0.10	4.14 ± 0.1
6	2.03 ± 0.04	4.73 ± 0.1	4.39 ± 0.10	4.53 ± 0.1

when the shift length increases. However, the off-line error of NGO is only slightly affected in comparison with SGA, HMGA and RIGA. This result shows that NGO is very robust to locate and track multiple optima even in severely changing environments.

2. Effect of Varying the Number of Peaks:

This set of experiments investigate how NGO scales with the number of peaks in the MPB problem. The number of peaks was set to different values in the range form 1 to 30. Table 3 presents the experimental results in terms of the off-line error. From Table 3, it can be seen that the performance of NGO is not influenced too much when the number of peaks is increased. Generally speaking, a fitness value over all the peaks increases monotonically with an increasing number of peaks and there are more local optima and it is increasingly difficult to track so many peaks. So, increasing the number of peaks makes it harder for algorithms to track the optima. Comparing the results of NGO with SGA, HMGA and RIGA algorithms, the off-line error achieved by NGO is much less than that achieved by SGA, HMGA and RIGA. From Table 3, it can also be seen that NGO outperforms SGA, HMGA and RIGA algorithms. As we know, a large number of peaks needs more niches to locate and track. It can be seen from Table 3 that NGO achieves much better results for many peaks problems.

3. Effect of Varying the Correlation of Shifts:

The effects on NGO of changing the MPB correlation of shifts parameter λ are presented in Table 4. Results are compared with average values reported by SGA, HMGA and RIGA. Results provided by NGO are significantly better than those reported by SGA, HMGA and RIGA for all values of λ .

4. Effect of Varying the Dimensionality:

If dimensionality increases, the problem becomes more complicated and its resolution calls more diversity. Numerical results for different dimensionality values for MPB are presented in Table 5. For all dimension instance of MPB average of off-line errors reported by NGO is significantly better than those obtained by SGA, HMGA and RIGA. The results confirm the general robustness of our approach, and show that our guidelines for parameter adjusting are appropriate also for higher dimensional problems. In order to measure the ability of the NGO to track multiple optima a new measure called Dynamic Maximum Peak Ratio

Table 3 Off-line error for varying the number of peaks

Number of peaks	NGO	SGA	HMGA	RIGA
1	0.13 ± 0.02	3.68 ± 0.04	1.43 ± 0.07	2.68 ± 0.10
5	0.63 ± 0.02	3.79 ± 0.04	2.31 ± 0.11	1.17 ± 0.07
10	1.05 ± 0.02	4.02 ± 0.06	1.68 ± 0.04	1.81 ± 0.04
15	1.26 ± 0.02	4.18 ± 0.06	2.95 ± 0.08	1.88 ± 0.07
20	1.48 ± 0.02	4.76 ± 0.07	3.38 ± 0.11	3.27 ± 0.11
25	1.55 ± 0.02	4.88 ± 0.1	3.68 ± 0.11	3.65 ± 0.10
30	1.75 ± 0.02	5.01 ± 0.1	4.08 ± 0.08	3.86 ± 0.07

Table 4 Off-line error for varying the correlation of shifts (λ)

λ	NGO	SGA	HMGA	RIGA
0.1	0.63 ± 0.02	2.48 ± 0.07	1.43 ± 0.05	2.13 ± 0.07
0.3	0.85 ± 0.02	3.06 ± 0.07	2.08 ± 0.05	3.18 ± 0.07
0.5	1.23 ± 0.02	4.12 ± 0.08	3.44 ± 0.07	4.14 ± 0.07
0.7	1.43 ± 0.03	4.32 ± 0.08	4.09 ± 0.07	4.48 ± 0.07
0.9	1.23 ± 0.03	4.37 ± 0.1	4.18 ± 0.07	4.53 ± 0.11
1	1.65 ± 0.04	4.56 ± 0.1	4.68 ± 0.10	4.24 ± 0.10

Table 5 Off-line error for varying the dimensionality

Dimension	NGO	SGA	HMGA	RIGA
5	0.64 ± 0.05	3.18 ± 0.07	3.06 ± 0.07	1.52 ± 0.07
10	1.87 ± 0.05	4.23 ± 0.07	4.18 ± 0.07	2.18 ± 0.07
15	2.05 ± 0.05	5.16 ± 0.07	4.49 ± 0.07	4.56 ± 0.07
20	2.74 ± 0.05	6.57 ± 0.07	5.09 ± 0.07	5.13 ± 0.08
25	3.01 ± 0.08	10.34 ± 0.1	9.14 ± 0.11	10.18 ± 0.10
30	3.65 ± 0.1	12.17 ± 0.2	10.28 ± 0.12	12.18 ± 0.11

(DMPR) is applied in the last experiment. DMPR is based on the maximum peak ratio (MPR).

In multi-modal static problems, MPR is used to indicate both the quality and the number of the identified optima. The maximum value for MPR is equal to 1, in this case the algorithm has identified every optimum [15]. MPR represents the sum of the fitness of the local optima identified divided by the sum of the fitness of the actual optima in the search space:

$$MPR = \frac{\sum_{i=1}^C F_i}{\sum_{j=1}^n f_j} \quad (12)$$

Table 6 Dynamic maximum peak ratio (DMPR) for varying the number of peaks

Number of peaks	NGO	GADNS	HMGA	RIGA
2	1	0.88	0.54	0.67
5	1	0.78	0.44	0.57
10	1	0.69	0.41	0.47
15	0.97	0.59	0.31	0.27
20	0.93	0.43	0.25	0.17
25	0.84	0.35	0.31	0.24
30	0.75	0.31	0.23	0.16

where F_i is the fitness of identified optimum i and f_j is the fitness of the optimum j . C represents the number of the identified optima, and n is the number of real optima.

ρ is the best MPR since the last change in the environment

$$\rho_t = \max(MPR_\tau, MP R_{\tau+1}, \dots, MP R_t) \tag{13}$$

where τ is the last iteration at which a change to the environment occurred. Therefore, DMPR is the average of all evaluations over the environmental changes:

$$DMPR = \frac{1}{T} \sum_{t=1}^T \rho_t \tag{14}$$

Table 6 shows that the GADNS is able to identify the multiple optima when the number of peaks is less than 10. However, it fails to locate all possible solutions for the other cases remaining. While other algorithms have low performance. The NGO displays the best results in terms of DMPR. In fact, the comparison between the results obtained using GADNS, RIGA, HMGA and the proposed model confirms the usefulness and the ability of the proposed approach to track multiple optima in MDO.

6 Conclusion

To solve MDO, this paper proposed an unsupervised fuzzy clustering genetic algorithm for multi-modal dynamic optimization problems. The NGO employs an iterative technique based on tree components. First, the dynamic niche sharing is used to encourage speciation. Second, an unsupervised fuzzy clustering method is allowed to identify clusters that correspond to niches. In addition, for each detected cluster, the UFL-FCM-VAL provides the prototypes which represent the expected optima. Third, SS permits to induce stable sub-population and to encourage local competition. Besides, a new genetic operator mechanism called fertilization has been

introduced to maintain diversity. Moving peaks problems are used to test the performance of the proposed algorithm. The experimental results show that NGO presents good performance on these test problems. The clustering method used is effective in ameliorating dynamic niche sharing. Also, the fertilization operator introduced maintains much better population diversity. Generally speaking, NGO can effectively locate and track multiple optima in dynamic environments. The experimental results indicate that NGO can be a good optimizer in dynamic environments. It would also be interesting to combine other techniques into NGO to further improve its performance in dynamic environments. For example, swarm particle to search solution in local region, and NGO to track solutions. Another idea is to use NGO with adaptive or self adaptive population size, crossover rate or/and mutation rate.

References

1. Bezdec J (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
2. Bouroumi A, Essaïdi A (2000) Unsupervised fuzzy learning and cluster seeking. *Intell Data Anal* 4(3):241–253
3. Branke J (2002) Evolutionary optimization in dynamic environments. Kluwer Academic, Dordrecht
4. Branke J, Schmidt L, Schmeck H (2000) A multi-population approach to dynamic optimization problems. In: Parmee IC (ed) 4th international conference on adaptive computing in design and manufacture (ACDM 2000). Springer, Berlin, pp 299–308
5. Cedeno W, Vemuri VR (2007) A self-organizing random immigrants genetic algorithm for dynamic optimization problems. *Genet Program Evol Mach* 8(3):255–286
6. Cobb HG (1990) An investigation into the use of hypermutation as an adaptive operator in genetic algorithms having continuous, time-dependent non-stationary environments. TIK-report 6760, NLR memorandum, Naval Research Laboratory, Washington, DC, USA
7. Deb K, Agrawal RB (1995) Simulated binary crossover for continuous search space. *Complex Syst* 9(2):115–148
8. Deb K, Kumar A (1995) Real-coded genetic algorithms with simulated binary crossover: studies on multimodal and multiobjective problems. *Complex Syst* 9(6):431–454
9. Goldberg DE, Wang L (1997) Adaptive niching via coevolutionary sharing. TIK-report 97007, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, 117 Transportation Building, 104 S. Mathews Avenue Urbana, IL 61801
10. Grefenstette JJ (2007) A hybrid immigrants scheme for genetic algorithms in dynamic environments. *Int J Autom Comput* 4(3):243–254
11. Jebari K, Bouroumi A, Ettouhami A (2012) New genetic operator for dynamic optimization. In: Lecture notes in engineering and computer science: proceedings of The world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012, pp 742–747. <http://www.iaeng.org/publication/WCE2012/>
12. Jebari K, Bouroumi A et al (2011) Unsupervised fuzzy tournament selection. *Appl Math Sci* 28(1):2863–2881
13. Kuncheva LI, Bezdec JC (1998) Nearest prototype classification: clustering, genetic algorithms or random search. *IEEE Trans Syst Man Cybernet* 28(1):160–164
14. Louis SJ, Xu Z (2008) An immigrants scheme based on environmental information for genetic algorithms in changing environments. In: The 2008 IEEE congress on evolutionary computation, Hong Kong. IEEE, pp 1141–1147

15. Miller BL, Shaw MJ (1995) Genetic algorithms with dynamic niche sharing for multimodal function optimization. TIK-report 95010, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, 117 Transportation Building, 104 S. Mathews Avenue, Urbana, IL 61801
16. Saareni B, Krähenbühl L (1998) Fitness sharing and niching methods revisited. *IEEE Trans Evol Comput* 2(3):97–106
17. Yang S, Yao S (2007) Population-based incremental learning with associative memory for dynamic environments. *IEEE Trans Evol Comput* 12(5):542–561

A Hardware Design for Binary Image Recognition

Saul Martinez-Diaz

Abstract Recently, nonlinear composite correlation filters have been proposed for distortion-invariant pattern recognition. The filter design is based on logical operations and the correlation is computed with a nonlinear operation called morphological correlation. In this paper a new implementation in parallel hardware of these kinds of filters for image recognition is proposed. The architecture is designed for a Field Programmable Gate Array (FPGA) device. The proposed design performs the most time consuming task of the recognition procedure. In consequence, it reduces the time required for the nonlinear operations in the spatial domain. Simulation results are provided and discussed.

Keywords FPGA · Morphological correlation · Nonlinear filters · Parallel processing · Pattern recognition · Programmable devices

1 Introduction

Image recognition applications have been increased considerably the last years. In order to achieve distortion invariance and robustness in the recognition systems, a lot of efforts have been undertaken. Since VanderLugt introduced the Matched Spatial Filter (MSF) in 1964 [1], correlation methods have been used extensively the last decades for this purpose [2–10]. Firstly, correlation methods exploit all information from images in the recognition process. Besides correlation is shift-invariant and has solid mathematical foundation. In this case the basic recognition procedure is:

- Design a template (filter) with one or several training images.
- Correlate the filter with an input test image.

S. Martinez-Diaz

División de Estudios de Posgrado e Investigación, Instituto Tecnológico de La Paz., Boulevard Forjadores de Baja California Sur No.4720 Apdo. Postal 43-B, 23080 La Paz, B.C.S, México
e-mail: saulmd@itlp.edu.mx

- Establish a threshold at the correlation output.

A correlation value greater than threshold indicates that target is located at coordinates of the correlation peak.

Correlation filters can be designed by optimizing some performance criteria with linear techniques. For example, it can be shown that by minimizing the mean squared error (MSE) between an input image and a shifted version of the target, under certain assumptions, the linear correlation between the image and the target is maximized. As well, information from several distorted training objects could be incorporated by using composite filters. Some popular composite filters are synthetic discriminant functions (SDF) [2, 3] and minimum average of correlation energy (MACE) [4] filters. Moreover, these techniques can be used for multiclass pattern recognition. In addition, in order to reject other objects from scenes, an adaptive approach for the filter design has been proposed [5]. A drawback of linear filters is its sensitivity to most kind of noise in real images.

On the other hand, several approaches of nonlinear filter design have been proposed too [6–10]. Recently, nonlinear composite filters for distortion-invariant pattern recognition were introduced [9, 10]. The filters are designed as a logical combination of binary objects. Correlation is computed among the filter and a test scene with a nonlinear operation called Morphological Correlation (MC) [11]. MC is applied among images and the result is normalized to a desired value. It can be shown that, under certain assumptions, maximizing the morphological correlation minimizes de mean absolute error (MAE). Additionally, MC produces sharper peaks in the correlation output and it is more robust than linear correlation in scenes corrupted by non-Gaussian noise. The proposed filters have demonstrated a good discrimination capability and noise tolerance. Moreover, with the help of threshold decomposition [12], this technique can be applied to greyscale images as well. A disadvantage of this process is the high computational cost for large images. The main reason is that non fast algorithms have been developed for this nonlinear process. Nevertheless, the calculation of the nonlinear correlation can be parallelized by using specialized hardware [13].

In this paper we propose the implementation of the nonlinear filtering process in parallel hardware. The aim is to reduce the processing time. The architecture is intended for a Field Programmable Gate Array (FPGA). FPGA's have been used in several applications of parallel processing executing the most time consuming tasks [13–17]. Simulation results of proposed system are provided and discussed. The paper is organized as follows: Sect. 2 describes composite nonlinear filters. Section 3 presents an FPGA-based architecture. In Sect. 4 computer simulations are provided and discussed. Section 5 summarizes the conclusions.

2 Nonlinear Filters

2.1 Filtering Procedure

The proposed technique is a locally adaptive processing of the signal in a moving window. The moving window is a spatial neighborhood containing pixels surrounding geometrically the central window pixel. The neighborhood is referred to as the W -neighborhood. The shape of the W -neighborhood is similar to the region of support of the target. The size of the neighborhood is referred to as $|W|$, and it is approximately taken as the size of the target. In the case of non-stationary noise or cluttered background (space-varying data), it is assumed that the W -neighborhood is sufficiently small and the signal and noise can be considered stationary over the window area.

2.2 Threshold Decomposition

Suppose a gray-scale image $I(m, n)$ with Q levels of quantization, where (m, n) are the pixel coordinates. According to the threshold decomposition concept [12], the image I can be represented as a sum of binary slices

$$I(m, n) = \sum_{q=1}^{Q-1} I^q(m, n) \tag{1}$$

Here $\{I^q(m, n), q = 1, \dots, Q - 1\}$ are binary images obtained by decomposition of the greyscale image with a threshold q , as follows

$$I^q(m, n) = \begin{cases} 1, & \text{if } I(m, n) \geq q \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

2.3 Filter Design

Now, assume that there are M reference objects to be recognized (true class) and N objects to be rejected (false class). We construct a filter as a logical combination of the training images:

$$H(m, n) = \sum_{q=1}^{Q-1} \left[\left(\bigcap_{i=1}^N T_i^q(m, n) \right) \cap \left(\bigcup_{j=1}^M \bar{F}_j^q(m, n) \right) \right], \tag{3}$$

where $\{T_i^q(m, n), q = 1 \dots Q - 1, i = 1 \dots N\}$ are the binary slices obtained by threshold decomposition of the true class images. $\{\bar{F}_j^q(m, n), q = 1 \dots Q - 1, j = 1 \dots M\}$ are the logical complement of binary images obtained by threshold decomposition of false class training images. \cap represents the logical intersection: the result at coordinates (m, n) is 1 if the corresponding pixels of both planes are equal to 1; otherwise, the result is 0. \cup represents the logical union: the result at coordinates (m, n) is 0 if the corresponding pixels of both planes are equal to 0; otherwise, the result is 1. The neighborhood W is taken as the region of support of the composite filter.

2.4 Morphological Correlation

Let $\{H(m, n)\}$ and $\{S(k, l)\}$ be a template and a test scene respectively, both with Q levels of quantization. The local nonlinear correlation (morphological correlation) between a normalized input scene and a shifted version of the target at coordinates (k, l) can be defined as

$$c(k, l) = \sum_{m, n \in W} \min(a(k, l) S(m+k, n+l) + b(k, l), H(m, n)) \quad (4)$$

where $c(k, l)$ is the local nonlinear correlation at the coordinates (k, l) . $\min(x, y)$ is the minimal value among x and y . The sum is taken over the W -neighborhood. $a(k, l)$ and $b(k, l)$ are local normalizing coefficients, which take into account unknown illumination and bias of the target, respectively. The coefficients estimates are given by:

$$a(k, l) = \frac{\sum_{m, n \in W} T(m, n) \cdot S(m+k, n+l) - |W| \cdot \bar{T} \cdot \bar{S}(k, l)}{\sum_{m, n \in W} (S(m+k, n+l))^2 - |W| \cdot (\bar{S}(k, l))^2} \quad (5)$$

$$b(k, l) = \bar{T} - a(k, l) \cdot \bar{S}(k, l) \quad (6)$$

It can be shown that the nonlinear correlation in (4) can be computed with the binary slices obtained from threshold decomposition of the input scene and the filter as

$$c(k, l) = \sum_{q=1}^{Q-1} \sum_{m, n \in W} (S^q(m+k, n+l) \cap H^q(m, n)) \quad (7)$$

where $\{S^q(k, l)\}$ and $\{H^q(m, n)\}$ are binary slices of the gray-scale images obtained by threshold decomposition of the normalized input scene $\{a(k, l) \cdot S(k, l) + b(k, l)\}$ and the template $\{H(m, n)\}$, respectively.

3 Hardware Architecture

As can be seen, the nonlinear correlation process is computationally expensive. Suppose a test scene of $K \times L$ pixels and a template of $M \times N$ pixels; then computation of correlation requires $K \times L \times M \times N$ operations for each binary slice. On the other hand, these operations can be computed in a parallel way.

3.1 Field Programmable Gate Arrays

In the last decades several electronic configurable devices have been introduced. This kind of hardware permits the design of logic circuits for a specific application. Modern configurable devices contain millions of gates, which allow the design of very complex circuits. For example, Application Specific Integrated Circuits (ASIC) offer high performance, high density and flexibility for designing digital circuits. In addition, Field Programmable Gate Arrays (FPGAs) and Complex Programmable Logic Devices (CPLD) can be programmed several times by the end user. These devices have radically changed the way digital logic is designed [18, 19].

As have been mentioned above, morphological correlation involves a high computational cost for a sequential computer; however, a faster response can be achieved by using parallel processing. A suitable device for this purpose is a field programmable gate array (FPGA). FPGA's are programmable devices based on an array of basic logic cells, an interconnect matrix surrounding the basic cells and a set of input/output cells. All of them, cells and interconnections can be reprogrammed by using a Hardware Description Language (HDL). Figure 1 shows the basic structure of an FPGA: the device consists of an array of programmable basic cells (shown in white), an interconnect matrix surrounding the basic cells (shown in gray) and a set of input/output programmable cells (shown in black) [19]. Modern devices could contain millions of logic cells for a low cost.

All kind of cells and the interconnections can be reprogrammed. Because of its capacity of reconfiguration, these kinds of device are a versatile choice for experimental designs. In addition, its high capacity and low cost make them suitable for implementation of specialized hardware. These kinds of devices have been used before in several applications [13–16].

3.2 Proposed Architecture

Figure 2 is a block diagram of the proposed architecture. RAM1 memory stores the entire binary test scene, RAM2 stores the binary template and RAM3 stores the correlation output. The size of RAM1 and RAM3 are equal to the size of the test scene. Dimensions of the template image are M rows and N columns. The size of

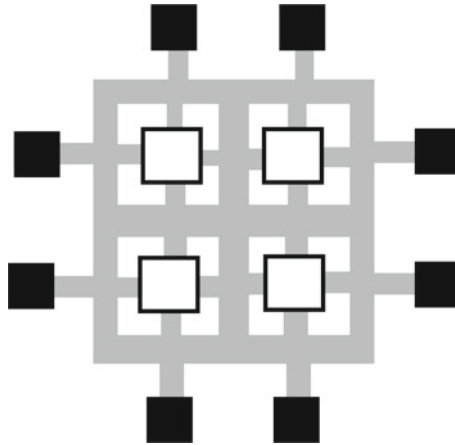


Fig. 1 Basic structure of a field programmable gate array (FPGA)

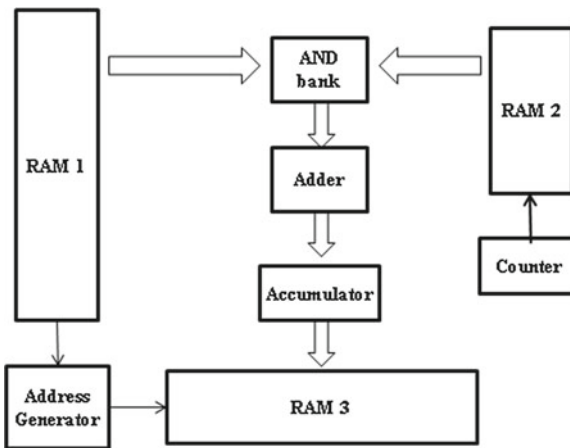
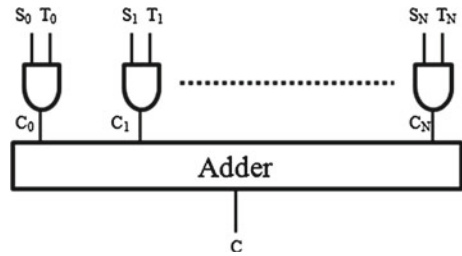


Fig. 2 Block diagram of the proposed architecture. All memory banks are clock synchronized

RAM2 equals the size of the template. All data buses are M bits wide. The AND gates bank contains M units that execute the operation of intersection among an entire row of both, the local test scene and the template images. Figure 3 shows detail of the AND bank. An adder unit calculates simultaneously the sum of correlation, which is entered to the accumulator unit. The final result is stored in the RAM3 memory. Besides, an address generator, which is clock synchronized, computes the memory allocation sequences to read data and store the results of the correlation process. Only a counter unit is needed to compute the address of RAM3 memory.

Fig. 3 Detail of AND bank



3.3 Operation

The basic operation of proposed hardware is as follows: the address of the first pixel of test scene is loaded and data is read from RAM1 memory. Since bus is M bits wide, an entire column is loaded in a single clock cycle from RAM1 and RAM2. In this way, the whole intersection among the local image and the template is processed in N clock cycles. The adder unit processes in parallel the M bits of each column. In consequence, not additional time is required for the sum. The partial result of each column sum is sequentially stored in the accumulator unit. Once processed a local image the address generator computes the address for storing the result in RAM3 memory and for reading the next local image from RAM1 memory. Besides, the counter unit is restarted to read the first column of template again. This procedure is repeated until all pixels of the test scene are processed.

Note that this procedure requires only $O(K \times L \times N)$ operations. In addition, not additional time is required for executing arithmetic operations and loading data. The proposed architecture can be easily modified for larger images.

4 Computer Simulations

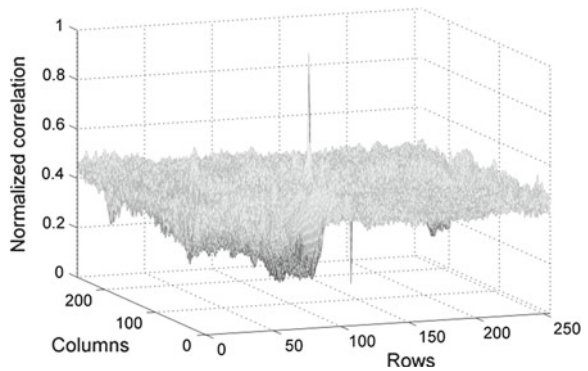
In this section computer simulations are provided. Figure 4 shows a test scene with two objects embedded. The template is 24×36 pixels and the test scene is 256×256 pixels. Both are greyscale images with 256 levels of quantization. The filter is designed including both butterflies. The target is the butterfly in the left side and the other butterfly is included to be rejected. First, threshold decomposition is executed by a computer. Next, each binary image is processed in the proposed architecture. Finally the correlation results are read from computer and normalized.

Figure 5 shows the final correlation output obtained. As can be seen, the correlation plane contains a sharp peak of magnitude equal to one at central coordinates of the target. On the other hand, at central coordinates of the false object the correlation value obtained is zero and no other sharp peak is distinguished. Note that the filter is able to detect the target embedded into the cluttered background and with a similar object in the same scene.

Fig. 4 Test scene including target (*left* butterfly) and an object to be rejected (*right* butterfly)



Fig. 5 Correlation plane obtained with proposed architecture for the test scene on Fig. 4



With the proposed architecture the number of clock cycles required to process the entire image is only 3 % of the cycles required with a sequential processor. In order to compare performance of proposed system with a sequential processor, several tests were executed in a personal computer. A 1.6GHz sequential processor was utilized. Results of processing time in the personal computer were averaged. A 200MHz FPGA was selected for the comparison. Because of the difference of speed among the processors, the interaction with the computer of FPGA and other factors, the time required to process an entire image with the proposed architecture was 50 % of that required with the sequential processor. However, for largest images this ratio can be improved at the cost of more memory and gates on the FPGA device.

5 Conclusion

In this paper was presented a new architecture designed to perform nonlinear correlation for pattern recognition. The kind of filters employed has demonstrated robustness to non-Gaussian noise and good discrimination capability. Besides, the proposed

architecture performs the most demanding time tasks of the correlation process. Moreover, the time consumption can be reduced by increasing the memory in such way that more correlations can be executed simultaneously. The proposed architecture is flexible and can be easily adapted to other sizes of images. Future work includes processing of the entire greyscale image into the FPGA device.

References

1. VanderLugt AB (1964) Signal detection by complex filtering. *IEEE Trans Inf Theory* 10:139–145
2. Hester CF, Casasent D (1980) Multivariant technique for multiclass pattern recognition. *Appl Opt* 19:1758–1761
3. Casasent DCW (1986) Correlation synthetic discriminant functions. *Appl Opt* 25:2343–2350
4. Mahalanobis A, Vijaya-Kumar BVK, Casasent D (1987) Minimum average correlation energy filters. *Appl Opt* 31:1823–1833
5. González-Fraga JA, Kober V, Álvarez-Borrego J (2006) Adaptive synthetic discriminant function filters for pattern recognition. *Opt Eng* 45:057005-1–057005-10
6. Doh YH, Kim JC, Kim JW, Choi KH, Kim SJ, Alam MS (2004) Distortion-invariant pattern recognition based on a synthetic hit-miss transform. *Opt Eng* 43:1798–1803
7. Wang Q, Deng Y, Liu S (2006) Morphological fringe-adjusted joint transform correlation. *Opt Eng* 45:087002-1–087002-9
8. Garcia-Martinez P, Tejera M, Ferreira C, Lefebvre D, Arsenault HH (2002) Optical implementation of the weighted sliced orthogonal nonlinear generalized correlation for nonuniform illumination conditions. *Appl Opt* 41:6867–6874
9. Martínez-Díaz S, Kober V (2008) Nonlinear synthetic discriminant function filters for illumination-invariant pattern recognition. *Opt Eng* 47:067201-1–067201-9
10. Martínez-Díaz S, Kober V () Morphological correlation for robust image recognition. In: *Proceedings of IEEE 2011 international conference on computational science and its applications*, Santander, Spain, 20–23 June 2011, pp 263–266
11. Maragos P (1989) Morphological correlation and mean absolute error criteria. In: *IEEE transactions on acoustics speech and signal processing*, pp 1568–1571
12. Fitch JP, Coyle EJ, NC Gallagher Jr (1984) Median filtering by threshold decomposition. In: *IEEE transactions on acoustics speech and signal processing*, pp 1183–1188
13. Martinez-Diaz S (2012) Parallel architecture for binary images recognition. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012*, WCE 2012, London, UK, 4–6 July 2012, pp 856–859
14. Vardavoulia MI, Andreadis I, Tsalides P (2002) Hardware implementation of soft color image morphological operations. *Opt Eng* 41:1536–1545
15. Bouchoux S, Bourenane E (2005) Application based on dynamic reconfiguration of field programmable gate arrays: JPEG 2000 arithmetic decoder. *Opt Eng* 44:107001-1–107001-6
16. Grivas E, Kyriakis-Bitzaros ED, Halkias G, Katsafourous SG, Morthier G, Dumon P, Baets R (2008) Wavelength division multiplexing based optical backplane with arrayed waveguide grating passive router. *Opt Eng* 47:025401-1–025401-7
17. Rakvic RN, Ives RW, Lira J, Molina C (2011) Case for a field-programmable gate array multicore hybrid machine for an image-processing application. *J Electron Imaging* 20:013015-1–013015-9
18. Hauck S, Dehon A (2008) *Reconfigurable computing*. Elsevier, Amsterdam
19. Zeidman B (2002) *Designing with FPGA's and CPLD's*. Elsevier, Amsterdam

In-Situ Vibrational Spectroscopies, BTEM Analysis and DFT Calculations

Feng Gao, Chuanzhao Li, Effendi Widjaja and Marc Garland

Abstract Reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with two conjugated dienes, namely, 2,3-dimethyl-1,3-butadiene (DMBD) and isoprene, were performed in anhydrous hexane under argon atmosphere with multiple perturbations of reagents. These reactions were monitored by in-situ FTIR (FIR and MIR) and/or Raman spectroscopies and the collected spectra were further analyzed with BTEM family of algorithms. The combined spectroscopic data seems to suggest that one organo-rhodium product $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-diene})$ (diene = DMBD, isoprene) was the main product during the reactions. DFT calculations further confirm that three carbonyls are bonded to one rhodium atom while the 4th carbonyl and a chelating diene ligand are bonded to the other rhodium atom. The possible coordination geometry was obtained with (1) the consideration of the coordination chemistry and (2) the consistence between the DFT predicted spectra in FTIR and Raman regions with the corresponding BTEM estimates. The present contribution shows that BTEM can be meaningfully applied to the reaction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ and DMBD/isoprene in order to provide enhanced spectroscopic analysis, especially in the FIR and Raman regions. Furthermore, the present results provide a better understanding of the coordination chemistry of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with conjugated dienes.

Keywords BTEM · Conjugated diene · DFT · FTIR/Raman · In-situ spectroscopies · Rhodium complexes

F. Gao (✉) · C. Li · E. Widjaja · M. Garland (✉)
The Institute of Chemical and Engineering Sciences, 1 Pesek Road,
Jurong Island, Singapore 627833, Singapore
e-mail: gao_feng@ices.a-star.edu.sg

C. Li
e-mail: chuanzhao.li@newcastle.ac.uk

E. Widjaja
e-mail: effendi.widjaja@merck.com

M. Garland
e-mail: marc_garland@ices.a-star.edu.sg

1 Introduction

Reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with simple ligands (L) usually produce substituted dinuclear complexes $\text{Rh}_2(\text{CO})_{4-m}\text{Cl}_2\text{L}_m$, ($m = 1-4$) as well as mononuclear complexes $\text{Rh}(\text{CO})_{3-n}\text{ClL}_n$, ($n = 1-3$) [1, 2]. These ligands include ethylene and phosphine etc. A cis-mononuclear complex $\text{Rh}(\text{CO})_2\text{Cl}(\text{py})$ was formed when pyridine (py) was used [3].

With respect to the rhodium complexes with conjugated dienes, Porri and Lionetti have reported some mononuclear bis(diene) complexes $[(\text{diene})_2\text{RhCl}]$ (diene = butadiene, isoprene, trans-penta-1,3-diene, 2,3-dimethyl-butadiene, and methyl hexa-2,4-dienoate) by reacting RhCl_3 with various dienes [4, 5]. Some mixed mononuclear complexes $[(\text{diene})(\text{L})\text{RhCl}]$ (L = cyclooctene or p-toluidine) have also been reported by them with the similar method. In addition, two dinuclear complexes $[(\text{diene})\text{RhCl}]_2$ (diene = 2,3-dimethylbutadiene and cyclohexa-1,3-diene) have been prepared by reacting RhCl_3 with the dienes [6]. Dinuclear diene complexes $\text{Rh}_2(\text{CO})_2\text{Cl}_2$ (COD) and $\text{Rh}_2\text{Cl}_2(\text{COD})_2$ are the predominant species formed when $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ reacts with 1,5-cyclooctadiene (COD) [7, 8]. An additional complex $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ has also been reported when $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ reacted with 2,3-dimethyl-1,3-butadiene (DMBD). The proposed structure suggested that the diene ligand was acted as a bridge and the core structure was unchanged compared to the precursor $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ [9]. In general, a five-coordinated square pyramidal structure was preferred for the mononuclear complexes where the centers of the four C=C bonds constituting the basal plane and the chlorine atom occupying the apical position [10]. However, a four-coordinate, chlorine-bridged structure was preferred for the dinuclear complexes. In both mononuclear and dinuclear complexes, the diene ligand was found to be bonded to the rhodium atom in a chelate mode.

In-situ vibrational spectroscopies have been successfully used in combination with BTEM spectral analysis and DFT calculations to better understand the detailed chemistry of many organometallic and homogeneous catalyzed reactions, particular those involving rhodium carbonyl complexes [11, 12]. In the present contribution, both FIR and low wavenumber MIR measurements are used, in addition to MIR and Raman in the carbonyl range, in order to better understand the coordination chemistry of conjugated dienes with $\text{Rh}_2(\text{CO})_4\text{Cl}_2$. The BTEM analysis of the data provided the spectral estimates of two dinuclear rhodium complexes $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-diene})$ (diene = DMBD, isoprene) in various interest regions. DFT calculations were performed to explore the possible coordination geometries and predict their corresponding FTIR and Raman vibrational frequencies.

2 Experimental

General information All solution preparations and transfers were performed with Schlenk techniques [13] under argon (99.999 %, Soxal, Singapore). The solvent hexane (99.6 %+, Fluka) was refluxed for ca. 5 h over sodium-potassium alloy under

argon. Rhodium carbonyl chloride (99 %, Strem) 1,3-dimethyl-1,3-butadiene (99 %, Sigma-Aldrich), and isoprene (99 %, Sigma-Aldrich) were used as received.

Equipment setup and in-situ spectroscopic measurements A Bruker FTIR spectrometer (Vertex 70) with deuterated triglycine sulfate (DTGS) detector was used. The spectral resolution was 2 cm^{-1} for the MIR region of $400\text{--}5000\text{ cm}^{-1}$. And the spectral resolution was 4 cm^{-1} for the FIR region of $30\text{--}700\text{ cm}^{-1}$. A cell with diamond windows was used for the FIR measurements. Purified compressed air was used to purge the FTIR spectrometer system. In-situ Raman spectra in the region of $100\text{--}2400\text{ cm}^{-1}$ were recorded with a dispersive-type Raman microscope (InVia Reflex Renishaw, UK). The laser source used was a 785 nm near-infrared diode laser with 100 % power (ca. 100 mW on the sample) and an exposure time of ca. 30 s. The in-situ spectroscopic measurements were performed at room temperature (ca. $24\text{ }^{\circ}\text{C}$).

The general experimental system has been reported elsewhere [11, 12]. The reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with DMBD/isoprene were performed under argon with multiple perturbations. A typical reaction was initiated by the injection of a certain amount of DMBD/isoprene solution through a rubber septum. Afterwards at pre-determined times, various perturbations of DMBD/isoprene solution and $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ powder were performed. It is well documented that multiple perturbation experiments greatly improve the quality of spectral reconstructions via BTEM.

Spectral reconstruction via BTEM family of algorithms Multi-reconstruction entropy minimization (MREM) [14, 15] was used to first survey the underlying spectral patterns. Subsequently, based on the obtained initial local spectral estimates via MREM, the pure component spectral estimates were further refined using either BTEM [16, 17] or tBTEM [18].

Density functional calculations Gaussian 09 [19] was used for all the calculations in this study. The geometric optimizations were performed using PBEPBE density function with DGDZVP basis set plus solvent effect of heptane and the FIR/MIR/Raman vibrational frequencies were calculated afterwards.

3 Results

3.1 Reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with 2,3-Dimethyl-1,3-Butadiene (DMBD)

The experimental spectra from the reaction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with DMBD were cut into three regions, namely, $200\text{--}650$, $800\text{--}1300$, and $1500\text{--}2200\text{ cm}^{-1}$, and further processed with the BTEM family of algorithms. The FIR/MIR BTEM estimates of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ in these three regions are shown in Fig. 1. Its corresponding FTIR-Raman BTEM estimate in the region of $1900\text{--}2150\text{ cm}^{-1}$ is shown in Fig. 2.

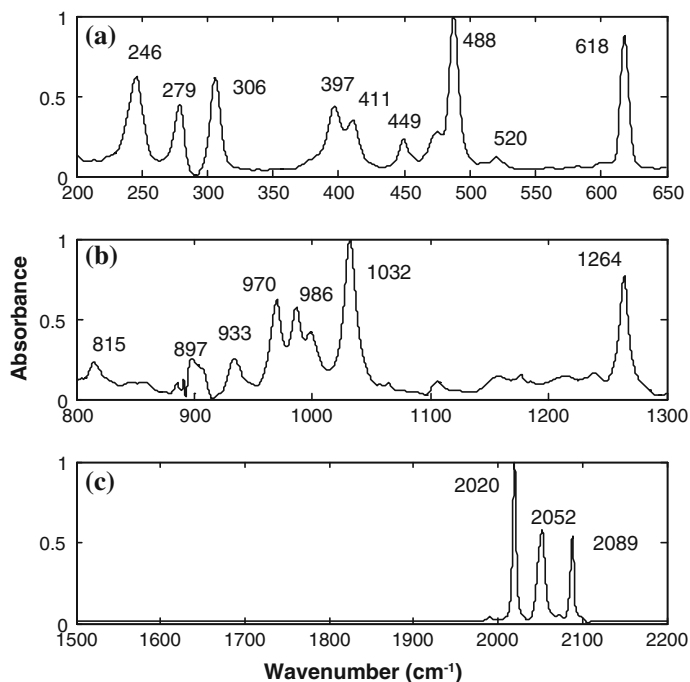
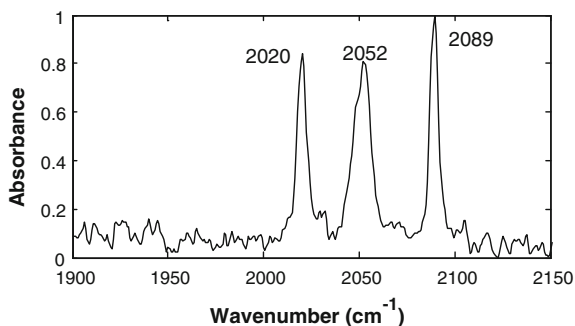


Fig. 1 The BTEM estimates of the organo-rhodium complex Rh₂(CO)₄Cl₂(η⁴-C₆H₁₀) in the FIR/MIR regions. (a) 200–650; (b) 800–1300; and (c) 1500–2200 cm⁻¹

Fig. 2 The Raman BTEM estimate of Rh₂(CO)₄Cl₂(η⁴-C₆H₁₀) in the region of 1900–2150 cm⁻¹



A combination of BTEM estimates of Rh₂(CO)₄Cl₂(η⁴-C₆H₁₀) and DMBD and experimental reference of cell plus hexane and Rh₂(CO)₄Cl₂ in the carbonyl region of 1500–2200 cm⁻¹ were used in the least squares fitting. The calculated relative concentrations are plotted in Fig. 3. Almost 100 % signal recovery was obtained.

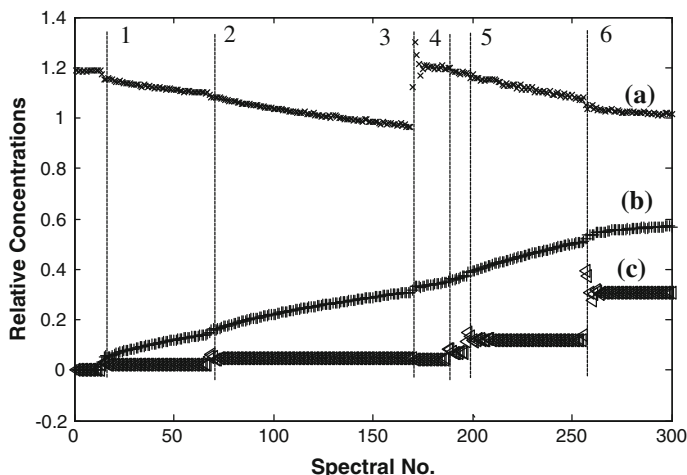


Fig. 3 The relative concentration profiles for the major species present in the system. (a) $\text{Rh}_2(\text{CO})_4\text{Cl}_2$; (b) $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$; and (c) DMBD. 1–2, 4: 20 μL DMBD injected in sequence; 3: 30 mg of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$; 5: 50 μL DMBD injected; 6: 200 μL DMBD injected

3.2 Reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with Isoprene

A similar spectral pattern was obtained when BTEM analysis was applied to the experimental spectra for the reaction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with isoprene. The BTEM estimate of the organo-rhodium product $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ in the MIR region of 1600–2200 cm^{-1} is plotted in Fig. 4. It should be noted here that one pure component spectrum with maxima at 2016, 2054, 2061, and 2086 cm^{-1} was also recovered from the reaction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with 1,5-cyclooctadiene (COD) although it was a minor species [20].

3.3 DFT Spectral Prediction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$

The DFT predicted FTIR spectra of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ in three FIR/MIR regions, namely, 200–650, 800–1300, and 1500–2200 cm^{-1} , are shown in Fig. 5. The corresponding Raman spectrum in the carbonyl region is shown in Fig. 6. It can be seen from this figure that the DFT predicted and the BTEM estimated wavenumbers (Figs. 1 and 2) show a systematic deviation of 0–2% in the MIR region and ca. 1–10% in the FIR region. Therefore the DFT predicted spectra are fairly consistent with the BTEM estimates in the three regions.

The optimized geometry of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ is plotted in Fig. 7. The core structure consists of two Rh atoms bridged by two Cl atoms. This core structure is obviously non-planar, the dihedral angle about the Cl...Cl hinge is ca. 141° which

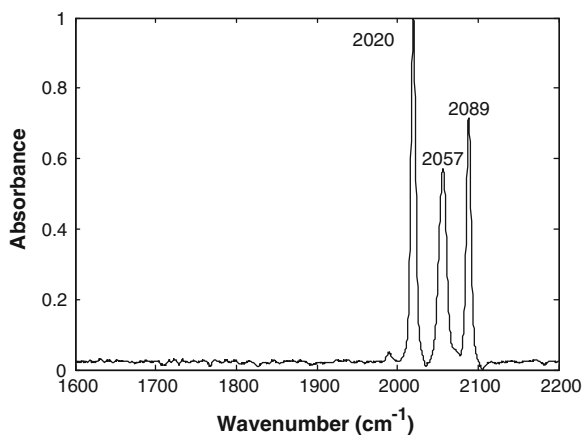


Fig. 4 The BTEM estimate of the organo-rhodium product $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ in the MIR region of $1600\text{--}2200\text{ cm}^{-1}$

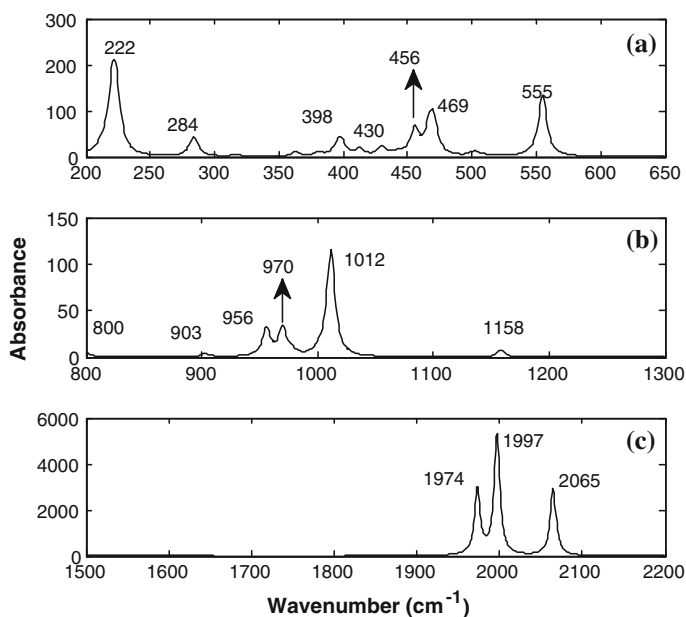


Fig. 5 The DFT predicted spectra of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ in three FIR/MIR regions. (a) $200\text{--}650$; (b) $800\text{--}1300$; and (c) $1500\text{--}2200\text{ cm}^{-1}$

Fig. 6 The DFT predicted Raman spectrum of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ in the carbonyl region of $1900\text{--}2150\text{ cm}^{-1}$

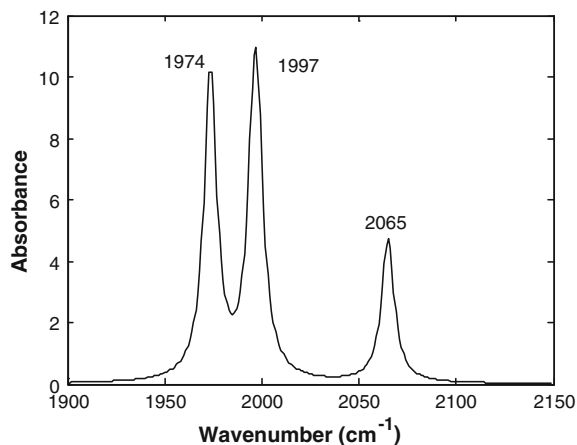
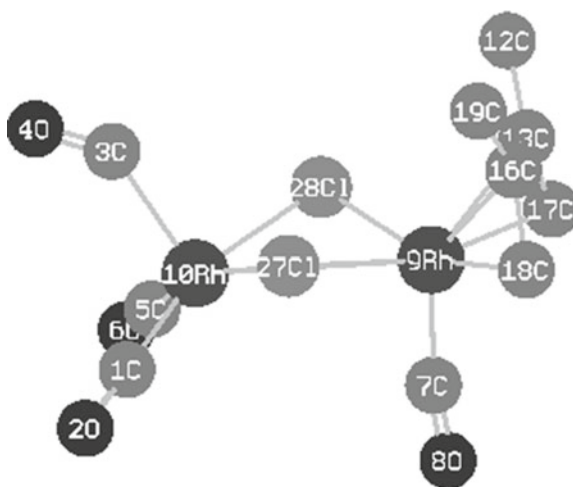


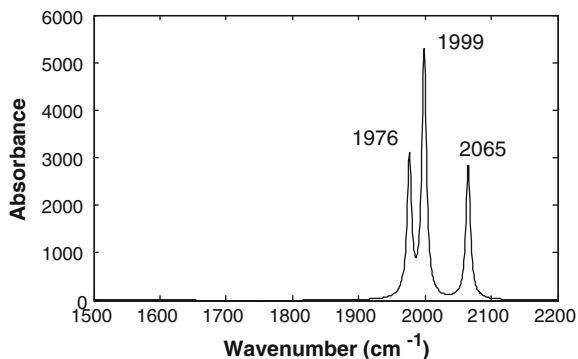
Fig. 7 The optimized geometry of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$



is larger than that of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ (ca. 124°). Three carbonyls are bonded to one rhodium atom while the 4th carbonyl and a chelating diene ligand are bonded to the other rhodium atom.

In addition, the DFT predicted wavenumbers (2001 and 2059 cm^{-1}) of the bridged-diene species are similar to those of the precursor $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ (2003 and 2060 cm^{-1}). This similarity may arise from the similar core structures of these two species.

Fig. 8 The DFT predicted spectrum of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ in the carbonyl regions of $1500\text{--}2200\text{ cm}^{-1}$



3.4 DFT Spectral Prediction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$

The DFT predicted spectrum of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ in the carbonyl region is shown in Fig. 8. A systematic deviation (0–2 %) from the BTEM estimated wavenumbers (Fig. 4) was also observed in the carbonyl region. It is worthwhile to mention here that the predicted wavenumbers and optimized geometry of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ are similar to those of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$.

4 Discussion

In-situ FTIR and/or Raman measurements were combined with signal processing and DFT calculations in order to better understand the reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with DMBD and isoprene [21]. The major species formed was a known diene complex $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-diene})$ (diene = DMBD, isoprene). The carbonyl band pattern shown in Figs. 1c, 2 and 4 suggest that the species may be a dinuclear with low symmetry and four carbonyl ligands. DFT calculations further confirm that the possible structure of this dinuclear species may be with three carbonyls bonded to one Rh atom and the 4th carbonyl and a chelating diene ligand bonded to the other Rh atom. And the two Rh atoms are bridged by two Cl atoms. This assignment was made based on the consistency between the DFT predicted spectra and the BTEM estimates, as well as consideration of the potential coordination chemistry.

Differences were observed between the DFT predicted wavenumbers and the BTEM estimates for the major organo-rhodium species. These differences may be due to the difficulties to accurately model the electron density associated with the Rh-Cl, Rh-Rh and Rh-C bonds.

5 Conclusion

A combination of in-situ FTIR and/or Raman spectroscopic measurements, BTEM reconstruction and DFT calculations was successfully used to study the reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with DMBD and isoprene. One known organo-rhodium complex $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-diene})$ (diene = DMBD, isoprene) was formed as the primary species. The pure component spectra of $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_6\text{H}_{10})$ and $\text{Rh}_2(\text{CO})_4\text{Cl}_2(\eta^4\text{-C}_5\text{H}_8)$ in various regions were successfully recovered with the application of BTEM family of algorithms and without the purification from the solution. This study demonstrate that with the combination of in-situ spectroscopic study, BTEM reconstruction and DFT calculations, some of the diverse and complex coordination chemistries between $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ and conjugated dienes can be addressed.

Acknowledgments We would like to thank Dr Chacko Jacob of the Institute of Chemical and Engineering Sciences in Singapore for his useful discussions.

References

1. Gallay J, de Montauzon D, Poilblanc R (1972) Comprehensive study of the systems μ -dichlorotetracarbonyldirhodium-phosphines and evidence of oxidative addition involving new dinuclear complexes. *J Organometal Chem* 38:179–197
2. Maisonnat A, Kalck P, Poilblanc R (1974) Dinuclear bridged d8 metal complexes. I. Preparation and structure of chlorocarbonyl substituted phosphine or phosphite rhodium dimer $[\text{RhCl}(\text{CO})\text{L}]_2$ compounds. *Inorg Chem* 13:661–667
3. Heaton BT, Jacob C, Sampanthar JT (1998) Rhodium carbonyl complexes containing pyridine; crystal structure of an unusual octahedral rhodium (I) complex $[\text{Rh}_2(\mu\text{-CO})_3\text{Cl}_2(\text{py})_4]$. *J Chem Soc Dalton Trans* 1403–1410
4. Porri L, Lionetti A, Allegra G, Immirzi A (1965) Dibutadienerrhodium(I) chloride. *Chem Commun (Lond)* 336–337
5. Porri L, Lionetti A (1966) New complexes of rhodium(I) with conjugated diolefins or other compounds containing activated double bonds. *J Organometal Chem* 6:422–424
6. Winkhaus G, Singer H (1966) *Chem Ber* 99:3602–3609
7. Chatt J, Venanzi LM (1957) Olefin co-ordination compounds. VI. Diene complexes of rhodium(I). *J Chem Soc* 4735–4741
8. Abou Rida M, Saikaili J, Smith AK, Thozet A (2001) Dicarboxyldi- μ -chloro-cis, cis- η^4 -1,5-cyclooctadienerrhodium(I). *Acta Cryst C* 57(4):352–353
9. Winkhaus G, Singer H (1966) Addition complexes of rhodium carbonyl chloride and olefins. *Chem Ber* 99:3593–3601
10. Immirzi A, Allegra G (1969) X-ray structure of $\text{Rh}^{\text{I}}\text{Cl}(\text{C}_4\text{H}_6)_2$. *Acta Cryst B* 25:120–124
11. Allian AD, Tjahjono M, Garland M (2006) Reaction of alkynes with $\text{Rh}_4(\text{CO})_{12}$. A mid-infrared vibrational and kinetic study of $(\mu_4\text{-}\eta^2\text{-alkyne})\text{Rh}_4(\text{CO})_8(\mu\text{-CO})_2$. *Organometallics* 25:2182–2188
12. Allian AD, Widjaja E, Garland M (2006) Experimental Raman spectra of dilute and laser-light-sensitive $[\text{Rh}_4(\text{CO})_9(\mu\text{-CO})_3]$ and $[(\mu_4\text{-}\eta^2\text{-3-hexyne})\text{Rh}_4(\text{CO})_8(\mu\text{-CO})_2]$. Comparison with theoretically predicted spectra. *J Chem Soc Dalton Trans* 4211–4217
13. Shriver DF, Drezdson MA (1986) *The manipulation of air-sensitive compounds*. Wiley, New York

14. Zhang HJ, Chew W, Garland M (2007) The multi-reconstruction entropy minimization method: unsupervised spectral reconstruction of pure components from mixture spectra, without the use of a Priori information. *Appl Spectrosc* 61:1366–1372
15. Gao F, Zhang HJ, Guo LF, Garland M (2009) Application of the BTEM family of algorithms to reconstruct individual UV-Vis spectra from multi-component mixtures. *Chemom Intell Lab Syst* 95:94–100
16. Chew W, Widjaja E, Garland M (2002) Band-target entropy minimization (BTEM): an advanced method for recovering unknown pure component spectra. Application to the FTIR spectra of unstable organometallic mixtures. *Organometallics* 21:1982–1990
17. Widjaja E, Li CZ, Garland M (2002) Semi-batch homogeneous catalytic in-situ spectroscopic data. FTIR spectral reconstructions using band-target entropy minimisation (BTEM) without spectral preconditioning. *Organometallics* 21:1991–1997
18. Zhang HJ, Garland M, Zeng YZ, Wu P (2003) Weighted two-band target entropy minimization for the reconstruction of pure component mass spectra: simulation studies and the application to real systems. *J Am Soc Mass Spectrometry* 14:1295–1305
19. Gaussian 09 (2009) Revision A.02. Gaussian, Wallingford CT
20. Gao F, Li CZ, Widjaja E, Jacob C, Garland M (2010) The reactions of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with 1,5-cyclooctadiene and tetramethylallene. A combined in-situ vibrational spectroscopies, spectral reconstruction and DFT study. *J Organometal Chem* 695:2394–2400
21. Gao F, Li CZ, Widjaja E, Garland M (2012) Investigation of the reaction of $\text{Rh}_2(\text{CO})_4\text{Cl}_2$ with 2,3-dimethyl-1,3-butadiene/isoprene by In-situ vibrational Spectroscopies. BTEM analysis and DFT calculations, Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012, pp 823–826

Convergence Speed of Generalized Longest-Edge-Based Refinement

José P. Suárez, Tania Moreno, Pilar Abad and Ángel Plaza

Abstract In the refinement of meshes, one wishes to iteratively subdivide a domain following geometrical partition rules. The aim is to obtain a new discretized domain with adapted regions. We prove that the Longest Edge n -section of triangles for $n \geq 4$ produces a finite sequence of triangle meshes with guaranteed convergence of diameters and review previous result when n equals 2 and 3. We give upper and lower bounds for the convergence speed in terms of diameter reduction. Then we fill the gap in the analysis of the diameters convergence for generalized Longest Edge based refinement. In addition, we give a numerical study for the case of $n = 4$, the so-called LE quartersection, evidencing its utility in adaptive mesh refinement.

Keywords Diameter · Longest-edge · Mesh refinement · n -section · Refinement · Triangulation · Triangle partition

J. P. Suárez (✉) · P. Abad · Á. Plaza
Division of Mathematics, Graphics and Computation (MAGiC), IUMA,
Information and Communication Systems, University of Las Palmas de
Gran Canaria, Canary Islands, Las Palmas de Gran Canaria, Spain
e-mail: jsuarez@dcegi.ulpgc.es

P. Abad
e-mail: mabad@dcegi.ulpgc.es

Á. Plaza
e-mail: aplaza@dcegi.ulpgc.es

T. Moreno
Faculty of Mathematics and Informatics, University of Holguin, Avenida XX Aniversario,
via Guardalavaca, Holguin, Cuba
e-mail: tmorenog@facinf.uho.edu.cu

1 Introduction

Since the apparition of Finite Element Method in the 60th, many mesh partitions methods became popular. Mesh Refinement algorithms use such partition methods to refine a given mesh. One wishes to construct a sequence of nested conforming meshes which are adapted to a given criterion. Nested sequences of triangles where each element in the sequence is a child of parent triangle of same sequence are of quite interest in many areas as Finite Element Multigrid Methods, Image Multiresolutions etc. [1].

Let t be a triangle in \mathbb{R}^2 . We find the Longest Edge (LE) of t , insert $n - 1$ equally-space points in the LE and connect them to the opposite vertex. This yields the generation of n new sub-triangles whose parent is t . Now, continue this process iteratively. Proficient algorithms for mesh refinement using this method are known when $n = 2$, but less known when $n = 3$ and completely unknown when $n \geq 4$.

Although Delaunay triangulations maximize the minimum angle of all the angles of the triangles in any triangulation, some other competitive methods have emerged in the last decade specially those with cheaper computational cost as Longest Edge (LE) based subdivision. LE n -section based algorithms are surprisingly cheap. They are linear in the number of elements, as the only necessary calculations are: (i) Longest Edges and (ii) insertion of n points in the LE side, both of constant-time.

1.1 Longest Edge Based Refinement

We first give a short overview of existing methods for LE bisection and LE trisection.

1.1.1 Longest Edge Bisection Partition

Possible the first method to repeatedly subdivide a triangle mesh was the Longest Edge bisection of a triangle. Rosenberg and Stenger showed the non-degeneracy property for LE-bisection: $\alpha_n \geq \frac{\alpha_0}{2}$ [2] where α_n is the minimum interior angle in new triangles appeared at iteration n , and α_0 the minimum angle of initial given triangle.

1.1.2 Four-Triangles Longest-Edge Partition

Further research on Longest-Edge bisection has been carried out since the ninety. Many other variants of LE-bisection have appeared in this period. For example, the *Four-Triangles Longest-Edge Partition (4T-LE)* bisects a triangle into four subtriangles as follows: the original triangle is first subdivided by its longest edge and then the two resulting triangles are bisected by joining the new midpoint of the longest edge to the midpoints of the remaining two edges of the original triangle. The 4T-LE

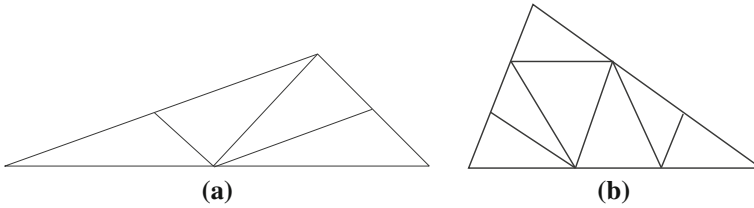


Fig. 1 **a** Scheme for the Four-Triangles Longest-Edge Partition; **b** scheme for the Seven-Triangles Longest-Edge Partition

partition of a given triangle t never produces an angle smaller than half the minimum original angle and besides, it shows a remarkable mesh quality improvement between certain limits, as recently studied in [3]. See Fig. 1a for a scheme of that partition. 4T-LE refinement improves angles and this improvement has been studied in depth [3] where sharper bounds for the number of dissimilar triangles arising from the 4T-LE are given.

1.1.3 Seven-Triangles Longest-Edge Partition

Superior quality improvement of the triangulation can be achieved by the 7-Triangle Longest-Edge (7T-LE) partition [4]. This partition is constructed by positioning two equally spaced points per edge and join them, using parallel segments, to the edges, at the points closest to each vertex. Then joining the two interior points of the longest edge of the initial triangle to the base points of the opposite sub-triangle in such a way that they do not intersect, and finally, triangulating the interior quadrangle by the shortest diagonal. See Fig. 1b for a scheme of that partition. Two of new triangles generated are similar to the new triangle also generated by the 4T-LE, and the other two triangles are, in general, better shaped. As a consequence, the area covered by better triangles is showed to be superior compared to the 4T-LE [4].

1.1.4 Longest Edge Trisection Partition

There is very little research so far on LE n -section methods other than bisection. Recently, a new class of triangle partitions based on the Longest-Edge Trisection has been presented by Plaza et al. [5]. It simply consists in inserting three equal points in the Longest Edge and then connecting them to the opposite vertex. Empirical evidence has been given of the non-degeneracy of the meshes obtained by iterative application of LE-trisection. In fact, if α_0 is the minimum interior angle of the initial triangle, and α_n , the minimum interior angle after n levels of LE-trisection, then $\alpha_n \geq \alpha_0/6.7052025350$, independently on the value of n [5]. To complete the study of non-degeneracy for LE-Trisection, it has been proved in [6] that for LE-trisection

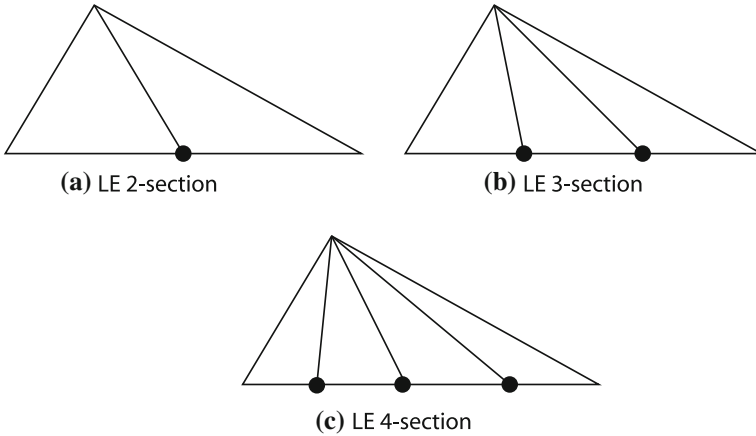


Fig. 2 Scheme for Longest Edge (LE) n -section of a triangle ($n = 2, 3, 4$)

$\alpha_n \geq \frac{\alpha_0}{c}$ where $c = \frac{\pi/3}{\arctan\left(\frac{\sqrt{3}}{11}\right)}$. This result confirms previous numerical research in [5]. In Fig. 2 we illustrate some schemes for LE n -section, $n=2,3,4$.

1.2 Convergence Speed for Longest Edge Refinement

For better understanding of repeated refinement of triangles, the Longest Edge of successive triangle generation has been studied. In such methods, it should be ensured the convergence of the triangle sequence to a singleton at a fast convergence rate which is often measured by means of the corresponding sequence of diameters in the Euclidean norm. For the LE bisection ($n = 2$), these studies began with the result of Kearfott [7] who proved a bound on the behavior of the length of the longest edge of any triangle (diameter) obtained. Later Stynes [8] presented a better bound for certain initial triangles. After that, Stynes [9] and Adler [10] improved this bound for all triangles. Other methods not based on the LE, for example, the Generalized Conforming Bisection algorithm [11] also has proved to fulfill the convergence of diameters.

In [12] a mathematical proof of how fast the diameters of a triangle mesh tend to zero after repeated LE trisection is presented. Again, bounds on the behavior of the diameters of the longest edge are known for the case of $n = 3$. However this study has not been carried out yet for LE n -section, $n \geq 4$ [13].

In this chapter we prove that the LE n -section of triangles for $n \geq 4$ of triangles produces a finite sequence of triangle meshes with guaranteed convergence of diameters. We give upper and lower bounds for the convergence speed in terms of diameter reduction. In addition, we explore in details the so-called LE quartersection ($n = 4$) of triangles by studying the triangles shapes emerging in that process.

The structure of this chapter is as follows: Sects. 2 and 3 introduces and proves the upper and lower bound respectively. Section 4 gives a numerical study where shape quality is studied for the case of $n = 4$ of LE quartersection, evidencing its utility in adaptive mesh refinement. We close with some final conclusions in Sect. 5.

2 Upper Bound of Diameters in LE n -Section ($n \geq 4$)

We prove following theorem:

Theorem 1. *Let d_k be the diameter in the k iterative application of Longest-Edge n -section ($n \geq 4$) to a given arbitrary triangle $\triangle ABC$, then:*

$$d_{2k} \leq \left(\frac{\sqrt{n^2 - n + 1}}{n} \right)^k d_0$$

where d_0 is the diameter of initial triangle and $k \geq 0$.

Before, we give some previous lemmas that are used in the proof:

Lemma 1. *(Theorem of Stewart.) Let $\triangle ABC$ be an arbitrary triangle and S be a point in \overline{BC} . Then:*

$$|\overline{AS}|^2 |\overline{BC}| = |\overline{AB}|^2 |\overline{SC}| + |\overline{AC}|^2 |\overline{BS}| - |\overline{BS}| |\overline{SC}| |\overline{BC}|. \tag{1}$$

Lemma 2. *Let $\triangle ABC$ be an arbitrary triangle where $|\overline{AB}| \leq |\overline{AC}| \leq |\overline{BC}|$. Then:*

1. $|\overline{AS}| \leq |\overline{AC}|$ for each $S \in \overline{BC}$, see Fig. 3a.
2. Let X and Y be points in segment \overline{BC} such that segments \overline{BX} and \overline{CY} are equal and have empty intersection, then $|\overline{AX}| \leq |\overline{AY}|$, see Fig. 3b.

Note that part two of Lemma 2 is straightforward as a consequence of part one, see Fig. 3.

If for an arbitrary triangle $\triangle ABC$ we have that $|\overline{AB}| \leq |\overline{AC}| \leq |\overline{BC}|$ then we say that length of shortest edge is $|\overline{AB}|$, length of medium edge is $|\overline{AC}|$ and length of longest edge is $|\overline{BC}|$.

Lemma 3. *Let $n \geq 4$ and $\triangle ABC$ such that $|\overline{AB}| \leq |\overline{AC}| \leq |\overline{BC}|$. Let X_1, X_2, \dots, X_{n-1} points of \overline{BC} such that $|\overline{BX}_1| = |\overline{X}_1 \overline{X}_2| = \dots = |\overline{X}_{n-1} \overline{C}| = \frac{1}{n} |\overline{BC}|$. Then the length of medium size of triangle $\triangle BAX_1, \triangle X_1AX_2, \dots, \triangle X_{n-1}AC$ is less or equal than $|\overline{AX}_{n-1}|$.*

Proof: From Lemma 2 we have that $|\overline{AX}_i| \leq |\overline{AX}_{n-1}| \leq |\overline{AC}|$ for each $i \in \{1, 2, \dots, n-1\}$. Thus, it is clear that $\angle AX_{n-2}X_{n-1} \geq \frac{\pi}{2}$, and so $|\overline{BX}_1| = |\overline{X}_1 \overline{X}_2| = \dots = |\overline{X}_{n-1} \overline{C}| \leq |\overline{AX}_{n-1}|$, see Fig. 4.

Fig. 3 a $|\overline{AS}| \leq |\overline{AC}|$ for each $S \in \overline{BC}$, b $|\overline{AX}| \leq |\overline{AY}|$

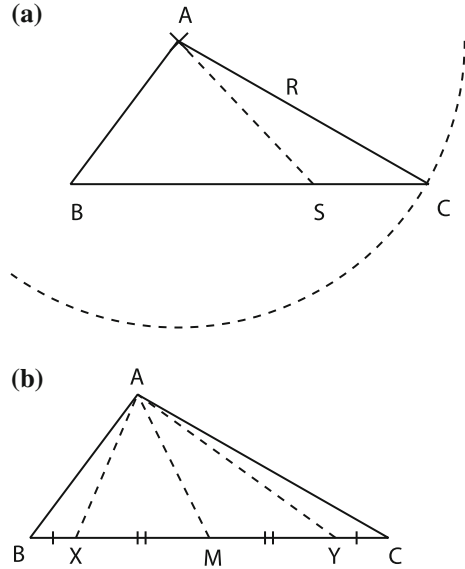
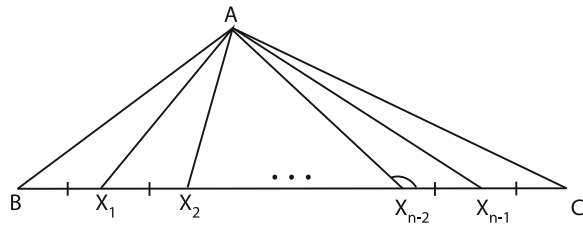


Fig. 4 $|\overline{BX_1}| = |\overline{X_1X_2}| = \dots = |\overline{X_{n-1}C}| \leq |\overline{AX_{n-1}}|$



Suppose that $|\overline{AB}|$ is the length of medium edge of $\triangle BAX_1$, then the length of its longest edge is, either $|\overline{AX_1}|$ or $|\overline{BX_1}|$. Note that last two segments are less or equal than $|\overline{AX_{n-1}}|$, as previously proven. \square

Lemma 4. Let a triangle $\triangle ABC$ such that $|\overline{AB}| \leq |\overline{AC}| \leq |\overline{BC}|$ and $P \in \overline{BC}$ such that $\frac{BP}{PC} = n - 1$ for $n \geq 4$. Then:

$$|\overline{AP}| \leq \frac{\sqrt{n^2 - n + 1}}{n} |\overline{AC}|. \tag{2}$$

Proof: Let $|\overline{AB}| = c$, $|\overline{BC}| = a$, $|\overline{AC}| = b$. Then $|\overline{BP}| = \frac{n-1}{n}a$ and $|\overline{CP}| = \frac{1}{n}a$. By Lemma 1 we have:

$$|\overline{AP}|^2 a = c^2 \cdot \frac{1}{n} a + b^2 \cdot \frac{n-1}{n} a - \left(\frac{n-1}{n^2} \right) a^3,$$

from where

$$|\overline{AP}|^2 = \frac{c^2 + (n - 1)b^2}{n} - \left(\frac{n - 1}{n^2}\right) a^2.$$

Note that $c \leq b \leq a$, and thus $s^2 \leq \frac{b^2(n-1)+b^2}{n} - \left(\frac{n-1}{n^2}\right)b^2 = \left(\frac{n^2-n+1}{n^2}\right)b^2$, from where we have inequality which proves the result of the Lemma. \square

At this point, we follow with the proof of main result, Theorem 1.

Proof of Theorem 1: Let us consider the sequence $\{I_k\}_{k=1}^\infty$ such that I_k is the longest of the two medium edges of each triangle obtained after iteration k of LE n -section at ($n \geq 4$). Then:

$$d_{k+1} \leq I_k$$

Note that at iteration k , each longest edge previously obtained at iteration ($k + 1$) is subdivided in n equal parts. Each of these parts is the shortest edge of at least one of the triangle obtained at iteration ($k + 1$).

Using Lemmas 3 and 4 we have that: $I_k \leq \frac{\sqrt{n^2-n+1}}{n} d_k$. It is clear that $\frac{\sqrt{n^2-n+1}}{n} < 1$, and so $d_{k+2} \leq \frac{\sqrt{n^2-n+1}}{n} d_k$. Thus, we follow that $d_{2k} \leq \left(\frac{\sqrt{n^2-n+1}}{n}\right)^k d_0$. \square

3 Lower Bound of Diameters in LE n -Section ($n \geq 4$)

We next provide a lower bound for the LE n -section of triangles.

Theorem 2. *Let d_k be the diameter in the k iterative application ($k \geq 1$) of Longest-Edge n -section ($n \geq 4$) to a given arbitrary triangle with edges a, b and c such that $c \leq b \leq a$. Then, there exists constants p, q, r, s, t and u only dependent on a, b, c and n , such that it holds:*

1. For $n = 4$, $d_k^2 \geq \left(\frac{1}{4}\right)^k (pk^2 + qk + r) > 0$.
2. For $n \geq 5$, $d_k^2 \geq s \frac{1}{n^k} + t \left(\frac{n^2-2n+n^{\frac{3}{2}}\sqrt{n-4}}{2n^2}\right)^k + u \left(\frac{n^2-2n-n^{\frac{3}{2}}\sqrt{n-4}}{2n^2}\right)^k$.

Proof: Let $n \geq 4$ and $\triangle ABC$ be an arbitrary triangle with $|\overline{AB}| \leq |\overline{AC}| \leq |\overline{BC}|$, $|\overline{AB}| = c$, $|\overline{BC}| = a$ and $|\overline{AC}| = b$. Let us consider the triangle sequence $\{\Delta_k\}_{k=0}^\infty$ such that $\Delta_0 = \triangle A_0 B_0 C_0$, $A_0 = A$, $B_0 = B$, $C_0 = C$, and for each $k \geq 0$ let $\Delta_{k+1} = \triangle A_{k+1} B_{k+1} C_{k+1}$ where $A_{k+1} \in \overline{B_k C_k}$ such that $|\overline{A_{k+1} C_k}| = \frac{1}{n} |\overline{B_k C_k}|$, $B_{k+1} = C_k$ y $C_{k+1} = A_k$. It can be noted that for each $k \geq 1$, $|\overline{A_k B_k}| \leq |\overline{A_k C_k}| \leq |\overline{B_k C_k}|$; from where we have that for each $k \geq 1$, Δ_k is one of the n triangles obtained by applying the LE n -section to triangle Δ_{k-1} . \square

Let now consider the sequence $\{a_k\}_{k=0}^\infty$ where $a_k = |\overline{B_k C_k}|$. Using Lemma 1, following recurrence equation can be obtained:

$$a_{k+3}^2 - \frac{n-1}{n} a_{k+2}^2 + \frac{n-1}{n^2} a_{k+1}^2 - \frac{1}{n^3} a_k^2 = 0,$$

where $a_0 = nc$, $a_1 = a$ y $a_2 = b$. Stating $y_k = a_k^2$, we have:

$$y_{k+3} - \frac{n-1}{n} y_{k+2} + \frac{n-1}{n^2} y_{k+1} - \frac{1}{n^3} y_k = 0,$$

where $a_0 = n^2 c^2$, $a_1 = a^2$ and $a_2 = b^2$. It can be noted that from the construction of sequence y_k it is deduced that each terms of the sequence is positive. The characteristic equation of such recurrence equation is as follows:

$$\lambda^3 - \frac{n-1}{n} \lambda^2 + \frac{n-1}{n^2} \lambda - \frac{1}{n^3} = 0.$$

At this point, two separated situations can be given: (i) $n = 4$, where a square root of multiplicity three appears, and (ii) $n \geq 5$ where three real roots appears.

- (i) Case $n = 4$. The solution of the characteristic equation is $\lambda = \frac{1}{4}$, of multiplicity 3, and then:

$$y_k = \left(\frac{1}{4}\right)^k (pk^2 + qk + r),$$

where p, q y r are real constants only dependent on a, b y c . Such constants are solutions of an equation system obtained from the initial conditions, omitted here for brevity.

- (ii) Case $n \geq 5$. The characteristic equation has three real roots: $\lambda_1 = \frac{1}{n}$, $\lambda_2 = \frac{n^2 - 2n + n^{\frac{3}{2}} \sqrt{n-4}}{2n^2}$ and $\lambda_3 = \frac{n^2 - 2n - n^{\frac{3}{2}} \sqrt{n-4}}{2n^2}$. Thus:

$$y_k = \left(\frac{1}{n}\right)^k s + \left(\frac{n^2 - 2n + n^{\frac{3}{2}} \sqrt{n-4}}{2n^2}\right)^k t + \left(\frac{n^2 - 2n - n^{\frac{3}{2}} \sqrt{n-4}}{2n^2}\right)^k u.$$

where s, t and u are real constants only dependent on a, b, c and n . Such constants are solutions of an equation system obtained from the initial conditions, omitted here for brevity.

It should be noted that $d_k \geq a_k$, and so $d_k^2 \geq y_k = (\frac{1}{n})^k s + (\frac{n^2-2n+n^{\frac{3}{2}}\sqrt{n-4}}{2n^2})^k t + (\frac{n^2-2n-n^{\frac{3}{2}}\sqrt{n-4}}{2n^2})^k u$. □

Among the LE n -section methods, the LE 4-section or *LE quartersection* this point forwards, has not been explored yet as far as we know. We are dealing with LE quartersection in the rest of the chapter. In Fig. 5 it is graphed the bound diameters evolution when repeated LE quartersection is applied to three initial triangles with initial diameter equals 1. The coordinates (x, y) of the targeted triangles are:

$$\begin{aligned} \Delta 1 &= (0, 0) \quad (0.5, \sqrt{3}/2) \quad (1, 0) \\ \Delta 2 &= (0, 0) \quad (0.1, 0.1) \quad (1, 0) \\ \Delta 3 &= (0, 0) \quad (0.4, 0.01) \quad (1, 0) \end{aligned}$$

4 A Mapping Diagram to Represent Triangle Shapes

A mapping diagram is constructed as follows, see [14], to visually represent triangle shapes in LE quartersection refinements: (1) for a given triangle or subtriangle the longest edge is scaled to have unit length. This forms the base of the diagram, (2) it follows that the set of all triangles is bounded by this horizontal segment (longest edge) and by two bounding exterior circular arcs of unit radius. The diagram is then defined by the set:

$$\{(x, y) : x^2 + y^2 \geq 1\} \cap \{(x, y) : (x - 1)^2 + y^2 \geq 1\} \cap \dots \\ \{(x, y) : x \geq 0, y \geq 0\}$$

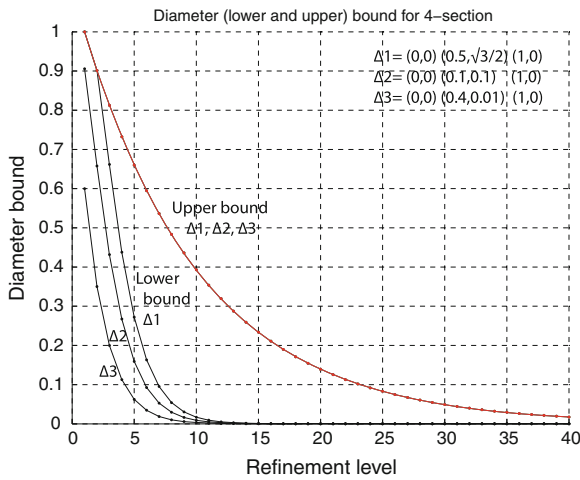


Fig. 5 Upper and lower bound for several triangles

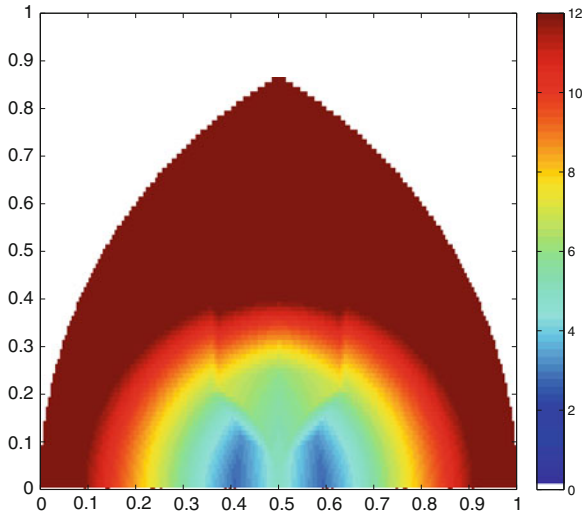


Fig. 6 Mapping diagram for two levels of quartersection. *Shaded colors* are $\frac{\tau_0}{\tau_2}$, being τ the minimum angle

In this manner, a point within the diagram univocally represents a triangle, whose apex is this point itself and the other two vertices are $(0, 0)$ and $(1, 0)$ respectively. This lead us to an easy and simple way to uniformly represent triangle shapes. For instance, a degenerated triangle in which its three vertices are collinear is represented by and apex over the base, the segment defined by coordinates $(0, 0)$ and $(1, 0)$, of the diagram. The equilateral triangle corresponds to the apex at $(\frac{1}{2}, \frac{\sqrt{3}}{2})$. As the vertex of a triangle moves from this point along either boundary arc, the maximum angle increases from $\frac{\pi}{3}$ to approach a right angle at the degenerate ‘needle triangle’ limit near $(0, 0)$ or $(1, 0)$.

In order to show the behaviour of LE quartersection for some refinement iterations, we employ the so described mapping diagram where shaded values within the diagram represents the quotient $\frac{\tau_0}{\tau_2}$, being τ_0 the minimum initial angle and τ_2 the minimum angle after two levels of refinement, see Fig. 6. Approximately five thousand of triangles are targeted for refinement, covering uniformly the interior area of the diagram. It can be noted in the diagram two interesting areas around coordinates $x = 0.4$ and $x = 0.6$. These focused areas are triangles with greatest minimum angles, whereas upper zones correspond to lower values.

The mapping diagram can be also used to analyze the shapes of subsequent triangles generated in iterative LE quartersection. For example, Fig. 7a, b shows separate cases for iterative refinement of given initial triangles, corresponding respectively to triangles $\Delta 1$ and $\Delta 3$. In these diagrams initial triangles $\Delta 1$ and $\Delta 3$ are marked with star-shape. It should be noted as new subdivided triangles are of different shapes as seen from the distribution of points within the diagram.

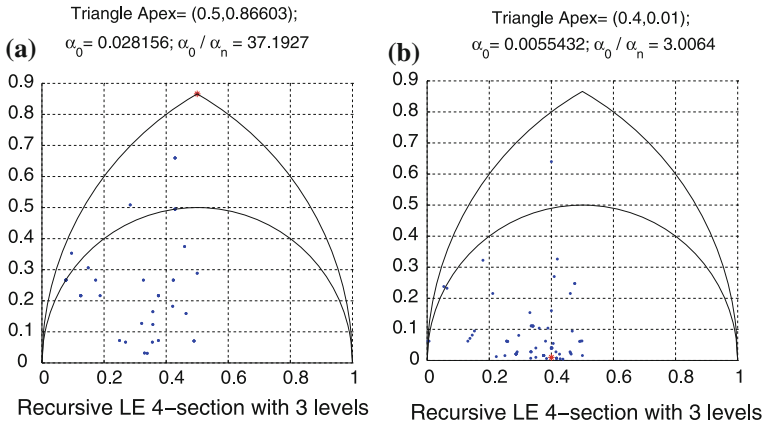


Fig. 7 Triangle generation resulting after 3 levels of LE quartersection applied to an initial triangle (marked in red). Studied cases for initial triangles: **a** $\Delta 1 = (0, 0)(0.5, \sqrt{3}/2)(1, 0)$, **b** $\Delta 3 = (0, 0)(0.4, 0.01)(1, 0)$. **a** Case of the regular initial triangle ($\Delta 1$) which marks the worst case for minimum angles. Shape of all the new triangles are worse than the regular, **b** A very bad initial triangle ($\Delta 3$) leading to some new improved subtriangles

Table 1 Minimum angles ratios α_0/α_n for three triangle cases and three refinement levels

	α_0/α_1	α_0/α_2	α_0/α_3
$\Delta 1$	11.5191	22.3865	37.1927
$\Delta 2$	5.0247	8.1670	12.0620
$\Delta 3$	1.8685	2.4038	3.0064

Finally, reported values of $\alpha_0/\alpha_n, n = 1, 2, 3$ for same triangles cases are reported in Table 1. Note that in LE quartersection shape quality is clearly dependent of the initial triangle. This behavior also occurs in other families of LE n -section. Thus, the regular triangle, as expected, poses the worst case $\alpha_0/\alpha_3 = 37.1927$, which is in agreement with the behaviour of LE bisection and LE trisection. In the case of initial triangles of poor quality, for example triangle $\Delta 3$, LE quartersection leads to reasonable output values for the minimum angles and then an adaptive mesh refinement algorithm that uses this method can be a valuable option for special narrowed or skinny triangle like this type, as those appearing in meshes from Fluid Dynamic, Electromagnetism etc.

5 Conclusions

Nested sequences of triangles where each element in the sequence is a child of parent triangle has shown to be critical in Multigrid Methods or Finite Element. In this chapter we generalized a class of triangle mesh refinement based on the Longest Edge and introduced the so-called Longest Edge n -section methods. Proficient algorithms for mesh refinement using this method are known when $n = 2$, but less known when

$n = 3$ and completely unknown when $n \geq 4$. LE n -section based algorithms are surprisingly cheap. They are linear in the number of elements, as the only necessary calculations are: (i) Longest Edges and (ii) insertion of n points in LE sides, which is of constant-time.

In this chapter we proved upper and lower bounds for the sequence of diameters generated by iterative application of LE n -section partition. We gave upper and lower bounds for the convergence speed in terms of diameter reduction. In addition, we have explored in details the LE quartersection ($n = 4$) of triangles by studying the triangles shapes that emerge in that process. We then evidence its utility in adaptive mesh refinement specially in meshes with narrowed or skinny triangles as those appearing in Fluid Dynamic and Electromagnetism.

References

1. Carey GF (1997) Computational grids: generation, adaptation, and solution strategies. Taylor & Francis, Bristol
2. Rosenberg I, Stenger F (1975) A lower bound on the angles of triangles constructed by bisecting the longest side. *Math Comput* 29:390–395
3. Plaza A, Suárez JP, Padrón MA, Falcón S, Amieiro D (2004) Mesh quality improvement and other properties in the four-triangles longest-edge partition. *Comput Aided Geomet Des* 21(4):353–369
4. Márquez A, Moreno-González A, Plaza A, Suárez JP (2008) The 7-triangle longest-side partition of triangles and mesh quality improvement. *Finit Elem Anal Design* 44:748–758
5. Plaza A, Suárez JP, Padrón MA (2010) On the non-degeneracy property of the longest-edge trisection of triangles. *Appl Math Comput* 216(3):862–869
6. Perdomo F, Plaza A, Quevedo E, Suárez JP (2011) A lower bound on the angles of triangles constructed by LE-trisection. In: *Proceedings of XIV Spanish meeting on computational geometry*, pp 201–204
7. Kearfott B (1978) A proof of convergence and error bound for the method of bisection in R^n . *Math Comp* 32(144):1147–1153
8. Stynes M (1979) On faster convergence of the bisection method for certain triangles. *Math Comp* 33(146):717–721
9. Stynes M (1980) On faster convergence of the bisection method for all triangles. *Math Comp* 35(152):1195–1201
10. Adler A (1983) On the bisection method for triangles. *Math Comp* 40(162):571–574
11. Hannukainen A, Korotov S, Krizek M (2010) On global and local mesh refinements by a generalized conforming bisection algorithm. *J Comput Appl Math* 235(2):419–436
12. Perdomo F, Plaza A, Quevedo E, Suárez JP (2012) A mathematical proof of how fast the diameters of a triangle mesh tend to zero after repeated trisection. *Math Comput Simulat* (in review)
13. Suárez JP, Moreno T, Abad P, Plaza A (2012) Convergence speed of longest edge n -section of triangles. *Lecture notes in engineering and computer science: proceedings of world congress on engineering, WCE 2012, London, UK, 4–6 July 2012*, pp 869–873
14. Plaza A, Suárez JP, Carey GF (2007) A geometric diagram and hybrid scheme for triangle subdivision. *Comp Aided Geom Des* 24(1):19–27

Labeling the Nodes in the Intrinsic Order Graph with Their Weights

Luis González

Abstract This chapter deals with the study of some new properties of the intrinsic order graph. The intrinsic order graph is the natural graphical representation of a complex stochastic Boolean system (CSBS). A CSBS is a system depending on an arbitrarily large number n of mutually independent random Boolean variables. The intrinsic order graph displays its 2^n vertices (associated to the CSBS) from top to bottom, in decreasing order of their occurrence probabilities. New relations between the intrinsic ordering and the Hamming weight (i.e., the number of 1-bits in a binary n -tuple) are derived. Further, the distribution of the weights of the 2^n nodes in the intrinsic order graph is analyzed.

Keywords Complex stochastic Boolean systems · Hamming weight · Intrinsic order · Intrinsic order graph · Subgraphs · Subsets

1 Introduction

Consider a system depending on an arbitrary number n of random Boolean variables. That is, the n basic variables, x_1, \dots, x_n , of the system are assumed to be stochastic (non-deterministic), and they only take two possible values (either 0 or 1). We call such a system a complex stochastic Boolean system (CSBS). CSBSs often appear in many different knowledge areas, since the assumption “random Boolean variables” is satisfied very often in practice.

Each one of the possible situations (outcomes) associated to a CSBS is given by a binary n -tuple of 0s and 1s, i.e.,

$$u = (u_1, \dots, u_n) \in \{0, 1\}^n$$

L. González (✉)

Research Institute IUSIANI, Department of Mathematics, University of Las Palmas de Gran Canaria, Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain
e-mail: luisglez@dma.ulpgc.es

and, from now on, we assume that the n random Boolean variables $\{x_i\}_{i=1}^n$ are mutually independent. Hence, denoting

$$\Pr \{x_i = 1\} = p_i, \quad \Pr \{x_i = 0\} = 1 - p_i \quad (1 \leq i \leq n),$$

the occurrence probability of each binary n -tuple, $u = (u_1, \dots, u_n)$, can be computed as the product

$$\Pr \{(u_1, \dots, u_n)\} = \prod_{i=1}^n \Pr \{x_i = u_i\} = \prod_{i=1}^n p_i^{u_i} (1 - p_i)^{1-u_i}, \quad (1.1)$$

that is, $\Pr \{(u_1, \dots, u_n)\}$ is the product of factors p_i if $u_i = 1$, $1-p_i$ if $u_i = 0$. Throughout this chapter, the binary n -tuples (u_1, \dots, u_n) of 0s and 1s will be also called binary strings or bitstrings, and the probabilities p_1, \dots, p_n will be also called basic probabilities.

One of the most relevant questions in the analysis of CSBSs consists of ordering the binary strings (u_1, \dots, u_n) according to their occurrence probabilities. For this purpose, in [2] we have established a simple, positional criterion (the so-called *intrinsic order criterion*) that allows one to compare two given binary n -tuple probabilities, $\Pr \{u\}$, $\Pr \{v\}$, without computing them, simply looking at the positions of the 0s and 1s in the n -tuples u, v . The usual representation for the intrinsic order relation is the *intrinsic order graph*.

In this context, the main goal of this chapter is to state and derive some new properties of the intrinsic order graph, concerning the Hamming weights of the binary strings (i.e., the number of 1-bits in each binary n -tuple). Some of these properties can be found in [9], where the reader can also find a number of simple examples that illustrate the preliminary results presented in this chapter.

For this purpose, this chapter has been organized as follows. In Sect. 2, we present some preliminary results about the intrinsic ordering and the intrinsic order graph, in order to make the presentation self-contained. Section 3 is devoted to present new relations between the intrinsic ordering and the Hamming weight. In Sect. 4, we study the distribution of the Hamming weights of the 2^n nodes in the intrinsic order graph. Finally, conclusions are presented in Sect. 5.

2 Intrinsic Ordering in CSBSs

2.1 The Intrinsic Partial Order Relation

The following theorem [2, 3] provides us with an intrinsic order criterion—denoted from now on by the acronym IOC—to compare the occurrence probabilities of two given n -tuples of 0s & 1s without computing them.

Theorem 2.1 *Let $n \geq 1$. Let x_1, \dots, x_n be n mutually independent Boolean variables whose parameters $p_i = \Pr \{x_i = 1\}$ satisfy*

$$0 < p_1 \leq p_2 \leq \dots \leq p_n \leq \frac{1}{2}. \tag{2.1}$$

Then the probability of the n -tuple $v = (v_1, \dots, v_n) \in \{0, 1\}^n$ is intrinsically less than or equal to the probability of the n -tuple $u = (u_1, \dots, u_n) \in \{0, 1\}^n$ (that is, for all set $\{p_i\}_{i=1}^n$ satisfying (2.1)) if and only if the matrix

$$M_v^u := \begin{pmatrix} u_1 & \dots & u_n \\ v_1 & \dots & v_n \end{pmatrix}$$

either has no $\binom{1}{0}$ columns, or for each $\binom{1}{0}$ column in M_v^u there exists (at least) one corresponding preceding $\binom{0}{1}$ column (IOC).

Remark 2.2 In the following, we assume that the parameters p_i always satisfy condition (2.1). The $\binom{0}{1}$ column preceding to each $\binom{1}{0}$ column is not required to be necessarily placed at the immediately previous position, but just at previous position. The term *corresponding*, used in Theorem 2.1, has the following meaning: For each two $\binom{1}{0}$ columns in matrix M_v^u , there must exist (at least) two *different* $\binom{0}{1}$ columns preceding to each other.

The matrix condition IOC, stated by Theorem 2.1 is called the *intrinsic order criterion*, because it is independent of the basic probabilities p_i and it only depends on the relative positions of the 0s and 1s in the binary n -tuples u, v . Theorem 2.1 naturally leads to the following partial order relation on the set $\{0, 1\}^n$ [3]. The so-called intrinsic order will be denoted by “ \preceq ”, and we shall write $u \succeq v$ ($u \preceq v$) to indicate that u is intrinsically greater (less) than or equal to v . The partially ordered set (from now on, poset, for short) $(\{0, 1\}^n, \preceq)$ on n Boolean variables, will be denoted by I_n .

Definition 2.3 *For all $u, v \in \{0, 1\}^n$*

$$\begin{aligned} v \preceq u \text{ iff } & \Pr \{v\} \leq \Pr \{u\} \text{ for all set } \{p_i\}_{i=1}^n \text{ s.t. (2.1)} \\ & \text{iff } M_v^u \text{ satisfies IOC.} \end{aligned}$$

2.2 A Picture for the Intrinsic Ordering

Now, the graphical representation of the poset $I_n = (\{0, 1\}^n, \preceq)$ is presented. The usual representation of a poset is its Hasse diagram (see [12] for more details about these diagrams). Specifically, for our poset I_n , its Hasse diagram is a directed graph (digraph, for short) whose vertices are the 2^n binary n -tuples of 0s and 1s, and whose



Fig. 1 The intrinsic order graph for $n = 1$

edges go upward from v to u whenever u covers v , denoted by $u \triangleright v$. This means that u is intrinsically greater than v with no other elements between them, i.e.,

$$u \triangleright v \iff u \succ v \text{ and } \nexists w \in \{0, 1\}^n \text{ s.t. } u \succ w \succ v.$$

A simple matrix characterization of the covering relation for the intrinsic order is given in the next theorem; see [4] for the proof.

Theorem 2.4 (Covering Relation in I_n) *Let $n \geq 1$ and let $u, v \in \{0, 1\}^n$. Then $u \triangleright v$ if and only if the only columns of matrix M_v^u different from $\binom{0}{0}$ and $\binom{1}{1}$ are either its last column $\binom{0}{1}$ or just two columns, namely one $\binom{1}{0}$ column immediately preceded by one $\binom{0}{1}$ column, i.e., either*

$$M_v^u = \begin{pmatrix} u_1 & \dots & u_{n-1} & 0 \\ u_1 & \dots & u_{n-1} & 1 \end{pmatrix} \tag{2.2}$$

or there exists i ($2 \leq i \leq n$) s.t.

$$M_v^u = \begin{pmatrix} u_1 & \dots & u_{i-2} & 0 & 1 & u_{i+1} & \dots & u_n \\ u_1 & \dots & u_{i-2} & 1 & 0 & u_{i+1} & \dots & u_n \end{pmatrix}. \tag{2.3}$$

The Hasse diagram of the poset I_n will be also called the *intrinsic order graph* for n variables, denoted as well by I_n .

For small values of n , the intrinsic order graph I_n can be directly constructed by using either Theorem 2.1 (matrix description of the intrinsic order) or Theorem 2.4 (matrix description of the covering relation for the intrinsic order). For instance, for $n = 1$: $I_1 = (\{0, 1\}, \leq)$, and its Hasse diagram is shown in Fig. 1. Note that $0 \succ 1$ (Theorem 2.1).

However, for large values of n , a more efficient method is needed. For this purpose, in [4] the following algorithm for iteratively building up I_n (for all $n \geq 2$) from I_1 (depicted in Fig. 1), has been developed.

Theorem 2.5 (Building Up I_n from I_1) *Let $n \geq 2$. The graph of the poset $I_n = \{0, \dots, 2^n - 1\}$ (on 2^n nodes) can be drawn simply by adding to the graph of the poset $I_{n-1} = \{0, \dots, 2^{n-1} - 1\}$ (on 2^{n-1} nodes) its isomorphic copy $2^{n-1} + I_{n-1} = \{2^{n-1}, \dots, 2^n - 1\}$ (on 2^{n-1} nodes). This addition must be performed placing the powers of 2 at consecutive levels of the Hasse diagram of I_n . Finally, the edges connecting one vertex u of I_{n-1} with the other vertex v of $2^{n-1} + I_{n-1}$ are given by the set of 2^{n-2} vertex pairs*

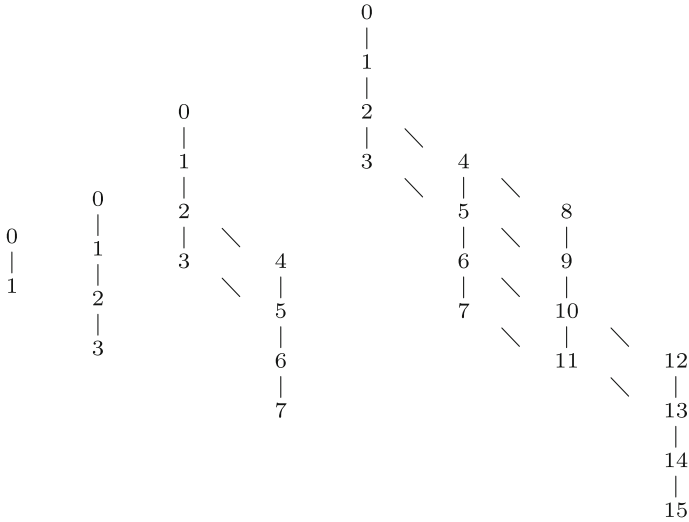


Fig. 2 The intrinsic order graphs for $n = 1, 2, 3, 4$

$$\left\{ (u, v) \equiv \left(u_{(10)}, 2^{n-2} + u_{(10)} \right) \mid 2^{n-2} \leq u_{(10)} \leq 2^{n-1} - 1 \right\}.$$

Figure 2 illustrates the above iterative process for the first few values of n , denoting all the binary n -tuples by their decimal equivalents.

Each pair (u, v) of vertices connected in I_n either by one edge or by a longer path, descending from u to v , means that u is intrinsically greater than v , i.e., $u \succ v$. On the contrary, each pair (u, v) of non-connected vertices in I_n either by one edge or by a longer descending path, means that u and v are incomparable by intrinsic order, i.e., $u \not\prec v$ and $v \not\prec u$.

The edgeless graph for a given graph is obtained by removing all its edges, keeping its nodes at the same positions [1]. In Figs. 3 and 4, the edgeless intrinsic order graphs of I_5 & I_6 , respectively, are depicted.

For further theoretical properties and practical applications of the intrinsic order and the intrinsic order graph, we refer the reader to e.g., [2–11].

3 Weights and Intrinsic Ordering

Now, we present some new relations between the intrinsic ordering and the Hamming weight. Let us denote by $w_H(u)$ the Hamming weight—or weight, simply—of u (i.e., the number of 1-bits in u), i.e.,

$$w_H(u) := \sum_{i=1}^n u_i.$$

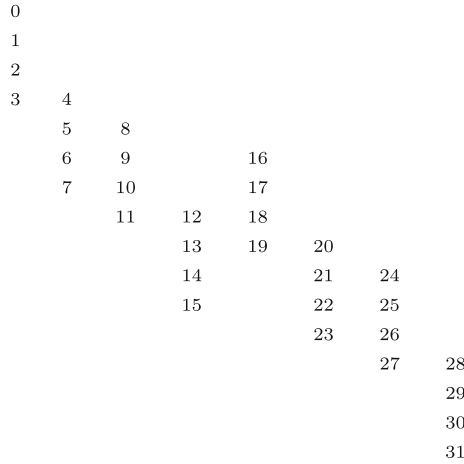


Fig. 3 The edgeless intrinsic order graph for $n = 5$

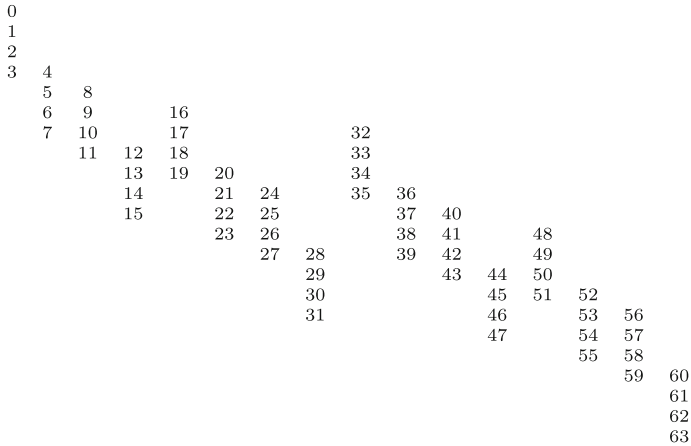


Fig. 4 The edgeless intrinsic order graph for $n = 6$

Our starting point is the following necessary (but not sufficient) condition for intrinsic order (see [3] for the proof).

$$u \succeq v \implies w_H(u) \leq w_H(v) \quad \text{for all } v \in \{0, 1\}^n. \tag{3.1}$$

However, the necessary condition for intrinsic order stated by Eq.(3.1) is not sufficient. That is,

$$w_H(u) \leq w_H(v) \not\Rightarrow u \succeq v,$$

as the following simple counter-example (indeed, the simplest one that one can find!) shows.

Example 3.1 For

$$n = 3, \quad u = 4 \equiv (1, 0, 0), \quad v = 3 \equiv (0, 1, 1),$$

we have (see the digraph of I_3 in Fig. 2)

$$w_H(4) = 1 < 2 = w_H(3).$$

However $4 \not\prec 3$, since matrix

$$M_3^4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

does not satisfy IOC.

In this context, two dual questions naturally arise. They are posed in the two subsections of this section. First, we need to set the following notations.

Definition 3.2 For every binary n -tuple u , C^u (C_u , respectively) is the set of all binary n -tuples v whose occurrence probabilities $\Pr\{v\}$ are always less (greater, respectively) than or equal to $\Pr\{u\}$, i.e., those n -tuples v intrinsically less (greater, respectively) than or equal to u , i.e.,

$$\begin{aligned} C^u &= \{v \in \{0, 1\}^n \mid \Pr\{u\} \geq \Pr\{v\}, \forall \{p_i\}_{i=1}^n \text{ s.t. (2.1)}\} \\ &= \{v \in \{0, 1\}^n \mid u \succeq v\}, \end{aligned}$$

$$\begin{aligned} C_u &= \{v \in \{0, 1\}^n \mid \Pr\{u\} \leq \Pr\{v\}, \forall \{p_i\}_{i=1}^n \text{ s.t. (2.1)}\} \\ &= \{v \in \{0, 1\}^n \mid u \preceq v\}. \end{aligned}$$

Definition 3.3 For every binary n -tuple u , H^u (H_u , respectively) is the set of all binary n -tuples v whose Hamming weights are less (greater, respectively) than or equal to the Hamming weight of u , i.e.,

$$\begin{aligned} H^u &= \{v \in \{0, 1\}^n \mid w_H(u) \geq w_H(v)\}, \\ H_u &= \{v \in \{0, 1\}^n \mid w_H(u) \leq w_H(v)\}. \end{aligned}$$

3.1 When Greater Weight Corresponds to Less Probability

Looking at the implication (3.1), the following question immediately arises.

Question 3.1: We try to characterize the binary n -tuples u for which the necessary condition (3.1) is also sufficient, i.e.,

$$u \succeq v \Leftrightarrow w_H(u) \leq w_H(v), \text{ i.e., } C^u = H_u.$$

The following theorem provides the answer to this question, in a very simple way.

Theorem 3.4 *Let $n \geq 1$ and $u = (u_1, \dots, u_n) \in \{0, 1\}^n$ with Hamming weight $w_H(u) = m$ ($0 \leq m \leq n$). Then*

$$C^u = H_u$$

if and only if either u is the zero n -tuple ($m = 0$) or the m 1-bits of u ($m > 0$) are placed at the m right-most positions, i.e., if and only if u has the general pattern

$$u = \left(\underbrace{0, \dots, 0}_{n-m}, \underbrace{1, \dots, 1}_m \right) \equiv 2^m - 1, \quad 0 \leq m \leq n, \quad (3.2)$$

where any (but not both!) of the above two subsets of bits grouped together can be omitted.

Proof.

Sufficient condition. We distinguish two cases:

- (i) If u is the zero n -tuple $0 \equiv \left(\underbrace{0, \dots, 0}_n \right)$, then u is the maximum element for the intrinsic order (see, e.g., [9]). Then

$$\begin{aligned} C^0 &= \{v \in \{0, 1\}^n \mid 0 \succeq v\} = \{0, 1\}^n \\ &= \{v \in \{0, 1\}^n \mid w_H(0) = 0 \leq w_H(v)\} = H_0. \end{aligned}$$

- (ii) If u is not the zero n -tuple, then u has the pattern (3.2) with $m > 0$. Let $v \in H_u$, i.e., let v let a binary n -tuple with Hamming weight greater than or equal to m (the Hamming weight of u). We distinguish two subcases:
 - (a) Suppose that the weight of v is

$$w_H(v) = m = w_H(u).$$

Then v has exactly m 1-bits and $n - m$ 0-bits. Call r the number of 1-bits of v placed among the m right-most positions ($\max\{0, 2m - n\} \leq r \leq m$). Obviously, v has r 1-bits and $m - r$ 0-bits placed among the m right-most positions, and also it has $m - r$ 1-bits and $n - 2m + r$ 0-bits placed among the $n - m$ left-most positions. These are the positions of the

$$r + (m - r) + (m - r) + (n - 2m + r) = m + (n - m) = n$$

bits of the binary n -tuple v .

Hence, matrix M_v^u has exactly $m - r$ $\binom{1}{0}$ columns (all placed among the m right-most positions) and exactly $m - r$ $\binom{0}{1}$ columns (all placed among the $n - m$ left-most positions). Thus, M_v^u satisfies IOC and then $u \succeq v$, i.e., $v \in C^u$.

So, for this case (a), we have proved that

$$\{v \in \{0, 1\}^n \mid w_H(v) = w_H(u) = m\} \subseteq C^u \quad (3.3)$$

(b) Suppose that the weight of v is

$$w_H(v) = m + p > m = w_H(u) \quad (0 < p \leq n - m).$$

Then define a new binary n -tuple s as follows. First, select any p 1-bits in v (say, for instance, $v_{i_1} = \dots = v_{i_p} = 1$). Second, s is constructed by changing these p 1-bits of v into 0-bits, assigning to the remainder $n - p$ bits of s the same values as the ones of v . Formally, $s = (s_1, \dots, s_n)$ is defined by

$$s_i = \begin{cases} 0 & \text{if } i \in \{i_1, \dots, i_p\}, \\ v_i & \text{if } i \notin \{i_1, \dots, i_p\}. \end{cases}$$

On one hand, $u \succeq s$ since

$$w_H(s) = w_H(v) - p = m = w_H(u)$$

and then we can apply case (a) to s .

On the other hand, $s \succeq v$ since matrix M_v^s has p $\binom{0}{1}$ columns (placed at positions i_1, \dots, i_p), while its $n - p$ remainder columns are either $\binom{0}{0}$ or $\binom{1}{1}$. Hence M_v^s has no $\binom{1}{0}$ columns, so that it satisfies IOC.

Finally, from the transitive property of the intrinsic order, we derive

$$u \succeq s \text{ and } s \succeq v \Rightarrow u \succeq v, \text{ i.e., } v \in C^u.$$

So, for this case (b), we have proved that

$$\{v \in \{0, 1\}^n \mid w_H(v) > w_H(u) = m\} \subseteq C^u \quad (3.4)$$

From (3.3) and (3.4), we get

$$\{v \in \{0, 1\}^n \mid w_H(v) \geq w_H(u) = m\} \subseteq C^u,$$

i.e., $H_u \subseteq C^u$, and this set inclusion together with the converse inclusion $C^u \subseteq H_u$ (which is always satisfied for every binary n -tuple u ; see Eq. 3.1) leads to the set equality $C^u = H_u$. This proves the sufficient condition.

Necessary condition. Conversely, suppose that not all the m 1-bits of u are placed at the m right-most positions. In other words, suppose that

$$u \neq \left(0, \overbrace{\dots}^{n-m}, 0, 1, \overbrace{\dots}^m, 1 \right).$$

Since, by assumption, $w_H(u) = m$ then simply using the necessary condition we derive that

$$\left(0, \overbrace{\dots}^{n-m}, 0, 1, \overbrace{\dots}^m, 1 \right) \succ u,$$

and then

$$\left(0, \overbrace{\dots}^{n-m}, 0, 1, \overbrace{\dots}^m, 1 \right) \in H_u - C^u$$

so that,

$$H_u \not\subseteq C^u.$$

This proves the necessary condition. □

Corollary 3.5 *Let $n \geq 1$ and let*

$$u = \left(0, \overbrace{\dots}^{n-m}, 0, 1, \overbrace{\dots}^m, 1 \right) \equiv 2^m - 1, \quad 0 \leq m \leq n,$$

where any (but not both!) of the above two subsets of bits grouped together can be omitted. Then the number of binary n -tuples intrinsically less than or equal to u is

$$|C^u| = \binom{n}{m} + \binom{n}{m+1} + \dots + \binom{n}{n}.$$

Proof. Using Theorem 3.4, the proof is straightforward. □

3.2 When Less Weight Corresponds to Greater Probability

Interchanging the roles of u & v , (3.1) can be rewritten as follows. Let u be an arbitrary, but fixed, binary n -tuple. Then

$$v \geq u \Rightarrow w_H(v) \leq w_H(u) \quad \text{for all } v \in \{0, 1\}^n. \tag{3.5}$$

Looking at the implication (3.5), the following dual question of Question 3.1, immediately arises.

Question 3.2: We try to characterize the binary n -tuples u for which the necessary condition (3.5) is also sufficient, i.e.,

$$v \succeq u \Leftrightarrow w_H(v) \leq w_H(u), \text{ i.e., } C_u = H^u.$$

The following theorem provides the answer to this question, in a very simple way. For a very short proof of this theorem, we use the following definition.

Definition 3.6 (i) *The complementary n -tuple of a given binary n -tuple $u \in \{0, 1\}^n$ is obtained by changing its 0s into 1s and its 1s into 0s*

$$u^c = (u_1, \dots, u_n)^c = (1 - u_1, \dots, 1 - u_n).$$

Obviously, two binary n -tuples are complementary if and only if their decimal equivalents sum up to

$$\left(1, \overset{n}{\dots}, 1\right)_{(10)} = 2^n - 1.$$

(ii) *The complementary set of a given subset $S \subseteq \{0, 1\}^n$ of binary n -tuples is the set of the complementary n -tuples of all the n -tuples of S*

$$S^c = \{u^c \mid u \in S\}.$$

Theorem 3.7 *Let $n \geq 1$ and $u = (u_1, \dots, u_n) \in \{0, 1\}^n$ with Hamming weight $w_H(u) = m$ ($0 \leq m \leq n$). Then*

$$C_u = H^u$$

if and only if either u is the zero n -tuple ($m = 0$) or the m 1-bits of u ($m > 0$) are placed at the m left-most positions, i.e., if and only if u has the general pattern

$$u = \left(1, \overset{m}{\dots}, 1, 0, \overset{n-m}{\dots}, 0\right) \equiv 2^m - 2^{n-m}, \quad 0 \leq m \leq n, \quad (3.6)$$

where any (but not both!) of the above two subsets of bits grouped together can be omitted.

Proof. Using Theorem 3.4 and the facts that (see, e.g., [5, 7])

$$(C_u)^c = C^{u^c}, \quad (H^u)^c = H_{u^c},$$

we get

$$\begin{aligned} C_u = H^u &\Leftrightarrow (C_u)^c = (H^u)^c \Leftrightarrow C^{u^c} = H_{u^c} \\ &\Leftrightarrow u^c \text{ has the pattern (3.2)} \Leftrightarrow u \text{ has the pattern (3.6),} \end{aligned}$$

as was to be shown. □

Corollary 3.8 *Let $n \geq 1$ and let*

$$u = \left(1, \overset{m}{\dots}, 1, 0, \overset{n-m}{\dots}, 0 \right) \equiv 2^n - 2^{n-m}, \quad 0 \leq m \leq n,$$

where any (but not both!) of the above two subsets of bits grouped together can be omitted. Then the number of binary n -tuples intrinsically greater than or equal to u is

$$|C_u| = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{m}.$$

Proof. Using the fact that $(C_u)^c = C^{u^c}$ and Corollary 3.5, the proof is straightforward. □

4 Nodes and Weights in the Intrinsic Order Graph

The results derived in Sect. 3, and more precisely those stated by Theorems 3.4 and 3.7, can be illustrated by labeling the nodes of the intrinsic order graph with their respective Hamming weights. In this way, due to Theorem 3.4 (Theorem 3.7, respectively), for a given binary n -tuple u with weight m whose m 1-bits are all placed among the right-most (left-most, respectively) positions, the set of nodes v with Hamming weight greater (less, respectively) than or equal to m will be exactly the set of nodes v connected to vertex u by a descending (ascending, respectively) path from u to v .

This suggests the analysis of the distribution of the Hamming weights of the 2^n nodes in the intrinsic order graph. The following Theorem provides only some basic consequences of such analysis.

Theorem 4.1 *Let $n \geq 2$. Label each of the 2^n nodes in the intrinsic order graph I_n , with its corresponding Hamming weight. Then*

- (i) *The weights (labels) of the 2^n nodes are (with repetitions): $0, 1, \dots, n$.*
- (ii) *The weights (labels) of the 4 nodes in each of the saturated chains $4k \triangleright 4k + 1 \triangleright 4k + 2 \triangleright 4k + 3$ are: $w_H(k), w_H(k) + 1, w_H(k) + 1, w_H(k) + 2$.*
- (iii) *The set of weights (labels) of the 2^n nodes of the graph $I_n = \{0, 1\}^n$ can be partitioned into the following two subsets: (a) *The weights of the nodes of the top subgraph $\{0\} \times \{0, 1\}^{n-1}$ of I_n , which one-to-one coincide with the respective weights of the nodes of the graph $I_{n-1} = \{0, 1\}^{n-1}$.* (b) *The weights of the nodes of the bottom subgraph $\{1\} \times \{0, 1\}^{n-1}$ of I_n , which one-to-one coincide with 1 plus the respective weights of the nodes of the graph $I_{n-1} = \{0, 1\}^{n-1}$.**

Fig. 5 Weights in the edgeless intrinsic order graph for $n = 5$

0									
1									
1									
2	1								
		2	1						
		2	2		1				
		3	2		2				
			3	2	2				
				3	3	2			
				3		3	2		
				4		3	3		
						4	3		
							4	3	
								4	3
									4
									4
									5

Proof.

- (i) Trivial.
- (ii) Use the fact that for all $k \equiv (u_1 \dots, u_{n-2}) \in \{0, 1\}^{n-2}$:

$$\begin{aligned}
 4k &\equiv (u_1 \dots, u_{n-2}, 0, 0), & 4k + 1 &\equiv (u_1 \dots, u_{n-2}, 0, 1), \\
 4k + 2 &\equiv (u_1 \dots, u_{n-2}, 1, 0), & 4k + 3 &\equiv (u_1 \dots, u_{n-2}, 1, 1).
 \end{aligned}$$

- (iii) Use Theorem 2.5 and the fact that for all $(u_1 \dots, u_{n-1}) \in \{0, 1\}^{n-1}$:

$$\begin{aligned}
 w_H(0, u_1 \dots, u_{n-1}) &= w_H(u_1 \dots, u_{n-1}) \\
 w_H(1, u_1 \dots, u_{n-1}) &= w_H(u_1 \dots, u_{n-1}) + 1.
 \end{aligned}$$

as was to be shown. □

Figure 5 illustrates Theorem 4.1, by labeling (and substituting) all the 32 nodes of the graph I_5 (depicted in Fig. 3) with their corresponding Hamming weights.

5 Conclusions

It is well-known that if a binary n -tuple u is intrinsically greater (less, respectively) than or equal to a binary n -tuple v then necessarily the Hamming weight of u must be less (greater, respectively) than or equal to the Hamming weight of v . We have characterized, by two dual, simple positional criteria, those n -tuples u for which each

of these necessary conditions is also sufficient. Further, motivated by these questions, we have presented some basic properties concerning the distribution of weights of the 2^n nodes in the intrinsic order graph. For future researches, additional properties of such distribution worth to be studied.

Acknowledgments This work was supported in part by the “Ministerio de Economía y Competitividad” (Spanish Government), and FEDER, through Grant contract: CGL2011-29396-C03-01.

References

1. Diestel R (2005) Graph theory, 3rd edn. Springer, New York
2. González L (2002) A new method for ordering binary states probabilities in reliability and risk analysis. *Lect Notes Comput Sci* 2329:137–146
3. González L (2003) N -tuples of 0s and 1s: necessary and sufficient conditions for intrinsic order. *Lect Notes Comput Sci* 2667:937–946
4. González L (2006) A picture for complex stochastic Boolean systems: the intrinsic order graph. *Lect Notes Comput Sci* 3993:305–312
5. González L (2007) Algorithm comparing binary string probabilities in complex stochastic Boolean systems using intrinsic order graph. *Adv Complex Syst* 10(Suppl 1):111–143
6. González L (2010) Ranking intervals in complex stochastic Boolean systems using intrinsic ordering. In: Rieger BB, Amouzegar MA, Ao S-I (eds) Machine learning and systems engineering. Lecture notes in electrical engineering, vol 68. Springer, New York, pp 397–410
7. González L (2012) Duality in complex stochastic Boolean systems. In: Ao S-I, Gelman L (eds) Electrical engineering and intelligent systems. Lecture notes in electrical engineering, vol 130. Springer, New York, pp 15–27
8. González L (2012) Edges, chains, shadows, neighbors and subgraphs in the intrinsic order graph. *IAENG Int J Appl Math* 42:66–73
9. González L (2012) Intrinsic order and Hamming weight. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, U.K., pp 783–788, 4–6 July 2012
10. González L (2012) Intrinsic ordering, combinatorial numbers and reliability engineering. *Appl Math Model* (in press)
11. González L, García D, Galván B (2004) An intrinsic order criterion to evaluate large, complex fault trees. *IEEE Trans Reliab* 53:297–305
12. Stanley RP (1997) Enumerative combinatorics, vol 1. Cambridge University Press, Cambridge

Solving VoIP QoS and Scalability Issues in Backbone Networks

Martin Hruby, Michal Olsovsky and Margareta Kotocova

Abstract Providing quality of service should be one of the main objectives when deploying sensitive applications into the network. Since network performance parameters are subject to frequent change, in this chapter we propose a novel approach to routing sensitive VoIP traffic in large networks. Our approach takes measured delay and jitter into consideration and we establish an overlay of the original network to route primarily VoIP traffic. This is achieved by first modeling the probability distributions of network performance parameters and then by calculating the best paths by means of graph algorithm utilizing aspects. Our approach also identifies weak network areas not suitable for VoIP deployment which can be subject to future network improvements.

Keywords Active measurement · Algorithm · Delay · Jitter · Network · QoS · Routing · VoIP

1 Introduction

This chapter is focused on sub-optimal traffic distribution in heterogeneous computer networks as this is the main issue we are facing when delivering bandwidth intensive services with QoS guarantees. Usually the decision on which paths to use to forward

M. Hruby (✉) · M. Olsovsky · M. Kotocova
Faculty of Informatics and Information Technologies, Institute of Computer
Systems and Networks, Slovak University of Technology, Ilkovicova 3,
842 16 Bratislava 4, Slovakia
e-mail: hruby@fiit.stuba.sk

M. Olsovsky
e-mail: olsovsky@fiit.stuba.sk

M. Kotocova
e-mail: kotocova@fiit.stuba.sk

traffic is left on the interior routing protocol [1, 2] and the process of adding IP telephony functionality to an existing network environment is often cumbersome [3–6].

Our objective is to optimize the flow of delay sensitive traffic by creating a VoIP backbone overlay in the network in a computationally efficient and effective manner. Two parameters which mostly impact the quality of VoIP traffic are delay and jitter (delay variance) [7, 8]. In our approach, we aim to identify network links which provide minimum delay and jitter by utilizing network probes as a method of active NPP measurements in production environments. The task of deploying active network measurement probes is non-trivial in large-scale networks [9, 10]. The number of logical links and load imposed on measurement endpoints (performing measurements and providing results) is a limiting factor when it comes to deciding which links to monitor. The value of measurement information however is considerable [11, 12]. While gathering results continuously, a multi-dimensional data structure is built wherein maxima can be found which represent optimal links for VoIP traffic and these are marked as VoIP backbone components. Further, these VoIP backbone components can be joined by well-known graph algorithms [13–15]. Having a VoIP backbone clearly identified in the topology, design considerations apply which can strongly improve the quality of service for VoIP traffic by placing crucial components of the IP telephony architecture (e.g. call managers, gatekeepers, Unity servers, PSTN gateways, border elements, etc.) on the VoIP backbone [16, 17].

2 Concept

We assume a transit IP network (e.g. service provider backbone) which consists of large number of routers and is using an interior gateway routing protocol (e.g. OSPF). Generally, VoIP flows will traverse the best path as chosen by the routing protocol together with ordinary data flows; however this path doesn't have to provide the best time-variables like delay and jitter which are necessary for VoIP traffic, as in most cases these parameters are not taken into consideration by the routing protocol [18].

Using standard routing voice-related devices can be placed arbitrarily as there is no specific VoIP backbone [19]. In our approach standard routing will be divided into two independent routing instances, one will be used for data flows (existing one) and the second one will be used inside the determined optimal VoIP backbone based on our proposed network modeling.

As the VoIP backbone has to be a continuous network [20], it will consist not only of the best links chosen by our algorithm but of an additional minimal amount of links which will make the backbone continuous. List of these additional links can be considered for a potential future upgrade.

Knowing the new VoIP backbone we are able to decide where to place and connect specific voice-related devices in the real network environment.

To be able to decide which links will form the VoIP backbone, we need to use some key elements like the proposed n-cube model, active variable discovery and

Multivariate normal distribution. These techniques are described in more detail in the following chapters.

Our objective is to optimize the flow of delay sensitive traffic by creating a VoIP backbone overlay in the network in a computationally efficient and effective manner. Two parameters which mostly impact the quality of VoIP traffic are delay and jitter (delay variance) [7, 8]. In our approach, we aim to identify network links which provide minimum delay and jitter by utilizing network probes as a method of active NPP measurements in production environments. The task of deploying active network measurement probes is non-trivial in large-scale networks [9, 10]. The number of logical links and load imposed on measurement endpoints (performing measurements and providing results) is a limiting factor when it comes to deciding which links to monitor. The value of measurement information however is considerable [11, 12]. While gathering results continuously, a multi-dimensional data structure is built.

2.1 Variable Discovery

Key component to time-variables' measurement is a periodic active measurement probe (SAA). This probe is small datagram sent over specific link towards a known destination. Sender of this probe knows the exact time when the probe was sent out as well as waiting period between consecutive probes. Once the probe reaches its destination, the probe is sent back towards to the original sender. As each probe contains unique identifier, original sender is able to calculate time-variables once the probe is received. Using the probes a network node is able to measure delay and jitter for a specific link in the network.

Every node in the network is capable to reply to delay-based probes; however every node is not suitable for jitter-measurements. For better measurements results we are using IP SLA responder which can be used for all suitable measurements and even subtract the processing delay which occurs in the node when the probe is being processed.

Final phase of the variable discovery is dedicated to the data collection. To facilitate measurement data exports, SNMP is used. Once the SNMP server collects a representative amount of the measurement results, these results can be released for further processing (n-cube model, minimal tree in the graph).

2.2 Multivariate Normal Distribution

Measurement results are only group of results. To be able to make decision about their properties, it is necessary to plot these results in a specific model. For our purposes we use a hypercube. As it is n -dimensional model, we are able to correlate n parameters. In our example here we will correlate 3 parameters (therefore a 3 dimensional cube will be created) where each of the 3 axes represents a separate attribute:

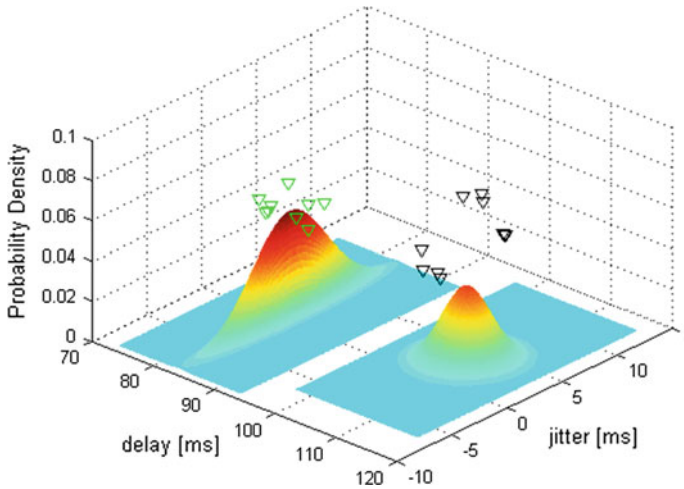


Fig. 1 Model with 2 links

- Axis X—variable jitter
- Axis Y—probability density
- Axis Z—variable delay.

Axes X and Z will represent the base of this distribution while axis Y will represent the probability density of delay-jitter pairs. Our objective is to identify specific areas of the plotted surface inside the n-cube model, which are characterized as peaks of a multivariate normal distribution that represent edges with discovered values of delay and jitter, as depicted in the Fig. 1. We model the multivariate normal distribution peaks by determining parameters of the probability density function of the d-dimensional multivariate normal distribution (see Formula 1) where the parameter μ is a 1-by-d vector of mean values of vectors of gathered variables (delays and jitters) and σ is determined as a covariance matrix of the vectors of gathered variables, as can be seen in Formula 2.

$$y = f(x, \mu, \sigma) = \frac{1}{\sqrt{|\sigma|} (2\pi)^d} e^{-\frac{1}{2}(x-\mu)\sigma^{-1}(x-\mu)'} \tag{1}$$

$$\text{cov}(x_1, x_2) = E[(x_1 - \mu_1) * (x_2 - \mu_2)] \tag{2}$$

In Formula 2, $E[x]$ is the expected value of x . Representation of the network links as edges with directly measured parameters in the n-cube model has several advantages.

First, the trustworthiness of a link can be directly determined by the σ parameter as depicted in Fig. 2. The σ parameter corresponds to the steepness of the distribution.

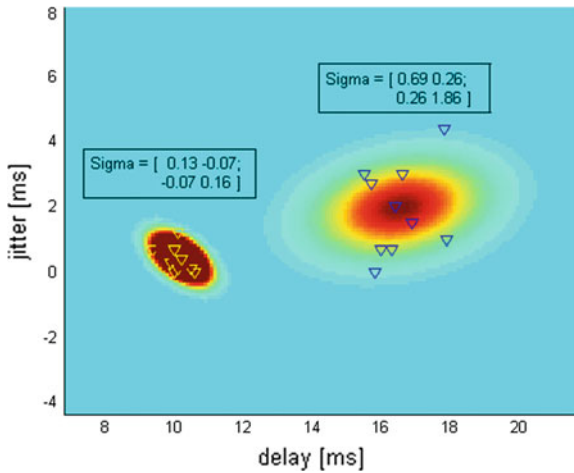


Fig. 2 Trustworthiness of a link in the n-cube model

Those edges with steeper modeled distributions experience low variation in delay and jitter as opposed to those with gradually distributed probability function.

3 Algorithm

Routing of VoIP traffic is operation based on many sub operations. First group of operations is responsible for gathering data via measurements and further processing. Later on, using the multivariate normal distribution new representative results are calculated for every single link and plot into one common model. Finally, we can apply specific rules to determine voice-capable links. However the deployment of this backbone into production network is conditioned with turning these links into continuous graph.

As this algorithm is still under development, it's necessary to provide functional and performance validation. These validations are described in [21] in the chapter dedicated to preliminary results.

3.1 Voice-Capable Links

The main idea of seeking voice-capable links is to take the cube with multivariate normal distributions of all links and do a three phase filtering. First phase will filter out all unstable links as stable links are important for VoIP backbone. We are able to apply some mechanisms to reduce the effect of worse time-variable in case we can

predict these variables. From the point of the n-cube model and multivariate normal distribution stable links are represented with small base, narrow distribution and high peaks. Small base means that measured values don't differ in significant way and high peak represents the probability density. To be able to find these stable links we have introduced new variable η_1 —threshold of probability density for voice-capable links (Formula 3 and 4). Based on this threshold, new layer (parallel to area of axes X-Z) is added to the final n-cube model. Every link whose peak exceeds this layer (threshold) has matched the first of three conditions to become voice-capable link.

$$\eta_{\max} = f(x', \mu', \sigma') \geq f(x, \mu, \sigma) \quad (3)$$

$$\eta_1 = 0.8 * \eta_{\max} \quad (4)$$

$$\eta_2 = 0.56 * \eta_{\max} \quad (5)$$

Last two conditions are related to time-variables. Matching just the first conditions isn't enough because this condition will guarantee only the stability of the link—it doesn't matter how high the delay or jitter is, once the values aren't varying excessively, the link is stable. To be able to filter out links with worse time-variables we have introduced another 2 variables (thresholds).

First variable φ_1 is jitter threshold (Formula 7). We don't keep exact jitter, this value represents the fraction of the worst measured result (Formula 6). Based on this threshold another layer (parallel to area of axes X-Z) is added to the final n-cube model. Every link whose multivariate normal distribution belongs to the area bordered with the new layer and the area of axes X-Z has matched the second condition to become voice-capable link.

$$\varphi_{\max} \in \Phi, \forall \varphi' \in \Phi; \varphi' \leq \varphi_{\max} \quad (6)$$

$$\varphi_1 = 0.8 * \varphi_{\max} \quad (7)$$

$$\varphi_2 = 0.5 * \varphi_{\max} \quad (8)$$

Second variable called χ_1 represents the delay. As for φ_1 , we don't keep exact delay value (Formula 10), this value represents the fraction of the worst measured result (Formula 9). Based on this threshold another layer (parallel to area of axes X-Y) is added to the final n-cube model. Every link who's Gaussian distribution belongs to the area bordered with the new layer and the area of axes X-Y has matched the last, third condition to become voice-capable link.

$$\chi_{\max} \in X, \forall \chi' \in X; \chi' \leq \chi_{\max} \quad (9)$$

$$\chi_1 = 0.8 * \chi_{\max} \quad (10)$$

$$\chi_2 = 0.5 * \chi_{\max} \quad (11)$$

To sum it up all three thresholds η_1 , φ_1 and χ_1 and appropriate layers created sub-cube model in the main n-cube model and all links with their peaks in this sub-cube model are voice-capable links.

With thresholds calculated it is now possible to separate measured data in the n-cube model into regions. Each region Ω is defined by upper and lower planes and links are considered to be part of a region when the link's probability function maximum is located between the regions planes (Formulas 12, 13, 14).

$$\Omega_1 = \begin{bmatrix} \varphi_1 & \chi_1 & \eta_{\max} \\ 0 & 0 & \eta_1 \end{bmatrix} \quad (12)$$

$$\Omega_2 = \begin{bmatrix} \varphi_2 & \chi_2 & \eta_1 \\ \varphi_1 & \chi_1 & \eta_2 \end{bmatrix} \quad (13)$$

$$\Omega_3 = \begin{bmatrix} \varphi_{\max} & \chi_{\max} & \eta_2 \\ \varphi_2 & \chi_2 & 0 \end{bmatrix} \quad (14)$$

A continuous sequence of links will be called a VoIP backbone component. To calculate a single VoIP backbone, all unique VoIP backbone components must be connected by a minimal series of other links (edges). To be able to make the connection of all VoIP backbone components into one single component with the best of remaining links, these links needs to be classified. As this connection process will be based on the theory of graphs, we need to assign each link one composite metric—weight of the edge.

We have decided to define another set of thresholds— η_2 , φ_2 and χ_2 which will be worse than the original set of threshold (Formula 5, 8 and 11). These thresholds will divide the n-cube model into 3 regions (Fig. 3). Green region contains voice-capable links (region Ω_1). For future processing, these links have metric of 1. Purple region (Ω_2) contains link which aren't suitable VoIP backbone but can be used for voice flows. All these links have metric of 2. Last region (Ω_3), pink region represents the rest of the links which aren't suitable for VoIP traffic and should be used only as a temporary solution.

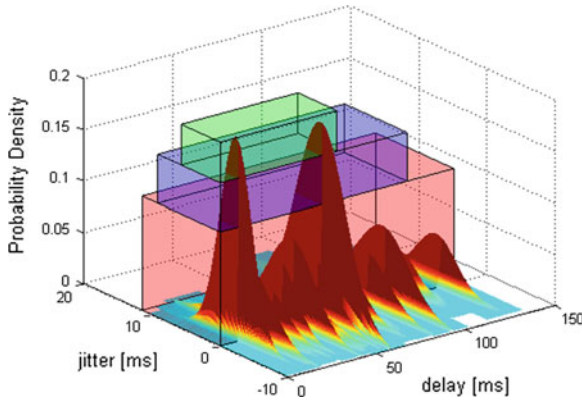


Fig. 3 Link classification for VoIP backbone

3.2 Minimal Tree

The process of fetching the minimal tree is based on graph theory. The graph as defined in [21] is represented as a set of edges stored in variable *EDGES* and, a set of vertices stored in variable *VERTICES*. Each edge stored in variable *EDGES* has the following properties:

- *Metric*—represents the weight of the edge (1, 2, 5)
- *Visited*—holds a Boolean value (*TRUE*, *FALSE*) to identify links which were already processed
- *Tree*—holds a Boolean value (*TRUE*, *FALSE*) to identify links which were chosen into the final tree.

Each vertex stored in variable *VERTICES* has the following properties:

- *Border*—holds a Boolean value (*TRUE*, *FALSE*) to identify whether this vertex represents a border router
- *Edges*—contains a list of adjacent edges.

We will illustrate our algorithm in steps on an example graph depicted in Fig. 4. Vertices are numbered and are connected by edges with associated weights. Vertex 1 has the *Border* parameter set to true (it must be included in the resulting tree because it is a border router).

1. Select all voice-capable links (edges with *Metric* = 1) and create continuous subgraphs by traversing neighboring edges and setting their *Visited* parameter to *TRUE*. Each edge added to a continuous subgraph has its *Tree* property set to *TRUE* and will hence be included in the final tree. All created subgraphs will be stored in variable *SUBGR* which contains a finite list of graphs represented by edges and vertices, see Fig. 5.

Fig. 4 Example graph with weighted edges

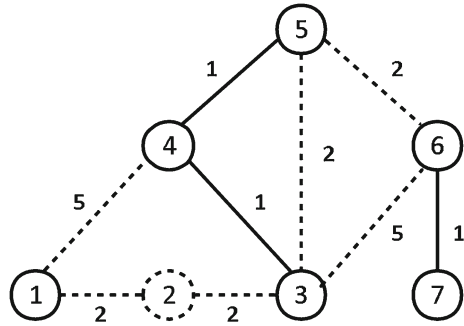


Fig. 5 Example list of graphs in SUBGR

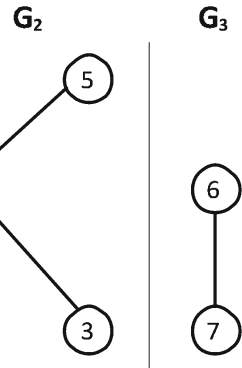
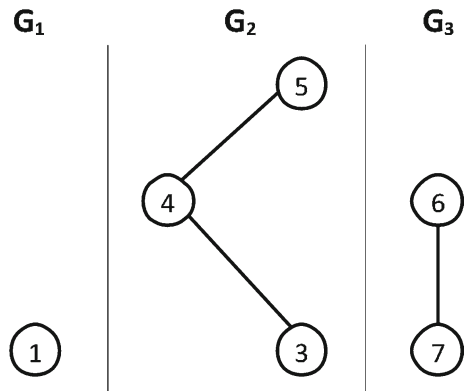


Fig. 6 Example list of graphs in SUBGR with added border vertex



2. Check if vertices which have the *Border* parameter set to *TRUE* are present in *SUBGR*. In case a vertex with *Border* set to *TRUE* is not present in *SUBGR*, create new subgraph in *SUBGR*. This new subgraph will contain only 1 vertex, see Fig. 6.

Fig. 7 Example graph with weighted edges

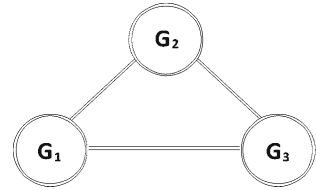
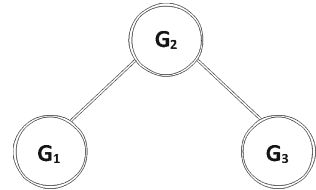


Fig. 8 Example graph with weighted edges



3. Use Dijkstra's algorithm to find the best (cheapest) path between every pair of vertices in the main graph [13, 14]. Store information for each vertex in separate variable `MAIN_GRAPH_SPF`.
4. Find the best path between every pair of subgraphs in `SUBGR`. As the graph is undirected, path between subgraph `G1` and `G2` is the same as path between subgraph `G2` and `G1`. Take first pair of subgraphs (`G1`, `G2`) and first vertex (1, 3) from each subgraph. Check the cost from 1 to 3 using tables calculated in step 3. Put all these values in to the variable `GR`, which will contain best paths between each pair of subgraphs. Figure 7 illustrates the scenario.
5. Complete list of paths is representation of another full-mesh graph. To get the minimum spanning tree for this connected weighted graph use the Kruskal's algorithm [15]. Mark the edges of the minimum spanning tree in the table of paths with `Tree = 1`. The resulting graph is depicted in Fig. 8.

We now define our algorithm in pseudo-code to better illustrate the process of selecting voice-capable links (those with *Metric* = 1) and creating a set of continuous subgraphs. These subgraphs are areas of the network which are best suited for routing voice traffic. Next we add border routers as autonomous subgraphs and finally compute shortest paths between each pair of subgraph, thus interconnecting

the areas. Unneeded edges are removed by utilizing the Kruskal algorithm. The below pseudo-code summarizes our algorithm.

```

void function fetch_subgraphs (VERTEX) {
    if VERTEX has no EDGE in VERTEX.EDGES with Metric == 1 return;
    if VERTEX has no EDGE in VERTEX.EDGES with EDGE.VISITED == false return;
    SUBGR[SUBGR_ID] += VERTEX;
    choose any EDGE from VERTEX.EDGES where Metric == 1 and EDGE.VISITED ==
false;
    EDGE.VISITED = true;
    EDGE.Tree = true;
    VERTEX_2 = OTHERSIDE(EDGE);
    fetch_subgraph (VERTEX_2);
}

void function fetch_tree (EDGES, VERTICES) {
    foreach VERTEX in VERTICES {
        if VERTEX.BORDER == true {
            SUBGR[SUBGR_ID] += VERTEX;
        } else fetch_subgraph(VERTEX);
        SUBGR_ID++;
    }
}

MAIN_GRAPH_SPF = Dijkstra(VERTICES);

foreach VERTEX_A from SUBGR[x] {
    foreach VERTEX_B from SUBGR[y] {
        GR = find min MAIN_GRAPH_SPF (VERTEX_A, VERTEX_B);
        y++;
    }
    x++;
}

RESULT_GR = Kruskal(GR);
}

```

4 Conclusion

In this chapter we proposed a novel method of routing VoIP traffic in large networks. By performing periodic measurements of network performance parameters (delay and jitter) we model the probability distributions of these parameters for each link and use the model as input for a path selection algorithm based on a modification of well known Dijkstra's and Kruskal's algorithms. In this way we establish a VoIP backbone which is a network overlay suitable for VoIP traffic deployment.

In [21] we list preliminary functional results proving the correctness of our approach and also performance results conducted on randomly generated networks with random statistically generated network performance parameters. Our modeling approach and path selection algorithm performed well, choosing an average of 10 % of all available network links into the VoIP backbone. We believe that our approach presents a useful method for VoIP implementation in planning and production phases of large network deployment.

Acknowledgments The support by Slovak Science Grant Agency (VEGA1/0676/12 “Network architectures for multimedia services delivery with QoS guarantee”) is gratefully acknowledged.

References

1. Gunnar A, Johansson A, Telkamp T (2004) Traffic matrix estimation on a large IP backbone: a comparison on real data. In: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement (IMC '04), pp 149–160, 25–27 Oct 2004
2. Lan K-C, Wu T-H (2011) Evaluating the perceived quality of infrastructure-less VoIP. In: IEEE international conference on multimedia and Expo (ICME), 2011, pp 1–6, 11–15 July 2011
3. Fortz B, Rexford J, Thorup M (2002) Traffic engineering with traditional IP routing protocols. In: IEEE communications magazine, vol 40, pp 118–124, Dec 2002
4. Wang X, Wan S, Li L (2009) Robust traffic engineering using multi-topology routing. In: Global telecommunications conference, 2009, pp 1–6
5. Hansen TJ, Morup M, Hansen LK (2011) Non-parametric co-clustering of large scale sparse bipartite networks on the GPU. In: IEEE international workshop on machine learning for signal processing (MLSP), 2011, pp 1–6, 18–21 Sept 2011
6. Wang H (2010) From a mess to graphic maps: visualization of large-scale heterogeneous networks. In: Second international conference on, computer modeling and simulation, 2010, ICCMS '10, vol 1, pp 531–535, 22–24 Jan 2010
7. Abrahamsson H, Bjorkman M (2009) Robust traffic engineering using l-balanced weight-settings in OSPF/IS-IS. In: Broadband communications, networks, and systems (BROAD-NETS), 2009, pp 1–8
8. Son H, Lee Y (2010) Detecting anomaly traffic using flow data in the real VoIP network. In: 10th IEEE/IPSJ international symposium on applications and the internet (SAINT), 2010, pp 253–256, 19–23 July 2010
9. Tebaldi C, West M (1998) Bayesian inference on network traffic using link count data. *J Am Stat Assoc* 93(442):557–576. ISSN 0162–1459
10. Cao J, Davis D, Vander Wiel S, Yu B (2000) Time-varying network tomography: router link data. *J Am Stat* 95:1063–1075
11. Goldschmidt O (2000) ISP backbone traffic inference methods to support traffic engineering. In: Internet statistics and metrics analysis, workshop (ISMA'00), 7–8 Dec 2000
12. Casas P, Vaton S, Fillatre L, Chonavel L (2009) Efficient methods for traffic matrix modeling and on-line estimation in large-scale IP networks. In: Proceedings of 21st international teletraffic congress, 2009. ITC 21 2009, pp 1–8, 15–17 Sept 2009
13. Liu B, Choo S-H, Lok S-L, Leong S-M, Lee S-C, Poon F-P, Tan H-H (1994) Finding the shortest route using cases, knowledge, and Dijkstra's algorithm. *IEEE Expert* 9(5):7–11
14. Noto M, Sato H (2000) A method for the shortest path search by extended Dijkstra algorithm. In: IEEE international conference on systems, man, and cybernetics, 2000, vol 3, pp 2316–2320
15. Guttoski PB, Sunye MS, Silva F (2007) Kruskal's algorithm for query tree optimization. In: Proceedings of 11th international database engineering and applications symposium, 2007. IDEAS 2007, pp 296–302, 6–8 Sept 2007

16. Butcher D, Li X, Guo J (2007) Security challenge and defense in VoIP infrastructures. In: IEEE transactions on systems, man, and cybernetics, part C: applications and reviews vol 37, no 6, pp 1152–1162
17. Narayan S, Yhi S (2010) Application layer network performance evaluation of VoIP traffic on a test-bed with IPv4 and IPv6 LAN infrastructure. In: IEEE Region 8 international conference on computational technologies in electrical and electronics engineering (SIBIRCON), 2010, pp 215–219, 11–15 July 2010
18. Xiao J, Boutaba R (2005) QoS-aware service composition in large scale multi-domain networks. In: Proceedings of 9th IFIP/IEEE international symposium on integrated network management, 2005. IM, pp 397–410, 15–19 May 2005
19. Okech JM, Hamam Y, Kurien A (2008) A cross-layer adaptation for VoIP over infrastructure mesh network. In: Third international conference on proceedings of broadband communications. Information Technology & Biomedical Applications, Nov, pp 97–102
20. Cao F, Malik S (2006) Vulnerability analysis and best practices for adopting IP telephony in critical infrastructure sectors. IEEE Commun Mag 44(4):138–145
21. Hruby M, Olsovsky M, Kotocova M (2012) Routing VoIP traffic in large networks. Lecture notes in engineering and computer science: proceedings of the world congress on engineering (2012) WCE 2012, London, UK, pp 798–803, 4–6 July 2012

Determining the Importance of Design Features on Usable Educational Websites

Layla Hasan

Abstract This research investigated the relative importance of specific design criteria developed for the purpose of this research, in the evaluation of the usability of educational websites from the point view of students. The results showed that content and navigation were the first and second preferred design categories to be considered while evaluating the usability of educational websites, while the architecture/organisation was the least important category. Also, the results showed that there was a statistically significant difference between males and females regarding only one category: the content. Females considered this to be the most important category while males considered it as the second most important. By contrast, the results showed that there were no statistically significant differences between the students of the two selected faculties (the Faculty of Information Technology and Science, and the Faculty of Economics and Administrative Sciences) concerning the relative importance of the developed criteria based on their majors/specialisations.

Keywords Design criteria · Educational · Students' preferences · Usability · User testing · websites

1 Introduction

Usability is one of the most important characteristics of any user interface; it measures how easy the interface is to use [1, 2]. Usability has been defined as: “a measure of the quality of a user’s experience when interacting with a product or system—whether a website, a software application, mobile technology, or any user operated device” [3].

L. Hasan

Department of Management Information Systems, Zarqa University,
132222, Zarqa 13132, Jordan
e-mail: l.hasan2@yahoo.co.uk

Research has offered some advantages that can be gained if the usability of websites is considered or improved. Agarwal and Venkatesh [4] and Nielsen [5] indicated that addressing the usability of sites could reduce the number of errors, enhance accuracy, and encourage positive attitudes toward the target interface. Furthermore, researchers indicated that addressing the usability of educational websites could help students to enjoy the learning experience, increase students' confidence, and encourage students to use the website [6].

Despite the importance of making educational websites usable, few studies were found in the literature that evaluated the usability of such sites [7, 8]. The studies that were found stressed the importance of usability in the design of educational websites and provided an outline of the design features that are important and that need to be included in the design of educational websites [6–11]. However, these studies did not investigate, and therefore consider, the relative importance of the design features in the usability of educational websites from the viewpoint of students. It is worth mentioning, however, that research has been conducted which has investigated the relative importance of design features for the usability of different types of website, such as: an e-commerce site [12]; portals and search engines, retail, entertainment, news and information, and financial services [13], online bookstores, automobile manufacturers, airlines and car rental agencies [4], financial, e-commerce, entertainment, education, government, and medical [14] from the viewpoint of users. However, no research has been conducted specifically to investigate educational websites. The research described here, which is an extended version of a paper published earlier [15], aims to address the gap noted in the literature by determining the relative importance of design criteria, which were specifically developed for the purpose of this research, for the usability of educational websites from the point view of students.

2 Literature Review

This subsection reviews studies that have investigated certain design criteria and shed light on the relative importance of design issues for different types of website from the point of view of users. For example, the study conducted by Pearson et al. [12] investigated the relative importance of five design criteria in the evaluation of the usability of an e-commerce site from the viewpoint of 178 web users. The criteria related to navigation, download speed, personalisation and customisation, ease of use, and accessibility. The results showed that these five criteria were significant predictors of website usability from the point of view of website users. Ease of use and navigation were the most important criteria in determining website usability, while personalisation and customisation were the least important. It was also found that males and females viewed these web usability criteria differently. The two usability criteria, navigation and ease of use, were found to have significant differences based on gender. Females placed greater emphasis on both of these web usability criteria than did males.

Similarly, Tarafdar and Zhang [13] investigated the influence of six web design issues on the usability of websites using different criteria related to information content, ease of navigation, download speed, customisation and personalisation, security, and availability and accessibility. The investigation was carried out by two web users only who evaluated a total of 200 websites using the six design factors. These sites were selected from five different domains: portals and search engines, retail, entertainment, news and information, and financial services (40 sites in each industry). Interestingly, the results showed that the four design factors that influenced website usability were: information content, ease of navigation, download speed, and availability and accessibility. However, the results showed that security and customisation did not influence a website's usability.

Agarwal and Venkatesh [4] also investigated the relative importance of evaluation criteria in determining the usability of websites for two types of user (consumers and investors) across four industry sectors: online bookstores, automobile manufacturers, airlines and car rental agencies. The criteria related to the Microsoft Usability Guidelines (MUG), which includes five categories: content, ease of use, promotion, made-for-the-medium, and emotion. The findings showed that content was the most important category in all eight groups (four products/industries, two types of user). Customers of all products deemed the content of a website to be equivalently important. The second category of ease of use was modestly moderately important across all eight groups.

Similarly, Zhang et al. [14] investigated user perception regarding the relative importance of website design features in six different website domains: financial, e-commerce, entertainment, education, government, and medical. The results indicated that an ease of navigation feature was a must-have for all six domains, while search tool was commonly ranked by the following four domains as important: education, government, medical, and e-commerce. The results showed that education and medical domains required comprehensiveness of information which was not ranked within the list of the five most important features in the other four domains.

Alternatively, Zhang and Dran [16] presented a two-factor model that can be used to distinguish website design factors into two types: namely, hygiene and motivator factors. Hygiene factors are those whose presence makes a website functional, useful and serviceable, and whose absence causes user dissatisfaction (i.e. live/broken links). Motivator factors, however, are those whose presence will enhance users' satisfaction with the website and motivate them to return, while their absence will leave users feeling neutral, but not necessarily dissatisfied, as long as the fundamentals or hygiene factors are in place (i.e. the use of multimedia). The results indicated that eighty-six percent of the participants believed that website types do affect the way they judge hygiene or motivator factors. For example, the participants specifically commented that they expected educational websites to have accurate, factual, nonbiased, and richer materials.

3 Aims and Objectives

The aim of this research is to determine the relative importance of specific design criteria in the evaluation of the usability of educational websites from the viewpoint of students.

The specific objectives for the research are:

- To develop evaluation criteria for assessing the usability of educational websites.
- To ask students to judge the relative importance (weights) of the different categories of the developed usability criteria.
- To determine if both gender and students' majors/specialisations have an impact on the relative importance of the developed usability criteria.

4 Methodology

This section describes the methodology employed in this research. It consists of five subsections which provide an idea regarding: the research instruments employed in this research; selections of the sample/participants; the pilot study; the procedure which was followed while collecting the data; and, finally, the analysis of the collected data.

4.1 Research Instruments

Usability criteria for assessing the usability of educational websites were developed based on an extensive review of the literature [4–14]. Section 5 presents the categories and subcategories of the criteria.

In order to collect information regarding the characteristics of students who participated in the research, a pre-test survey was developed. In order to obtain the relative importance (weights) of the different categories of the developed usability criteria, Agarwal and Venkatesh's [4] method for the assessment of usability was adopted. Based on this, a survey was developed. It aimed to collect the relative importance (weights) of the different categories and subcategories of the developed usability criteria by asking students to distribute 100 points across the five major categories of the criteria, and then to distribute the points assigned to each category across the corresponding subcategories.

4.2 Participants/Sample

The participants in this study were undergraduate students enrolled on twelve classes related to two faculties (the Faculty of Information Technology and Science, and the

Table 1 Demographic information of the research participants

		Faculty		Total
		Information technology and science	Economics and administrative sciences	
Sex	Male	82	67	149
	Female	35	35	88
Computer experience	<one year	0	4	4
	From one to three years	4	13	17
	>three years	113	103	216
Internet experience	<one year	4	11	15
	From one to three years	20	40	60
	>three years	92	70	162
Frequently use of internet	Daily	101	84	185
	Weekly	13	24	37
	Monthly	3	6	9
	By semester	0	4	4
	Yearly	0	2	2

Faculty of Economics and Administrative Sciences) at one of the universities in Jordan. Six classes were selected from each faculty. The total number of students was 237; the number of males was 149, while the number of females was 88 (Table 1). 237 provided usable responses. Unusable responses, which were ten, were primarily related to incomplete information. In cases where some students were enrolled onto more than one of the classes included in the sample, they were asked to leave the session and to participate only once. Demographic information concerning the students is shown in Table 1.

4.3 Pilot

A pilot study was conducted before the experiment/main test to test the method of assigning weights. Before conducting the pilot study, the surveys were translated into Arabic. The surveys were pilot tested using ten Jordanian undergraduate students using the Arabic language version. The pilot study identified some ambiguity in the surveys. Results from the pilot test were taken into consideration and minor changes were made to the surveys.

4.4 Procedure

All data collection sessions followed the same procedure. Data were gathered using two surveys in a university in Jordan where all students had access to the Internet.

The session began with the researcher welcoming the students and explaining the objectives of the study; the number of surveys that needed to be filled in; and students' right to withdraw from the session at any time. The students were then asked to fill in the pre-test questionnaire in order to obtain information regarding their background and experience. Then, the students were asked to provide their perceptions of the relative importance (weights) of the developed usability criteria (5 categories) using the relative importance survey. Following this, students were asked to distribute the points across the various subcategories.

4.5 Analysis

The data collected were analysed in several ways. Descriptive analysis was used to analyse the data collected from the pre-test questionnaire to describe the characteristics of the students. In order to find the relative importance (weights) for the developed criteria (the five categories and their corresponding subcategories) from the viewpoint of students, the average weight was calculated. Descriptive analysis (the mean and standard deviation) of the weights (i.e. the relative importance) of the developed criteria based on gender and faculty (majors/specialisations) was carried out. To determine if there was a statistically significant difference in the relative importance of the web usability criteria based on gender and faculty, the one-way analysis of variance (ANOVA) was used for each category and the corresponding subcategories of the developed usability criteria.

5 Criteria for Evaluating the Usability of Educational Websites

Specific criteria for evaluating the usability of educational websites were developed based on a literature review. The developed criteria consisted of five main categories. This section presents the categories and their corresponding subcategories:

- **Navigation:** this assesses whether a site includes the main tools (i.e. navigation menu, internal search facility) and links which facilitate the navigation of users through a site. Research showed that navigation was one of the design factors that influenced website usability [4, 12, 13]. Navigation comprised five subcategories. These were: *navigation support*: navigational links are obvious in each page so that users can explore and find their way around the site and navigate easily; *effective internal search*: internal search is effective: e.g. it is fast, accurate and provides useful, concise and clear results; *working links*: links are discernible, working properly and not misleading so that the user knows what to expect from the destination page; *no broken links*: the site has no broken links; and *no orphan pages*: the site has no dead-end pages.

- **Architecture/organisation:** this criterion relates to the structure of a site's information which should be divided into logical, clear groups; each group should include related information. Architecture/organisation consists of three subcategories. These are: *logical structure of a site:* the structure of the site is simple and straightforward; related information is grouped together; *not deep architecture:* architecture is not too deep so that the number of clicks to reach goals is not too large, e.g. it does not require clicking more than 3 links; and *simple navigation menu:* the navigation menu is simple and straightforward.
- **Ease of use and communication:** this relates to the cognitive effort required to use a website [4], and to the existence of basic information which facilitates communications with the university in different ways. Research has found that ease of use is an important factor/issue in determining web usability [4, 12–14]. Ease of use and communication comprises four subcategories. These are: *quick downloading of web pages:* the download time of the pages is appropriate; *easy interaction with a website:* interaction with the website is easy for different groups of users, e.g. navigating through the site's pages is easy; *contact us information:* useful information to enable easy communication with the university is displayed, e.g. contact us; and *foreign language support:* the site's content is displayed in different languages.
- **Design:** this relates to the visual attractiveness of a site's design; the appropriate design of a site's pages, and the appropriate use of images, fonts and colours in the design of a site. Design comprises six subcategories: *aesthetic design:* the site is attractive and appealing so that it impresses the potential customer; *appropriate use of images:* the quality of images is adequate, there are no broken images, image size is relevant so that it has minimal effect on loading time; *appropriate choice of fonts:* font types are appropriate and easy to read; *appropriate choice of colours:* choice of colours for both fonts and background is appropriate, the combination of background and font colours is appropriate; *appropriate page design:* pages are uncluttered, page margins are sufficient, the page title is appropriate; and *consistency:* page layout or style is consistent throughout the website: e.g. justification of text, font types, font sizes.
- **Content:** this assesses whether a site includes the information users require. Research stresses the importance of this factor and shows that it is one of the most important factors that influence web usability [4, 13]. Content consists of seven subcategories. These are: *up-to-date information:* the information is up-to-date, current and often updated; *relevant information:* the information is sufficient and relevant to user needs; *no under-construction pages:* there are no 'under construction' pages; *accurate information:* the information is accurate; *information about the university:* basic facts about the university are displayed, e.g. university overview, academic calendar, registration, description, etc.; *information about the faculties:* adequate information about the faculties is displayed, e.g. overview, departments, specialisations, etc.; and *information about the departments:* adequate information about the departments is displayed, e.g. overview, academic staff, outlines, etc.

Table 2 The relative importance (weights) for the categories and subcategories of the developed usability criteria and the total weight for each category

Categories	Subcategories	Weight	Total weights for each category
Navigation	Navigation Support	5.11	20.75
	Effective Internal Search	5.01	
	Working Links	4.49	
	No Broken Links	2.96	
	No Orphan Pages	3.17	
	Logical Structure of a Site	7.16	
Architecture/Organisation	Not Deep Architecture	5.73	18.66
	Simple Navigation Menu	5.77	
	Quick Downloading of Web Pages	6.20	
Ease of Use and Communication	Easy Interaction with a Website	5.38	19.88
	Contact Us Information	4.43	
	Foreign Language Support	3.86	
	Aesthetic Design	4.27	
Design	Appropriate Use of Images	3.16	21.56
	Appropriate Choice of Fonts	2.57	
	Appropriate Choice of Colours	2.74	
	Appropriate Page Design	3.35	
	Consistency	3.06	
	Up-to-date Information	4.74	
	Relevant Information	3.23	
	No Under Construction Pages	2.07	
Content	Accurate Information	3.20	100
	Information about the University	2.79	
	Information about the Faculties	2.51	
	Information about the Departments	3.01	
Total weights			

6 Results

The results showed that the most important design category for the usability of educational websites from the viewpoint of users was the content as it has the highest weight (Table 2). The results also showed that the navigation was the second most important category for the usability of educational websites. The results showed that ease of use, and communication and design were the third and fourth important categories, respectively in the usability of educational websites from the viewpoint of students. Finally, the results showed that the architecture/organisation was the least important category for the usability of educational websites from the viewpoint of students.

Interestingly, the results showed that the weights of the subcategories varied; the highest weight assigned to a single subcategory was 7.16, while the lowest weight was 2.07. Specifically, the results showed, with regard to each design category, the design feature that students preferred the most for a usable educational website, and the design feature which was the least important from the viewpoint of students. The following summarises the specific design features with regard to each category:

- Navigation: *navigation support* was the most preferred design feature for usable educational websites, while *no broken links* was the least preferred one.
- Architecture/organisation: *logical structure of a site* was the most important subcategory from the viewpoint of students for usable educational website; while *not deep architecture* was the least important one.
- Ease of use and communication: *quick downloading of web pages* was preferred the most by students, while *foreign language support* was the least preferred design feature by them for usable educational websites.
- Design: *aesthetic design* was the most preferred design feature for usable educational websites, while *appropriate choice of fonts* was the least preferred one.
- Content: the most important design feature regarding this category was *up-to-date information*, while *no under construction pages* was the least important subcategory from the viewpoint of students.

In addition to the results summarised above, Table 2 provides more information regarding the order of the subcategories of each category based on their relative importance according to the students. For example, the students considered information about the departments to be more important than information about the faculties and the university, as they gave it a higher weight (3.01) compared to the other two subcategories (2.51 and 2.79, respectively).

The ANOVA test revealed no statistically significant differences between males and females regarding the relative importance of four categories of the criteria: navigation, architecture/organisation, ease of use and communication, and design [15]. However, the ANOVA test showed that there was a statistically significant difference between males and females regarding the relative importance of the content category. The females considered this category as the most important and gave it therefore the highest weight (23.58), while the males considered this category as the second most important category and therefore gave it the weight of 20.37.

The descending order of the usability categories based on their relative importance according to males was: navigation, content, ease of use and communication, design, and architecture/organisation. However, the descending order of the usability categories based on their relative importance according to females was: content, navigation, ease of use and communication, architecture/organisation, and design.

The ANOVA test showed that there were no statistically significant differences between the students of the two faculties (the Faculty of Information Technology and Science, and the Faculty of Economics and Administrative Sciences) concerning the relative importance of the five categories of the criteria [15]. The descending order of the categories based on their relative importance according to the students of the

Faculty Information Technology and Science was: navigation, content, ease of use and communication, design, and architecture/organisation.

However, the descending order of the categories based on their relative importance according to the students of the Faculty of Economics and Administrative Sciences was: content, navigation, ease of use and communication, architecture/organisation, and design.

7 Discussion

This section discusses the results of this research in the light of the existing literature. As indicated in Sect. 2, nearly all earlier research which investigated users' preferences on the relative importance of website design features, did not investigate educational/academic websites. The study of Zhang et al. [14] is worth mentioning as this was the only study found in the literature which included an educational domain as one of the six domains that were investigated concerning the importance of website features.

However, this research has addressed the gap noted in the literature and focused primarily on investigating the relative importance of website design features on the usability of educational websites from the viewpoint of students.

Despite the fact that earlier research employed different design criteria and used them in the investigation of their importance on the different domains investigated, there are common design categories among the criteria suggested by the different studies and the design criteria suggested and used in this research. The common categories relate to: content, navigation and ease of use/navigation. Therefore, the results of this research are comparable with the results of earlier research from the point of view that similar categories of criteria were used. This subsection discusses the results of this research and compares them with the results of earlier research based on the similar categories used.

The results of this research, which showed that the content, navigation, and ease of use were the first, second, and third most important categories, respectively that influence the usability of educational websites from the point of view of students, are comparable with other research [4, 13, 14]. Therefore, the results stressed the importance of the content and ease of use design categories, not only in the domain of e-commerce websites, as shown by [4, 14] and other domains (portals and search engines, entertainment, news and information, financial services, government, and medical) [4, 13, 14], but also in the educational website domain. Also, the results stressed the importance of considering navigational issue (i.e. search tools) when designing educational websites, as well as e-commerce, education, government, medical websites as shown by earlier research [4, 12–14].

Interestingly, the results of this research were comparable with other research [12] regarding the rating of the design categories of the suggested criteria differently by males and females. The results of this research showed that content was the only category which showed significant differences based on gender, since females

placed a greater emphasis on this than did males, while the results of Pearson et al. [12] showed that the navigation and ease of use categories had significant differences based on gender, where females placed greater emphasis on them than did males. The differences between the results might relate to the fact that the research conducted by Pearson et al. [12] concerned e-commerce websites, while this research considered educational websites. This suggests that universities and/or academic institutions which are especially for females should give the content category first priority when designing usable educational websites, or when evaluating the usability of their websites. However, universities and/or academic institutions for males should give the navigation category the first priority.

This research, unlike earlier research, also investigated whether the relative importance of the design categories of the suggested criteria differed from the viewpoint of students based on the differences in their majors/specialisations. The results, as discussed in Sect. 6, showed that all the design categories did not have statistical significant differences based on faculty (majors/ specialisations). However, the order of the design categories, from the first to the least important from the point of view of students, was different based on faculty, as discussed in Sect. 6. This provides evidence for universities and/or academic institutions to consider the preferences of design categories from the viewpoint of students based on their majors/specialisations.

Furthermore, this research, unlike earlier research, shed light on the specific design features students most preferred for usable educational websites, such as: *navigation support, logical structure of a site, quick downloading of web pages, aesthetic design, and up-to-date information*. Also, this research shed light on the specific least important design features for usable educational websites from the viewpoint of students, such as: *no broken links, not deep architecture, foreign language support, appropriate choice of fonts, and no under construction pages*. These results, together with the previous ones, shed light on the design categories and subcategories that must be taken into consideration when designing and/or evaluating the usability of educational websites, as well as the design categories and subcategories which should have less focus when designing and/or evaluating the usability of such websites (Table 2).

8 Conclusions

This research provides empirical evidence for academic institutions and universities regarding the relative importance of specific design features on which to focus when designing and/or evaluating the usability of their educational websites. The results showed that content and navigation were the most and second most important design categories, respectively for the usability of educational websites from the viewpoint of students. The results also showed that the third, fourth and least important categories for educational websites were: ease of use and communication; design; and architecture/organisation, respectively.

This research also investigated whether gender and majors/specialisations had an impact on the relative importance of the developed usability criteria. The results showed that there was a statistically significant difference between males and females regarding only one category: the content. Females considered it as the most important category while males considered it as the second most important category. By contrast, the results showed that there were no statistically significant differences between the students of the two selected faculties concerning the relative importance of the developed criteria based on majors/specialisations.

The next step is to ask students to provide ratings for nine Jordanian university websites on the developed usability criteria and their categories. Then to use the weights (the relative importance of the different categories of the developed usability criteria obtained from this research) and ratings together to assess the overall usability of each Jordanian university website.

Acknowledgments This research was funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan.

References

1. Murugappan A, Ferdin JJ, Shamika M, Manideep V, Mridul M (2009) Metric based architecture to enhance software usability. In: Proceedings of the international multiconference of engineers and computer scientists 2009, IMECS 2009, vol I, Hong Kong, pp 18–20
2. Nielsen J (2009) Usability 101: Introduction to usability. Useit.com. <http://www.useit.com/alertbox/20030825.html>. Accessed 14 Feb 2006
3. Anonymous, Step-by-Step Usability Guide, (2006). <http://www.usability.gov> Accessed 20 Sept 2011
4. Agarwal R, Venkatesh V (2002) Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability. *Inf Syst Res* 13(2):168–186
5. Nielsen J (2000) Designing web usability: the practice of simplicity. New Riders Publishing, Indianapolis
6. Lencastre J, Chaves J (2008) A usability evaluation of educational websites. In: Proceedings of the EADTU conference
7. Toit M, Bothma C (2010) Evaluating the usability of an academic marketing department's website from a marketing student's perspective. *Int Retail and Mark Rev* 5(1): 15–24
8. Mustafa S, Al-Zoua'bi L (2008) Usability of the academic websites of Jordan's universities. In: Proceedings of the international Arab conference on information technology, Tunisia
9. Gonzalez M, Granollers T, Pascual A (2008) Testing website usability in Spanish-speaking academia through heuristic evaluation and cognitive walkthrough. *J Univ Comput Sci* 14(9):1513–1527
10. Kostaras N, Xenos M (2006) Assessing educational web-site usability using heuristic evaluation rules. In: Proceedings of the 11th Panhellenic conference in informatics
11. Papadopoulos T, Xenos M (2008) Quality evaluation of educational websites using heuristic and laboratory methods. In: Proceedings of the 2nd Panhellenic scientific student conference on informatics, related technologies, and applications, pp 43–54
12. Pearson JM, Pearson A, Green D (2007) Determining the importance of key criteria in web usability. *Manag Res News* 30(11):816–828
13. Tarafdar M, Zhang J (2005) Analyzing the influence of website design parameters on website usability. *Inf Resour Manag J* 18(4):62–80

14. Zhang P, von Dran G, Blake P, Pipithsuksunt V (2000) A comparison of the most important website features in different domains: an empirical study of user perceptions. In: Proceedings of Americas conference on information systems (AMCIS'2000), Long Beach, CA, pp 1367–1372
15. Hasan L (2012) Investigating the relative importance of design criteria in the evaluation of the usability of educational websites from the viewpoint of students. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012, pp 832–837
16. Zhang P, von Dran G (2000) Satisfiers and dissatisfiers: a two-factor model for website design and evaluation. *J Am Soc Inform Sci* 51(14):1253–1268

A Flexible Dynamic Data Structure for Scientific Computing

Josef Weinbub, Karl Rupp and Siegfried Selberherr

Abstract We present an approach for a generic, multi-dimensional run-time data structure suitable for high-performance scientific computing in C++. Our concept for associating meta-information with the data structure as well as different underlying datatypes is depicted. High-performance, multi-dimensional data access is realized by utilizing a heterogenous compile-time container generation function. The generalized data structure implementation is discussed and performance results are given with respect to reference implementations. We show that our approach is not only highly flexible but also offers high-performance data access by simultaneously relying on a small code base.

Keywords C++ · Data structure · Dynamic · Generic programming · Meta programming · Multi-Dimensional

1 Introduction

The plethora of applications in the field of scientific computing introduces different requirements for data structures. A matrix for the representation of a linear system of equations is a prominent example of a two-dimensional datastructure [1]. On the contrary, an application may require a set of second-order tensors to describe, for instance, the stresses in the field of continuum mechanics [2]. Typically, the dimen-

J. Weinbub (✉) · K. Rupp · S. Selberherr
Institute for Microelectronics, Technische Universität Wien,
Gußhausstraße 27-29, 1040 Wien, Austria
e-mail: weinbub@iue.tuwien.ac.at

K. Rupp
e-mail: rupp@iue.tuwien.ac.at

S. Selberherr
e-mail: selberherr@iue.tuwien.ac.at

sionality is known during compile-time, thus the data structure can be optimized by the compiler [3]. However, if the dimensionality of the data structure is not known during compile-time, no presumptions with respect to the storage layout can be made thus imposing challenges on the data structure implementation. A typical challenge is to provide a high-performing, dimension agnostic access mechanism, as a utilization of a dynamic container for the index-tuple, which does not result in poor performance due to run-time overhead.

Such a scenario arises when, for instance, dealing with input routines, where the data structure dimension depends on the data read from an input stream, inherently being a run-time process. Another case would be a unified data structure interface, where unified relates to a single datatype used to reflect data structures of arbitrary dimensionality. Such can be the case for plugin interfaces within a software component framework. Each plugin provides and receives data, though the dimensionality is not known in advance. Overall, an approach is required to support multiple dimensions during run-time.

An additional challenge arises with respect to the underlying datatype of the data structure elements. C++, as a statically typed programming language, enables type checks during compilation. This not only allows for the compiler to optimize code, but also to detect errors at the earliest possible stage of the development. However, such a system imposes restrictions in regard to the run-time handling of datatypes. For example, a floating-point number of type `double` can only hold a double-precision value, but not a `string` object. This limitation is in principle desired, but introduces challenges for the implementation of a generic data structure, where the type is not known in advance. Related to the previously introduced examples, the data structure cannot only vary in its dimension but also in the type it can hold.

The field of scientific computing not only processes sets of values, but typically also sets of quantities. A quantity is referred to as a value which is associated with a unit. Supporting or even enforcing units is a vital part for ensuring the correctness of scientific simulations [4]. However, units may not be the only additional meta-information. For example, data values can be related to measurements carried out at a specific temperature. Overall, the need for a flexible property system arises, which should not only reside in the run-time domain, but also be orthogonal to the data structure. In this context, orthogonality refers to exchanging the data structure without influencing the attached meta-information. Such an approach is highly versatile, as it introduces exchangeability.

The continually growing demand for increased simulation performance introduces the need to parallelize simulation tools. Ideally, the individual computations should scale beyond a multi-core processor, namely to a distributed computing environment. Typically, the Message Passing Interface (MPI) is utilized for communication within a distributed environment. The data structure should support seamless integration into such an MPI based environment, to ease the integration process. Therefore a serialization approach for the data structure should be available, which allows out-of-the-box transmission by an MPI communication channel.

We present a revision of our previous work, introducing our approach for a flexible dynamic data structure [5].

The data structure handles multiple dimensions, run-time generation, and supports different underlying datatypes. Additionally, we support direct transmission capabilities over MPI and an orthogonal and flexible coupling of meta-information with the data structure. We achieve this by utilizing modern programming techniques, in particular generic [6] and meta-programming [7], and the Boost Libraries [8]. We show that our approach does not only provide a high degree of flexibility, but also offers high-performance data access. Additionally, due to the heavy utilization of libraries in conjunction with the application of modern programming techniques, the required code base can be kept to a minimum. This fact significantly improves the maintainability of our implementation.

This work is organized as follows: Sect. 2 puts the work into context. Section 3 introduces our approach in detail and Sect. 4 depicts performance results.

2 Related Work

This section provides a short overview of other approaches for data structure implementations, each being briefly discussed and differences to our work are outlined.

A flexible run-time data structure for multi-dimensional problems is provided by Marray [9]. Marray is a C++ header-only library and publicly available under the MIT License. The library provides not only the generation of multi-dimensional arrays during run-time, but also views on sub-spaces of the generated data structures. A C++98 compliant implementation is available as well as a C++11 version, which utilizes, for example, the variadic template mechanism [10] to provide dimension independent access to the data structure. Marray is a feature-rich library supporting the dynamic generation of arrays of arbitrary dimension. However, our performance evaluations depict that our approach offers a significantly increased access performance (Sect. 4).

Several multi-dimensional array libraries are available for the case of fixed dimensions during run-time. For example, the Boost MultiArray Library [11] and the Blitz++ Library [12] provide the generation of multi-dimensional arrays during compile-time. Additionally, views are provided to access a specific subset of the generated data structures. Superior performance is obtained by these libraries due to the use of static array dimensions for compile-time optimizations, yet they lack the ability to change the storage layout during run-time. This fact renders these approaches unfit for pure run-time problems, as it would require to instantiate the types for all possible dimensions, which is obviously not realizable within a finite time frame.

However, implementing even the most likely cases not only increases the compilation time as well as the executable size, but is also highly inflexible, as adding additional dimensions requires source code extensions and therefore recompilation.

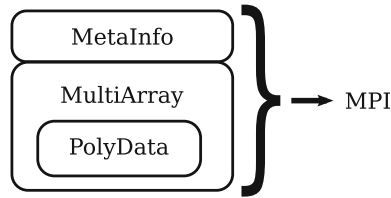


Fig. 1 Our approach is based on a polymorphic datatype which is used by the multi-dimensional array data structure. Meta-information is orthogonally coupled with the data structure. The overall approach is serializable, thus any object can be transmitted over an MPI communication channel

3 Our Approach

Our approach focuses on several key-aspects, being:

1. A Polymorphic Datatype
2. Data Structure Generalization
3. Attaching Meta-Information
4. Serialization.

Figure 1 depicts an overview of our approach. First, a polymorphic data-type supporting different datatypes during run-time is introduced. Note that polymorphy denotes the ability to represent different datatypes. Second, the polymorphic entries are embedded in a multi-dimensional run-time array data structure. Third, run-time meta-information is attached to the data structure. Fourth, our implementation is serialized to enable convenient transfer by the MPI. Throughout this section we discuss each of these aspects in detail and provide our implementation approach.

3.1 A Polymorphic Datatype

One of the core aspects of our approach is the ability to support different datatypes during run-time. The challenge is to provide one datatype which can in fact hold several different types. This is a peculiar task for statically typed languages, like C++, as the type system only allows to assign objects of the same or convertible type. If the types are not the same, cast operations have to be performed. However, applying casts can result in information loss, for example, when a datatype of higher precision, like `double`, is transformed to a datatype with lower precision, like `float`.

We utilize the Boost Variant Library (BVL) [13] for supporting different datatypes during run-time. Informally, a BVL datatype can be seen as an enum for datatypes. A set of possible, supported datatypes has to be provided during compile-time. During run-time, the instantiated BVL object can be associated with any of these datatypes.

We identify four different categories of datatypes, which are listed in the following:

- signed integer
- unsigned integer
- floating-point
- string.

A meta-function for the generation of the polymorphic datatype based on the introduced four categories is provided. A meta-function is a class or a class template which provides a nested `type` typedef [14]. In the following, this mechanism is introduced in detail. First, the set of supported types is generated by utilizing an associative heterogeneous container provided by the Boost Fusion Library (BFL) [15], as depicted in the following.

```

1 typedef make_map <
2   Signed,  Unsigned,      Float,  String,
3   int,    unsigned int, double, string
4   >::type

```

The `make_map` meta-function is utilized to generate an associative BFL container (Lines 1–4). Note that the tags in Line 2 represent the individual categories, and the datatypes in Line 3 relate to the corresponding datatypes. Tags are typically implemented by so-called empty structures, for example, `struct Signed{};`. Generally, in our approach the datatypes can be set non-intrusively, meaning that the underlying datatypes can be exchanged, thus significantly improving the applicability and extendability of our approach. For example, instead of the floating-point datatype, a multi-precision datatype provided by the GNU Multiple Precision Arithmetic Library (GMP) [16] can be used, which would significantly improve the accuracy of subsequent floating-point operations.

In the following, the associative `Types` container is converted into a Boost Metaprogramming Library (MPL) [14] vector container by the following meta-function.

```

1 typedef generate_typeset<Types>::type TypeSet;

```

This step is necessary, as the subsequent step of utilizing the BVL is eased, when the supported datatypes are available as an MPL sequence. A default implementation is available, which allows convenient generation of this typeset and only relies on built-in datatypes as shown in the following.

```

1 typedef generate_typeset<>::type TypeSet;

```

The typeset is then used to generate the actual polymorphic datatype based on the BVL. Again a meta-function is used to generate the polymorphic datatype as depicted in the following.

```

1 typedef generate_polyvalue<
2   TypeSet>::type PolyValue;

```

Internally, the BVL `make_variant_over` meta-function is utilized to generate the actual polymorphic datatype.

Finally, due to the BVL, it is possible to provide a generic way to support different datatypes during run-time:

```

1 PolyValue signed_integer =
2   value_at_key<Types, Signed>::type(4);
3 PolyValue floating_point =
4   value_at_key<Types, Float>::type(4.0);

```

A type-safe approach for a signed integer and a floating-point datatype instantiation is implemented by using the BFL meta-function for key-based element access (`value_at_key`). Type safety is accomplished in this case, by accessing the actual type in the previously provided `Types` container.

3.2 Data Structure Generalization

Based on the previously introduced polymorphic datatype the actual array data structure can be implemented. The implementation has two goals: First, multiple dimensionality should be supported during run-time. Second, the data access should be as fast as possible. In this work we do not focus on advanced functionality, as, for example, provided by the Marray library. Instead, we aim for a straightforward data structure, coordinate-based access, and a high-performance implementation.

The following code snippet outlines the creation of a two-dimensional data structure, where the first and second dimension holds three and four elements, respectively.

```

1 typedef MultiArray<PolyValue>      MultiArrayT;
2 MultiArrayT::dimensions_type      dim;
3 dim.push_back(3);
4 dim.push_back(4);
5 MultiArrayT                        multiarray(dim);

```

The `MultiArray` implementation can be configured to hold arbitrary value types. In this case, the previously introduced polymorphic value type based on the BVL is used (Line 1). The dimensions are formulated by utilizing a Standard Template Library (STL) `vector` container [17], where the type is accessed by the member-type `dimensions_type` (Line 2). Each element of the dimensions container holds the number of elements of the respective dimension (Lines 3–4). The number of dimensions is therefore inherently provided by the size of the container. A `MultiArray` object is instantiated with the dimension configuration (Line 5). Internally, an STL `vector` container, which represents a linear memory block, is used. Data Structures of arbitrary dimensionality are mapped on this linear container, as depicted in Fig. 2 for the two-dimensional case. The individual columns of the respective domains are stored consecutively. This approach minimizes the allocation time, as only one memory allocation step is necessary. However, a linear storage approach requires index handling to map the coordinate index-tuple on the corresponding position within the linear data structure. This is the performance critical part, as the data access implementation is likely to be called on a regular basis.

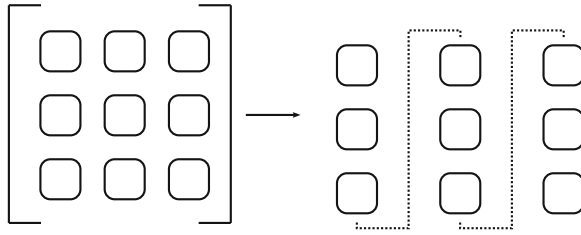


Fig. 2 A two-dimensional array is mapped on our internal, one-dimensional data structure

The central challenge of providing coordinate index access is the handling of data structures of arbitrary dimension, as the number of access-indices corresponds to the number of dimensions. Typically, the elements of a two dimensional array are accessed in coordinates, like `array(i, j)`. From the software development point of view, the challenge is to implement an access mechanism which is both high-performing and can be used for arbitrary dimensionality.

Several approaches for the access implementation and the related index computation have been investigated. One approach is based on so-called variadic functions provided by the C programming language [18]. The primary drawback of this approach is the fact that the number of indices has to be provided explicitly. In addition, this approach is not type-safe, as it is based on macros. Another approach is based on utilizing an STL `vector` for the index container. The number of indices can easily vary during run-time. However, this approach suffers due to run-time overhead for the creation and the traversal of the index-vector for each data access. In the end, a BFL `vector` sequence is utilized, which offers superior performance due to the fact that the sequence is a compile-time container. The run-time generation is performed by a generation function provided by the BFL, like depicted in the following.

```
1 multiarray(make_vector(2,3)) = Numeric(3.5);
```

Note that the BFL generation function `make_vector` is utilized to generate the compile-time index container in-place. This can be considered a drawback with respect to usability, as it requires additional coding. However, convenience specializations can be implemented to hide the vector generation step from the user. Due to restrictions of the C++98/03 standard, these specializations can only be provided for a finite set of dimensions, thus such an approach cannot be considered truly multi-dimensional. Internally, the BFL `make_vector` function relies on a macro, for the generation of arbitrary dimensional compile-time data structures. However, we consider this to be an excellent compromise, at least until the C++11 standard is broadly available. This standard introduces the aforementioned variadic template mechanism, which is also applied by `Marray` in its C++11 extension. Our investigations revealed, that with variadic templates the same performance as with our current approach can be achieved but simultaneously the required access interface (`multiarray(2,3)`) for arbitrary dimensions can be realized.

Our high-performance index computation is implemented based on the BFL algorithms which allow partial compile-time computation, as depicted in the following.

```

1 template<typename IndexSequ >
2 Element & operator () (IndexSequ const & indices) {
3   return container[
4     accumulate(pop_front(indices), at_c<0>(indices),
5       make_index<dimensions_type>(dimensions)
6     ) ]; }

```

Lines 4–5 compute the actual index, which is then used to access the element in the linear container in Line 3. The index computation is based on various BFL mechanisms, like, `accumulate`. For example, for a two-dimensional problem, the index is evaluated as follows: $i = I_0 + I_1 \cdot D_0$, where I_0 and I_1 refer to the first and second index, respectively. D_0 relates to the number of elements in the first dimension. This procedure can be extended to arbitrary dimensionality.

The introduced data structure generalization approach can also be applied for the tensor-valued elements. Therefore, a two-level hierarchy of the proposed MultiArray data structure supports tensor datasets of arbitrary dimension and varying datatypes.

3.3 Attaching Meta-Information

Scientific computations do not merely process values, but physical quantities. This is a subtle difference, as the former indicate simple values, like, a `double`, but the latter associates the respective value with a unit, promoting the value to a physical quantity. This is rather important, as scientific computations should not just imply a unit system, they should enforce it to eliminate unit-related errors [4]. Keeping in mind that units might not be the only additional property which can be associated with a dataset, a flexible approach is required to associate additional meta-information with the data structure.

Another important aspect, however, is to ensure extendability and exchangeability. As such, the approach has to support the exchange of the data structure as well as the meta-information package. For example, it should be possible to exchange our data structure with the Marray implementation [9], without changing the associated meta-information package. Such an approach is considered orthogonal, as the exchange of one part does not influence the behavior of another part. Obviously, the implementation of the meta-information package has to be non-intrusive with respect to the data structure. More concretely, the package should not be placed inside the data structure class, but externally associated with it.

We implement an approach for storing arbitrary meta-information during runtime, which is straightforwardly based on the STL `map` container.

```

1 typedef map <string, string>           MetaInformation;
2 MetaInformation minfo;
3 minfo["unit"] = "kg";

```

This approach is very flexible, as arbitrary properties can be added. Most importantly, though, the implementation effort is kept to a minimum, as already available functionality is utilized.

Finally, the data structure and the meta-information is coupled by the associative container of the BFL.

```

1 typedef make_map <
2   data,          metainf,
3   MultiArrayT,  MetaInformation
4 >   QuantityDataset;

```

Note the orthogonal and extendable association of the data with additional properties. Orthogonality can be identified when, for example, exchanging `MultiArrayT` with the corresponding `Marray` datatype, which neither has an impact on the associated meta-information package `MetaInformation` nor on the overall handling of the `QuantityDataset`. Extendability refers to the fact, that, by adding additional tags, further data can be associated with the dataset.

If the availability of a unit should be enforced, then the unit information should be moved from the `minfo` container to the `QuantityDataset`.

By utilizing a new tag and a string value, a `QuantityDataset` expects the unit information, like depicted in the following.

```

1 typedef make_map <
2   data,          metainf,          unit,
3   MultiArrayT,  MetaInformation,  string
4 >   QuantityDataset;
5 QuantityDataset quantity_dataset = make_map <
6   data, metainf, unit > (multiarray, minfo, "kg");

```

This also outlines the flexibility of our approach, as different setups of the `QuantityDataset` can be enforced.

Our approach presumes that the unit applies to the complete dataset. In case heterogeneous units should be supported, an additional layer has to be introduced assigning a unit to a specific value.

3.4 *Serialization*

Serialization refers to the process of storing and retrieving the elements of a data structure. Typically, input/output mechanisms utilize serialization processes, as, for example, a matrix is written to a file. This is usually performed by implementing dedicated file writer functions. However, a major disadvantage is, that for each new file format a new writer has to be implemented. One approach to ease the burden of serialization is to introduce an additional layer, which provides a common ground between the data structure and the target storage format. Thus, it is possible to implement the serialization mechanism for a data structure once, and then access the already available functionality based on the additional layer. However, serialization cannot only

be used for file input/output processes, but also for MPI communication [19]. For this purpose we utilize the Boost Serialization Library (BSL) [20], which provides a serialization facility for arbitrary data structures. Based on our previously introduced quantity dataset, serialization extensions have been implemented. In the following, Process 0 transmits an available quantity dataset to Process 1.

```
1 if (world.rank() == 0) {
2     for_each(quantity_dataset, send(comm, 1));
3 }
```

Process 1 receives the quantity dataset from Process 0:

```
1 if (world.rank() == 1) {
2     QuantityDataset quantity_dataset;
3     for_each(quantity_dataset, recv(comm, 0));
4 }
```

Note that the unary auxiliary functor `send/recv` gets an element of the quantity dataset, being a BFL pair data structure, and sends/receives the data element of the respective pair.

The BFL `for_each` algorithm is utilized to traverse the elements of the quantity dataset. Finally note, that additional convenience levels can be implemented to further wrap code away from the user. For example, a generic serialization implementation can be provided, which is capable of handling arbitrary BFL data structures.

4 Performance

This section presents performance results for our data structure, especially our BFL based index computation approach. The tests have been carried out on an AMD Phenom II X4 965 with 8 GB of memory running a 64-Bit Linux distribution. The GNU GCC compiler in version 4.4.5 has been used with the flags `-O2 -DNDEBUG`. Benchmarks have been averaged over five runs to reduce noise. The element access performance for various problem sizes and different array dimensions is depicted, based on storing `double` values. The reference implementation is based on a hierarchy of STL vectors, as no index computation is required for the element access procedure. Additionally, we compare our approach with the already mentioned, publicly available Marray library [9]. Furthermore, we investigate the influence of optimal and non-optimal traversal, identified with OPT and NOPT, respectively. In the optimal case, the element access is as sequential as possible, meaning that the elements are accessed in the same consecutive manner as they are stored in the memory. Sequential access is favored by the so-called prefetching mechanism [21]. We investigate the non-optimal case, by exchanging the traversal loops for the two- and three-dimensional problems.

Figure 3 depicts the results for a one-dimensional array. Our approach is equally fast as the reference implementation, and takes around 0.15s for writing data on

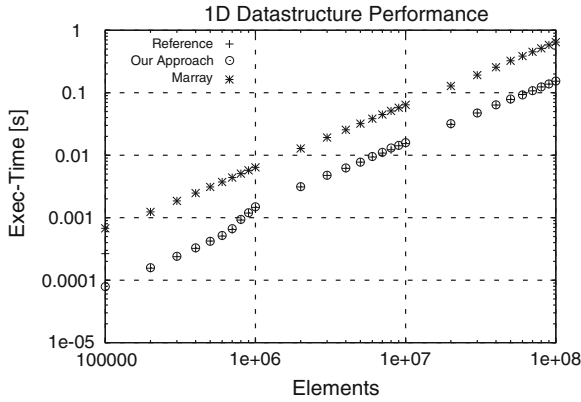


Fig. 3 A one-dimensional array structure is benchmarked. Our approach is equally fast than the reference implementation, whereas Marray is a factor of 2.9 slower

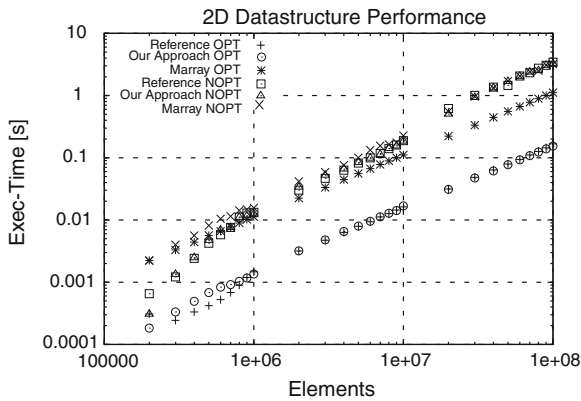


Fig. 4 A two-dimensional array structure is benchmarked. OPT and NOPT refers to optimal and non-optimal traversal, respectively. For 10^8 elements, our approach and Marray is a factor of 1.6 and 6.8, respectively, slower than the reference implementation. All approaches are equally slower in the non-optimized case, namely around 3.3 s for 10^8 elements

all 10^8 elements. The Marray implementation is about a factor of 2.9 slower. The two-dimensional results are depicted in Fig. 4. For 10^8 elements and the optimal traversal case our approach is again equally fast as the reference implementation, whereas Marray is a factor of 7 slower. In the non-optimal case, all implementations are significantly slower, and take approximately equally long (around 3.3 s for 10^8 elements). Figure 5 shows the results for a three-dimensional problem. Our optimal traversal implementation is a factor of 1.5 faster than the reference. For 10^8 elements Marray is a factor of 9.9 slower than our approach. As expected, the non-optimized traversal implementations are significantly slower, for instance, our optimal traversal approach is a factor of 48 faster than the non-optimized one.

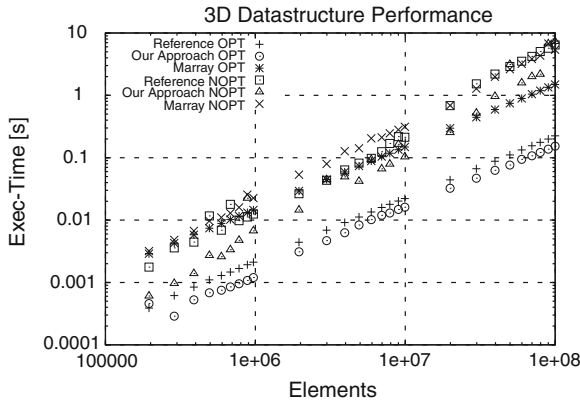


Fig. 5 A three-dimensional array structure is benchmarked. OPT and NOPT refers to optimal and non-optimal traversal, respectively. For 10^8 elements, our approach and Marray is a factor of 1.4 and 6.5, respectively, slower than the reference implementation. Non-optimal traversal significantly reduces the performance for all implementations

Execution times of our implementation for all presented dimensions are approximately equal, which is not only due to the utilization of a linear storage but also due to the compile-time based index evaluation algorithm. This fact underlines the applicability for high-dimensional data storage applications.

5 Conclusion

We have introduced a flexible, multi-dimensional run-time data structure. Our approach offers high extendability and can be applied in MPI based computing environments. The presented performance results depict that our access mechanism offers excellent performance for different dimensions and problem-sizes. The drawback of additional coding at the user-level access code will be rendered obsolete with the availability of variadic templates provided by the C++11 standard. Finally, our approach offers a small code base, as only around 100 code lines are required to implement the introduced functionality.

Acknowledgments This work has been supported by the European Research Council through the grant #247056 MOSILSPIN. Karl Rupp acknowledges support by the Austrian Science Fund (FWF), grant P23598.

References

1. Meyer Carl D (2001) Matrix analysis and applied linear algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, USA
2. Nemat-Nasser S (2004) Plasticity. Cambridge University Press
3. Alexandrescu A (2001) Modern C++ Design. Addison-Wesley Professional, Boston, USA
4. Stroustrup B (2012) Software development for infrastructure. Computer 45(1):47–58
5. Weinbub J, Rupp K, Selberherr S (2012) A generic multi-dimensional run-time data structure for high-performance scientific computing. In lecture notes in engineering and computer science: proceedings of the world congress on engineering (2012) WCE 2012 U.K , London, pp 1076–1081
6. Reis GD et al (2005) What is Generic Programming? In: Proceedings of the first international workshop on library-centric software design (LCSD), OOPSLA 2005, San Diego, CA, USA, pp 1–10
7. Abrahams D, Gurtovoy A (2004) C++ Template Metaprogramming. Addison-Wesley Professional, Boston, USA
8. Boost (2012) The Boost C++ Libraries.<http://www.boost.org/>
9. Andres B, Köthe U, Kröger T, Hamprecht FA (2010) Runtime-Flexible Multi-dimensional Arrays and Views for C++98 and C++0x. ArXiv e-prints, Technical Report.<http://www.andres.sc/marray.html>
10. Gregor D, Järvi J, Maurer J, Merrill J (2007) Proposed wording for variadic templates. Technical Report N2152=07-0012, ANSI/ISO C++ Standard Committee
11. Garcia R, Lumsdaine A (2005) MultiArray: A C++ library for generic programming with arrays. Softw Pract Exper 35(2):159–188
12. Veldhuizen TL (1998) Arrays in Blitz++. In: Proceedings of the second international symposium on computing in object-oriented parallel environments (ISCOPE), Santa Fe, NM, USA, pp 223–230
13. Friedmann E, Maman I (2012) The boost variant library.<http://www.boost.org/libs/variant/>
14. Gurtovoy A (2012) The boost metaprogramming library.<http://www.boost.org/libs/mpl/>
15. Guzman J, Marsden D, Schwinger T (2012) The boost fusion library.<http://www.boost.org/libs/fusion/>
16. GNU (2012) GNU Multiple precision arithmetic library (GMP).<http://gmplib.org/>
17. Stroustrup B (2000) The C++ programming language. Addison-Wesley, Boston, USA
18. Banahan M, Brady D, Doran M (1991) The C book. Addison Wesley, Boston, USA
19. Gregor D, Troyer M (2012) The boost MPI library.<http://www.boost.org/libs/mpl/>
20. Ramey R (2012) The boost serialization library.<http://www.boost.org/libs/serialization/>
21. Drepper U (2007) What every programmer should know about memory. Linux Weekly News

Application of Curriculum Design Maturity Model at Private Institution of Higher Learning in Malaysia: A Case Study

Chee Ling Thong, Yusmadi Yah Jusoh, Rusli Abdullah
and Nor Hayati Alwi

Abstract Capability Maturity Model (CMM) is applied as a process improvement model not only in software industry but also in education sector. This study proposes a maturity model, which constructed based on CMM, in guiding curriculum designers in Institution of Higher Learning (IHL) in Malaysia to design quality curriculum. The proposed maturity model possesses process and product elements; and it contains a set of key process areas and best practices. A case study is carried out in a private IHL in Malaysia to perform a pilot test on the proposed maturity model. The results may also help the institution to be informed of current as well as future improvement process, finally aid in producing quality curriculum for IHL in Malaysia.

Keywords Capability maturity model · Curriculum design · Curriculum design process · Curriculum designer · Institution of higher learning · Quality curriculum

1 Introduction

In software development, the process of developing software product is important as it helps to produce quality product. Similarly, the process of designing curriculum is

C. L. Thong (✉)

School of IT, Faculty of Business and Information Science, UCSI University, No. 1, Jalan Menara Gading, UCSI Heights,
46500 Kuala Lumpur, Cheras, Malaysia
e-mail: chloethong@ucsi.edu.my

Y. Y. Jusoh and R. Abdullah

Department of Information System, Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
e-mail: yusmadi@fsktm.upm.edu.my

N. H. Alwi

Department of Educational Foundation, Faculty of Educational Studies,
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
e-mail: nalwi@putra.upm.edu.my

crucial, as it helps to produce quality curriculum. Curriculum design and software development bear some similar common features. Both possess complex activities and development life cycle, which emphasize on design quality. The success of both domains is attributed to the good structure and used of best practices which consider a process that helps us to structure and do things right [1, 2]. In this study, a process is defined as a domain with a set of activities, practices and transformation that faculty may use to improve the quality of the curriculum [3].

2 Background

2.1 Related Work

Quality assurance in higher education is a relatively relevantly recent development in most ASEAN countries [4]. One aspect of quality assurance of higher education which is teaching-learning particularly for curriculum design and maintenance has been categorized as field/area critical to the development of Malaysia [5]. With reference to the Malaysian higher education sector, the implementation of Malaysian Qualification Framework (MQF) and establishment of Malaysian Qualification Agency (MQA) are seen as measures taken to improve the overall quality of higher education in meeting global standards [6]. The quality assurance has been a critical agenda stated in the document namely National Higher Education Action plan [7]. An increase in quality programmes has become a key item that being emphasized in the National Higher Education Action Plan Phase II (2011–2015) and designing of quality programme (curriculum) has become the responsibility of educators who are also known as curriculum designers. In this study, a process improvement model is constructed and it is used for providing guidance to curriculum designers in the areas of improving the design process looking from both the institutional (faculty) and programme aspects; subsequently produce quality curriculum. Once the model is constructed, it is applied in a private IHL as pilot testing of the model. The results are presented at other sections in this chapter.

2.2 The Curriculum Design Maturity Model

CMM is a maturity model used in software engineering and it was originally developed in the 1980s by the U.S. Department of Defense Software Engineering Institute (SEI) at Carnegie Mellon University as a method for objective evaluation of contractors for military software projects [8]. The key to CMM model is it is designed to provide good engineering and organizational management practices “for any project in any environment” [9]. It achieves this through a structure that breaks each level into a number of process areas. Each of these areas is in turn organized into a number

Table 1 Capability maturity model [11]

Maturity level	Brief description
Initial	The software process is characterized as ad hoc, and occasionally even chaotic. Few processes are defined, and success depends on individual effort and heroics.
Repeatable	Basic project management processes are established to track cost, schedule, and functionality. The necessary process discipline is in place to repeat earlier successes on projects with similar applications.
Defined	The software process for both management and engineering activities is documented, standardized, and integrated into a standard software process for the organization. All projects use an approved, tailored version of the organization’s standard software process for developing and maintaining software.
Managed	Detailed measures of the software process and product quality are collected. Both the software process and products are quantitatively understood and controlled.

of sections called common features, which are used to organize the key practices that accomplish the goals of relative process areas [5]. Refer to Table 1 for CMM.

There are many maturity models constructed based on CMM in the past to alleviate the shortage of quality standard in education. Among the various maturity models, some are process improvement models used by higher education to support online course design and curriculum design for Information system education, while others are process improvement models for other areas such as e-learning [3, 10, 11]. Even though there are many maturity models constructed to support and improve quality design process, there is a shortage of maturity model providing self-guidance to curriculum designers so that they know which maturity level they are in and whether the processes has reached maturity level and guided to complete the process before moving to next level.

The construction of Curriculum Design Maturity Model (CDMM) is based on two models, which are online course design maturity model (OCDMM) and curriculum redesign process improvement model for Information system education proposed by Dennis and Minnie [10] and Neuhauser [3] respectively. CDMM defines five levels of process capability and each process is broken up into a set of key practices, which are assessed at each level using a five-point scale (not adequate, partially adequately, moderately adequate, adequate and fully adequate). The process capability of CDMM is presented in Table 2 and the assessment criteria are stated in Table 5. For the interest of this study, the processes are assessed using a holistic approach and a single result is obtained instead of detailed analysis.

CDMM consists of key process areas (KPA) which further breakdown into managerial and technical processes (refer to Table 3). This study assumes that the institution has arrived at repeatable maturity level after assessing their curriculum (re)design processes using criteria set based on benchmark standards [12, 13] and the technical process of this level i.e. curriculum assurance quality is also the focus of this study.

Table 2 Curriculum design maturity model—levels of process capability [2]

Maturity level	Brief description
Initial	The curriculum (re)design processes are characterized as ad-hoc and occasionally is even chaotic. Lacking in policies and practices for controlling curriculum (re)design process. Although a few processes are defined, the success depends on individual effort.
Repeatable	Basic curriculum (re)designs processes are established such as develop key practices that allow it to repeat success and discard those that hindered success. It is restricted to course level rather than broader programme level. The focus of this level is to design a clear and measurable learning outcome for each course.
Defined	The curriculum (re)design process for course-level activities is documented, standardized and integrated into a standardized design process for the program. All courses use an approved, tailored version of programme's standard curriculum design process for designing and redesigning curriculum. The focus of this level is to define standardized process at course level, alignment of key process areas within the courses itself and between streamlined courses to the programme.
Managed	Detailed measure of curriculum (re)design process and curriculum quality is collected. Both curriculum (re)design process and curriculum (product) are quantitatively understood and controlled. The focus of this level is to ensure both the quality (alignment) of course learning outcomes within the courses and programme learning outcome.
Optimized	Continuous process improvement is enabled by quantitative feedback from the (re)design process and from piloting innovative ideas. The focus of this level is continual improvement in the KPAs of the curriculum (re)design process.

2.3 The Curriculum Design Maturity Model-1 (CDMM-1)

The technical process at repeatable level of CDMM can be further developed to a stand-alone staged-based improvement model with a five-maturity levels structure and it is named as Curriculum Design Maturity Model-1 (CDMM-1). This name is derived from CDMM. The five levels of process capability are shown in Table 4.

Table 2 presents the process portion of CDMM and it is used to assess maturity level in (re)designing a programme of an academic department i.e. faculty. CDMM was structured in five maturity levels, each of which associated key process areas. However, the product portion of a CDMM-1 is inverted. It contains product attribute named key product quality attribute (KPQA). Figure 1 shows the key activities and goals of KPA and its relationship.

There are some best practices and design principles used to design curriculum particularly in setting learning outcomes. This study introduces the standardized way of setting learning outcomes of study areas. A learning outcome contains an “*action key verbs*” followed by “*content/topics*” of the study area (domain). For example, the course learning outcome (CO) of a software engineering module entitled “object-oriented modeling” is: upon completion of the module, students are able to *describe the basic concept of elements of use cases and class diagrams*. One of the

Table 3 Key process areas: managerial and technical processes

Maturity level	Key process area (KPA)
Initial	Initial (Ad Hoc, Chaotic)
Repeatable	Repeatable (Disciplined Process) Technical processes: <ul style="list-style-type: none"> ● Curriculum Quality Assurance Managerial processes: <ul style="list-style-type: none"> ● Curriculum (Re)-Design Tracking ● Curriculum (Re)-Design Planning ● Requirement Management
Defined	Defined (Standard, Consistent Process) Technical processes: <ul style="list-style-type: none"> ● Peer review Managerial processes: <ul style="list-style-type: none"> ● Intergroup coordination ● Integrated curriculum management
Managed	Managed (Predictable process) Managerial processes: <ul style="list-style-type: none"> ● Curriculum quality management ● Quantitative process management
Optimized	Optimizing (Continually improving process) Technical processes: <ul style="list-style-type: none"> ● Defect detection Managerial processes: <ul style="list-style-type: none"> ● Process change management

characteristics of a good CO should be mapped to the learning domains in Bloom or other taxonomy [13].

3 Methodology

The proposed CDMM-1 was developed based on three basis: (1) review of existing CMM that related to curriculum design, (2) experience from the authors of the paper, (3) document analysis and literature review. The main aim of constructing the model is to provide guidance to curriculum designers in order to produce quality curriculum during the (re)design process. The domain used in this study is software engineering curricular. In order to ensure the proposed model remains up-to-date, each maturity level is descriptive and requires iterative refinement.

Table 4 Curriculum design maturity model-1

Maturity level	Key product quality attribute Curriculum alignment	Key activities /Decision points
Initial	Intended learning outcomes (ILO) lack of clarity and partially based on Bloom's taxonomy (or other taxonomy)	Identify ILO which lack of clarity and do not based on Bloom's taxonomy (or other taxonomy)
Repeatable	Clarity of ILO increases, and key action verbs are applied at this level; learning domains are also included to reflect different aspects of student learning.	Identify ILO which includes MQA LO domains, and course learning outcomes (CO). Ensure MQA LO domains, and CO are aligned using key action verbs
Defined	Bloom's taxonomy has fully applied in designing ILO; alignment of TLAs and ATs to ILO happened at this level	Identify possible TLAs and ATs based on other taxonomy (i.e. Solo Taxonomy). Align TLAs and ATs with ILO based on taxonomy table
Managed	Measure of alignment of ILO, TLAs and ATs using template is collected	Alignment of ILO, TLA and ATs using alignment matrix
Optimized	Continuous improvement of the alignment of ILO, TLAs and ATs and students performance at instructional level to gauge the effectiveness of alignment	Monitor of achievement of learning outcome using alignment matrix

4 The Results

4.1 Results to Pilot Test of CDMM

The model was tested to a bachelor degree level programme, which is offered nationally and internationally. This programme introduced in year 2003 and has been upgraded on regular basis. In recent years, the faculty also looked at the constructive recommendation given by the industry advisory board and external examiners so that the programme is instituting better coverage of the lacking areas in the programme. These are several programme improvement opportunities surfaced with this exercise. The results from the pilot test of CDMM are presented in Table 5.

A classification table is set in order to measure the various levels of maturity the IHL is in and whether curriculum redesign process has reached the maturity in that particular level (Table 6).

The results of the pilot test shows that initial level reflects the ad-hoc nature of curriculum redesign process and the support around it. For example, curriculum designers redesign courses independently and the guidance provided by the institution is limited during the redesign process. However, some relevant policies and practices are provided and explained by institution for controlling the curriculum redesign process. This shows that the institution is still at the initial maturity level in

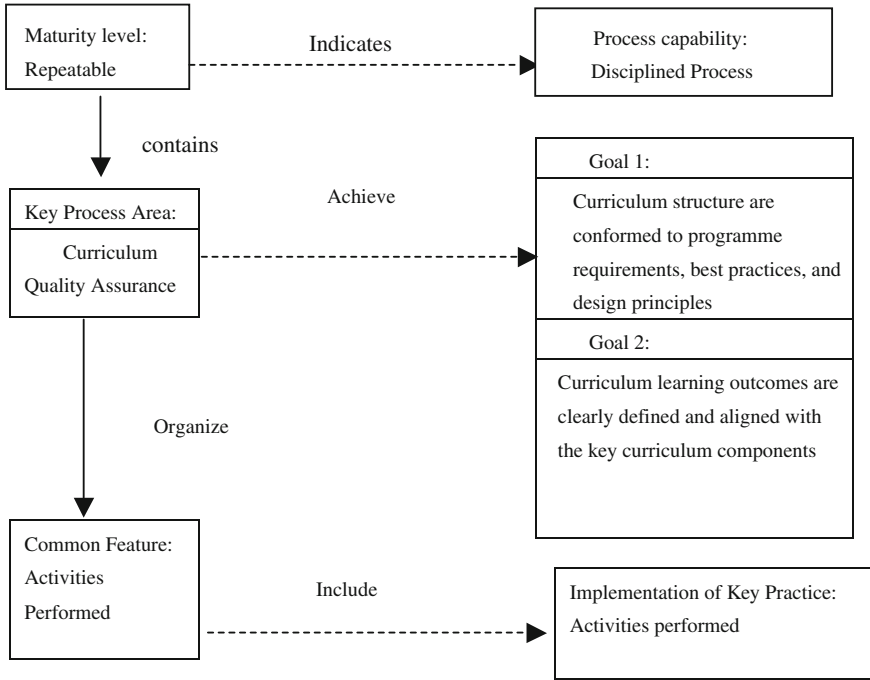


Fig. 1 CDMM-1- Technical Process of Key Process Area for Level 2: Curriculum Quality Assurance

curriculum redesign process. In order to move to the next level, initial level needs to reach maturity level by fulfilling the assessment criteria set at this level.

At the repeatable level, basic curriculum redesign process may establish. The institution is rather weak in repeatable level as basic management of curriculum redesign process has yet to be established such as course level redesign guidelines and compilation of best practices for the process. At this level of maturity, the process was still centered on the individual courses. According to the guidelines of CMM, in order to reach the next level, it is necessary to encompass the entire programme.

In order to fill up the limitation in repeatable level, defined level is introduced. Defined level is not focus on the success of individual courses, but the entire programme. It is also addressed the concern about alignment of programme learning outcomes to educational objectives and vision and mission of the IHL. Faculty is more involved at this level to the programme-wide coordination to meet the educational goals set by the IHL. At the same time, peer review process may be introduced to ensure the alignment among the streamlined courses in terms of course learning outcomes and content. For example, the first object-oriented modeling course (OOM1) may introduce initial object-oriented concepts used in analysis and design; and second object-oriented modeling course (OOM2) may provide students with an

Table 5 Template for assessment of process capability and results from pilot test: level 1 [2]

Initial: Ad-hoc or chaotic management of curriculum redesign process		
Key practices at initial level		Scale
A1.	Courses are redesigned based on systematic and planned process	3
A2.	Institution clearly communicate how curriculum re-design processes should be used during the courses or programme redesign	2
A3.	Guidelines in managing and defining curriculum redesign process are provided by institution	2
A4.	Curriculum designers are guided in a proper methods of redesigning curriculum by faculty	2
A5.	Curriculum designers are person who have the subject knowledge and experience to deal with curriculum (re) design	3
A6.	Courses reviewed and redesigned interdependently	2
A7.	Courses are redesigned to reflect the programme objectives and learning outcomes	3
A8.	Quality and scope of each course are based on objectives set collectively by curriculum designers	2
A9.	Use of quality manual to ensure revision of existing programmes is accordance with quality requirements of the institution	2
A10.	Policies and practices are provided and clearly explained by institution for controlling the curriculum redesign process	3

Note 1 = not adequate, 2 = partially adequate, 3 = moderately adequate, 4 = largely adequate, 5 = fully adequate

Table 6 Classification table

Average rating	Description
Less than or equal to 3.0 (≤ 3.0)	Key practices are moderately, partially and not adequate. Therefore, key practices need to be established and not able to move on to next maturity level
More than 3.0 (> 3.0)	Key practices are largely and fully adequate. Therefore, key practices are established and able to move on to next maturity level

understanding of more advanced object-oriented analysis and design concepts and principles, with particular reference to Unified Modeling Language (UML).

Next, we began moving towards attained the “Managed” level of CDMM. Metrics can be used as a statistical quality control measures. Based on the study, quantitative evaluation of programme quality has not been practicing by the institution. It is suggested this process be evaluated based on the achievement of programme learning outcomes and present in the form of metrics quantitatively. Some measurement instruments such as standard evaluation form is developed for normalizing alignment of courses.

At optimized level, the institution will be focusing on continuous process improvement. At this level, the curriculum committee has been formed to identify three to five areas that will be targeted for improvement in the academic year [3]. The

Table 7 Object-oriented modeling 1 (*Source: SE programme in a private university*)

Curriculum element (ILO)	Description
Course learning outcomes (CO) - expressed in MQA LO Domain)	Upon successful completion of this course, students are able to:
Knowledge	To <u>identify</u> the basic components of use case and class diagrams
Practical skills	To <u>draw</u> a series of unified modeling diagrams such as class diagram, sequence diagram, state chart, activity diagram and use case diagram
Social skills and responsibility	NA
Ethics, professionalism and humanities	To <u>describe</u> the need for a standard professional approach in the use of modeling techniques To <u>participate</u> in applications development to specify non-trivial but non-complex application domains using use case and class diagrams
Communication, leadership and team skills	To <u>involve</u> in group assignment to allow team management and collaborative work
Scientific methods, critical thinking and problem solving skills	To <u>determine</u> when, how and where use cases and class diagrams are used in the modeling of information systems
Lifelong learning and information management	To <u>judge</u> the importance of analysis in the process of software development
Entrepreneurship and managerial skills	To <u>understand</u> the importance of project management and the different challenges in change management

institution involved in this study, has instituted some measures across programme learning outcomes in response to accreditation standard.

4.2 Results to Pilot Test of CDMM-1

The CDMM-1 is further tested to the same bachelor degree programme offer by the private IHL. This model is applied in one of the modules entitled object-oriented modeling 1 (OOM1) which is a year one module in software engineering (SE) curricular. The key product quality attribute presented in this study is alignment. As presented in Table 4, key activities are served as decision point for curriculum designers to know what to do and how to create an aligned curriculum.

At initial level, intended learning outcomes (ILO) of OOM1 is lack of clarity and partially based on Bloom’s taxonomy. The course learning outcomes (CO) are stated in Table 7.

Key activities at initial level are identifying ILO that lack of clarity and not basing on Bloom’s taxonomy. CO are examined by asking question such as can it be measured? It is discovered that not all key action verbs in CO are measurable and

only partially based Bloom's taxonomy. The unclear and immeasurable key action verbs could be fixed by using Bloom's taxonomy.

After applying Bloom's taxonomy at initial level, moving on to the next level i.e. repeatable level. At this level, clarity of ILO increases, and key action verbs are applied, learning domains are also included to reflect different aspects of student learning. Key activities at this level are identifying ILO, which includes MQA LO DOMAINS, and CO. According to Bloom, if taxonomy is regarded as a useful and effective tool, it must be accepted and used by the curriculum workers (designers) [14].

The goal set for this level is Bloom's taxonomy has to be fully applied in designing ILO; alignment of TLAs and ATs to ILO is happened at defined level. Key activities at defined level are identifying possible TLAs and ATs based on other taxonomy (i.e. Solo Taxonomy). At this level, curriculum designers need to review the TLAs and ATs stated in OOM1.

Then at managed level, alignment of ILO, TLAs and ATs are in place. Curriculum designers are able to view the entire course whether CO is aligned to TLAs and ATs using the alignment matrix. Detailed measures of product quality are collected and quality of the product is quantitatively understood and controlled at this level.

The optimizing level requires continuous improvement of alignment between ILO, TLAs and ATs is needed and student's performance is used at instructional level to gauge the effectiveness of alignment. Continuous quality improvement (CQI) results need to be communicated to the academic staff for further improvement. Student performance tells the effectiveness of the alignment and continuous improvement process iteratively [15].

5 Discussion

This study covers two pilot tests to CDMM and CDMM-1 respectively. Through the first pilot test, it is discovered the institution is still at initial maturity level. Although the institution has implemented some key practices such as courses are reviewed based on programme standard and guidelines provided, it is not yet fully adequate. However, CDMM introduces in this study could help to provide a roadmap for IHL to improve curriculum redesign processes. Most curriculum designers use ad-hoc approach where the redesign process has more to do with individual work rather than institutional planning. The advantage of this report is that it outlines an overview of overall curriculum redesign process without examining the detailed report. Furthermore, CDMM enables the faculty to enhance the ability to gauge its ability and prioritize necessary improvements in its current practices. In summary, it is discovered that there might be a great need for the institution to establish quality review process for curriculum redesign and ensure quality curriculum is developed.

It is believe that in the absence of guidance, authors do what they think is right [16]. Through the pilot test of CDMM-1, it is discovered that OOM 1 is not considered as a quality module embedded in SE programme. The reasons are some of

the CO statements are unclear and need to be fixed according to Bloom's taxonomy; furthermore CO does not cover all the three main learning domains introduced by Bloom (cognitive, affective and psychomotor domains). TLAs and ATs are also chosen based on curriculum designers' personal experience and belief on what is the best, but this practice is insufficient. However, by using CDMM-1 in guiding curriculum designers step-by-step, quality work at design level may be produced.

6 Conclusion and Future Work

The proposed models are tested on the private IHL as a pilot study. The results might vary as time goes by as some key practices might be adopted overtime by the institution. This study is a starting point and continuous improvement is needed to ensure that the model helps the curriculum designers to develop quality curriculum. For future work, participation from other IHLs may be needed to validate the model and also to identify more practices and improvement processes.

Acknowledgments This work was supported by Universiti Putra Malaysia, under the Research University Grant (RUGS:9308600).

References

1. Christof L, Andrew LR, Gilian D, John H (2007) A maturity model for computing education conference in research in information technology (CRPIT). Australian Computer Society, Inc. Ninth Australasian computing education conference (ACE2007), Ballarat, Victoria, Australia, Feb 2007
2. Thong CL, Jusoh YY, Abdullah R, Alwi NH (2012) Applying capability maturity model to curriculum design: a case study at private institution of higher learning in Malaysia, lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2012, WCE 2012, U.K., London, 4–6 July, pp 1070–1075
3. Neuhauser C (2004) A maturity model: does it provide a path for online course design? *J Interact Online Learn* 3(1):17 <http://ncolr.org>
4. Ng MLY, Mustafa R (2012) Higher Education and Human Capital Development Malaysia and CLMV: Towards Strategic Partnerships and Alliance. Higher Education Monograph 17/2012. Chap. 3:39–70
5. Ng MLY, Tan C, Rahman SA, Abdulllah NA, Kaur S (2012) Higher Education and Human Capital Development Malaysia and CLMV: Towards Strategic Partnerships and Alliance. Higher Education Monograph 17/2012. Chap. 5:145–157
6. Morshidi S (2008) Trends in international higher education and regionalism: issues and challenges for Malaysia. Paper presented at the international symposium on Asian cooperation, integration and human resources, Waseda University, Tokyo, Japan
7. MOHE (2011) National higher education action plan 2007–2010: triggering higher education transformation. Kementerian Pengajian Malaysia, Malaysia
8. Paulk M, Weber C, Curtis B, Chrissis MB (1995) The capability maturity model: guidelines for improving the software process. MA: Addison-Wesley
9. Hafeez M (1999) Application of SPICE (ISO/IEC 15504) in an Academic Environment. <http://citeseer.nj.nec.com/499756.html>

10. Dennis D, Minnie YY (2008) Controlling curriculum redesign with a process improvement model. *J Inf Syst Educ* 19(3):331–342
11. Marshall S, Mitchell G (2002) An e-learning maturity model. In: Proceedings of EDUCAUSE '02: 19th annual conference of Australian society for computers in learning in tertiary education (Auckland, Australian Society for Computers in Learning in Tertiary Education, 2002)
12. Malaysian Qualification Agency (2008) Code of practice for programme accreditation (COPPA). Petaling Jaya, Malaysia
13. Malaysian Qualification Agency (2010) Programme standards: computing. Petaling Jaya, Malaysia
14. Bloom BS (1956) Taxonomy of educational objectives, handbook I: the cognitive domain. David McKay Co Inc., New York
15. Squires DA (2005) Aligning and balancing the standard-based curriculum. Corwin Press, California, Thousand Oaks, pp 57–59
16. Maier AM, Moultrie J, Clarkson JP (2012) Assessing organizational capabilities: reviewing and guiding the development of maturity grids. *IEEE Trans Eng Manage* 59(1):138–159

Reducing Job Failure Due to Churn in Dynamics Grids

K. Abdelkader and J. Broeckhove

Abstract The utilization of desktop grid computing in large-scale computational applications is an important issue at present for solving compute-intensive problems. However, such large-scale distributed systems are subject to churn, i.e., continuous hosts arrival, leaving and failure. We address the problem of churn in dynamic grids, and evaluate the impact of reliability-aware resource allocation on the performance of the system.

Keywords Auction market · Desktop grid · Grid economics · P2P · Resource management · Spot markets

1 Introduction

The utilization of desktop grid computing in large-scale computational applications is an important issue at present. Platforms such as BOINC [1] and SZTAKI [2] that used to provide large-scale intensive computing capability have attracted recent research interest. Such platforms are referred to as volunteer grids or public resource grids [3] where the hosts or providers are typically end-users' public PCs (e.g. homes, offices, universities or institutions) located at the edge of the Internet. Studies aimed at evaluating host availability have shown that providers connect to and disconnect from the grid without any prior notification. This effect which is called churn [4]. To optimize the performance of the system that is subject to churn, we shall consider the churn characteristics of providers i.e. the rate of connection/disconnection and the

K. Abdelkader (✉)

The Higher Institute of Comprehensive Professions, Anahda 1, Ghadames, Libya
e-mail: abdelkader.khalid@gmail.com

J. Broeckhove (✉)

University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium
e-mail: jan.broeckhove@ua.ac.be

duration of the corresponding time periods. Also, it is significant from the perspective of the grid user, to consider the number of jobs failing and succeeding without resubmission being required [5]. These effects have to be taken into account when devising an appropriate allocation policy for such systems.

Our main contribution is to focus on how to reduce the effect of churn by reducing jobs failure due to system downtime prior to job completion. We look for allocation strategies that are economically based, but that also take into account the apparent reliability of the providers. We aim to do this in a manner as simple as possible, by taking into account historical information of the providers and using this to screen the bids they submit to the market.

2 Modeling Churn

The model adopted for the churn in the dynamic grid system plays a key role. Basically one models the distribution of the time durations during which a resource, in this case a computing system, is available or unavailable. This is similar to the system-based churn model as described in [6]. One can also look at the availability of the CPU, which might differ, due to the provider policies or preferences, from the machine availability.

The model adopted for the churn in the dynamic grid system plays a key role. Basically one models the distribution of the time durations during which a resource, in this case the provider's computing system, is available or unavailable. This is similar to the system-based churn model as described in [6]. One can also look at the availability of the CPU, which might differ, due to the provider policies or preferences, from the machine availability.

The provider availability is a binary value that indicates whether a provider is reachable and responsive. This corresponds to the definition of availability in [7, 8]. By contrast, the authors in [9] addressed the problem of *CPU* availability instead of provider or host availability. Of course provider unavailability implies *CPU* unavailability, but the converse is not true, particularly when multi-processor and or multi-core machine are involved. The state transition of provider availability to unavailability and back is depicted in Fig. 1 that represents a timeline with the following time intervals:

- **uptime** stage: a provider is available and when a job is allocated to the provider it uses all the CPU power of that provider
- **downtime** stage: a provider has withdrawn from the grid system because of policy decision, shutdown, . . . etc

Churn is modeled with two provider-level characterizations. Firstly, the *uptime* length distribution, which is one of the most basic properties of churn. It captures the period of time that the providers remain active in the grid system each time they appear. Secondly, the *downtime* can be defined as the interval between the moment

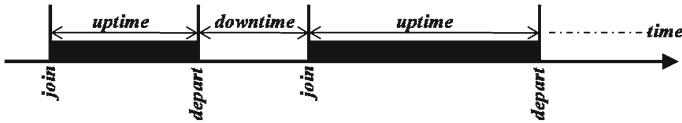


Fig. 1 Provider availability and unavailability time periods indicated in the *uptime* and *downtime* intervals

a provider departs from the grid system and the moment of its next arrival in the system (see Fig. 1). Most churn studies use these two distributions to model a churn. A provider’s *lifetime* is the time between the moment that a provider first participates in the grid system and the time that a provider finally and permanently exits from the grid system.

We will in the remainder of this chapter refer to the provider’s reliability over an interval of elapsed time $[t_a, t_b]$. Given uptime periods $uptime(k)$ for $k = 0, 1, 2, \dots$ starting respectively at time t_k^i and ending at t_k^f , the reliability is defined as the aggregate of the uptime periods within that time interval, divided by the elapsed time

$$R = \frac{\sum_k \max(t_b, t_k^f) - \min(t_a, t_k^i)}{t_b - t_a} \tag{1}$$

The reliability index R thus represents the fraction of the elapsed time that the provider was available in the time interval from t_a to t_b . Obviously, the higher R the more available the provider. This is a narrow definition that equates reliability with provider availability.

2.1 Simulating Churn

Many types of distributions have been used in the literature for the uptime and downtime: exponential, Pareto, Weibull, Log-Normal, etc. In our simulation work, we will model churn using Weibull distributions. The Weibull distribution function (“WD” for short henceforth) is given by

$$F_w(x) = 1 - e^{-(x/\lambda)^\alpha} \tag{2}$$

The α and λ are called the *shape* parameter and *scale* parameter respectively.

We have performed simulations using parameters taken from [10]. These authors have analysed large number of data sets representing availability/unavailability in different settings. They have then fitted various probability distributions for uptime and downtime to the data. They also have performed goodness of fit (GOF) analyses on those fitted distributions. These indicate that by and large the Weibull distributions

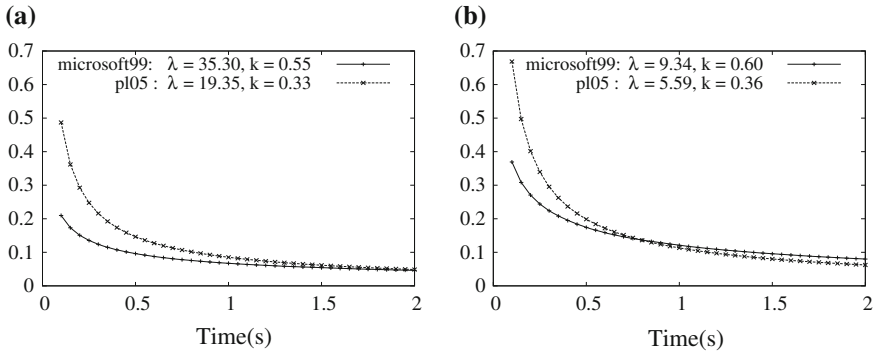


Fig. 2 Weibull probability density function of the host up- and downtimes. **a** Uptime. **b** Downtime

equation in (2) provides the best fit according to the GOF metric. We therefore have selected the fits with the Weibull distributions for use in our simulations.

We have used fits that correspond to two data sets. The `microsoft99` is a data set containing log files of thousands desktop PCs at Microsoft Corporation [11]. The authors of [10] have measured machine availability by pinging these desktop every hour. This data set corresponds to desktop grid setting, where a large organisation harnesses the combined computing power of its desktop systems for large scale computing. The `pl05` data set contains trace data measured between all pairs of PlanetLab nodes [12] using pings [13]. This corresponds to a peer-to-peer (P2P) setting where a loose amalgamation of systems, belonging to different organisations, coordinates computing tasks.

Figure 2 shows the plot of the Weibull probability density function with the k and λ parameters indicated in Table 2 for `microsoft99` (Fig. 2a) and `pl05` (Fig. 2b) data sets respectively.

Table 1 Notations used in Sect. 3

Notation	Meaning
μ_i	<i>CPU</i> speed
c_n	<i>CPU</i> cost
$R_{n,j}$	Resource of provider n
U_m	The i^{th} consumer in the grid
P_n	The n^{th} provider in the grid
$J_{j,m}$	j th job from m^{th} consumer
$b_{j,m}$	The bidding price for $J_{j,m}$
$l_{j,m}$	Job length for $J_{j,m}$
$T(J_{j,m,n}, R_i, n)$	Time required for $J_{j,m,n}$ at resource n
$B(J_{j,m,n}, R_i, n)$	Cost required for $J_{j,m,n}$ at resource n

3 GES Model

We present the model and scenarios used in the Grid Economic Simulator to analyse methods for alleviating the effects of churn in dynamic grid systems. In Table 1 one finds some of the notations that are used.

The simulation model consists of three key elements. Firstly, a set of N geographically distributed grid providers denoted by P_1, P_2, \dots, P_N , each of which is committed to deliver computational power. In addition, there is a group of potential providers, providers that are in a waiting state but that are ready to join and deliver computational resources in the grid.

Secondly, a market for resource allocation and job scheduling. All resource owners follow the same pricing strategy for determining the outcome of the bidding process in the market. The market will be discussed in more detail in Sect. 3.

Thirdly, M resource consumers or “users” denoted by U_1, U_2, \dots, U_M , that are also geographically distributed and each has a queue of jobs to be executed. A job is specified by its job length and budget and is used to acquire resources for the job execution. In our simulation, the job lengths are randomly selected from a uniform distribution. The jobs are a CPU-bound computational task. The consumers are subdivided into four groups such that each has different deadlines for the jobs to be finished. That is, the jobs of each group have to be completed before the deadline of that particular group with the initial budget that has been allocated to the job.

The consumers interact with resource brokers that hide the complexities of grids by transforming user requirements into schedules for jobs on the appropriate computing resources. The brokers also manage the job submission and collect the results when they are finished. For instance, the consumer U_m , where $(m = 1, 2, \dots, M)$ sends the job $J_{j,m}$ with its bidding value $b_{j,m}$ to the broker. In accordance to consumer’s request, one of the available resource providers P_n , where $(n = 1, 2, \dots, N)$ will receive this request. The broker plays a complex role of a wide range of tasks, i.e. it is an intermediary between resource providers and consumers. Beside this functionality, the broker provides information about the status of CPU usage in the grid system.

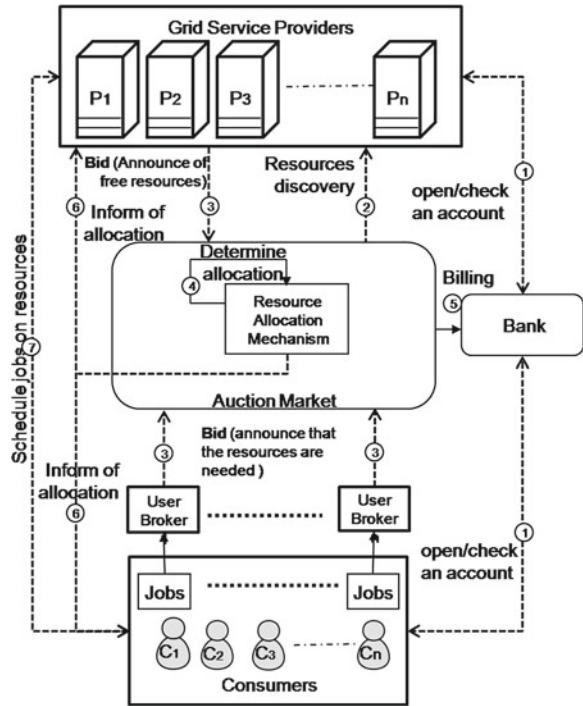
Finally, an entity that functions as a bank is used. When a job failure occurs, the broker will send a report to the bank. The bank refund the money that had already been prepaid by U_m to the account number of P_n . This enables the consumer to recover the money and use it to resubmit the failed jobs.

In Fig. 3 one finds a graphical representation of all the entities involved in the model and the key steps in the flow of control.

3.1 Decentralized Marketplace

As mentioned previously, GES applies market-based principles for resource allocation to application scheduling. In this section we describe how resources are priced.

Fig. 3 An overview of the auction market architecture in GES



In effect, we adopt an auction market for the pricing of computational resources. In contrast to the previous work [14], where the motivation was focused on price stability using a commodity market, the auction market has been engineered to be more realistic in geographically distributed systems.

The decentralized auction is a first-price sealed-bid (FPSB) auction with no reserve price (see the diagram 3), where the auctioneer collects all received bids and subsequently computes the outcome. The bidders can submit only one bid. The resource is allocated to the highest bidder at the end of the auction. Consumer U_m typically, has its own valuation “ $v(R_{i,n})$ ” for resource $R_{i,n}$ to bid. The consumer U_m communicates its willingness to pay ($b_{j,m}$) for resource $R_{i,n}$ and the required processing time for job, $J_{j,m}$. The resource information concerning resource $R_{i,n}$ of provider P_n consists of information about the *CPU* such as the *CPU speed*, μ_i . The information on job J_j consists of the job length, l_j , the job deadline d_j and the budget B_j available for execution of the job.

When the auction ends by awarding the highest bid, the auctioneer charges the winner an amount of c_n (i.e. $c_n = (b_{j,m})$) per time step of the job for resource usage. That is, the charges are determined at allocation time and remain fixed throughout job execution. The required time for the $J_{j,m}$ to execute on $R_{i,n}$ and the associated cost are computed using the Eqs. 3 and 4 respectively.

$$T(J_{j,m}, R_{i,n}) = \frac{l_{j,m}}{\mu_i} \quad (3)$$

$$B(J_{j,m}, R_{i,n}) = c_m * T(J_{j,m}, R_{i,n}) \quad (4)$$

3.2 The Role of Churn

When churn is incorporated in the above model, leading to a dynamic grid system, jobs may fail because the provider withdraws from the grid system. Also users can withdraw from the grid system. The churn itself is modeled through system uptimes and downtimes as explained in Sect. 2. As a matter of fact, we have only included provider churn in the model. In case a consumer withdraws from the market, their results do not get recovered by them and they will have to resubmit their jobs. There is however no impact on the functioning of resource allocation of the grid system as a whole.

When providers withdraw from the market, jobs execution fails. Consequently, consumers have to be reimbursed; they have to resubmit their jobs. In an effort to meet their deadlines the consumer may have to increase their bids. Thus it is necessary to try and submit the jobs to a provider that is not likely to fail. In the next section we propose such an approach.

3.3 Alleviating the Effect of Churn

In order to alleviate the adverse effect of churn on the usefulness of the grid system we propose a simple, straightforward algorithm that only exploits the historical information on uptimes and downtimes. This information is maintained in the grid information system and is certainly not privileged information. The basic quantity that is used in the algorithm is the reliability index introduced previously (Eq. 1) for provider P_i in the time interval from t_a to t_b

$$R(P_i) = \frac{\sum_k \max(t_b, t_k^f) - \min(t_a, t_k^i)}{t_b - t_a} \quad (5)$$

where t_k^i and t_k^f are the start and end times respectively of the k^{th} uptime period of P_i in that time interval. The reliability index thus represents the fraction of the elapsed time that the provider was available in the time interval from t_a to t_b , and characterizes each provider individually. We also consider a threshold number which we denote γ and refer to as the `reliability threshold`. This threshold is applied in a very straightforward manner: consumer will only bid with providers P_i such that

$$R(P_i) \geq \gamma \quad (6)$$

That is, providers whose reliability index exceeds gamma. This has the effect of screening the less reliable providers. It is of course a simple heuristic rule, based on information of the provider's track records. It does not involve any statistical evaluation nor does it evolve any quality of service elements. It does however lead to quantitative improvements of the job failures rates due to churn as we shall see from the simulation results in the next section.

4 Simulation Results

We evaluate the effect of churn in the above model through two simulation scenarios. The key parameters that apply in all scenarios discussed in this section are listed in Table 2. The relative workloads on the grid system that are assigned are roughly the same for those scenarios that we compare with and without churn. The relative workload, ϕ , can be defined as the ratio of aggregate workload, l , to the aggregate computational capacity, α .

$$\phi = \frac{\sum_m l_m}{\sum_n \alpha_n} \quad (7)$$

This relative workload ϕ highly affects the system performance. Because of this, scenarios where we directly compare the effects of churn by executing the scenario with and without churn have been designed to run under the same relative workloads.

Table 2 Simulation parameters

Simulation steps	2000 _s	
Number of consumers	6000	
Number of providers	1000	
Initial budget	1000000	
$d_i(\text{Group1})$	{1, ..., 100}	
$d_i(\text{Group2})$	{1, ..., 1300}	
$d_i(\text{Group3})$	{1, ..., 1700}	
$d_i(\text{Group4})$	{1, ..., 2000}	
microsoft99 Scenario	K	λ
uptime	0.55	35.30
downtime	0.60	9.34
pl05 Scenario	K	λ
uptime	0.33	19.35
downtime	0.36	5.59

In all experiments we have divided the users into four groups, each characterized by a different range of deadlines to be met for the jobs that the user needs to have executed (see Table 2). In some of the experiments we have also subdivided the providers into groups with different parameters for the uptime churn distribution. Simulation parameters that differ per scenario are reported in the appropriate subsection.

4.1 Performance Metric

The performance objective for consumers is a high rate of successfully executed jobs within the deadline. Jobs may fail to meet this objective because of provider churn but also simply because the aggregate computational load on the grid system is such that the consumer fails to acquire the necessary computational resources within deadline and within their budget constraint. We are interested in studying the former effect.

In order to discriminate between both effects we define the *failure rate* (F_{rate}) with the following equations:

$$F_{rate} = \frac{FJ}{TJ} - F_{rate_{noChurn}} \quad (8)$$

$$FJ: \text{number of jobs fail} \quad (9)$$

$$TJ: \text{number of total jobs} \quad (10)$$

$$F_{rate_{noChurn}}: \text{failure rate without churn} \quad (11)$$

Thus, the *failure rate* represents the fraction of jobs that fail, but we do not count those that would fail even if there were no churn in the system. That is to say, the *failure rate* measures the job failures directly attributable to provider churn in the grid system.

4.2 Scenario I: The Threshold Algorithm

In this scenario, the experiments were run using the parameters of Table 3. In the experiments we explore the effect of applying the threshold algorithm. The differences

Table 3 Scenario I simulation parameters

microsoft99 case	
Nr. of jobs per user at injection step	7
Job duration in time steps	{40 ··· , 50}
pl05 case	
Nr. of jobs per user at injection step	3
Job duration in time steps	{100 ··· , 110}

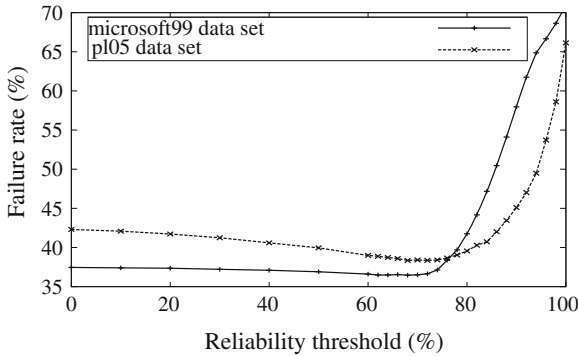


Fig. 4 Scenario I: the job failure rate for different thresholds γ

in job duration between the `microsoft99` and `p105` cases are related to the differences in the average uptime in the two churn distributions. They have been set to generate roughly identical relative workloads in the grid system in both cases.

For each provider, at every point in time, the reliability index is computed. Users screen providers i.e. they submit bids for computational resources only with those providers whose reliability index exceeds a preset threshold value. We explore the effect of changing the threshold on the job failure rate observed in the grid system.

Figure 4 shows the failure rate as a function of the threshold, both for the `microsoft99` and `p105` churn distributions. For the latter the rate decreases slightly up to $\gamma = 70\%$ while the former hardly changes at all. From $\gamma = 70\%$ on the failure rate steadily climbs.

The first observation is somewhat disappointing because it basically indicates that the threshold algorithm does not really improve the failure rate compared to the situation with $\gamma = 0\%$ i.e. when no reliability screening is applied. The second observation is rather easy to understand: when one increases γ , the pool of providers available for bidding shrinks. Thus, the higher γ the smaller the set of available providers. Even though these might be the more reliable ones, the computational capacity shrinks in such a manner these jobs simply fail to find computational resources to run on and fail to meet deadline.

In the panels in Fig. 5 the results of the same experiments are shown, but now differentiated per user group. Four subgroups of users have been introduced that each differs in the deadlines that are set for their jobs. The relevant parameters are listed in see Table 2. The deadlines for *Group1* are the most stringent, those of *Group4* are the least stringent.

The results in these figures indicate that the failure rate in the `p105` case is more responsive to application of threshold algorithm than that in the `microsoft99` case. This observation is consistent with the aggregate (over the user groups) failure rate shown in Fig. 4.

The results also indicate that the improvement in the failure rate, even if slight, is most pronounced when the deadline is the least stringent and vice versa. We have

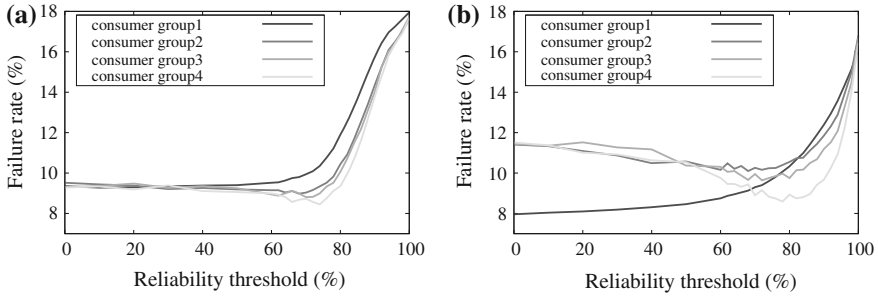


Fig. 5 Scenario I: job failure rates per consumer group for different γ for `microsoft99` (a), and `p105` (b) distributions

no clear, unambiguous indications as to the cause of this relation because of the complexity of the system. However, we conjecture that the effect is most probably due to the fact that the more stringent the deadline, the less effective the mechanism of resubmissions and because of that also the less effective the screening of unreliable providers.

4.3 Scenario II: Provider Groups

In this scenario we repeat the experiments of the previous scenario, but now introduce subgroup in the provider population each characterized by different uptime parameter λ in the churn distribution. To be precise: both for the `p105` and `microsoft99` distributions we subdivide the provider populations in subgroups with uptime parameter λ equal to 200, 150, 100, 50 and 25 percent of the original λ -value. This represents the situation of different coupled subgrids with different churn characteristics.

In Fig. 6 we show the effect of applying the threshold algorithm in this case. The figure plots the failure rate as a function of the threshold γ in the algorithm. A comparison with Fig. 4 reveals for both the `p105` and `microsoft99` cases the failure rate is somewhat more responsive to the impact of the provider screening based on the reliability threshold in the present case. This can be understood in the following sense. The provider population is more diverse in terms of reliability in the case of this scenario compared to the previous one. Consequently the impact of screening for reliable provides via the threshold requirement is more effective in this case and thus improvement of the failure as a consequence of the provider screening is more pronounced.

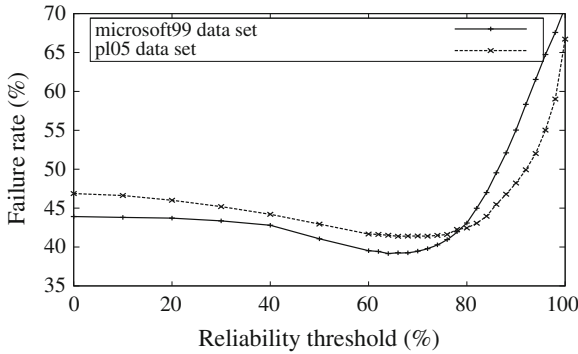


Fig. 6 Scenario II: the job failure rates with respect to different (γ) with five provider groups with different uptime parameter λ

Table 4 Scenario III simulation parameters

microsoft99 case	
Nr. of jobs per user for load_1	7
Nr. of jobs per user for load_2	4
Nr. of jobs per user for load_3	2
Job duration in time steps	{40 ··· , 50}
pl05 case	
Nr. of jobs per user for load_1	3
Nr. of jobs per user for load_2	2
Nr. of jobs per user for load_3	1
Job duration in time steps	{100 ··· , 110}

4.4 Scenario III: Relative Workload

In this scenario we study the effect of the system workload on the failure rate. To do so, we run the same experiments under different workload conditions by changing the number of jobs per consumer (see Table 4 for a listing the simulation parameters) [15].

Figure 7 shows the failure rates as a function of γ , the reliability threshold, for each of the workload conditions, both for the `microsoft99` (top panel) and `pl05` (lower panel) churn distributions. It is obvious here that the lighter the effective workload on the grid system, the more pronounced the improvement of the failure rate due to the threshold algorithm. This can be understood in the following sense. The application of the threshold criterion has two inherent, counteracting effects: the higher the threshold is set, the more reliable the providers satisfying the criterion but also the fewer the number of providers passing it. The first effect will lower the failure rate because providers are less likely to churn out and cause the jobs to fail. The second effect will increase the failure rate because there are fewer providers available, which lowers the total computational capacity in the grid and causes jobs not to finish before deadline or not be run at all. At a certain threshold value γ an

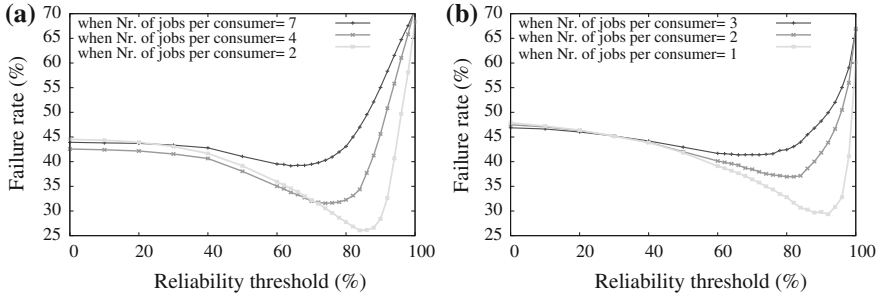


Fig. 7 Scenario III: The effect of workload on the failure rate for different (γ). The *microsoft99* (a), and *p105* (b) churn distributions and 5 provider groups

optimum equilibrium between these opposing effects is achieved and the failure rate is at its lowest value, given the other parameters of the problem.

As we lower the total effective workload in the grid system, the second effect i.e. decreasing the number of available providers due to the mismatch in the reliability criterion, has less impact. Thus the optimum failure rate is lower at lower workloads and occurs at higher γ -values. This is indeed what we observe, for both churn distributions.

5 Conclusion

In desktop and peer-to-peer dynamic grids job failure due to churn in the grid system are inevitable. There is a need to minimize the impact of these job failures on the quality of service provided by such grids. In order to find effective approaches that can alleviate the impact of churn we need to model churn. In our work we have used models that are available in the literature.

With the Grid Economics simulator we have programmed these churn models and developed a resource allocation scheme based on first-price-sealed-bid auctions. We have subsequently investigated an algorithm, which we refer to as the threshold algorithm, to alleviate the impact of churn. It is based on a simple criterion which limits bids to reliable providers, where reliability is defined as having a reliability index above a certain threshold. The reliability index is a simple measure of availability based on average aggregate uptime.

We analyse experiments in a number of scenarios and arrive at the conclusions that firstly the effect of the threshold algorithm is fairly transparent and secondly it has the potential of improving failure rates significantly if the effective workload in the grid system is not too high.

References

1. Anderson DP (2004a) A system for public-resource computing and storage. In: Proceedings of the 5th IEEE/ACM international workshop on grid, computing, p 4–10
2. Kacsuk P, Terstyanszky G (2008) Special section: Grid-enabling legacy applications and supporting end users. *Future Gener Comp Syst* 24(7):709–710
3. Anderson DP (2004b) BOINC: A system for public-resource computing and storage. Proceedings of the fifth IEEE/ACM international workshop on grid computing, IEEE Computer Society, In, pp 4–10
4. Stutzbach D, Rejaie R (2006) Understanding churn in peer-to-peer networks. In: IMC '06 Proceedings of the 6th ACM SIGCOMM conference on internet measurement, New York, NY, USA. ACM, p 189–202
5. Stutzbach D, Rejaie R (2005) Characterizing churn in peer-to-peer networks. Technical Report CIS-TR-2005-03, University of Oregon
6. Ko SY, Hoque I, Gupta I (2008) Using tractable and realistic churn models to analyze quiescence behavior of distributed protocols. In: SRDS '08 Proceedings of the (2008) symposium on reliable distributed systems. IEEE Computer Society, Washington, DC, USA, pp 259–268
7. Kondo D, Taufer M, Brooks III CL, Casanova H, Chien AA (2004) Characterizing and evaluating desktop grids: An empirical study. In: Proceedings of the 18th international parallel and distributed processing symposium (IPDPS'04), IEEE Computer Society, p 1–10
8. Bhagwan R, Savage S, Voelker GM (2003) Understanding availability. In: Proceedings of the 2nd international workshop on Peer-to-Peer systems (IPTPS '03) p 256–267
9. Arpaci RH, Dusseau AC, Vahdat AM, Liu LT, Anderson TE, Patterson DA (1995) The interaction of parallel and sequential workloads on a network of workstations. *SIGMETRICS Perform Eval Rev* 23(1):267–278
10. Kondo D, Javadi B, Iosup A, Epema D (2010) The failure trace archive: Enabling comparative analysis of failures in diverse distributed systems. In: Proceedings of the (2010) 10th IEEE/ACM international conference on cluster, cloud and grid computing, CCGRID '10, Washington, DC, USA, IEEE Computer Society, pp 398–407
11. Bolosky W, Douceur J, Ely D, Theimer M (2000) Feasibility of a serverless distributed file system deployed on an existing set of desktop pcs. In: SIGMETRICS '00 Proceedings of the (2000) ACM SIGMETRICS international conference on measurement and modeling of computer systems. NY, ACM, USA, New York, pp 34–43
12. PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services (2002) <http://www.planet-lab.org/>
13. Stribling J (2005) Planetlab all pairs ping. <http://pdos.csail.mit.edu/strib/>
14. Abdelkader K, Broeckhove J (2009) Pricing computation resources in a dynamic grid. *Int J Grid Util Comput (IJGUC)* 1:205–215
15. Abdelkader K, Broeckhove J (2012) Alleviating the impact of churn in dynamic grids. In lecture notes in engineering and computer Science: Proceedings of the world congress on engineering 2012, London, U.K., 4–6 July, 2012, pp 1113–1118

Parallel Algorithm for Multiplying Integer Polynomials and Integers

Andrzej Chmielowiec

Abstract This chapter aims to develop and analyze an effective parallel algorithm for multiplying integer polynomials and integers. Multiplying integer polynomials is of fundamental importance when generating parameters for public key cryptosystems, whereas their effective implementation translates directly into the speed of such algorithms in practical applications. The algorithm has been designed specifically to accelerate the process of generating modular polynomials, but due to its good numerical properties it may surely be used to multiply integers. The basic idea behind this new method was to adapt it to parallel computing. Nowadays, it is a very important property, as it allows us to fully exploit the computing power offered by modern processors. The combination of the Chinese Remainder Theorem and the Fast Fourier Transform made it possible to develop a highly effective multiplication method. Under certain conditions our integer polynomial multiplication method is asymptotically faster than the algorithm based on Fast Fourier Transform when applied to multiply both: polynomials and their coefficients. Undoubtedly, this result is the major theoretical conclusion of this chapter.

Keywords CRT · Fast multiplication · FFT multiplication · Integer multiplication · Multiplication algorithm · Parallel multiplication · Polynomial multiplication

Polish National Science Centre grant N N516478340.

A. Chmielowiec (✉)

Institute of Fundamental Technological Research, Polish Academy of Sciences,
Pawinskiego 5B, 02-106 Warszawa, Poland
e-mail: achmielo@iptt.gov.pl; andrzej.chmielowiec@cmmsigma.eu

1 Introduction

In 1971 Schönhage and Strassen [15] proposed a new algorithm for large integer multiplication. Since that time, methods based on the Fast Fourier Transform (FFT) have been continuously developed and upgraded. Now we have many multiplication algorithms which are based on the FFT. They are used to multiply integers [4, 13] or power series [10, 14, 17, 18]. Some of them are architecture independent and some are dedicated to a specific processor. The algorithms serve as black boxes which guarantee the asymptotic complexity of the methods using them. However, practical implementation often works in the case of such numbers for which it is ineffective to apply a fast multiplication method. The determination of modular polynomials is a good illustration of this problem. The latest methods for generating classic modular polynomials were developed by Charles, Lauter [1] and Enge [6]. Moreover, Müller [12] proposed another family of modular polynomials which may also be used in the process of counting points on an elliptic curve. The Müller's polynomials are characterized by a reduced number of non-zero coefficients and lower absolute values of coefficients, compared to classic modular polynomials. All the aforesaid authors give the computational complexity of algorithms used to determine modular polynomials based on the assumption that both polynomials and their coefficients are multiplied with the use of the Fast Fourier Transform. The complexity of such a multiplication algorithm is

$$O((n \log n)(k \log k)),$$

where n is the degree of the polynomial, and k is the number of bits of the largest coefficient. However, the application of an asymptotically fast algorithm to multiply numbers becomes effective only when the numbers are of considerable length. According to Garcia's report [7], fast implementation of multiplication in GMP (GNU Multiple Precision Arithmetic Library) becomes as effective as classic multiplication algorithms only for numbers of at least $2^{17} = 131072$ bits. That is why it would be worth to develop a multiplication algorithm which operates fast for polynomials with relatively small coefficients. In order to achieve that, we decided to use the Chinese Remainder Theorem (CRT). This chapter presents results described in our WCE'12 article [2] and extends it to show that the proposed method can also be used to implement fast parallel integer multiplication. In general our idea fits into the scheme proposed in the work [8].

The chapter is organized as follows.

In Sect. 2 for completeness we briefly recall the general idea of the Fast Fourier Transform. The FFT may be implemented in many forms and a choice of proper implementation depends on the problem we want to solve and the processor we are using.

In Sect. 3 we show in detail how to use the CRT to distribute polynomial arithmetic between many processors. Our new method is very simple both in concept and implementation. It does not need any communication between processors, which is

an additional advantage. This algorithm may use any implementation of the FFT. Particularly it may be used with parallel FFT, which reduces the total time of computation.

In Sect. 4 we present numerical results of our 32-bit implementation based on OpenMP parallel programming standard. We compare the proposed method with algorithm based on the FFT over large finite field.

In Sect. 5 we show how fast the 64-bit implementation dedicated and optimized for x86-64 processors is. We compare this implementation to GMP integer multiplication algorithm.

To summarize, to multiply polynomials developer combines two independent techniques to achieve the best performance from a machine or processor:

1. distribution of computations between smaller domains being polynomial rings (apply CRT),
2. optimization of FFT operations within these smaller domains.

The whole idea is illustrated on the following scheme.

$$\begin{array}{ccccccc}
 & & & \mathbb{F}_{p_1}[X] & \xrightarrow{FFT} & \mathbb{F}_{p_1}[X] & \\
 & & & \vdots & & \vdots & \\
 \mathbb{Z}[X] & \xrightarrow{CRT} & \mathbb{F}_{p_i}[X] & \xrightarrow{FFT} & \mathbb{F}_{p_i}[X] & \xrightarrow{CRT^{-1}} & \mathbb{Z}[X] \\
 & & & \vdots & & \vdots & \\
 & & & \mathbb{F}_{p_k}[X] & \xrightarrow{FFT} & \mathbb{F}_{p_k}[X] &
 \end{array}$$

It means that multiplications in $\mathbb{Z}[X]$ can be distributed between k independent rings $\mathbb{F}_{p_i}[X]$ and each such multiplication can be done independently in parallel.

2 Fast Fourier Transform and its Implementations

A Fast Fourier Transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform. The basic idea of DFT is to represent polynomials as sequences of values rather than sequences of coefficients. Computing DTF of n values using the definition takes $O(n^2)$ arithmetic operations, while FFT can compute the same result in only $O(n \log n)$ operations. This is the reason why the Fast Fourier Transform plays a very important role in efficient computations and is considered in many publications. Some of them give a general description of the FFT [3, 5, 9, 11], others contain details about very fast implementations [10, 16–18]. In our numerical experiments in the last section a classic algorithm of FFT has been used. However for practical purposes we suggest application of the *cache-friendly truncated FFT*

recently developed [10]. This new FFT method reduces the computational cost and is optimized against modern processor architecture.

3 Using Chinese Remainder Theorem to Distribute Computations Between Many Processors

In the rest of this chapter we will assume that the largest absolute value of polynomial coefficients is less than B . To multiply integer polynomials we have to find a family of finite fields \mathbb{F}_{p_i} in which computations will be done. It is clear that the product $\prod_{i=1}^k p_i$ should be large enough to eliminate modular reduction during the multiplication process.

Definition 1.1 Let $f(X) = f_{n-1}X^{n-1} + \dots + f_1X + f_0 \in \mathbb{Z}[X]$ and $M \in \mathbb{Z}$. We define $f(X) \bmod M$ as follows

$$f(X) \bmod M = (f_{n-1} \bmod M)X^{n-1} + \dots + (f_0 \bmod M),$$

where

$$f_i \bmod M \in \left\{ \left\lfloor \frac{-M+1}{2} \right\rfloor, \dots, -1, 0, 1, \dots, \left\lfloor \frac{M-1}{2} \right\rfloor \right\}.$$

Lemma 1.2 Let $f(X) = f_{n-1}X^{n-1} + \dots + f_1X + f_0$, $g(X) = g_{n-1}X^{n-1} + \dots + g_1X + g_0$ be polynomials with integer coefficients such that $|f_i| < B$ and $|g_i| < B$. If integer M satisfies the following condition

$$2nB^2 < M$$

then $f(X)g(X) \bmod M = f(X)g(X)$.

Proof: If $f(X)g(X) = h(X) = h_{2n-2}X^{2n-2} + \dots + h_1X + h_0$ then

$$\begin{aligned} h(X) &= \left(\sum_{i=0}^{n-1} f_i X^i \right) \left(\sum_{j=0}^{n-1} g_j X^j \right) \\ &= \sum_{i=0}^{n-1} \sum_{j=0}^i f_j g_{i-j} X^i + \sum_{i=1}^{n-1} \sum_{j=0}^{n-1-i} f_{i+j} g_{n-1-j} X^{n-1+i} \\ &= \sum_{i=0}^{n-1} X^i \sum_{j=0}^i f_j g_{i-j} + \sum_{i=1}^{n-1} X^{n-1+i} \sum_{j=0}^{n-1-i} f_{i+j} g_{n-1-j} \end{aligned}$$

Based on the assumption that $|f_i| < B$ and $|g_i| < B$ we have

1. for all i from 0 to $n - 1$ we have

$$|h_i| = \left| \sum_{j=0}^i f_j g_{i-j} \right| \leq \sum_{j=0}^i |f_j| |g_{i-j}|$$

$$< \sum_{j=0}^i B^2 = (i + 1)B^2,$$

2. for all i from 1 to $n - 1$ we have

$$|h_{n-1+i}| = \left| \sum_{j=0}^{n-1-i} f_{i+j} g_{n-1-j} \right| \leq \sum_{j=0}^{n-1-i} |f_{i+j}| |g_{n-1-j}|$$

$$< \sum_{j=0}^{n-1-i} B^2 = (n - i)B^2.$$

It means that $|h_i| < nB^2$ for all i from 0 to $2n - 2$. If $M > 2nB^2$, then all coefficients (represented as in Definition 0) of $f(X)$, $g(X)$ and $h(X)$ can be represented in residue system modulo M without reduction. This leads to the formula $f(X)g(X) \bmod M = f(X)g(X)$ and ends proof. \square

Theorem 1.3 Let $f(X) = f_{n-1}X^{n-1} + \dots + f_1X + f_0$, $g(X) = g_{n-1}X^{n-1} + \dots + g_1X + g_0$ be polynomials with integer coefficients such that $|f_i| < B$ and $|g_i| < B$. If prime numbers p_i satisfy the following conditions:

- $p_i \neq p_j$,
- $M = \prod_{i=1}^k p_i$,
- $2nB^2 < \prod p_i = M$,
- $p_i = 2^{m+1}r_i + 1$ for some $2^{m+1} \geq 2n$ and $r_i \in \mathbb{Z}$,

then

$$f(X)g(X) = f(X)g(X) \bmod M$$

$$= (f(X) \bmod M)(g(X) \bmod M) \bmod M$$

and fields \mathbb{F}_{p_i} can be used to parallel multiplication of polynomials f and g with FFT method.

Proof: Since operation mod M is a natural homomorphism of \mathbb{Z} then we have

$$(f(X) \bmod M)(g(X) \bmod M) \bmod M =$$

$$f(X)g(X) \bmod M$$

Based on Lemma 1.2 we achieve the second equality

$$f(X)g(X) \bmod M = f(X)g(X).$$

It means that the multiplication of $g(X), f(X) \in \mathbb{Z}[X]$ gives the same result as the multiplication of $g(X) \bmod M, f(X) \bmod M \in (\mathbb{Z}/M\mathbb{Z})[X]$ if elements of ring $\mathbb{Z}/M\mathbb{Z}$ are represented by $\{-\frac{M-1}{2}, \dots, -1, 0, 1, \dots, \frac{M-1}{2}\}$. But M is a product of different primes p_i and the Chinese Remainder Theorem implies the following isomorphism:

$$\mathbb{Z}/M\mathbb{Z} \simeq \mathbb{F}_{p_1} \times \dots \times \mathbb{F}_{p_k}.$$

It is clear that the above isomorphism can be extended to isomorphism of polynomial rings, more precisely we have:

$$(\mathbb{Z}/M\mathbb{Z})[X] \simeq \mathbb{F}_{p_1}[X] \times \dots \times \mathbb{F}_{p_k}[X].$$

It means that multiplications in $(\mathbb{Z}/M\mathbb{Z})[X]$ can be distributed between k independent rings $\mathbb{F}_{p_i}[X]$ and each such multiplication can be done independently in parallel. Moreover all prime numbers $p_i = 2^{m+1}r_i + 1$ were chosen in the way to be well suited for FFT because each field \mathbb{F}_{p_i} contains primitive root of unity of degree 2^{m+1} . \square

In practice it is quite easy to find primes satisfying the assumptions of Theorem 1.3. For example, there exist 56 primes of the form

$$p_i = r_i \cdot 2^{22} + 1,$$

where $512 < r_i < 1024$. This set of primes allows us to multiply polynomials for which

- $\deg f + \deg g < 2^{22}$,
- $\max\{|f_i|, |g_i|\} \leq 2^{871}$.

If we want to use the proposed algorithm to multiply polynomials with larger degrees and coefficients then we can use 64-bit primes. For example the following set

$$\mathcal{P}_{64} = \{p_i : p_i \in \mathcal{P}, p_i = r_i \cdot 2^{32} + 2^{63} + 1, 1 < r_i < 2^{31}\}$$

can be used to multiply polynomials for which

- $\deg f + \deg g < 2^{32}$,
- $\max\{|f_i|, |g_i|\} < 2^{6099510377}$.

Suppose now that we have k prime numbers p_i that have the same bit length and satisfy the conditions described in Theorem 1.3. We have the following theorem:

Theorem 1.4 *If $\lfloor \log_2(p_i) \rfloor = \lfloor \log_2(p_j) \rfloor$ and formal power series have precision n , then the multiplication algorithm described in Theorem 1.3 consists of*

$$c_1 k^2 n + kn(2 + 3 \log(n)) + c_2 k^2 n$$

multiplications in \mathbb{F}_{p_i} . Where c_1, c_2 are some constants.

Proof: Since $\lfloor \log_2(p_i) \rfloor = \lfloor \log_2(p_j) \rfloor$ for each i, j , then we can assume that the cost of multiplication in every \mathbb{F}_{p_i} is the same. Single FFT multiplication consists of three basic steps:

1. Reduction modulo every chosen prime requires $c_1 k^2 n$ multiplications in \mathbb{F}_{p_i} . Each coefficient can be reduced modulo p_i using $c_1 k$ multiplications in \mathbb{F}_{p_i} . We have n coefficients and k small primes. It means that the total cost of this step is equal to $c_1 k \cdot n \cdot k = c_1 k^2 n$.
2. We perform the FFT multiplication for all $i \in \{1, \dots, k\}$:
 - (a) Fourier transform of two power series with n coefficients requiring $2n \log(n)$ multiplications in \mathbb{F}_{p_i} ,
 - (b) scalar multiplication of two vectors with $2n$ coefficients, which requires $2n$ multiplications in \mathbb{F}_{p_i} ,
 - (c) inverse Fourier transform of the vector to the power series with $2n$ coefficients requiring $n \log(n)$ multiplications in \mathbb{F}_{p_i} .
3. Application of the Chinese Remainder Theorem to get back final coefficients requires $c_2 k^2 n$ multiplications in \mathbb{F}_{p_i} . Each solution of the system $x \equiv a_i \pmod{p_i}$ can be reconstructed using $c_2 k^2$ multiplications in \mathbb{F}_{p_i} . Since we have to reconstruct n coefficients, the total cost is equal to $c_2 k^2 \cdot n = c_2 k^2 n$.

Thus the multiplication algorithm described in Theorem 1.3 consists of

$$c_1 k^2 n + kn(2 + 3 \log(n)) + c_2 k^2 n$$

multiplications in \mathbb{F}_{p_i} . □

Finally, let us see how the new algorithm compares with the method using the Fast Fourier Transform for multiplying both: polynomials and coefficients. If we assume that numbers p_i are comprised within a single register of the processor, then the complexity of the algorithm which multiplies the polynomial and its coefficients using FFT is

$$O((n \log n)(k \log k)).$$

The complexity of our algorithm is equal to

$$O(kn \log n + k^2 n).$$

If we assume that $k = O(n)$, it is clear that the algorithm based totally on FFT is much faster. Its complexity is equal to $O(n^2 \log^2 n)$, whereas our algorithm works in time $O(n^3)$. But what happens when the polynomial coefficients are reduced? Let us assume that $k = O(\log n)$. Under this assumption, the complexity of the algorithm based totally on FFT is $O(n \log^2 n \log \log n)$, whereas the asymptotic complexity of our method is $O(n \log^2 n)$. Although the difference is not significant, we definitely managed to achieve our goal which was to develop an effective algorithm for multiplying polynomials with coefficients of an order much lower than the degree.

Table 1 Multiplication of two polynomials of degree $n/2 - 1$ with coefficients less than 2^{256}

Polynomial degree	FFT $\mathbb{F}_{p_{544}}$ (1 core)	FFT-CRT $\bigotimes_{i=1}^{18} \mathbb{F}_{p_i}$ (1 core)	FFT-CRT $\bigotimes_{i=1}^{18} \mathbb{F}_{p_i}$ (4 cores)		
$n/2 - 1$	$T_1(s)$	$T_2(s)$	$T_3(s)$	T_1/T_2	T_2/T_3
511	0.0423	0.0183	0.0052	2.3	3.5
1023	0.0930	0.0383	0.0111	2.4	3.4
2047	0.2020	0.0803	0.0259	2.5	3.1
4095	0.4360	0.1705	0.0481	2.6	3.5
8191	0.9370	0.3575	0.1012	2.6	3.5
16383	2.0100	0.7444	0.2161	2.7	3.4
32767	4.2700	1.5491	0.4283	2.8	3.6
65535	9.0700	3.2168	0.9339	2.8	3.4
131071	19.1700	6.6716	1.8919	2.9	3.5

Corollary 1.5 *If $k = O(\log n)$, the complexity of the proposed algorithm is lower than the complexity of the multiplication algorithm based on FFT only, and equals to*

$$O(n \log^2 n),$$

whereas the complexity of the FFT-based algorithm is

$$O(n \log^2 n \log \log n).$$

However, in practice we managed to achieve much more than this. The numerical experiments showed that the new algorithm brings obvious benefits already in the case of polynomial coefficients consisting of several hundred bits. It means that its application is effective already for small values of k and n .

4 Results of Practical Implementation for 32-bit Processors

The implementation of the fast algorithm for multiplying polynomials has been prepared for 32-bit architecture with the use of OpenMP interface. The obtained time results turned out exceptionally good. They confirmed that in practice, the combination of the Fast Fourier Transform with the Chinese Remainder Theorem considerably accelerates computations. Tables 1 and 2 present the performance times of the algorithm for multiplying polynomials of the same degree with coefficients ranging from $[0, 2^{256})$ and $[0, 2^{512})$.

Our 32-bit implementation is fully compatible with ANSI C99 and was not optimised against any special architecture. The numerical experiments were done on Intel Core 2 processor (2.4 GHz) and confirmed that the simultaneous application

Table 2 Multiplication of two polynomials of degree $n/2 - 1$ with coefficients less than 2^{512}

Polynomial degree	FFT $\mathbb{F}_{p_{1088}}$ (1 core)	FFT-CRT $\bigotimes_{i=1}^{36} \mathbb{F}_{p_i}$ (1 core)	FFT-CRT $\bigotimes_{i=1}^{36} \mathbb{F}_{p_i}$ (4 cores)		
$n/2 - 1$	$T_1(s)$	$T_2(s)$	$T_3(s)$	T_1/T_2	T_2/T_3
511	0.1598	0.0511	0.0136	3.1	3.7
1023	0.3500	0.1055	0.0280	3.3	3.8
2047	0.7600	0.2203	0.0608	3.4	3.6
4095	1.6420	0.4562	0.1210	3.6	3.8
8191	3.5310	0.9430	0.2527	3.7	3.7
16383	7.5500	1.9412	0.5254	3.9	3.7
32767	16.0900	3.9944	1.0960	4.0	3.6
65535	34.1300	8.2184	2.1926	4.1	3.7
131071	72.2100	16.9245	4.5895	4.3	3.7

of CRT and FFT is very efficient. To the end of this section we will assume that: $p_{544} = 2^{544} - 2^{32} + 1$, $p_{1088} = 2^{1088} - 2^{416} + 2^{256} + 1$ and $2^{31} < p_i < 2^{32}$. We compare our FFT-CRT based implementation with multiplication algorithm based on FFT over fields $\mathbb{F}_{p_{544}}$ and $\mathbb{F}_{p_{1088}}$.

We use OpenMP standard to implement parallel version of the proposed algorithm. In Tables 1 and 2 fraction T_2/T_3 gives us information about how many of our 4 cores are on average used by a single multiplication. We can see that the algorithm based on the FFT and CRT uses between 80% and 90% computational power. It is a very good result for arithmetic algorithm.

5 Fast 64-bit Implementation and its Application to Integer Multiplication

To demonstrate the speed of our method we decided to compare it with GMP (The GNU Multiple Precision Arithmetic Library) implementation of integer multiplication. Every algorithm dedicated to multiply integer polynomials can be easily adapted to multiply integers. Suppose we have two integers $a, b \in \mathbb{Z}$ such that

$$\begin{aligned}
 a &= a_0 + a_1R + a_2R^2 + \dots + a_{n-1}R^{n-1}, \\
 b &= b_0 + b_1R + b_2R^2 + \dots + b_{n-1}R^{n-1}.
 \end{aligned}$$

These integers can be converted to integer polynomials

Table 3 Speed comparison of our algorithm and GMP FFT integer multiplication

Factor bits	GMP-FFT	FFT-CRT			
		512-bit digits		1024-bit digits	
		1 core (s)	4 cores (s)	1 core (s)	4 cores (s)
2^{21}	0.040	0.069	0.023	0.073	0.022
2^{22}	0.082	0.145	0.050	0.153	0.047
2^{23}	0.176	0.302	0.106	0.318	0.097
2^{24}	0.395	0.612	0.224	0.663	0.203
2^{25}	0.858	1.268	0.419	1.365	0.403
2^{26}	1.837	2.628	0.971	2.820	0.873

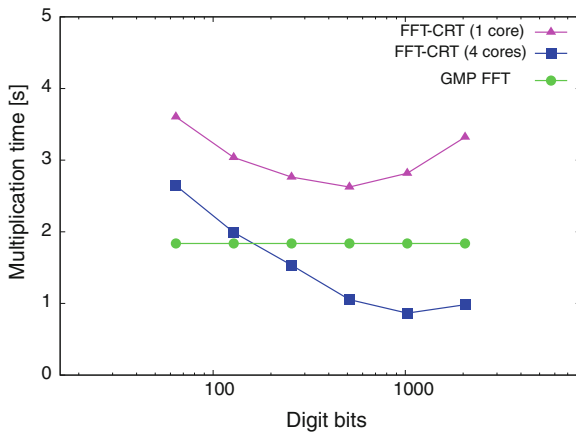


Fig. 1 Multiplication time of two integers with 2^{26} bits each

$$A(X) = a_0 + a_1 X + a_2 X^2 + \dots + a_{n-1} X^{n-1},$$

$$B(X) = b_0 + b_1 X + b_2 X^2 + \dots + b_{n-1} X^{n-1}$$

and multiplied to get polynomial $C(X) = \sum_{i=0}^{2n-2} c_i X^i$. If $nR < R^2$, then we can compute $c = ab$ as follows

$$c = \sum_{i=0}^{2n-2} (c_i \bmod R) R^i + R \sum_{i=0}^{2n-2} \left(\left\lfloor \frac{c_i}{R} \right\rfloor \bmod R \right) R^i + R^2 \sum_{i=0}^{2n-2} \left(\left\lfloor \frac{c_i}{R^2} \right\rfloor \bmod R \right) R^i.$$

One can see that if we have coefficients c_i and R is power of 2 then c can be computed using two multiple precision additions.

To properly compare the proposed algorithm with 64-bit implementation of integer multiplication in GMP we decided to optimise our implementation against x84-64 architecture (including inline assembler functions). Numerical tests show that we can achieve the best performance for a single thread for $R = 2^{512}$. Unfortunately it is about 1.45 times slower than in GMP. We have a better situation in the case of parallel computing. If we can use 4 parallel threads, then R should be equal to 2^{1024} and our implementation is about 2 times faster than GMP which can not be run in parallel.

6 Summary

We present an analysis of a new algorithm for multiplying integer polynomials and integers. It has been designed so as to exploit fully the computing power offered by modern multicore processors. Thanks to using the Chinese Remainder Theorem, it is possible to easily allocate tasks between the available threads. Moreover, under the adopted approach there is no need to synchronize the computations and to ensure communication between individual threads, which is an additional asset. For that reason the algorithm can be easily implemented with the use of a parallel programming standard OpenMP. The ratio T_2/T_3 in Tables 1 and 2 shows how many processors out of the four ones available were used on average during a single multiplication. The measurements show that the algorithm uses from 80 to 90% of the available computing power. In the case of an arithmetic algorithm, this should be considered a very good result. Therefore, we may conclude that the goal which consisted in designing a parallel algorithm for multiplying polynomials has been achieved.

As far as the theoretical results of the chapter are concerned, the analysis conducted in Sect. 3 and Corollary 1.5 being its essence, are of key importance. If we assume that the degree of the polynomial is n and the accuracy of its coefficients is k , then the asymptotic complexity of the proposed algorithm is

$$O(kn \log n + k^2n).$$

Owing to the two essential components of the asymptotic function, it is impossible to determine explicitly whether the new solution is better or worse than the method based on FFT only. It is due to the fact that if we use the Fast Fourier Transform to multiply both the polynomial and its coefficients, the complexity is equal to

$$O((n \log n)(k \log k)).$$

Therefore, one can see that if $k = O(n)$, the proposed algorithm performs worse than the method based on FFT only. However, if $k = O(\log n)$, the complexity of the new algorithm is lower. The computational complexity ratio is $O(\log n)$ to the advantage of the method presented in the chapter. This reasoning allows us to conclude that the algorithm based on CRT and FFT should be used when the number of coefficients of a polynomial exceeds greatly their accuracy. This is often the case

when computations use long polynomials or power series with a modular reduction of coefficients.

The results of numerical tests presented in Sect. 4 show that the proposed method has numerous practical applications. In this section the algorithm has been intentionally compared with the implementation using the classic algorithm for multiplying coefficients in large fields \mathbb{F}_p . It results from the fact that in the case of numbers p consisting of 500 or 1000 bits, multiplication based on the Fourier Transform is completely ineffective. The measurement results came as a great surprise, as it turned out (Tables 1 and 2) that the proposed algorithm is several times faster even when its application is not parallel.

In Sect. 5 we prove that our algorithm can be also used to multiply integers. Table 3 and Fig. 1 show that our four-thread parallel implementation is much faster than single-thread implementation of GMP.

References

1. Charles D, Lauter K (2005) Computing modular polynomials. *J Comput Math* 8:195–204
2. Chmielowiec A (2012) Fast, parallel algorithm for multiplying polynomials with integer coefficients. *Lecture notes in engineering and computer science: Proceedings of the world congress on engineering WCE 2012, 4–6 July 2012 UK, London*, pp 1136–1140
3. Cormen TH, Leiserson CE, Rivest RL, Stein C (2003) *Introduction to algorithms*. MIT Press, New York
4. Crandall R, Fagin B (1994) Discrete weighted transforms and large integer arithmetic. *Maths Comput* 62:305–324
5. Crandall R, Pomerance C (2001) *Prime Numbers— a computational perspective*. Springer, New York
6. Enge A (2009) Computing modular polynomials in quasi-linear time. *Math Comp* 78:1809–1824
7. Garcia L (2005) Can Schönhage multiplication speed up the RSA encryption or decryption?. *University of Technology, Darmstadt*
8. Gorlatch S (1998) Programming with divide-and-conquer skeletons: a case study of FFT. *J Supercomput* 12:85–97
9. Grama A, Gupta A, Karypis G, Kumar V (2003) *Introduction to parallel computing*. Addison Wesley, London
10. Harvey D (2009) A cache-friendly truncated FFT. *Theor Comput Sci* 410:2649–2658
11. Knuth DE (1998) *Art of computer programming*. Addison-Wesley Professional, London
12. Müller V (1995) Ein Algorithmus zur Bestimmung der Punktzahlen elliptischer Kurven über endlichen Körpern der Charakteristik grösser drei. Ph.D. Thesis, Universität des Saarlandes
13. Nussbaumer HJ (1980) Fast polynomial transform algorithms for digital convolution. *IEEE Trans Acoust Speech Signal Process* 28(2):205–215
14. Schönhage A (1982) Asymptotically fast algorithms for the numerical multiplication and division of polynomials with complex coefficients. *Lecture notes in computer science*, vol 144. Springer, Berlin, pp 3–15
15. Schönhage A, Strassen V (1971) Schnelle Multiplikation grosser Zahlen. *Computing* 7:281–292
16. Takahashi D, Kanada Y (2000) High-performance radix-2, 3 and 5 parallel 1-D complex FFT algorithms for distributed-memory parallel computers. *J Supercomput* 15:207–228
17. Van der Hoeven J (2004) The truncated Fourier transform and applications. *ISSAC 2004 ACM*, pp 290–296
18. Van der Hoeven J (2005) Notes on the truncated Fourier transform. Unpublished, available on <http://www.math.u-psud.fr/vdhoeven/>

A Model to Improve Reliability in Cloud Computing

P. Srivaramangai and Rengaramanujam Srinivasan

Abstract The cloud computing offers dynamically scalable resources provided as a service over the Internet. It promises the drop in capital expenditure. But practically speaking if this is to become reality there are still some challenges which is to be still addressed. Amongst, the main issues are related to security and trust, since the user's data has to be released to the Cloud and thus leaves the secured area of the data owner. The users must trust the providers. There must be a strong trust relationship exist between the service providers and the users. This paper provides a model based on reputation which allows only reliable providers to provide the computing power and the resources which in turn can provide a reliable infrastructure for cloud computing.

Keywords Activity · Computability · Cloud computing · Credibility · Malicious · Reliable · Reputation · Security · Specificity · Trust

1 Introduction

Cloud computing has a lot of common features as Grid computing. It can be argued that the cloud computing has evolved from Grid computing. Grid computing actually provides infrastructure to cloud computing which includes computing and storage resources where as cloud aims at economic based delivering of resources and services. Data Security is of prime importance for any business. A cloud service provider needs to secure its infrastructure, its applications, as well as the stored business data. With cloud computing all the data is stored and processed remotely in another machine.

P. Srivaramangai (✉)
Botho college, 501564 Gaborone, Botswana
e-mail: srivara.padma@gmail.com

R. Srinivasan
BSA University, Vandaloor, Chennai, India
e-mail: rs9966@gmail.com

The infrastructure that supports the platform from which the user interacts is unseen by the user. The consumer of services fear that the information stored in cloud may be accessed by hackers. The service providers must also expected to be committed to the local privacy policies of the customers.

Cloud computing contributes a lot to business world. But still Enterprises have a lot of concerns about the reliability and security of these remote clouds. People are not comfortable with the data being stored under the control of third party provider. There is no assurance that the cloud providers will not go out of business. Reliability is the major concerns of the customers of cloud computing.

Cloud computing and its related technologies will only be adopted by users, if they are confident that their data and privacy are secured, and the system is as scalable, robust and reliable as of their own, in their places. Trust and reputation systems have been recognized as playing an important role in decision making on the internet. Reputation based systems can be used in a Grid to improve the reliability of transactions. Reliability is the probability that a process will successfully perform its prescribed task without any failure at a given point of time. Hence, ensuring reliable transactions plays a vital role in cloud computing. To achieve reliable transactions, mutual trust must be established between the initiator and the provider. Trust is measured by using reputation, where as the reputation is the collective opinion of others.

This paper provides a model which introduces a new factor called compatibility which is based on Spearman's rank correlation. The feed backs of the recommenders which are incompatible with those of the initiator are eliminated by using the compatibility factor. Few other factors are also included for measuring the direct trust. This model effectively evaluates the trustworthiness of different entities and also it addresses various malicious behaviors. Two important factors—context and size, are incorporated in evaluating the trustworthiness of entities.

Section 1 of this chapter describes the cloud environment and has brought out the importance of the trust mechanism on the successful operation of the cloud. The scope of the research work is defined and the contributions are listed. Section 2 provides an overview of the related work. Section 3 introduces a new factor called compatibility, which is evaluated using Spearman's rank correlation coefficient and also gives a brief overview of the model. It is shown that using the compatibility factor, eliminates the biased and otherwise incompatible feedbacks and leads to reliable transactions in the cloud. Section 4 presents details about the experiments conducted and also the analysis of the results obtained. Section 5 concludes the chapter by summing up the findings and suggesting the scope for future work.

2 Related Work

A number of disciplines have looked at various issues related to trust, including the incremental values assigned by people in transactions with a trusted party and how trust affects people's beliefs and decision making. Considerable work has been done on trust in computer science, most of them being focused in the area of security.

Advanced models in this category compute a weighted average of all the ratings, where the rating weight can be determined by factors such as the raters' trustworthiness/reputation, the age of the rating, the distance between the rating and current score, etc. Xiong and Liu [13] used an adjusted weighted average of the amount of satisfaction that a user gets for each transaction. The parameters of the model are the feedbacks from transactions, the number of transactions, the credibility of feedbacks and the criticality of the transaction.

Stakhanova et al. [10] proposed a decentralized reputation based trust model for selecting the best peer. A local table is maintained for each entity to store the transaction records of all the other entities. Each entity table stores the id of all the other entities in the network, their reputation values, the number of bad transactions that occurred and the total number of transactions performed. A concrete formula is presented for calculating the Trust value of the entities willing to provide the resource. Stakhanova et al. [10] actually calculates the mistrust value, and if the value is above a given threshold value, reject the resource.

Tajeddine et al. [11, 12] proposed an impressive reputation based trust model. This model was extended, and they developed a comprehensive model called PATROL. Their works are based on the TRUMMAR model which was developed by Derbas et al for mobile agents.

Sonnek et al. [8] proposed a model which addresses the unreliability of nodes in a larger scale distributed system. In this model they say that the reliability is not a property but it is a statistics based on a node's performance and behavior. They propose algorithms which employ estimated ratings for reputation.

Luo and Ni [4] introduced a trust model which incorporate VO trust relationship in to traditional Grid entities. This model used the clustering analysis to evaluate trust for Grid entities.

Luo et al. [5] proposed a model for trust degree based access control in Grid environments. It analyzes the differences between intro domain and inter domain trust. Benjamin Linder, Scalent System's CEO, says: *"What I find as CEO of a software company in this space, Scalent Systems, is that most enterprises have a hard time trusting external clouds for their proprietary and high-availability systems. They are instead building internal "clouds", or "utilities" to serve their internal customers in a more controlled way."*

In articles the security issues with Google Docs different issues are discussed [2]. The Google response to one of them is given in article Google docs blog spot. There is nothing new in the nature of these vulnerabilities; only their setting is novel. In fact, IBM has repositioned its Rational AppScan tool, which scans for vulnerabilities in web services as a cloud security service in Blue Cloud Initiative [3].

Alhamad et al. [1] proposed a trust model for cloud users to select the reliable resources. It is based on a particular SLA frame work. Zhao-xiong proposes a weighted trust model for cloud. He used a Weighted Trust Information Transfer Algorithm (WTIT Algorithm) and Weighted Trust Information Combination Algorithm (WTIC Algorithm) for making the decision about the trust.

Priyank et al. [6] proposed a mobile agent based trust model for cloud computing. In this chapter they give a model for cloud architecture. The model uses mobile

agent as security agents to acquire useful information from the virtual machine. This information can be utilized by the users to keep track of privacy of their data and virtual machines.

Srivaramangai and Srinivasan [9] proposed a reliable infrastructure for cloud computing.

3 The Model for Reliable Providers

In this approach, the initiator host (client) calculates the reputation value of the target host (provider) based on its previous experiences and gathered feedbacks from other hosts (here the recommenders are the clients). The recommenders who give feed backs can be from the same administrative control (neighbor) or from different trusted domain (friends) or from a completely strange domain (stranger). Direct trust is calculated by using the parameters context and size of the job. Indirect trust is calculated by considering the feedbacks from all other hosts and the feed backs are multiplied by corresponding credibility factors. Total trust comprises of direct trust and indirect trust in which higher weightage is given for direct trust. If the total trust is greater than the minimum prescribed threshold value the model accepts the resource. The provider can be the trusted provider.

In order to allocate weightage to feed backs given by different recommenders, credibility factor is defined. The factor takes values between zero and one; they are based on three parameters, compatibility, activity and specificity. The credibility factor is given by the expression (1) where a, b and c are fractions with $a > b > c$ and $a + b + c = 1$.

$$\text{Credibility} = a * \text{compatibility} + b * \text{activity} + c * \text{specificity} \quad (1)$$

where compatibility is given by expression (2).

$$\text{Compatablity} = 1 - \frac{6 \sum_{i=1}^n \sum dr_i^2}{n(n^2 - 1)} \quad (2)$$

where dri gives the difference in ranks.

$$\text{activity} = \frac{\text{number of transactions of the recommender as a user}}{\text{Total number of transactions of all recommenders a users}} \quad (3)$$

$$\text{specificity} = \frac{\text{number of transactions of recommender as providers}}{\text{Total number of transactions of all recommenders as providers}} \quad (4)$$

The expression (3) and (4) give the activity and specificity.

After this calculation the indirect trust is calculated by using the expression (5) and (6). If there are more than one domain the IT1 represents the trust from the nodes in the same domain and IT2 represents the trust from the different domain.

$$IT1 = \frac{\sum_{i=1}^n \delta_{1i} rep \frac{y}{z_i}}{\sum_{i=1}^n \delta_{1i}} \tag{5}$$

$$IT2 = \frac{\sum_{i=1}^n \delta_{2i} rep \frac{y}{t_i}}{\sum_{i=1}^n \delta_{2i}} \tag{6}$$

Direct trust is calculated by using the expression in (7).

where δ_1 and δ_2 are credibility factors.

For calculating the direct trust, the model assumes that the feedback values given by the user for one kind of job provided by an entity, are different from another kind of job by the same entity. So the model uses three types of trusts, namely, DT1, DT2 and indirect trust. DT1 represents the trust of the user on the provider as a result of the same kind of transactions, and DT2 for different types of transactions. Indirect trust is calculated by the same expression as that of the previous models. Further, this model considers the fact that the reputation values are not always constant. When there is no transaction between two entities for a long period of time then the value of reputation is brought down. Thus this model adopts a function called the decay function, which decreases the value of reputation when there is no transaction, over a given interval. After the elapse of a specific period with out any transaction this decrement is done.

3.1 Computation of Trust

In this model three types of jobs are considered. The jobs can be the transfer of files, printing or computing. Further, the size of the jobs can fall under three categories—small, medium and large. The system assigns the complexity factor based upon context and size (Table 1). Nine different combinations of contexts and sizes of jobs are considered and a complexity factor is assigned for each of the combinations. Thus there are nine types of transactions; from Table 1, it follows that the complexity factor is highest (=1) for large computational jobs, and the smallest (=0.25) for simple file transfer jobs.

Table 1 Complexity table

Job type	Context	Size	Complexity factor
1	C1	S	0.25
2	C1	M	0.4
3	C1	L	0.5
4	C2	S	0.4
5	C2	M	0.5
6	C2	L	0.6
7	C3	S	0.6
8	C3	M	0.8
9	C3	L	1

C1: File transfer, C2: Printing, C3: Computing

Let us consider a scenario where A is the user and wants to use the resource, say the printer of the provider P. Let the job size be medium. Thus, from Table 1, the transaction type is 5. Before submitting the job to P, the user A has to be satisfied about the trust worthiness of P. The system refers to all the previous transactions between the user A and the provider P. If there are any transactions of the same type-s, context and size being the same as per the current requirement, then the average of the reputation values of all these transactions is taken as DT1. Thus $DT1_{x,y,s}$ the direct trust of the user x on y based on the same type of transactions as the present requirement, is given by expression (7).

$$DT1_{x,y,s} = \frac{\forall i \in type\ s \sum_{i=1} r_i}{f_s} \tag{7}$$

where f_s refers to the frequency of the same type of transactions and r_i corresponds to the reputation value based on the i_{th} transaction.

The direct trust between x and y based on differing type of transactions $DT2_{x,y,d}$ is given by expression (8).

$$DT2_{x,y,d} = \frac{\sum_{k=1}^n \sum_{i=1}^m c_i r_i / f_i}{n} \tag{8}$$

where n is the number of differing transaction types. If A and P have transacted all the types of transactions, n will be $(9 - 1 =)8$. However, if P is not the provider for computational jobs, then n will be $(6 - 1 =)5$.

3.2 Decaying Function

As time passes, entity reputation with respect to other entities typically changes to an unknown state, if little or no interaction occurs between them. When an entity Z receives a request (from entity X) for reputation information about entity Y, it modifies its reputation information relative to Y by using a decaying factor, and then sends the result to the requesting entity.

$$rep\ y_{z,t} = rep\ y_{z,t_0} * \gamma \tag{9}$$

where γ depends on time. If t is the current time and t_0 is the time at which the last transaction took place, then the calculation of γ is as follows.

$\gamma = 1$ if $t - t_0 < 1$ unit of time, month week etc.

$\gamma = 0.75$ if $1 < t - t_0 < 2$

$\gamma = 0.5$ if $2 < t - t_0 < 3$

$\gamma = 0$ if $t - t_0 > 3$.

3.3 Updating the Reputation

The reputation value in the data base is updated after each transaction is successfully completed. The updation is done by using the following rules.

IF a new reputation value > The existing reputation value

Update the existing value by $(\text{new reputation} * 0.3 + \text{old reputation} * 0.7)$

ELSE IF

$(\text{old reputation} > \text{new reputation})$ and $(\text{old reputation} - \text{new reputation}) > 1$ then go by the new reputation.

ENDIF

Else

Update the reputation by $(\text{new reputation} * 0.7 + \text{old reputation} * 0.3)$

ENDIF.

Thus the above formulation has a higher penalty for the current poor performance of an entity.

4 Experiments and Results

The Model has been tested by simulation for applicability in Grid. Since Grid computing can provide a powerful infrastructure for cloud computing this model can be applied for cloud also to provide a reliable infrastructure. The model is compared with one of existing model Patrol Model [2007] and the results are found to be productive. The results are given in the following table. In the simulation, 50 users and 50 providers are taken in to account. For the simulation study users 1–5 and providers 1–5 are malicious. A transaction table is also maintained to keep track of all the transactions. A transaction table is also maintained to keep track of all the transactions. Table 2 gives a summary of the results. In Table 2, column ‘YY’ refers to the situation, where the Patrol model and proposed model allow transactions to proceed, while column ‘NN’ corresponds to the denial of transactions by both. Columns ‘YN’ and ‘NY’ correspond to disagreement cases. In all there are 15 disagreement cases and Table 3 details them. In all these cases the disagreement is due to the malicious providers or initiators.

Table 2 Result summary for study 1

Simulation	YY	NN	YN	NY	Total
No of transactions	54	81	7	8	150
Percentage	36	54	4.6	5.4	100

Models compared: Patrol model and Proposed model taken in order

Table 3 Details of disagreement cases for study 1

S.No	User	Provider	PATROL model	Proposed model
1	33	1	YES	NO
2	26	5	YES	NO
3	25	2	YES	NO
4	19	3	YES	NO
5	2	14	YES	NO
6	34	3	YES	NO
7	22	33	NO	YES
8	21	18	NO	YES
9	13	23	NO	YES
10	22	33	NO	YES
11	11	21	NO	YES
12	28	8	NO	YES
13	42	12	NO	YES
14	23	16	NO	YES
15	13	22	NO	YES

Fig. 1 Comparison of the results of the PATROL model and Proposed model

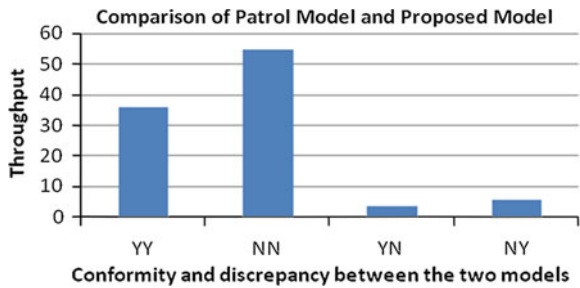


Figure 1 shows the allocation by the two models. The agreement between the existing model and the proposed model is found to be 90% and the disagreement is 10%, and each of the disagreement cases has been analyzed.

Here, the throughput specifies the percentage of the number of reliable successful transactions. Since this model considers the two way reputation along with the context and size of the job, the accuracy of the output is further increased. This model decides whether to grant the transactions or not, based upon the previous transactions and referrals from the other entities.

The simulation study 2 is conducted by varying the number of transactions. The model was initially tested with 150 transactions. Since the Grid consists of a large number of resources with a large number of transactions, the model was tested by increasing the number of transactions. The number of entities was fixed at 100. The percentage of malicious entities is 10%. The number of transactions was varied from 10 to 7,000, and the results were noted. From Fig. 2 it can be seen that the percentage of reliable and successful transactions is higher with the proposed model as compared to the Patrol model.

Fig. 2 Comparison of the models by varying the numbers of transactions

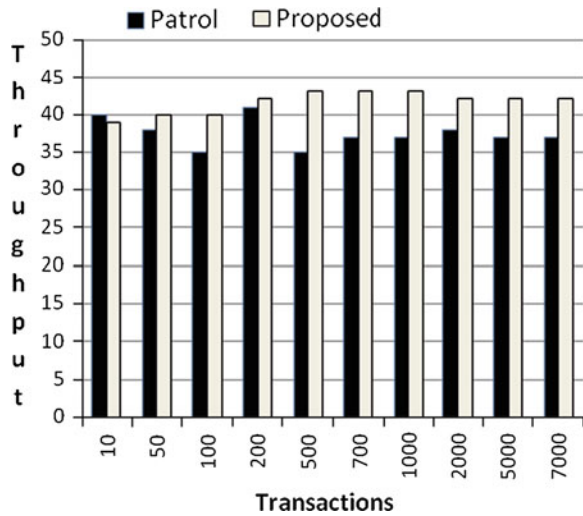


Table 4 Comparison of the accuracy values for the two models

Malicious nodes (%)	Patrol model (%)	Proposed model (%)
0	98	100
10	94.6	100
20	93	100

The accuracy of the model can be defined as:

$$\text{Accuracy} = \frac{\text{Number of correct Decisions} \times 100}{\text{Total number of Decisions}} \tag{10}$$

From Table 4 it follows that:

- (i) Of the 2 models the Patrol model has the lowest accuracy.
- (ii) Accuracy decreases as the percentage of malicious nodes increases.

A note of caution has to be issued at this stage. The accuracy so evaluated is the percentage of ‘correct’ decisions under specified conditions: (i) malicious user (ii) malicious provider (iii) specified contexts. These are the accuracy values obtained under the simulation study. The accuracy under ‘field conditions’ may be lower since field constraints may not exactly fit with the models considered. However, in all the cases, the relative ordering of the accuracy levels among the models will remain unaltered. Table 5, presents a comparison of the throughputs for the 2 models with various percentages of malicious nodes.

From Table 5 it can be concluded that the throughput decreases with an increasing percentage of malicious nodes; the throughput remains essentially at the same level, when the number of nodes is zero percent.

Table 5 A comparison of the throughput of the two models

Malicious nodes (%)	Patrol model (%)	Proposed model (%)
0	43	44
10	40.6	41.4
20	39	41

5 Conclusion

This paper suggests a model for improving the reliability in Grid computing. Since cloud computing is evolved from Grid computing and security is one of most burning issue in cloud computing this model can be very well used in cloud computing to improve the reliability. It is shown that the reliability of the transaction is improved with the inclusion of different parameters. The experimental results have established the usefulness of this model.

References

1. Alhamad M, Dillon T, Chang E (2010) SLA based trust model for cloud computing. In: Proceedings of 13th the international conference on network-based, information systems, pp 321–324
2. BlueCloud. www-03.ibm.com/press/us/en/pressrelease/26642.wss
3. GoogleDocs. <http://googledocs.blogspot.com/2009/03/just-to-clarify.html>
4. Luo J, Ni X (2007) A clustering analysis and agent-based trust model in a grid environment supporting virtual organizations. *Int J Web Grid Serv (IJWGS)* 5(1): 3–16
5. Luo J, Wu Z, Cao J, Tian T (2012) Dynamic multi-resource advance reservation in grid environment. *J SuperComput* 60(3):420–436, Springer, Netherlands
6. Priyank SH, Ranjita S, Mukul M (2011) Security agents: a mobile agent based trust model for Cloud Computing. *Int J Comput Appl* 36(12):12–15
7. Security issues with Google Docs. <http://www.peekay.org/security-issues2009/03/.../security-issues-with-google-doc>
8. Sonnek J, Chandra A, Weissman JB (2007) Adoptive reputation based scheduling on unreliable distributed infrastructure. *IEEE Trans Parallel Distrib syst* 18(11):1551–1564
9. Srivaramangai P, Srinivasan R (2012) A model to provide a reliable infrastructure for cloud computing. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE, U.K*, pp 1125–1129, 4–6 July 2012
10. Stakhanova N, Ferrero S, Wong J, Cai Y (2004) A reputation-based trust management in peer-to-peer network systems. In: *International workshop on database and expert systems applications*, pp. 776–781
11. Tajeddine A, Kayssi A, Cheab A, Artail H (2007) A comprehensive reputation-based trust model for distributed systems. In: *IEEE workshop on the value of security through collaboration (SECOVAL) vol 1(3–4)*, Athens, Greece, pp 416–447, 5–9 Sept 2007
12. Tajeddine A, Kayssi A, Chehab A, Artail H (2007) PATROL: a comprehensive reputation-based trust model. *Int J Internet Technol Secur Trans* 1(1/2):108–131
13. Xiong L, Liu L PeerTrust: supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Trans Knowl Data Eng* 16(7):843–857

Natural Gas Price Forecasting: A Novel Approach

Prerna Mishra

Abstract Earlier discarded as an irritant by-product of crude oil exploration, Natural gas is considered as world's most important fuel due to environmental considerations. It plays an important role in meeting global energy demand and has significant share in the international energy market. Natural Gas is emerging as an alternative to crude oil and coal as the main energy source and the global energy consumption pattern has transformed from preeminence of crude oil and gas to escalating share of gas. Accordingly, there is a spur in demand of natural gas and business entities across the world are interested to comprehend natural gas price forecast. The forecast is likely to meet different objectives of producers, suppliers, traders and bankers engaged in natural gas exploration, production, transportation and trading as well as end users. For the supplier the objective is to meet the demand with profit and for the trader it is for doing business. Of late researchers have exercised different approaches to forecast price by developing numerical models in terms of specific parameters which have relationship with Natural Gas price. This chapter examines application of contemporary forecasting techniques—Time Series Analysis as well as Nonparametric Regression invoking Alternating Conditional Expectations (ACE) to forecast Natural Gas price. Noticeable predictor variables that may explicate statistically important amount of inconsistencies in the response variable (i.e. Natural Gas price) have been recognized and the correlation between variables has been distinguished to model Natural Gas price.

Keywords ACE · ARIMA · Crude Oil · Econometrics · Natural Gas · Natural Gas price · Nonparametric Regression · Time Series

P. Mishra (✉)

Nanyang Technological University, 50 Nanyang Avenue, Singapore, Republic of Singapore
e-mail: prernamishr1492@yahoo.com

1 Introduction

Petroleum is a vital energy source and natural gas is progressively playing an increasingly important role in meeting global energy demand. Globally, natural gas is considered a clean energy source. Environmental legislation and energy policies have played critical role in intensive exploitation of natural gas. Today, natural gas is an important ingredient to the global energy-market, and is the preferred fuel of contemporary civilization. Natural gas is either consumed locally or exported via terrestrial pipelines or sea routes to meet the demand of importing countries. But, commercialization of natural gas is a great challenge because the regions of production and consumption are faraway. Supply of natural gas to a country depends increasingly on trade relationships and supply lines crossing the globe. Most of these imports are transported along sea-lanes and complex pipeline routes. Transportation cost is the impediment to gas trade. There are technological, legal, pecuniary and logistical problems associated with the construction, management, security and surveillance of cross-country pipelines against pilferage and sabotage owing to jagged and craggy topography and diplomatic as well as trade impasse with nations which provides corridor for placing terrestrial pipeline. Of late, natural gas markets previously considered being isolated, with prices varying from country to country, has taken global dimension due to technological interventions. The emergence of the new cost-cutting technologies in gas transportation and liquefaction has provided impetus to Liquefied Natural Gas (LNG) industry and there is a shift from crude oil to natural gas and LNG. LNG is natural gas that is condensed into a liquid under high pressure by reducing it to temperatures below -150°C . LNG enables long distance maritime transportation of natural gas. The global LNG trade allows producers to bypass the constraints of transport by pipeline that coupled gas with local or regional market.

Natural gas is explored and exploited as crude oil (Fig. 1). But, its transportation, storage and delivery to end user are significantly different. Carrying crude oil and

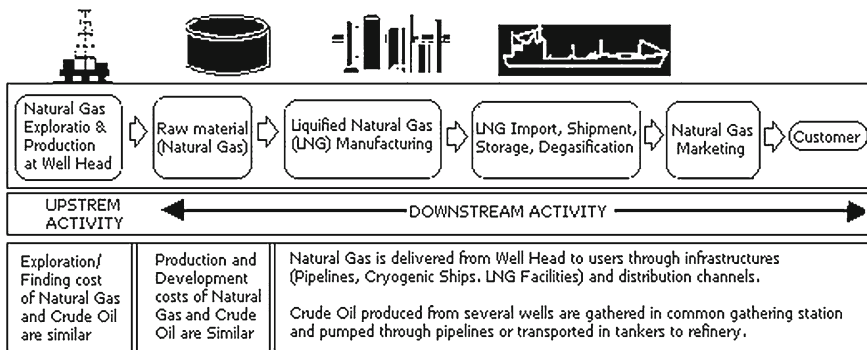


Fig. 1 Natural Gas Value Chain

natural gas from the well head to end user is capital intensive. The end user cost of crude oil and natural gas are not similar. The exploration cost which governs the well head price is identical in both the cases but the production, transportation, and marketing costs varies considerably. Pragmatic approximation of natural gas price at point of exploration is essential for due diligence of any project dealing with natural gas exploration, exploitation, transportation and trade. Consequently, with the spiraling use and international trade, the several vintages of pricing mechanisms have evolved.

In the following section an attempt is made to estimate price of natural gas at well head i.e. the point of production. This, in essence is the most fundamental price for natural gas and industry engaged in other activities can work out the project economics based on this price.

2 Literature Review

A methodology for prediction of short-term natural gas prices using econometrics and neural network has been provided by Doris [1]. The author suggested that developing a quantitative model to predict natural gas price based on market information is central for price risk management because of volatile nature of natural gas spot prices, and because prevalent practice in price prediction focuses on long term equilibrium price forecast with no concern for daily fluctuation in spot prices. The model based on long term equilibrium price may hold well for long-term contracts, but not for day-to-day prediction of spot price. Doris (1999) has dealt with this aspect with econometric and neural network model. Critical analysis of the impact of oil and natural gas prices on oil industry activities invoking linear regression analysis has been made by Solomon [2] and an attempt have been made to forecast crude oil and natural gas prices. An econometrics forecasting model of short term oil spot prices has been proposed by Zamani [3]. The relationship between stocks and prices in the short run for copper, lumber and heating oil has been attempted by Pindyck [4]. An attempt to forecast real oil price has been made by Stephane et al. [5]. The chaotic behaviour of historical time series oil and gas spot prices has been recognized by Agbon [6] and suggestion has been put forward for the use of nonlinear models for its prediction. A decision-analytic model to value crude oil price forecast has been proposed by Jablonowski et al. [7]. An equitable gas pricing model has been proposed by Ogwo [8]. The literature review provides an idea about various methodologies proposed to forecast natural gas price and their merits. But certainly each method has inherent advantage and limitations. In order to address these limitations, techniques invoking econometrics and nonparametric regression to develop models to estimate natural gas has been envisaged and this constitute the theme of this chapter. Nonparametric Regression techniques such as Alternating Conditional Expectations (ACE) [9] identify the relationship between dependent and independent variables. It overcomes the limitations of conventional regression which require *a priori* assumption

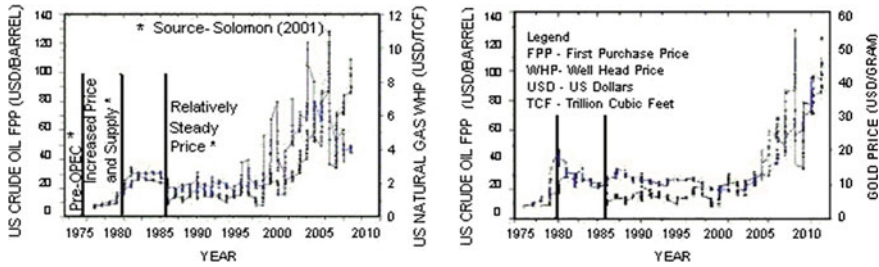


Fig. 2 Historical crude oil, natural gas, and gold price

of functional relationship between variables. Methodology to estimate optimal transform for multiple regression and correlation by application of ACE has been provided by Briemann et al. [9] and detail description of ACE algorithm is provided by Duolao et al. [10] and Guoping et al. [11, 12]. Description of Time Series Analysis is provided by Gujarati [13] and Ramanathan [14]. The application of ACE in modeling is provided by Jon Quah et al. [15]. A procedure for estimation of natural gas price by modeling the historical time series data of US Crude Oil First Purchase Price (USCOFPP) [16], US Natural Gas Well Head Price (USNGWHP) [17], and Gold Average Price (GP) [18] by invoking time series econometrics and nonparametric regression techniques has been discussed by Perna Mishra [19].

3 Factors of Consideration

The historical USCOFPP, USNGWHP, and GP data corroborates resemblance (Fig. 2). It is documented that events in the world history have impacted these prices which are depicted by spikes in the curves.

It is recognized that gold is the most commonly traded commodity and its price has influenced the global market and national currencies. Therefore, in order to derive a relationship between historical crude oil price and natural gas price, the price of gold has been introduced and an attempt has been made to establish the relationship between prices of Natural Gas, Crude Oil and Gold by invoking Alternating Conditional Estimation (ACE) [9].

4 Modeling Parameters

The USNGWHP is considered as dependent variable, and the USCOFPP and GP have been considered as independent variables (Fig. 3).

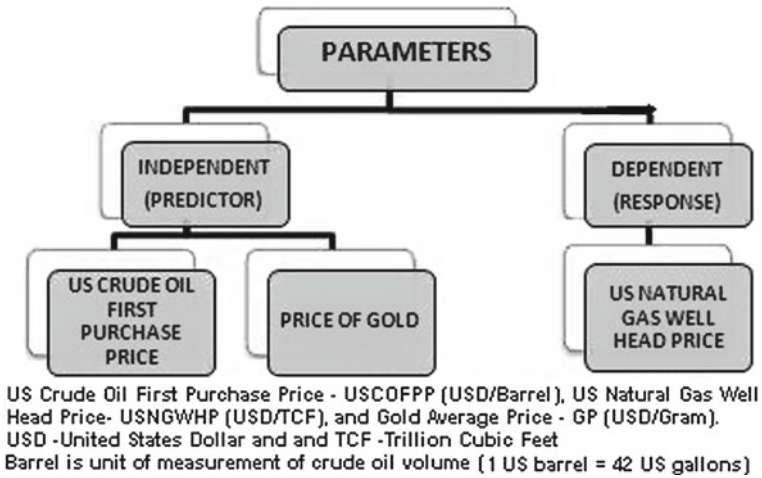


Fig. 3 Predictors and response parameters

5 Modeling Methodology

Scouting, scanning, collation and integration of historical statistical (monthly/ quarterly time series) data of variables is key to forecasting. Time Series data (monthly data) for the period January 1976–September 2011 constitute the input parameters for modeling. It is recognized that on the whole the natural gas price has increased over the years, but, there are occasional “abrupt departure” from the general trend. These fluctuations in natural gas price may be explained by several factors- endogenous or exogenous. However, the magnitude of impact of each factor is difficult to estimate and some major cause may be assigned to observed “sharp deviation”. Endogenous factors arise from activities within the petroleum industry. Exogenous factors include issues like seasonality in demand and supply arising due to swelling of gas demand during winter or waning during summer, natural calamities interrupting the supply chain, and geopolitical issues etc. It is unlikely to account for the relative influence of these factors on natural gas price at any particular moment. However some notable cause may be assigned to the “sharp deviation”. But, our objective is not to recognize these parameters which may be categorized as—secular movements illustrating general tendency of the data to increase or decrease; seasonality of price fluctuations due to weather conditions, cyclical variations or economic fluctuations (benchmarked with respect to changes in standard currency), irregular variations due to natural calamities, geopolitical issues.

5.1 Correlation and Regression Analysis

Correlation and regression analysis identify the nature, degree and type of relationship between multiple dependent and independent variables and associated functional forms.

5.2 Time Series Analysis

Time Series represents a set of observations recorded at equal intervals, viz. the values Y_1, Y_2, \dots, Y_n of a variable Y at a time t_1, t_2, \dots, t_n . In a time series analysis any dependent variable (Y) is modeled as a function of a trend component (T), a seasonal component (S) and a random (or stochastic) component (U), expressed by equation $Y = T * S * U$; where $*$ represents an arithmetic operator (addition, multiplication etc.) and S is due to regularly occurring seasonal phenomenon (Ramanathan) [14]. Thus, while selecting requisite model due consideration needs to be given to impact due to the elements leading to seasonality in price behavior as evidenced in time series (Fig. 4). Analyzing time series may facilitate in comprehending past behavior and thus predict within limits the future behavior by recognizing the pattern of regularities of occurrences of any feature over a period of time. Comparative evaluation of different time series may lead to significant logical derivations by relating variables in each series to their past values. Time Series Econometrics Models invoking Unit Root Test, Cointegration, Augmented Dickey-Fuller Test, VAR Models, ARIMA Model, Granger causality test etc. may be applied to understand the behavior of impacting parameters in the past and their forecast in future time period as well as causal relationship of factors.

These models provide logical understanding of the pattern of behavior of an event either singly or in combination with other events (in time, panel and cross sectional

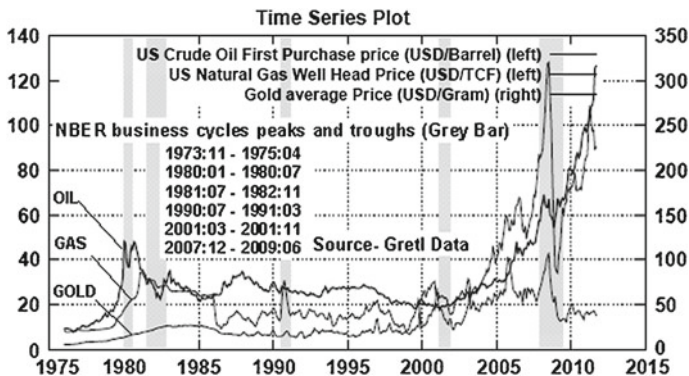


Fig. 4 Time series data with seasonal elements depicted

data pattern) leading to a predictability and forecast of that element and identification of causal relationship.

5.3 The Alternating Conditional Expectation (ACE) Algorithm

Alternating Conditional Expectation provides relationship between response and multiple predictor variables without a priori assumption of functional relationship which is fundamental requirement for simple regression analysis. ACE (Breiman and Friedman [9] and Guoping Xue et. al. [11]) resolves this shortcoming of parametric regression. In non-parametric regression *a priori* knowledge of the functional relationship between the dependent variable (say “A”) and independent variables (say “ $a_1, a_2, a_3, \dots, a_n$ ”) is not required. ACE generates an optimal correlation between a dependent variable and multiple independent variables through non-parametric transformations of the dependent and independent variables without assuming a functional form between the dependent and independent variables and the transformations are derived exclusively with the data set [9–12, 18]. In fact, one of the main features of such regression is to identify the functional relationship between dependent and independent variables.

Thus, natural gas price can be estimated by modeling various parameters. Conventional regression has generally been accomplished parametrically using multiple linear or nonlinear models that require *a priori* assumptions of functional forms. It is intended to employ econometrics and non-parametric regression to derive functional relationship between natural gas price and crude oil price or gold price by applying the Time Series Model and Alternating Conditional Expectation (ACE) algorithm for estimation of the natural gas price from crude oil and gold price.

6 Design of Experiment

The experiment to forecast natural gas price has been designed on the hypothesis that it may be pragmatic to conceive a model which integrates impacting parameters under consideration viz. crude oil price and gold price. The relationship may be recognized by applying different mathematical techniques including parametric or non parametric regression. In parametric regression, a model is fitted to data by assuming a functional relationship. But, it is not always feasible to distinguish the underlying functional relationship between response and multiple predictor variables. Nonparametric regression is an exceptional data analysis tool for exploring the underlying relationships between response and multiple predictor variables.

It is observed that the historical crude oil and natural gas price are non-linear. The non-linear nature of spot prices make prediction very difficult (Agbon) [6]. Linear models hypothesizing that prices increase monotonously and considering price and demand elasticity have been proposed to forecast crude oil and natural gas prices [6].

However, it may be hypothesized that the predictor variables under consideration i.e. crude oil and gold prices do not hold linear relationship with response variable (i.e. natural gas price). It is proposed to model these parameters by invoking nonparametric regression techniques such as ACE. Nonparametric method is preferred against parametric, because, in parametric regression a model is fitted to data by assuming a functional relationship. But, it may not be possible to characterize the underlying functional construct between dependent (response) and multiple independent (predictors) variables. Methodologies employing nonparametric regression explore such relationship without *a priori* understanding of the dependencies and the functional form is derived solely from data.

Following these premises, models employing ACE and Time Series Econometrics have been developed and the agreement or disagreement of forecast with observed values of dependent parameter has been investigated. It is implied that the degree of agreement between in-sample forecast with observed values of dependent variable may provide a clue to the efficacy of model for future time periods. The models have been evaluated or compared by taking into account the Mean Squared Error (MSE), Akaike information Criteria (AIC) etc.

6.1 Correlation and Regression Analysis

The Pearson correlation coefficient with 2-tailed test of significance between USCOFPP and USNGWHP is 0.78 and between USCOFPP and GP is 0.84 which indicate perfect positive correlation coefficient indicating high degree of linear relationship. The simple linear regression model between USCOFPP and USNGWHP and USCOFPP and GP has $R^2 = 0.6068$ and 0.5409 respectively (Fig. 5). Multiple linear regressions of USNGWHP and USCOFPP and GP have R-Square of 0.73.

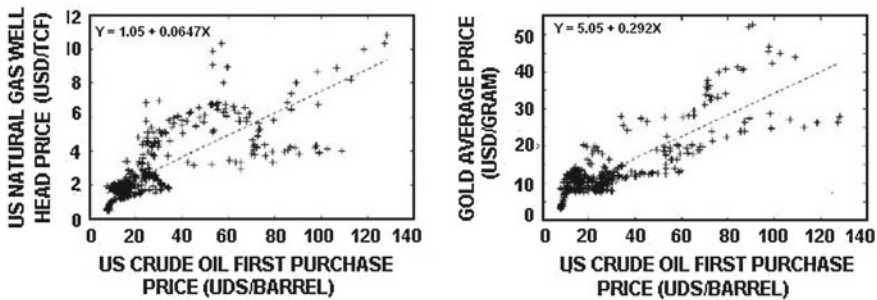


Fig. 5 Linear regression models with least square fit

6.2 Time Series Analysis

It is intended to use time series data of monthly price of USCOFPP, USNGWHP and GP. Prior to invoking time series analysis the data is required to be checked for stationarity and autocorrelation due to non-stationarity. It is imperative to find out whether the regression relationship between variables is spurious, because very often the spurious regression results due to non-stationary time series. It is needed to know now if a meaningful relationship exists between variables. Because, very often very high R^2 value is obtained even though there exists no meaningful relationship. It is of importance to find out whether the time series exhibit Random Walk Phenomena. Moreover, it is prudent to know if such forecasting is valid if the time series under consideration is non-stationary.

6.2.1 Test of Stationarity of Time Series

In order to test for stationarity, Augmented Dickey Fuller (ADF) Test is conducted. The more negative the Dickey Fuller Test statistic value and lesser the p value, there is more chance for the series to be stationary. The output of ADF test (Table 1) suggests that the natural gas well head price time series is stationary, but, the time series of crude oil price exhibits random walk process and non stationarity. The time series of gold price exhibits non stationary behavior. ADF test of difference of log prices of crude oil and natural gas, and crude oil and gold provide the p value as 0.702209 and 0.176501 respectively, which indicate that the series are not mean reverting.

It is essential to stationarize time series data. In order to address this aspect Auto Regressive Moving Average (ARIMA) model has been used. ARIMA is combination of Auto Regressive (AR) and Moving Average (MA) models, and can transform non-stationary time series into stationary time series. AR (p) and MA (q) represent models of p and q order respectively. By differencing “d” times a non-stationary series can be converted to stationary form, and the model is termed ARIMA (p, d, q) model.

6.2.2 ARIMA Model

Reasonable match has been observed between observed and forecasted values of USNGWHP with ARIMA (1, 2, 1) model using Standard error based on Hessian, Cochrane-Orcutt, Prais-Winsten, Hildreth-Lu, and GARCH (Fig. 6 and Tables 2, 3,

Table 1 Output of ADF test

	Crude oil	Natural gas	Gold
Dickey fuller test statistic	-1.68366	-3.47299	2.48024
p-value	0.711171	0.045213	0.99

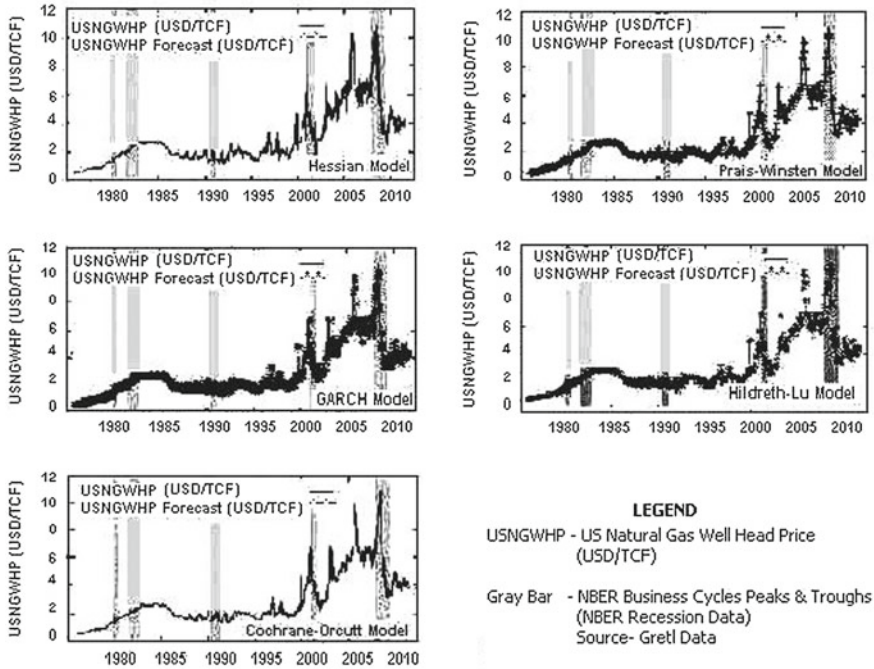


Fig. 6 USNGWHP (USD/TCF)—observed versus forecast

Table 2 USNGWHP—forecast parameters

Standard errors		Coeff.	Std. error	z	p-value
Hessian	GP	0.027	0.023024	1.1865	0.23544
	USCOFPP	0.035	0.006604	5.3234	<0.00001
GARCH Model	GP	0.050	0.00485	10.456	<0.00001
	USCOFPP	0.069	0.001475	47.415	<0.00001
Cochrane-Orcutt	GP	Coeff.	Std. error	t-ratio	p-value
	USCOFPP	0.046	0.022414	2.0802	0.03811
Prais-Winsten	GP	0.034	0.006405	5.4635	<0.00001
	USCOFPP	0.046	0.022371	2.0854	0.03763
Hildreth-Lu	GP	0.034	0.006396	5.4712	<0.00001
	USCOFPP	0.046	0.02244	2.0742	0.03866
		0.034	0.006406	5.4595	<0.00001

and 4). Simultaneous equation models very much in vogue during 1970’s may not be of relevance due to its limitations and “Lucas Critique” which advocated that the response parameters are not insensitive to policy changes (Gujarati) [13]. In fact, policy changes may impact the time series data under consideration.

Table 3 USNGWHP—forecast parameters (ARIMA model)

Parameters	Standard errors based on Hessian	GARCH model
Mean dependent var	2.89	2.89
Schwarz criterion	510	767
S.D. dependent var	1.92	1.92
S.D. of innovations	0.42	
Akaike criterion	485	743
Hannan-Quinn	495	753
Log-likelihood	-237	-365
Mean of innovations	0.002629	

Table 4 USNGWHP—forecast parameters (ARIMA model)

Parameters	Standard errors based on		
	Cochrane-Orcutt	Prais-Winsten	Hildreth-Lu
Mean dependent var	2.896192	2.89069	0.896192
Sum squared resid	75.91233	75.9125	5.91233
R-squared	0.952007	0.95217	0.952007
F(2, 426)	2.31038	22.4049	2.226
rho	0.009027	0.009038	0.008868
S.D. dependent var	1.916398	1.917536	1.916398
S.E. of regression	0.422135	0.421641	0.422135
Adjusted R-squared	0.951894	0.952062	0.951894
P-value(F)	6.10E-10	5.59E-10	6.58E-10
Durbin-Watson	1.978968	1.978952	1.97929

Foregoing describes various modelling techniques which have been exercised to forecast natural gas price. Each model provides with varying degree of certainty of the forecast, and reasonable match has been observed between observed and forecasted values of dependent variable.

6.3 ACE Model

Modeling of time series data with ACE provides Optimal Regression Correlation = 0.95, Optimal Inverse Transform $R^2 = 0.96$, Predicted Standard deviation = 0.9435 and Fitted Standard deviation = 1.29. The value USNGWHP may be estimated from the equation iii (Fig. 7 inset). The parameters considered in the model have been observed to have significant correlation with response variable as evidenced by strong regression correlation.

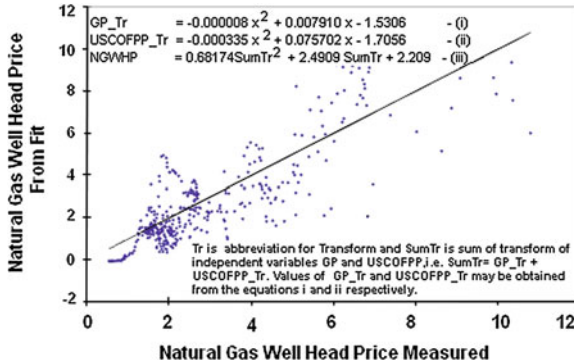


Fig. 7 Transform obtained from ACE

7 Analysis of Results

The estimate of US Natural Gas well head price has been attempted using values of US Crude Oil first purchase price and Gold average price by applying Time Series and ACE algorithm. Time Series models provide comparable values of forecast against the observed values. The transform obtained by ACE has an optimal regression coefficient of 0.95 and the satisfactory fit is observed between measured and that derived from transform.

8 Conclusion

Several methodologies have been proposed to forecast Natural Gas price. It is logical to devise methodologies which avoid any *a priori* underlying assumptions about relationship between the factors, and the model thus derived respect the historical values of predictor and response variables (Mishra) [19]. ARIMA and ACE offer solution to the problem and the result obtained by these techniques suggests that these may be used with reasonable degree of confidence.

References

1. D F. Reiter (1999) Prediction of short-term natural gas prices using econometric and neural network models, SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas 21–23 March 1999. SPE Publication # 52960
2. Solomon (2001) The responsiveness of global E&P industry to changes in petroleum prices: evidence from 1960–2000, SPE hydrocarbon economics and evaluation symposium, Dallas, Texas, 2–3 April 2001. SPE publication # 68587

3. Zamani, An econometrics forecasting model of short term oil spot prices. 6th IAEE European Conference, 2004, pp 1–7
4. Pindyck R S (1994) Inventories and the short-run dynamics of commodity prices. *RAND J Econ* 25(1):141–159 (The RAND Corporation), Spring.
5. Stephane D, Marcelo, Karadeloglou, Pavlos, Kaufmann, Robert K, Sanchez (2007) Modeling the world oil market: Assessment of a quarterly econometric model. *energy policy* 35(1):178–191
6. Agbon (2003) Predicting oil and gas spot prices using chaos time series analysis and fuzzy neural network model. SPE hydrocarbon economics and evaluation symposium, Dallas, Texas, 5–8 April 2003. SPE publication # 82014
7. Jablonowski and MacAskie (2007) The value of oil and gas price forecasts, hydrocarbon economics and evaluation symposium, Dallas, Texas, U.S.A, 1–3 April 2007. SPE, Publication # 107570
8. Ogwo (2007) Equitable gas pricing model, Nigeria annual international conference and exhibition, Abuja, Nigeria, 6–8 August 2007. SPE Publication # 11897
9. Briemann L, Friedman J H (1985) Estimating Optimal Transform for multiple regression and correlation. *J Amer Stat Asso* 79:321–328
10. Wang D (2004) Estimating optimal transformations for multiple regression using the ACE algorithm. *J Data Sci* 2(2004):329–346
11. Xue G (1997) Optimal transformations for multiple regression: application to permeability estimation from well logs. *SPE Form Eval* 12(2):85–93
12. Xue G, Datta-Gupta A, Valkó P, Blasingame TA (1997) Optimal transformations for multiple regression, application to permeability estimation from well logs. *SPE Form Eval* 12(2):85–93
13. Gujarati (2007) Basic econometrics, 4th edn. The Tata McGraw-Hill Companies Publishing Company Ltd, New Delhi
14. Ramu Ramanathan (2002) Introductory Econometrics with applications, 5th edn. Thomson South-Western, Cincinnati, Ohio
15. Quah J and Mishra P (2011) Application of Nonparametric Regression network to model risk parameters for ranking countries to carry out business in water, electronics, education, pharmaceuticals, and infrastructure sectors. Proceedings of the 2011 International Conference on Information & Knowledge engineering, WORLDCOMP'2011, USA. 199–204
16. U.S. Crude Oil First Purchase Price (Monthly data). http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=F000000_3&f=M
17. U.S. Natural Gas Well Head Price (Monthly data). <http://www.eia.gov/dnav/ng/hist/n9190us3M.htm>
18. Gold Average Price (Monthly data). <http://www.kitco.com>
19. Mishra P (2012) Forecasting Natural Gas price-Time Series and Nonparametric approach, lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering WCE 2012, 4–6 July, 2012U.K, London, pp 490–497

Structured Data Mining for Micro Loan Performance Prediction: The Case of Indonesian Rural Bank

Novita Ikasari and Fedja Hadzic

Abstract The ability to predict small businesses' future loan performance based on submitted loan applications is crucial for Indonesian rural banks. The small capacity of these particular banks requires an efficient approach to extract knowledge from structured (quantitative) and unstructured (qualitative) type of credit information. The eXtensible Markup Language (XML) is used to organize this complementary credit data from an Indonesian rural bank. The credit performance evaluation application presented utilizes a mapping approach to preserve structural aspects of data within a format on which wider selections of data mining techniques are applied. Results from decision tree and association rule mining algorithms demonstrate the potential of the approach to generate reliable and valid patterns useful for evaluation of existing lending policy.

Keywords Credit performance evaluation · Data mining techniques · Database structure model · Indonesian rural bank · Loan performance · XML

N. Ikasari (✉)
School of Economics and Finance, Curtin Business School,
Curtin University, Perth, Australia
e-mail: novita.ikasari@postgrad.curtin.edu.au

N. Ikasari
Faculty of Social and Political Science, University of Indonesia,
Depok, Indonesia

N. Ikasari
Enterprise Unit 4, Technology Park, De Laeter Way,
Bentley, WA 6102, Australia

F. Hadzic
Department of Computing, Faculty of Science and Engineering,
Curtin University, Perth, Australia
e-mail: f.hadzic@curtin.edu.au

F. Hadzic
Building 314-New Technologies, Bentley Campus, Curtin University,
GPO Box U1987, Perth, WA 6845, Australia

1 Introduction

Decades before international community marched on to recognize the positive economic contributions of Micro, Small and Medium Enterprises (MSMEs), household and micro businesses had to depend on rural financial setup to fulfill any funding needs. Approach to micro funding was sporadic and localized, with funds made available mostly by donors under development cooperation. Sustainability became critical issue for the funded, and even more for loan providers. In 1992, a theoretical approach for sustainable micro financial institutions was developed, highlighting determinants such as the role of savings and loans, self-sufficiency capacity, types of institutional setup and promotion modes [1]. The labors on micro and small business financing reached its pinnacle when the United Nations declared 2005 as the International Year of Microcredit and Muhammad Yunus received a Nobel Prize in 2006 for his work on Grameen Bank.

The dynamics of micro business financing were found to be constructive for policy development in developing countries [2, 3]. This is due to the socio-economic impacts that these institutions contribute through provision of loans. In Indonesia, as many as 1,669 rural banks are actively supporting small businesses [4]. These banks have capitalization of IDR 100 billion, are located in proximity of many small businesses, and offering only saving and loan services. Such features present advantage and disadvantage reflected by banks' loan performance. The small size and close proximity with the customers suggest that rural banks have an incentive to disburse smaller loans to familiar debtors, thus limit their exposure to default payments. However, risk is increased when banks are incapable to develop a reliable credit profile due to asymmetric information that stems from the nature of micro businesses' management [5]. This increases the need to make informed lending decision using knowledge from relationship lending practice.

The lack of and erratic documentations of micro businesses' financial transactions have forced banks to construct credit profile from structured (quantitative) and unstructured (qualitative) type of data. Although the Indonesian central bank has enforced the 5Cs good lending concept (of character, capacity (to make payments), capital, collateral and relevant conditions of economy), rural banks are not equipped with efficient approach to assess loan applications where a combination of these types of data is available. Furthermore, existing credit risk assessment methods for MSMEs have pivoted around quantitative data [6, 7], which leaves academic as well as practical gap in this domain to be fulfilled.

Our work in this area is characterized by using eXtensible Markup Language (XML) [8] to combine structured and unstructured data in one template and use tree mining techniques for analysis. Due to structural complexities impeding in-depth analysis via tree mining, our focus has recently changed to the use of Database Structure Model (DSM) approach [9]. It allows direct application of standard data mining techniques on XML, and in context of data provided by our industry partner (The Bank), the findings provide practical support in terms of MSMEs lending strategy

formulation [10]. This paper is an extended version of our work [10], including more discussions on the practical implications of the results.

In the next section we highlight existing studies on credit scoring method for MSMEs as well as the method used by The Bank. In Sect. 3, credit data in XML is presented and the structural aspects of XML illustrated. The applied method is discussed in Sect. 4. Section 5 presents the credit data profile, experiments and discussions of findings related to credit policy development. The paper is concluded in Sect. 6.

2 Review of Existing Works and Five Cs' Good Lending Concept

The need for banks to operate in an efficient and effective manner has been satisfied by statistically developed scoring, providing an edge on decision making [11] and outcome [12, 13]. Regardless of MSMEs' long existence, studies on relevant credit risk assessment methods are scarce. It is economically irrational for The Bank to exercise scoring since the initiation and maintenance of such system will take away significant amount of resources. The Bank is obliged to conform to 5Cs as set by the central bank and has applied them in a more straightforward manner where information on each of the Cs is captured irrespective of the designated label. In the next section we highlight the works on credit scoring method for MSMEs and shed light on 5Cs lending principle practiced by The Bank.

2.1 Credit Scoring Methods for MSMEs

The initial work on credit scoring using multiple discriminant analysis showed a promising 92 % and 80 % accuracy level on train and test dataset, respectively [14]. With growing criticism on application of discriminant analysis on credit scoring [15], logistic regression has emerged as preferred statistics method and proven to give high prediction accuracy [16]. The latter is used to overcome normality and equal variance assumptions of discriminant analysis. However, when it was compared to machine learning methods of decision tree and neural network, it was outperformed by neural network [17]. Another study used machine learning to develop a flexible credit scoring, whereby determinant factors for good and bad loans are automatically updated respective of movement in macroeconomic indicators [6].

The existing literature has documented the dominant role of structured financial information [6, 7] and the supplementary role of structured non-financial information [11, 18] to arrive at loan granting decision for small businesses. Unfortunately, available methods are yet to be well-established to provide satisfactory credit risk assessment methods for MSMEs that incorporates structured (categorical) and unstructured (text) types of data.

2.2 Implementation of 5Cs

Indonesian banks are given latitude to interpret the 5Cs principle to assess credit risk. In this context, The Bank perceives the first C, “Character”, as customers’ willingness to fulfill his/her loan obligations. Provided the quite wide recognition of the surrounding community, many customers are recurring debtors with loan history being systematically recorded and manually filed. The Bank practically has built professional and personal rapport with debtors through years of lending transactions. Since many customers are local inhabitants who run the business from their home, “Capacity (to pay)” takes a form of a simple but informative financial report, constructed from interviews and observations. Next, “Capital” is simplified by using the value of cost of goods sold as the proxy. Safety net for banks, “Collateral” is categorized into four types, namely land and building, motorcycles, cars and bank’s deposit. The last C, “Conditions (of economy)”, is constructed implicitly through information on how debtor oversees the chain of supply within industry competition and services.

3 XML Representation of Credit Data

In domains where the nature of the data is more complex and a domain-specific way of organizing the available data is required, semi-structured documents such as XML are often used [19]. Our motivation to use XML is to capture structured and unstructured data in a domain-specific way and effectively contextualize the information. Following data preprocessing, credit data is then populated into an XML document based on a pre-defined XML template. Figure 1 shows one credit application of the resulting XML document with selected attributes (subset) and values.

```
<?xmlversion="1.0"encoding="utf-8"?>
<CreditApplication id="PSP01">
  <loanapplication>
    <industry1>trade</industry1>
    <industry2>nr</industry2>
    <loanscheme>
      <principal>[100000000-249999000]</principal>
      <dailyprincipal>[140277.7778 -324074.0741]</dailyprincipal>
      <dailyinstallment>[225000-277370] </dailyinstallment>
      <dailyinstallmentdeposit>[4500-23580] </dailyinstallmentdeposit>
      <loanduration>720</loanduration>
      <interestrate>[14.4 - 19.0]</interestrate>
    </loanscheme>
  </loanapplication>
  <creditperformance>performing</creditperformance>
</CreditApplication>
...
</xml>
```

Fig. 1 A fragment of XML template

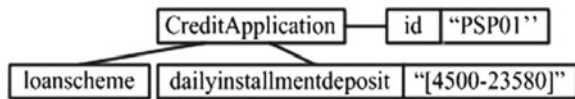
An XML document has a hierarchical document structure, where an element may contain further embedded elements, and each element can be attached with a number of attributes. It is therefore frequently modeled using a rooted ordered labeled tree. Several works [20–22] have been proposed for mining of semi-structured documents such as XML. Initially, the focus was mainly on values associated with the tags, which is by and large no different from traditional association rule mining. However, for certain application domains and to maximize the information content of discovered knowledge patterns, it is necessary to take the structural information of the document into account. In the remainder of this section, definitions and concepts have been reproduced and readapted from Hadzic et al. [19].

3.1 XML Document Entities

Nodes can be categorized as simple and complex [21]. Simple or basic nodes have no edges emanating from them. Complex nodes are also called internal nodes. From Fig. 2, representative simple nodes examples would be <industry1>, <industry2>, <principal>, <dailyprincipal>, etc. The complex nodes examples are <CreditApplication>, <loanapplication>, <loanscheme>.

The element-attribute relationship in XML is of significant value. On tree structure, this is more or less a depiction of a node with multi-labels and the level of relationships among them is of equal value. Relationships between elements in XML are the basic construct for hierarchical relationships. The relationships of two elements are either parent-child relationships or ancestor-descendant relationships. The two elements that are connected by one edge are the basis for a parent-child relationship. The two elements that are connected by more than one edge are the basis for ancestor-descendant relationships. The element-element relationship must be constructed only by two elements from different levels. Examples of both element-element relationships that is of parent-child and element-element relationships that is of ancestor-descendant are shown in Fig. 2.

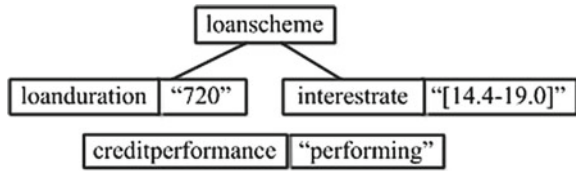
Fig. 2 Illustration of element-element and element-attribute relationships



3.2 Tree-Structured Items

Figure 3 shows examples of tree-structured items from Fig. 1. Elements that form sibling relationships may have ordering imposed on them. Each element of an XML document has a *name* (e.g. *loanduration*) and can have a *value* (e.g. “720”). Given

Fig. 3 Tree-structured items with size 3 (*top*) and with size 1 (*bottom*)



such parallelisms, an XML document can therefore be modeled as a rooted ordered labeled tree [19].

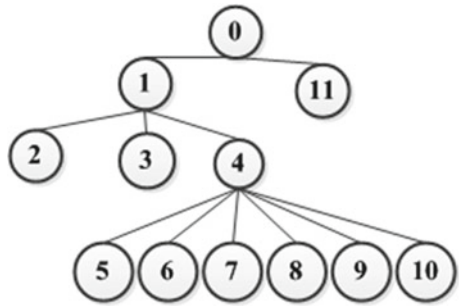
If structures and values are to be considered, XML can be transformed into a string where elements’ names and their inner text (value) are appended as a single string label (i.e. node/vertex label). From Fig. 2, the element <dailyinstallmentdeposit> and its value can be transformed into a single string ‘dailyinstallmentdeposit[“4500–23580”]’. The data representation/mining approach adopted in this work and described in Hadzic [9] utilizes the pre-order (depth-first) string encoding [22]. Since the processing of long strings can be very expensive, a common strategy used to expedite the processing of XML documents is to transform the string encoding into an integer-based form [9]. With this approach, the textual content of each element node will be mapped into an integer number and the mapping is stored in an index table where the original string can be looked up at later time for reporting purposes.

Table 1 is an example of a mapping between the strings (elements (and values)) and unique integers for the XML fragment in Fig. 1. One can use any hash function to map such strings into integer indexes. With this, the pre-order string encoding representation of the underlying tree structure of the credit application of Fig. 1 is transformed to “0 1 2 -1 3 -1 4 5 -1 6 -1 7 -1 8 -1 9 -1 10 -1 -1 -1 11” with corresponding tree shown in Fig. 4.

Table 1 Example of string to integer mapping for XML fragment of Fig. 1

CreditApplication	0
loanapplication	1
industry1[“trade”]	2
industry2[“nr”]	3
loanscheme	4
principal[“[100000000-249999000]”]	5
dailyprincipal[“[140277.7778-324074.0741]”]	6
dailyinstallment[“[225000-277370]”]	7
dailyinstallmentdeposit [“[4500-23580]”]	8
loanduration [“720”]	9
interestrate [“[14.4-19.0]”]	10
creditperformance[“performing”]	11

Fig. 4 Integer-indexed tree of XML fragment of Fig. 1



4 Method

In this work, we utilize a method for effectively representing tree-structured data into a structure preserving flat format with the main motivation of enabling direct application of well-established data mining/analysis techniques to tree-structured data [9]. It is promising for our purpose in the sense that many of the complexity issues [8] caused by the structural properties in the document can be overcome, and class distinguishing criteria can be directly sought after. The technique converts the string like representation into a flat data structure format (henceforth referred as table) so that both structural and attribute-value information is preserved. As the first step, extraction of a structure according to which all the instances/transactions are organized takes place. Thus, each of tree instances are a subtree of this assumed structure, referred to as the Database Structure Model (DSM) [9]. This DSM becomes the first row of the table; the labels of a particular tree instance are placed in the column, corresponding to the position in the DSM where this label was matched to. The string encoding is used to represent the DSM and since the order of the nodes (and backtracks ('-1')) is important, the nodes and backtracks are labeled sequentially according to their occurrence in the string encoding. For nodes (labels in the string encoding), x_i is used as the attribute name, where i corresponds to the pre-order position of the node in the tree, while for backtracks, b_j is used as the attribute name, where j corresponds to the backtrack number in the string encoding.

In our example, from Fig. 4, the string encoding of DSM becomes ' $x_0 x_1 x_2 b_0 x_3 b_1 x_4 x_5 b_2 x_6 b_3 x_7 b_4 x_8 b_5 x_9 b_6 x_{10} b_7 b_8 b_9 x_{11}$ '. This is the first row in the table and to fill in the remaining rows the string encoding of each record is traversed from the tree database. When a label is encountered, it is placed to the matching column under matching node (x_i) in the DSM structure. When a backtrack ('-1') is encountered, a value '1' is placed to the matching backtrack (b_j). Remaining entries are assigned a value of '0' (non-existence). To indicate structural characteristics of the knowledge patterns discovered, they can be re-mapped to the DSM. Hence, structural complexity is avoided and structural characteristics of the data are preserved. For pseudo code of the process and illustrative examples, please refer to Hadzic⁹.

5 Experimental Findings

Banks in Indonesia are not allowed to keep rejected loan applications; therefore it limits the knowledge to only success applications. In this study we focus on the loan performance (performing or non-performing) aspect with the purpose that the insights will benefit The Bank's to limit its credit risk exposure by identifying payment delay early in the stage. Furthermore, it supports lending policy revision upon identification of high risk factors implying non-performance loans.

5.1 Credit Data Profile

The Bank provided 96 credit application instances comprising of 58 performing and 38 non-performing loans at the time of collection. For The Bank it is customary to have debtors with more than one business and this is the case with 35 debtors (36%) in our dataset. The multiple ventures suggest borrower's strong professional portfolio and increase the likelihood of meeting his/her loan obligations on time due to extra source of income. However, lack of focus in business also implicates higher demand on financing. Majority of local community are traders with amount of loan principle concentrated on the low to middle range (IDR 1M up to IDR 50M).

As part of data pre-processing, the numerical values were converted into defined categories, using discretization methods [23]. This is necessary to reduce the number of unique values of an attribute and detect similarities (form generalizations) during the data mining process, in spite of subtle differences between original numerical values. In performing assessment on loan application, The Bank categorized borrower's credit information into Objective and Subjective Analysis to have a straightforward understanding of borrower's business. Objective Analysis contains factual numeric information (80%) on cash inflow and outflow incurred by the business in one financial year. Subjective Analysis contains supplementary and non-business related information in text format. Discretization on attributes that is displayed on a separate part of the loan application document, such as loan principal, loan duration and type of collateral, is done by domain expert in accordance with internal directive memorandum. Data preprocessing on Subjective Analysis was done manually by identifying implicit structure within the text of each sub section.

5.2 Experiments

The structural characteristics of the XML data are as follows: 701 unique labels (attributes and their values when applicable), 82 nodes in each tree instance (transaction), and the maximum height and fan-out of each tree instance were 4 and 12, respectively. The conversion approach described in Sect. 4 produces a relational table

format of this XML data, consisting of a total of 82 attributes (as 81 backtrack attributes (b_j) are omitted but kept in DSM to preserve the structure), each mapped to a particular element (and value if that element has one) from the original XML document. The class to be predicted is “creditperformance” with possible values of “performing” and “nonperforming”. We have used the association rule mining and decision tree learning algorithms from the publicly available machine learning workbench Weka [24] as a case in point.

5.2.1 Association Rule Mining

The Apriori algorithm is applied to generate association rules [25] on 70% of the data (training set) with minimum confidence of 90% and 10% support. A total of 116,654 rules were generated and evaluated for their classification accuracy on the same dataset and predictive accuracy using the “unseen” 30% of the data (testing set). The rule set had high accuracy level on both training (98.03%) and testing (89.71%) data with 100% coverage rate. Below are two examples of rules representing good and bad case of loan payment. For the first rule we also provide an example of its subtree based representation in Fig. 5 when nodes are matched to the extracted DSM. $X_3 = \text{debtorstatus(old)}$ AND $X_{23} = \text{typeofcollateral(vehicle)}$ AND $X_{51} = \text{otherinstallmentexpense(NR)}$ AND $X_{69} = \text{paymentofsales(credit)}$ AND $X_{71} = \text{salessustain(high)} \rightarrow X_{80} = \text{creditperformance(performing)}$ (7).

The first rule is an exemplar of the role of relationship lending in micro lending. With majority customers being in the trade sector, goods and cash turnovers are regarded as an indication of business longevity. Therefore, a business that has credit

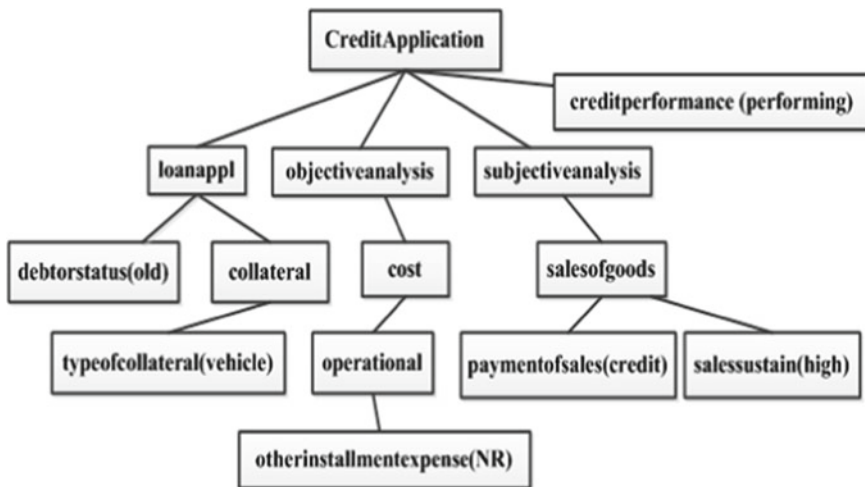


Fig. 5 Subtree based representation with contextual nodes indicated

policy on sales is perceived as risky since it disrupts daily operations. In addition, vehicle is not a preferred proxy for loan risk due to its high transferability as elaborated in the next rule. These two attributes suggest a risky loan. However, the other two attributes of high sales sustainability and the non-existent of other third party liabilities have contributed to applicant's ability to repay the loan. When such ambiguous profile appears, loan officer should examine the determining attribute(s) further, that is attribute(s) with most subjectivity element attached. Of the four attributes, sales sustainability deemed as a result of professional assessment of the domain expert through a collection of qualitative data on applicants' relationship with their suppliers. Thus, status of debtor plays a significant role where trust and previous knowledge do not only contribute to the ease of data collection, but also to the assessment quality. Within this context, the Bank needs to be mindful of loan staff turnover and should develop knowledge management system.

$X_{21} = \text{purposeofloan}(\text{additional capital})$ AND $X_{23} = \text{typeofcollateral}(\text{vehicle})$ AND $X_{57} = \text{totalinstallmentanddepositperday}([17500-149000])$ AND $X_{77} = \text{character}(\text{no information}) \rightarrow X_{80} = \text{creditperformance}(\text{nonperforming})$ (7)

Purpose of loan and characters of borrowers appear to influence loan performance in a proportional manner. From an interview with the owner of The Bank, it was clarified that a moving asset is less preferred collateral than a permanent and secured asset such as land and building. This is based on bad encounters with debtors who "misplaced" or have transferred the ownership of the asset to someone else when loan payment deteriorated, leaving The Bank without any means to recover the loan principal. One striking fact according to this pattern is that the lowest amount of daily installment and deposit is deemed to lead to a delayed loan. The Bank should be concerned with this since it contradicts common perception in micro lending, where micro borrowers are renowned for their daily loan payment track record. The argument behind this abnormality is more of ease of procedure rather than ill-intention. The pattern brings about an issue of price-floor, and the necessity to apply this to MSME lending. Small business entrepreneurs are also notorious for their financial mismanagement whereby business' liquid resources are used to finance non-business activities, hence the importance of collecting information on non-operating expenses such as household expenses, children's education and health expenses, etc. Given the low amount of money to be paid to The Bank, both parties (borrowers and The Bank) have adopted relaxed and flexible approach under the assumption that it is easily recovered in the next payment.

5.2.2 Decision Tree Learning

In general credit risk assessment domain, decision tree learning methods have had popular use since the underlying rules are easily interpreted and the method gives good accuracy results. The C4.5 decision tree algorithm [26] used in this experiment generates a decision tree of size 77 with 68 leaves (rules) in 0.01 seconds; with 94.34% accuracy evaluated using 10-fold cross-validation. We included only the main business sector, since multiple ventures is not a concern at the time of applica-

tion. The root node of the decision tree is “otherinstallmentexpense” having four split nodes reflecting the four categories of “otherinstallmentexpense”. As this attribute influences debtors’ loan repayment ability, the generated rules need to be cautiously considered in particular for non-performing loans. Below are two exemplary rules to explain this purpose.

$X_{42} = \text{otherinstallmentexpense}(\text{NR}) \text{ AND } X_{78} = \text{riskonpayment}(\text{high}) \text{ AND } X_{65} = \text{paymentofpurchase}(\text{cash}) \rightarrow X_{80} = \text{creditperformance}(\text{nonperforming})$ (27)

Intuitively, non-existence of other installment leads to good loan performance due to less financial obligations to be met to other parties. However, the high risk on payment and purchase style obstruct the flow of payment as indicated by the 27 non-performing loans. Since installment is collected daily by The Bank’s collector, high risk of payment indicates a failure to entrust money for loan payment to employees to be forwarded to collector when debtor is not present at the business location. Following this, purchase of trade goods made on cash basis should be noted in relations to sales payment mode. If sales method is credit and purchase method is cash, most likely debtor faces cash flow issue, which might be the reason for loan in the first place. Overall, this instigates The Bank to further examine ways to acquire comprehensive view on debtor’s management skill.

$X_{42} = \text{otherinstallmentexpense}(\text{NR}) \text{ AND } X_{78} = \text{riskonpayment}(\text{medium}) \text{ AND } X_5 = \text{sector1}(\text{delivery}) \rightarrow X_{80} = \text{creditperformance}(\text{nonperforming})$ (3)

The parent node of “riskonpayment(medium)” is fanned out to more than 20 child nodes of X_5 . This is particularly beneficial for The Bank because it provides knowledge on “risky” sectors. Aside from delivery services, other sectors that lead to non-performing loans are fisheries, cultivation, governmental projects, traditional market management, garment and rural groceries shops. Taking into account the location of The Bank, these sectors are not in line with the majority in the area; therefore loan staff should consider specific features related to each sector when developing debtor’s credit risk profile. For example, agribusiness depends heavily on customers’ demands, while these demands are influenced by macroeconomic condition. The fishmongers in this area deliver their catch to local restaurants. With the downturn of the economy, people have tightened their budget on luxury things such as dining out. This reduces the demand of fishery products and created a backlog on product circulation. Since these debtors are not established businessmen, they do not have the needed apparatus to maintain freshness of the fish and face difficulties in making wage payment to their employees, i.e. fishermen. Following these findings, The Bank is best advised to assess carefully sensitivity of different sectors to credit risk.

6 Conclusion

In this paper, we applied a systematic approach to distinguish characteristics that indicate performing and non-performing loans. Credit data obtained from the industry partner are structurally represented in a domain-specific way using an XML template

and a suitable technique is utilized to handle the complex nature of credit data to directly arrive at class discriminating factors. The capability to extract specific rules of high accuracy demonstrates the theoretical and practical contribution of the proposed approach. Some interesting rules have been presented to show different perspectives for internal policy refinement. The significant role of domain expert knowledge was pointed out in regard to a discovered rule. Thus, our future work will explore the role of domain expert knowledge and the task of predicting the periods of delayed payment.

References

1. Rhyne E, Otero M (1992) Financial services for microenterprises: principles and institutions. *World Dev* 20(11):1561–1571
2. Braverman A, Guasch JL (1986) Rural credit markets and institutions in developing countries: lessons for policy analysis from practice and modern theory. *World Dev* 14(10–11):1253–1267
3. Prior F, Argandona A (2009) Credit accessibility and corporate social responsibility in financial institutions: the case of microfinance. *Bus Ethics A Eur Rev* 18(4):349–363
4. Indonesian Bank Statistics (2011) Bank Indonesia, Jakarta, Indonesia
5. Berger AN, Klapper LF, Udell GF (2001) The ability of banks to lend to informationally opaque small businesses. *J Bank Finance* 25(12):2127–2167
6. Tsaih R, Liu Y-J, Liu W, Lien Y-L (2004) Credit scoring system for small business loans. *Decis Support Syst* 38(1):91–99
7. Wu C, Wang X-M (2000) A neural network approach for analyzing small business lending decisions. *J Rev Quant Finance Account* 15(3):259–276
8. Ikasari N, Hadzic F, Dillon TS (2011) Incorporating qualitative information for credit risk assessment through frequent subtree mining for XML. In: Tagarelli A (ed) *XML data mining: models, method, and applications*. IGI Global, Philadelphia, pp 467–503
9. Hadzic F (2012) A structure preserving flat data format representation for tree-structured data. In: Cao L, Huang JZ, Bailey J, Koh YS, Luo J (eds) *Lecture notes in computer science*, vol 7104. Springer, Heidelberg, pp 221–233
10. Ikasari N, Hadzic F (2012) Assessment of micro loan payment using structured data mining techniques: the case of Indonesian people' credit bank. In: Ao SI, Gelman L, Hukins DW, Hunter A, Korsunsky AM (eds) *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012*. London, UK, pp 511–517, 4–6 July 2012
11. Dinh THT, Kleimeier S (2007) A credit scoring model for Vietnam's retail banking market. *Int Rev Financial Anal* 16(5):471–495
12. Abdou H, Pointon J, El-Masry A (2008) A Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Syst Appl* 35(3):1275–1292
13. Chye KH, Chin TW, Peng GC (2004) Credit scoring using data mining techniques. *Singap Manag Rev* 26(2):25–47
14. Edminster RH (1971) Financial ratios and credit scoring for small business loans. *J Commer Bank Lend* September:10–23
15. Eisenbeis RA (1978) Problems in applying discriminant analysis in credit scoring models. *J Bank Finance* 2(3):205–219
16. Altman E, Sabato G (2007) Modelling credit risk for SMEs: evidence from the U.S. market. *Abacus* 43(3):332–357
17. Bencic M, Sarlija N, Zekic-Susac M (2005) Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intell Syst Account Finance Manag* 13:133–150

18. Lehmann B (2003) Is it worth the while?. The relevance of qualitative information in credit rating, SSRN eLibrary
19. Hadzic F, Tan H, Dillon T (2011) Mining of data with complex structures. In: Studies in computational intelligence series, vol 333. Springer, Berlin
20. Chi Y, Nijssen S, Muntz RR, Kok JN (2005) Frequent subtree mining—an overview. *Fundamenta Inform Special Issue Graph Tree Min* 66(1–2):161–198
21. Wang K, Liu H (1998) Discovering typical structures of documents: a road map approach. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval—SIGIR '98, Melbourne, Australia, pp 146–154, 24–28 August 1998
22. Zaki MJ (2005) Efficiently mining frequent trees in a forest: algorithms and applications. *IEEE Trans Knowl Data Eng* 17(8):1021–1035
23. Han J, Kamber M (2006) *Data mining: concepts and techniques*, 2nd edn. Morgan Kaufmann Publishers, California
24. Holmes G, Donkin A, Witten IH (1994) WEKA: a machine learning workbench. In: *Intelligent information systems, 1994. Proceedings of the 1994 second Australian and New Zealand conference*, pp 357–361, 29 Nov–2 Dec 1994
25. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In *VLDB '94 Proceedings of the 20th international conference on very large data bases*, San Francisco, pp 487–499
26. Quinlan R (1993) *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc, San Francisco

Financial Forecasting Using the Kolmogorov–Feller Equation

Jonathan Blackledge, Marc Lamphiere, Kieran Murphy and Shaun Overton

Abstract An approach to analysing a financial time series using the Kolmogorov-Feller Equation is considered, in particular, the Generalised Kolmogorov-Feller Equation (GKFE), subject to variations in the Stochastic Volatility. Using the Mittag-Leffler memory function, we derive an expression for the Impulse Response Function associated with a short time window of data which is then used to derive an algorithm for computing a new index using a standard moving window process. It is shown that application of this index to financial time series, subject to a low volatility condition, correlates with the start, direction and end of a trend depending on the sampling rate of the time series and the look-back window or ‘period’ that is used. An example of this is provided in the chapter using MetaTrader4.

Keywords Generalised Kolmogorov-Feller equation · Impulse response function · MetaTrader4 · Mittag-Leffler memory function · Time series analysis · Trend analysis · Stochastic volatility

J. Blackledge (✉) · M. Lamphiere
Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland
e-mail: jonathan.blackledge@dit.ie

M. Lamphiere
e-mail: marclamphier@gmail.com

K. Murphy
TradersNow Limited, 15 Cabinteely Way, Cabinteely, Dublin 18, Ireland
e-mail: kieran@tradersnow.com

S. Overton
2665 Villa Creek Dr, Suite 125, Dalla, TX 75234, USA
e-mail: soverton@onestepremoved.com

1 Introduction

This chapter follows the work of [1], extending the material to include an implementation of the results using *MetaTrader4*. Beta tested *MetaTrader4* functions are provided for those readers interested in implementing the approach considered, analysing other financial time series data and extending the algorithms further.

Price models involve the derivation and solution of a variety of stochastic differential and partial differential equations. Consider the ‘standard model’ for the price of a stock as a function of time $s(t)$ given by (e.g. [2])

$$\frac{d}{dt}s(t) = \mu s(t) + \sigma s(t)u(t) \quad (1)$$

where μ is the ‘Drift’, σ is the ‘Volatility’ and $u(t)$ is a stochastic function. This model is based on the idea that prices appear to be the previous price plus some random change and that these price changes are independent, i.e. asset price changes appear to be random and independent, prices being taken to follow some random walk-type behaviour. This is the basis for including a stochastic function $u(t)$. However the size of price movements also depends on the size of the price itself. The model is therefore revised to include this effect, the stochastic term $u(t)$ being replaced by $u(t)s(t)$ where σ determines the degree of randomness taken to influence a price change. In general, μ and σ vary with time, and, in the context of Eq. (1), $\sigma(t)$ is referred to as the ‘Stochastic Volatility’, e.g. [3–5]. The drift function $\mu(t)$ tends to vary over longer periods of time reflecting the long term trends associated with a price index.

In principle, $u(t)$ could be any stochastic function with statistical behaviour conforming to a range of Probability Density Functions. A conventional model is to assume that the log price changes are Gaussian distributed so that $u(t)$ is taken to be a zero-mean Gaussian distributed function. If this function is taken to have a fixed standard deviation of 1, then the volatility becomes a measure of the standard deviation, at least, for a (zero-mean) Gaussian model. The stock price model given by Eq. (1) then provides a method for estimating the volatility σ in terms of a lower bound as discussed in the following section.

In this chapter, we consider a solution to the Generalised Kolomogorov-Feller Equation to model the stochastic behaviour of a financial time series. By defining an Impulse Response Function which is based on a parameter associated with the Mittag-Leffler memory function used to construct the KFE, we consider an algorithm for analysing the trends of the time series.

2 Evaluation of the Stochastic Volatility

Consider the rate equation

$$f(t) = \mu + \sigma u(t)$$

where

$$f(t) = \frac{1}{s(t)} \frac{d}{dt} s(t) = \frac{d}{dt} \ln s(t)$$

and μ and σ are taken to be constant. We first obtain an estimate of the Drift by noting that, if the mean of $u(t)$ is approximately zero over $t \in [0, T]$, then

$$\int_0^T f(t) dt = \int_0^T \mu dt + \sigma \int_0^T u(t) dt \sim \mu T$$

so that

$$\mu \sim \frac{1}{T} \int_0^T f(t) dt \tag{2}$$

To obtain an estimate for the volatility, we now consider the case when the stochastic function $u(t)$ is a phase only function, i.e. given that

$$\tilde{u}(\omega) = \int_{-\infty}^{\infty} u(t) \exp(-i\omega t) dt$$

where ω is the (angular) frequency, we consider

$$\tilde{u}(\omega) = A \exp[i\theta(\omega)] \tag{3}$$

where the amplitude spectrum A is taken to be a constant for all values of ω . We also consider $u(t)$ to be a band-limited function $\omega \in [-\Omega/2, \Omega/2]$ with bandwidth Ω and a function of compact support $t \in [-T/2, T/2]$. Using Minkowski’s identity for Euclidean norms,

$$\|f(t)\|_2 \leq \|\mu\|_2 + \|\sigma u(t)\|_2$$

where

$$\|x(t)\|_2 := \left(\int |x(t)|^2 dx \right)^{\frac{1}{2}}$$

so that we can write

$$\sigma \|u(t)\|_2 \geq \|f(t)\|_2 - \mu\sqrt{T}$$

where μ is given by Eq. (2). Using Parseval’s Theorem (Rayleigh’s Energy Theorem), the condition expressed by Eq. (3) allows us to write

$$\int_{-T/2}^{T/2} |u(t)|^2 dt = \frac{1}{2\pi} \int_{-\Omega/2}^{\Omega/2} |\tilde{u}(\omega)|^2 d\omega = \frac{\Omega A^2}{2\pi}$$

We can therefore consider the equation

$$\sigma_{\min} = \frac{1}{A} \sqrt{\frac{2\pi}{\Omega}} (\|f(t)\|_2 - \mu\sqrt{T}) \tag{4}$$

which yields an expression for the lower bound of the volatility.

3 Numerical Computation of the Stochastic Volatility

Consider a discrete signal denoted by the array $f_n, n = 1, 2, 3, \dots, N$ where a uniform sampling interval of Δt is assumed. In this case, the discrete version of Eq. (4) becomes

$$\sigma_{\min} = \frac{1}{A} \sqrt{\frac{2\pi}{\Omega}} (\sqrt{\Delta t} \|f_n\|_2 - \mu\sqrt{T})$$

where we invoke the usual definition for a vector (Euclidean) norm, i.e.

$$\|f_n\|_2 := \left(\sum_{n=1}^N |f_n|^2 \right)^{\frac{1}{2}}, \quad \mu = \frac{\Delta t}{T} \sum_{n=1}^N f_n$$

The sampling interval Δt of f_n is related to the sampling interval $\Delta\omega$ of the Discrete Fourier Transform of f_n by the equation

$$\Delta t \Delta\omega = \frac{2\pi}{N}$$

and since the bandwidth of the discrete spectrum of f_n is $N\Delta\omega$ is clear that $\Delta t = 2\pi/\Omega$. Thus, given that the support of the signal is $T = N\Delta t$, we note that

$$T = \frac{2\pi N}{\Omega}$$

and therefore obtain

$$\sigma_{\min} = \frac{2\pi}{A\Omega} (\|f_n\|_2 - \sqrt{N}\mu), \quad \mu = \frac{1}{N} \sum_{n=1}^N f_n$$

The scaling constant $2\pi/(A\Omega)$ can then be used to define a re-scaled Stochastic Volatility given by

$$\hat{\sigma} := \sigma_{\min} \frac{A\Omega}{2\pi}$$

thereby yielding the expression

$$\hat{\sigma} = \|f_n\|_2 - \sqrt{N}\mu$$

Writing this result explicitly in terms of the price value s_n we obtain the equation

$$\hat{\sigma} = \left(\sum_{n=1}^{N-1} \left| \ln \left(\frac{s_{n+1}}{s_n} \right) \right|^2 \right)^{\frac{1}{2}} - \frac{1}{\sqrt{N-1}} \sum_{n=1}^{N-1} \ln \left(\frac{s_{n+1}}{s_n} \right) \tag{5}$$

To compute the ‘Stochastic Volatility’ σ_m , N is taken to determine the size of the data sampling window or ‘look-back’ window which is moved along the time series one element at a time so that we can write

$$\hat{\sigma}_m = \left(\sum_{n=1}^{N-1} \left| \ln \left(\frac{s_{m+n+1}}{s_{m+n}} \right) \right|^2 \right)^{\frac{1}{2}} - \frac{1}{\sqrt{N-1}} \sum_{n=1}^{N-1} \ln \left(\frac{s_{m+n+1}}{s_{m+n}} \right) \tag{6}$$

Equation (5) may be compared with other estimates for the Stochastic Volatility such as the Maximum Likelihood (ML) estimate given by, [6]

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{N-1} \sum_{n=1}^{N-1} \left[\ln \left(\frac{s_{n+1}}{s_n} \right) \right]^2 - \frac{1}{(N-1)^2} \left[\ln \left(\frac{s_N}{s_1} \right) \right]^2$$

The phase only condition used to derive Eqs. (5) and (6) is equivalent to modelling the stochastic function $u(t)$ in terms of a random walk in the (complex) Fourier domain where the amplitude of each step is the same.

4 Derivation of the Generalised Kolmogorov-Feller Equation

For an arbitrary Characteristic Function $P(k)$ with Probability Density Function (PDF) $p(x)$, Einstein’s evolution equation is, [7]

$$u(x, t + \tau) = u(x, t) \otimes_x p(x)$$

where $u(x, t)$ is a ‘density function’ representing the concentration of a canonical ensemble of particles undergoing elastic collisions. Consider a Taylor series for the function $u(x, t + \tau)$, i.e.

$$u(x, t + \tau) = u(x, t) + \tau \frac{\partial}{\partial t} u(x, t) + \frac{\tau^2}{2!} \frac{\partial^2}{\partial t^2} u(x, t) + \dots$$

For $\tau \ll 1$

$$u(x, t + \tau) = u(x, t) + \tau \frac{\partial}{\partial t} u(x, t)$$

and we obtain the ‘Classical KFE’ ([8, 9])

$$\tau \frac{\partial}{\partial t} u(x, t) = -u(x, t) + u(x, t) \otimes_x p(x) \tag{7}$$

Equation (7) is based on a critical assumption which is that the time evolution of the field $u(x, t)$ is influenced only by short term events and that longer term (historical) events have no influence on the behaviour of the field, i.e. the ‘system’ described by Eq. (7) has no ‘memory’. This statement is the physical basis upon which we introduce the condition $\tau \ll 1$ thereby allowing the Taylor series expansion of the $u(x, t + \tau)$ to be made to first order. The question then arises as to how longer term temporal influences can be modelled, other than by taking an increasingly larger number of terms in the Taylor expansion of $u(x, t + tau)$ which is not of practical analytical value. For arbitrary values of τ ,

$$\tau \frac{\partial}{\partial t} u(x, t) + \frac{\tau^2}{2!} \frac{\partial^2}{\partial t^2} u(x, t) + \dots = -u(x, t) + u(x, t) \otimes_x p(x)$$

We can model the effect on a solution for $u(x, t)$ of the series on the left hand side of this equation in terms of a ‘memory function’ $m(t)$ and write

$$\tau m(t) \otimes_t \frac{\partial}{\partial t} u(x, t) = -u(x, t) + u(x, t) \otimes_x p(x) \tag{8}$$

where \otimes_t is taken to denote the causal convolution integral over t . This is the Generalised KFE (GKFE) which reduces to the Classical KFE when

$$m(t) = \delta(t)$$

Note that for any memory function for which there exists a function or class of functions of the type $n(t)$, say, such that

$$n(t) \otimes_t m(t) = \delta(t)$$

then we can write Eq. (8) in the form

$$\tau \frac{\partial}{\partial t} u(x, t) = -n(t) \otimes_t u(x, t) + n(t) \otimes_t u(x, t) \otimes_x p(x) \tag{9}$$

where the Classical KFE is recovered when $n(t) = \delta(t)$.

Any solution obtained to the GKFE will be dependent upon the choice of memory function $m(t)$ used. There are a number of choices that can be considered, each or which is taken to be a ‘best characteristic’ of the stochastic system in terms of the influence of its time history. However, it may be expected that the time history of physically significant random systems is relatively localised in time. This includes memory functions such as the Mittag-Leffler function, [10]

$$m(t) = \frac{1}{\Gamma(1 - \beta)t^\beta}, \quad 0 < \beta < 1$$

where

$$n(t) = \frac{1}{\Gamma(\beta - 1)t^{2-\beta}}$$

given that

$$\int_0^\infty \frac{\exp(-st)}{\Gamma(\beta)t^{1-\beta}} dt = \frac{1}{s^\beta} \quad \text{and} \quad \int_0^\infty \delta(t) \exp(-st) dt = 1$$

5 Solution to the GKFE Using the Green’s Function Method

Consider Eq. (9) which can be written in the form

$$\begin{aligned} \tau \frac{\partial}{\partial t} u(x, t) + u(x, t) &= u(x, t) - n(t) \otimes_t u(x, t) \\ &+ n(t) \otimes_t u(x, t) \otimes_x p(x) \end{aligned}$$

so that the Green’s function solution is given by

$$\begin{aligned} u(x, t) &= g(t) \otimes_t u(x, t) - g(t) \otimes_t n(t) \otimes_t u(x, t) \\ &+ g(t) \otimes_t n(t) \otimes_t u(x, t) \otimes_x p(x) \end{aligned} \tag{10}$$

where the Green’s function is given by

$$g(t) = \frac{1}{\tau} \exp(-t/\tau), \quad t > 0$$

which is the solution to

$$\tau \frac{\partial}{\partial t} g(t - t_0) + g(t - t_0) = \delta(t - t_0)$$

and we assume the initial conditions $u(x, t = 0) = 0$ and $g(t = t_0) = 0$. We can now analyse this solution in Fourier-Laplace space by taking the Fourier transform

and the Laplace transform of Eq. (10) and using the convolution theorems for the Fourier and Laplace transform, respectively, to obtain

$$\tilde{u}(k, s) = \bar{g}(s)\tilde{u}(k, s) - \bar{g}(s)\bar{n}(s)\tilde{u}(k, s) + \bar{g}(s)\bar{n}(s)\tilde{u}(k, s)\tilde{p}(k) \tag{11}$$

where

$$\begin{aligned} \tilde{u}(k, s) &= \int_0^\infty \int_{-\infty}^\infty u(x, t) \exp(-ikx) dx \exp(-st) dt \\ \bar{g}(s) &= \int_0^\infty g(t) \exp(-st) dt, \quad \bar{n}(s) = \int_0^\infty n(t) \exp(-st) dt \end{aligned}$$

and

$$\tilde{p}(k) = \int_{-\infty}^\infty p(x) \exp(-ikx) dx$$

From Eq. (11) it is clear that we can write

$$\begin{aligned} \tilde{u}(k, s) &= -\frac{\bar{g}(s)}{1 - \bar{g}(s)} \bar{n}(s)\tilde{u}(k, s) + \frac{\bar{g}(s)}{1 - \bar{g}(s)} \bar{n}(s)\tilde{u}(k, s)\tilde{p}(k) \\ &= -\frac{\bar{n}(s)}{\tau s} \tilde{u}(x, t) + \frac{\bar{n}(s)}{\tau s} \tilde{u}(k, s)\tilde{p}(k) \end{aligned}$$

given that $\bar{g}(s) = (1 + \tau s)^{-1}$ and thus we obtain the equation

$$\tilde{u}(k, s) = \bar{h}(s)\tilde{u}(k, s)\tilde{p}(k) \tag{12}$$

where

$$\bar{h}(s) = \frac{\bar{n}(s)}{\tau s + \bar{n}(s)}$$

or, upon inverse transformations

$$u(x, t) = h(t) \otimes_t u(x, t) \otimes_x p(x) \tag{13}$$

with

$$h(t) \leftrightarrow \frac{\bar{n}(s)}{\tau s + \bar{n}(s)}$$

where \leftrightarrow denotes the Laplace transformation, i.e. mutual transformation from t -space to s -space.

Consider the iteration of Eq. (13) defined by

$$u_{n+1}(x, t) = h(t) \otimes_t u_n(x, t) \otimes_x p(x) \tag{14}$$

for an initial solution $u_0(x, t)$ where $n = 1, 2, \dots, N$. The equivalent iteration in Fourier-Laplace space is, from Eq. (12)

$$\tilde{u}_{n+1}(k, s) = \bar{h}(s)\tilde{u}_n(k, s)\tilde{p}(k) \tag{15}$$

with initial solution $\tilde{u}_0(k, s)$. From Eq. (15) it is clear that, after N iterations, we can write

$$\tilde{u}_N(k, s) = [\bar{h}(s)]^N [\tilde{p}(k)]^N \tilde{u}_0(k, s)$$

so that upon inverse Fourier-Laplace transformation, Eq. (14) becomes

$$u_N(x, t) = \prod_{j=1}^N p(x) \prod_{k=1}^N h(t) \otimes_x \otimes_t u_0(x, t) \tag{16}$$

where

$$\prod_{j=1}^N f(t) \equiv f(t) \otimes_t f(t) \otimes_t f(t) \otimes_t \dots$$

denoting the N th convolution of $f(t)$. The convergence criterion required for the iteration defined by Eq. (14) is given in the Appendix.

6 Mittag-Leffler Impulse Response Function

Form Eq. (16), if the initial solution is an impulse (i.e. $u_0(x, t) = \delta(x)\delta(t)$) then the Impulse Response Function (IRF), denoted by $r(x, t)$, is given by

$$r(x, t) = \prod_{j=1}^N p(x) \prod_{k=1}^N h(t)$$

with ‘transfer function’

$$\tilde{r}(k, s) = [\bar{h}(s)\tilde{p}(k)]^N$$

For a memory function $m(t)$ modelled by the Mittag-Leffler function (for $0 < \beta < 1$)

$$m(t) \leftrightarrow \frac{1}{s^{1-\beta}}, \quad \bar{h}(s) = \frac{1}{1 + \tau s^\beta} \sim \frac{1}{\tau s^\beta}$$

so that

$$h(t) \sim \frac{1}{\tau \Gamma(\beta) t^{1-\beta}}$$

Similarly, suppose we consider a Mittag-Leffler PDF of the form

$$p(x) = \frac{1}{\Gamma(1-\gamma) |x|^\gamma}, \quad 0 < \gamma < 1$$

so that the IRF becomes

$$r(x, t) \sim \prod_{j=1}^N \frac{1}{\Gamma(1-\gamma) |x|^\gamma} \prod_{k=1}^N \frac{1}{\tau \Gamma(\beta) t^{1-\beta}}$$

Note that, from Appendix, if $\|h(t)\| \times \|p(x)\| \ll 1$ then $r(x, t) \sim p(x)h(t)$, and, in the case of the Mittag-Leffler function used here, this will occur when $\tau \gg 1$. Also, note that $r(x, t) \rightarrow 0$ as $\gamma \rightarrow 1$ and as $\beta \rightarrow 0$.

7 Trend Analysis Using MetaTrader4

On the basis of the results discussed in the previous section, we consider a short time series model given by (for an arbitrary PDF p)

$$\hat{u}(t) \equiv \int_{-\infty}^{\infty} p(x)h(t)dx = \frac{a}{t^{1-\beta}}, \quad \beta > 0$$

where a is a scaling constant. This model represents the IRF associated with a random scaling fractal signal $u(t)$, [2]. For the discrete case when $\hat{u}_n \equiv \hat{u}(t_n)$ (for $n = 1, 2, \dots, N$) is taken to represent a window of data taken from an input data stream,

$$\hat{u}_n = at_n^\alpha, \quad t_n > 0$$

where $\alpha = \beta - 1$. Estimates of the parameters a and α are then chosen to minimise the error function

$$e(a, \alpha) = \|\ln \hat{u}_n - \ln u_n\|_2^2 \equiv \sum_{n=1}^N (\ln \hat{u}_n - \ln u_n)^2$$

where u_n is data which is taken to be normalised, i.e. $\|u_n\|_\infty = 1$. Differentiating with respect to $A = \ln a$ and α , it is trivial to show that

$$\alpha = \frac{\sum_{n=1}^N \ln u_n \sum_{n=1}^N \ln t_n - N \sum_{n=1}^N \ln u_n \ln t_n}{\left(\sum_{n=1}^N \ln t_n\right)^2 - N \sum_{n=1}^N (\ln t_n)^2} \tag{17}$$

and

$$a = \exp\left(\frac{\sum_{n=1}^N \ln u_n - \alpha \sum_{n=1}^N \ln t_n}{N}\right)$$

given that

$$\frac{\partial e}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial e}{\partial A} = 0$$

Note that in general, $\alpha = \beta - 1$ may be greater than (for $\beta > 1$) or less than (for $0 < \beta < 1$) zero thereby providing a measure of any (long term) ascending or descending trends in the data u_n , respectively. An example of this characteristic, coupled with the corresponding Stochastic Volatility of the same financial times series is given in Fig. 1, the version of Metatrader4 used for this application being available from Alpari Limited, [11]. This figure shows the results of computing the Stochastic Volatility using Eq. (6) and the α -index using Eq. (17) for daily sampled ‘Gold Spot’ data (XAGUSD, Daily) from 29 August to 20 September, 2012. This figure shows a window of data over which there are both downward and upward trends of roughly equal range and rate. Note that the Stochastic Volatility decays over the latter upward trend when the XAUUSD, Daily increases from approximately 1573 (on 25 July, 2012) to 1772 (on 1 October, 2012) USD/oz. This result is typical of observations indicating that an increase in the value of the α -index coupled with a decrease in the volatility provides an appropriate ‘investment signature’.

A further example is given in Fig. 2 which shows the α -index, and, as a comparator, the Lyapunov exponent λ computed via the equation

$$\lambda = \frac{1}{N} \sum_{i=1}^N \log \left| \frac{u_{n+1}}{u_n} \right| \tag{18}$$

for EUA (European Union Allowance) Carbon Dioxide Emissions using hourly sampled data (EUA, H1) from 17 August to 5 October, 2012. It is noted that the Lyapunov exponent exhibits similar behaviour to the α -index but that the latter index is smoother (for the same period used).



Fig. 1 The Gold Spot Price (XAUUSD, Daily) from 27 January to 1 October, 2012 (top window), the Stochastic Volatility (centre window) computed using Eq. (6) for a period of 50 days and the α -index (lower window) computed using Eq. (17) also for a period of 50 days



Fig. 2 European Union Allowance CO_2 emissions (EUA, H1) from 17 August to 5 October, 2012 (top window), the Lyapunov exponent (centre window) computed using Eq. (18) for a period of 30 and the α -index (lower window) using Eq. (17) also computed for a period of 30

8 Conclusion

Compared to equations such as the Classical Diffusion and Fractional Diffusion Equations, [12], the GKFE given by Eq. (8) represents a more accurate model for a density function describing random motion that conforms to Einstein’s evolution equation. We have considered the Green’s function solution of the GKFE as a model for a financial time series (or a derived index). The time dependence of this solution

depends upon the memory function used to model the higher order terms in the Taylor series expansion of the evolution equation, and, in this chapter, we have used the Mittag-Leffler memory function. It has been shown that this choice provides a temporal solution that scales at t^α where $\alpha = \beta - 1$, $0 < \beta < 1$. For $\beta > 0$ the parameter α provides an index that identifies the start, direction and end of a trend depending on the position in time where the polarity of α changes from being positive (indicating an upward trend) to negative (indicating a downward trend). Coupled with knowledge of the Stochastic Volatility being relative low, this index therefore provides a quantitative measure for implementing a trading strategy that is predicated on forecasting the type and extent of a trend. For readers interested in further evaluating this approach to trend analysis, the MetaTrader4 .mq4 modules used to obtain the results given in the previous section are available from [13].

Acknowledgments The authors acknowledges the support of the Science Foundation Ireland and Enterprise Ireland.

Appendix: Condition for Convergence of Equation (14)

Consider the error function $\epsilon_n(x, t)$ at any iteration n so that $u_n(x, t) = u(x, t) + \epsilon_n(x, t)$ From Eq. (15) we can then write

$$\tilde{\epsilon}_{n+1}(k, s) = \bar{h}(s)\tilde{p}(k)\tilde{\epsilon}_n(k, s)$$

so that

$$\tilde{\epsilon}_n(k, s) = [\bar{h}(s)\tilde{p}(k)]^n \tilde{\epsilon}_0(k, s)$$

and it is clear that, since we require $\tilde{\epsilon}_n \rightarrow 0$ and $n \rightarrow \infty$, $[\bar{h}(s)\tilde{p}(k)] < 1 \quad \forall(k, s)$. The condition for convergence therefore becomes

$$\|\bar{h}(s)\tilde{p}(k)\| \leq \|\bar{h}(s)\| \times \|\tilde{p}(k)\| < 1$$

or, for Euclidian norms, and, using Rayleigh’s theorem,

$$\|\bar{h}(s)\|_2 \times \|p(x)\|_2 < \frac{1}{\sqrt{2\pi}}$$

In (k, t) -space

$$\tilde{\epsilon}_n(k, t) = \prod_{k=1}^n h(t)[\tilde{p}(k)]^n \otimes_t \tilde{\epsilon}_0(k, t)$$

so that, using Hölder’s inequality

$$\begin{aligned} \|\tilde{\epsilon}_n(k, t)\| &\leq \left\| \prod_{k=1}^n h(t)[\tilde{p}(k)]^n \right\| \times \|\tilde{\epsilon}_0(k, t)\| \\ &\leq \|h(t)\|^n \times \|\tilde{p}(k)\|^n \times \|\tilde{\epsilon}_0(k, t)\| \end{aligned}$$

and the condition for convergence becomes

$$\|h(t)\|_2 \times \|p(x)\|_2 < \frac{1}{\sqrt{2\pi}}$$

References

1. Blackledge JM, Lamphere M, Murphy K, Overton S, Panahi A (2012) Stochastic volatility analysis using the generalised Kolmogorov-Feller equation. In: Lecture notes in engineering and computer science. Proceedings of the world congress on engineering 2012, WCE 2012, UK, London, 4–6 July, 2012, pp 453–458
2. Blackledge JM (2010) The fractal market hypothesis: applications to financial forecasting. Centre for Advanced Studies, Warsaw University of Technology, Poland, ISBN: 978-83-61993-01-83
3. Lamoureux CG, Lastrapes WD (1993) Forecasting stock-return variance: toward an understanding of stochastic implied volatilities. *Rev Financ Stud* 6:293–326
4. Wiggins JB (1987) Option values under stochastic volatilities. *J Financ Econ* 19:351–372
5. Melino A, Turnbull S (1990) The pricing of foreign currency options with stochastic volatility. *J Econ* 45:239–265
6. Stein EM, Stein JC (1991) Stock price distributions with stochastic volatility: an analytic approach. *Rev Financ Stud* 4:727–752
7. Einstein A (1905) On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der Physik (German)* 17:549–560
8. Kolmogorov AN (1992) On analytic methods in probability theory, selected works of A. N. Kolmogorov. In: Shiryaev AN (ed) *Probability theory and mathematical statistics*, vol II. Kluwer, Dordrecht, p 61–108 (From the original: Kolmogorov AN (1931) *Über die Analytischen Methoden in der Wahrscheinlichkeitsrechnung*. *Math Ann* 104:415–458)
9. Feller W (1957) On boundaries and lateral conditions for the Kolmogorov differential equations. *Ann Math 2nd series* 65(3):527–570
10. Olver FW, Maximon LC (2010) Mittag-Leffler function. In: Frank WJ et al (eds) *Handbook of mathematical functions* in Olver. NIST, Cambridge University Press, Cambridge
11. Alpari (2012) Limited is a broker headquartered in London for forex, spread betting, precious metals, futures and energy commodities markets and provides a range of trading platforms including MetaTrader 4. <http://www.alpari.co.uk/>
12. Gorenflo R, Mainardi F, Raberto M, Scalas E (2000) Fractional diffusion in finance: basic theory, a review paper based on a talk given by F. Mainardi at MDEF2000—Workshop ‘Modelli Dinamici in Economia e Finanza’, Urbino (Italy), September 28–30, 2000. <http://www.econ.uniurb.it/bischi/MDEF2000/MainardiMDEF.pdf>
13. MT4 Indicators URL (2012) The MetaTrader4 indicators used to generates the results given in this paper are available from <http://eleceng.dit.ie/jblackledge/Indicators.zip> which provides the.mq4 modules requires to compute the Stochastic Volatility, the α -index and the Lyapunov exponent

Surface Quality Improvement in CNC End Milling of Aluminum Alloy Using Nanolubrication System

Mohd Sayuti Ab Karim, Ahmed Aly Diao Mohammed Sarhan
and Mohd Hamdi Abd Shukor

Abstract Aerospace applications and energy saving strategies in general raised the interest and study in the field of lightweight materials, especially on aluminum alloys. Aluminum Al2017-T4 and Al6061-T6 alloy which are used in this research work have low specific weight and high strength. The (CNC) milling machine facilities provides a wide variety of parameters setup, making the machining process of the aluminum alloy excellent in manufacturing complicated special products. However, the demand for high quality focuses attention especially on the roughness of the machined surface. The key solution for this issue is by introducing the nanolubrication system since it could produce much less friction in the tool-chip interface. In this research work, the Al2017-T4 and Al6061-T6 is machined by using the carbon onion nanoparticle and SiO₂ nanoparticles, respectively when it mixed with ordinary mineral oil at various concentrations as a nanolubrication system. The reduction of surface roughness could be obtained when carbon onion and SiO₂ nanolubricant are used compared with the case of using ordinary lubricant due to the tribological properties of the carbon onion and SiO₂ nanolubricant to reduce the coefficient of friction in the tool-chip interface.

Keywords Al2017-T4 alloy · Al6061-T6 alloy · Carbon onion nanolubrication · End milling · Morphological surface · SiO₂ nanolubrication · Surface quality

M. S. A. Karim (✉) · A. A. D. Mohammed Sarhan · M. H. A. Shukor
Centre of Advanced Manufacturing and Material Processing, Department of Engineering
Design and Manufacturing, Engineering Faculty, University of Malaya,
50603 Kuala Lumpur, Malaysia
e-mail: mdsayuti@um.edu.my

A. A. D. Mohammed Sarhan
e-mail: ah_sarhan@um.edu.my

M. H. A. Shukor
e-mail: hamdi@um.edu.my

1 Introduction

Aluminum and its alloys are today considered one of the most practical of metals for a variety of reasons. Its low cost, light-weight, and modern appearance are among the primary reasons for its widespread use. Furthermore, it is non-sparking, electrically conductive, thermally conductive, non-magnetic, reflective, and chemically resistant. Aluminum Al2017-T4 and Al6161-T6 are among of the highest strength and hardest aluminium alloys with excellent fatigue strength available. Heat-treating increases its strength considerably. It is used for various applications from high strength structural components, aircraft, machine construction, military equipment, rivets. Aluminum Al2017 and Al6061-T6 also has very good machining characteristics and it is best to perform machining with the alloy in the T4 and T6 condition [1].

After the machining process, the existence of clean surfaces and high hydrostatic stresses favors the formation of strong adhesive friction junctions; the extent of these can be limited by the provision of a suitable lubricant [2–4]. Correct application of lubricants has been proven to greatly reduce friction. This results in surface quality improvement [5].

Although the significance of lubrication in machining is widely recognized, the usage of conventional flooding application in machining processes has become a huge liability. Not only does the Environmental Protection Agency regulate the disposal of such mixtures, but many countries and localities also have classified them as hazardous wastes. Beside that economically, the cost related to the lubrication and cutting fluid is 17 % of total production cost which is normally higher than that of cutting tool equipments which incurs only 7.5 % of total cost. At present, many efforts are being undertaken to develop advanced machining processes using less lubrications [6]. Promising alternatives to conventional flood coolant applications are the minimum quantity lubrication (known as MQL) [7]. Klocke and Eisennblatter (1997) state that MQL refers to the use of lubrication of only a minute amount—typically of a flow rate of 50–500 ml/h which is about three to four orders of magnitude lower than the amount commonly used in flood cooling condition. This has been reported to improve tool life due to its ability to penetrate into the tool-chip interface, this results in improving surface quality [8].

For more develop advanced machining processes for better surface quality using less lubrication, it is clear that a multi-pronged approach must be used, including innovation in technology [9]. In this chapter, authors will explore the development of nanolubrication in machining. It has been reported that, by introducing the nanolubrication system in machining process, the reduction of friction component could be achieved as it is working of billions of rolling elements in the tool-chip interface and consequently produce much better surface quality [10]. Nanolubricant is defined as new engineering material consisting of nanomaterial sized particles dispersed in base fluid. The nanolubricant is developed to sustain the high machining temperatures present in machining process, non-toxic, easy to be applied and effective in term of cost [11]. Over a decade, carbon onion has been successfully developed with high tribology performance. It consist of concentric graphitic shells and it is

one of the fullerene-related materials together with C60 and carbon nanotubes [12]. It has been proved that it can provide the similar lubrication with the graphite when tribologically tested at ambient air. It is expected to have good properties suitable for nanolubrication system due to its unique structure. It also has been proved that it could be used as a solid additive to grease replacing MoS₂ in several commercially available lubricants for use in ambient air [13]. On the other hand, silicon dioxide (SiO₂) nanoparticle is a hard and brittle material. This nanoparticle has very good mechanical properties especially in term of hardness (Vickers hardness—1,000 kgf mm⁻²) and in very small size range from 5 nm up to 100 nm.

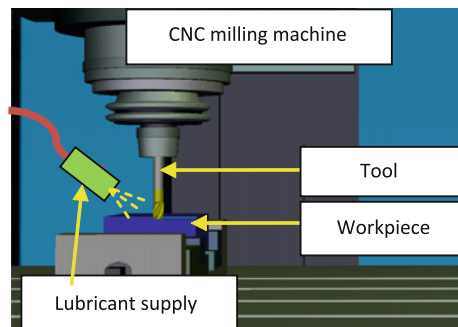
Following the review above, in this research work, the surface quality improvement is investigated in end milling of Al2017-T4 alloy using carbon onion nanolubrication. Also, the surface quality and surface morphology of Al6061 material was investigated when using SiO₂ nanoparticles at different concentration in milling process.

2 Experimental Set Up and Procedure

The experimental set-up used in this study is illustrated in Fig. 1. The machine used in this study is a vertical type machining centre (Sakai CNC MM-250 S3), in which the spindle has constant position preloaded bearings with oil-air lubrication and the maximum rotational speed is 5,000 min⁻¹. The tool used in the experiments is SEC-ALHEM2S8 end mill having a diameter of 8 mm, as shown in Fig. 2.

The cutting processes of rectangular Al2017-T4 and Al6061-T6 workpiece (118 HV) with dimension of 50 × 20 × 10 mm³ are selected as the case study. Table 1 shows the mechanical properties of Al2017-T4 and Al6061-T6, and Fig. 3 shows the workpiece and tool path in the cutting tests. The slot-milling test is carried out and the tool moves in the +X direction to cut a stroke of 50 mm. The cutting speed, feed rates and depths of cut are set at 75.408 m/min, 100 mm/min and 1.0 mm, respectively, and are selected based on the recommendations given by the tool manufacturer.

Fig. 1 Experimental set-up



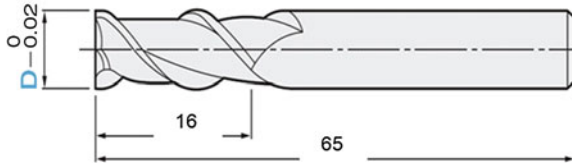
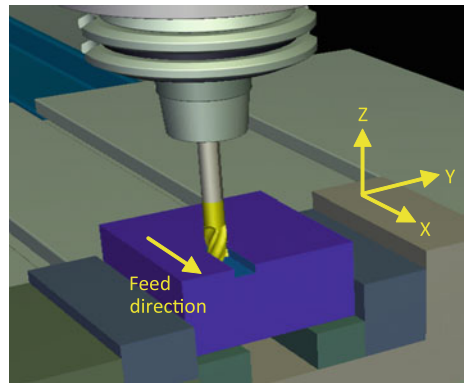


Fig. 2 The tool geometry

Table 1 The mechanical properties of Al2017 and Al6061

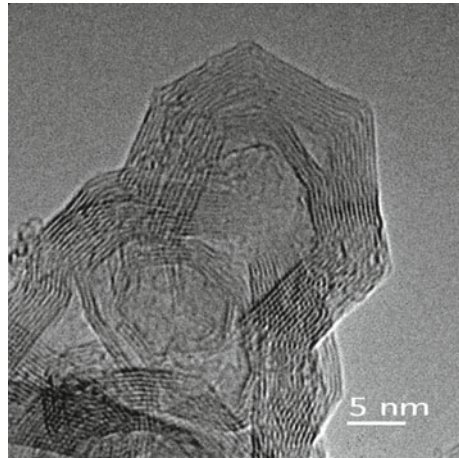
Mechanical properties	Al2017-T4	Al6061-T6
Hardness, Vickers	118	107
Ultimate tensile strength (MPa)	427	310
Tensile yields strength (MPa)	276	276
Modulus elasticity (GPa)	72.4	68.9
Poisson ratio	0.33	0.33
Fatigue strength (MPa)	124	96.5
Shear modulus (GPa)	27	26
Shear strength (MPa)	262	207

Fig. 3 Workpiece and tool path



In case of using carbon onion nanoparticle, the Alomicut lubricant type is chosen to reduce friction at the tool-chip interface due to its favorable lubrication characteristics. There are four different lubrication modes used in this study which consist of 0.0, 0.5, 1.0 and 1.5 %wt of carbon onion mixed with Alomicut oil, followed by sonification using Sono Bright ultrasonic vibration (240 V, 40 kHz, 500 W) for 30 min in order to suspend the particles homogeneously in the mixture. In case of using more than 1.5 %wt concentration, the mixing process of carbon onion in oil is challenging as the onion particles tend to collect together, having high weight and finally agglomerate. In future, more investigation is needed to solve this mixing problem.

Fig. 4 TEM picture of carbon onion [7]



Fundamentally, carbon onion nanoparticles are produced by heat-treatment of carbon black (Cabot R250 from Cabot Corporation) in a resistance-heated furnace using a graphite crucible under Helium gas at a pressure of 1 atm. The nano-size carbon onion is obtained by inductive heating at 2000 °C for 15 min and is used without further treatment (i.e. purification). Figure 4 shows the TEM image of carbon onion with an average size of 5–20 nm.

In the mean time, for SiO₂ nanolubrication, four different concentration of 0, 0.2, 0.5 and 1.0 wt% were used in order to investigate its effect to the morphological surface. The nanolubricant were prepared by mixing SiO₂ nanoparticles with an average size of 5–20 nm to the mineral oil followed by sonification (240 W, 40 kHz, 500 W) using Sono Bright ultrasonic device for 5 h in order to suspend the particle homogeneously in the mixture.

To ensure the consistent lubrication supply into the systems for both cases of nanolubricant, the experimentation is carried out using MQL with a thin-pulsed jet nozzle that is developed in laboratory and controlled by a variable speed control drive. The diameter of the nozzle orifices is 1 mm and the nozzle system is attached to a flexible portable fixture fixed on the machining spindle without interfering with the tool or workpiece during the machining process. The surface roughness (R_a) is measured using Nanofocus roughness tester equipped with μ sufr Software under a magnification of 10 \times , in accordance to the ISO 11562 standard with a 0.6 mm cut-off distance. Surface roughness measurements are performed and repeated at three different spots for each measurement, one in the middle and the other two on the edge, were used to measure the surface roughness of the cut. Following this, the mean of the three readings is recorded. Furthermore, the Field Emission Scanning Electron Microscopy (FESEM) was utilized to examine the morphology of machined surface.

3 Results

3.1 The surface Roughness Results

3.1.1 Surface Roughness Results Using Carbon Onion Nanolubricant

In case of using carbon onion nanolubrication system, the slot-milling test is carried out to investigate the surface quality improvement of the machined Al2017-T4. Figure 5a–d show the measured surface roughness for different carbon onion concentrations. The average measured surface roughness R_a at different carbon onion concentrations plotted in Fig. 6. As can be seen in Fig. 6, the lowest surface roughness values are obtained at the highest carbon onion concentration. In addition, the surface roughness improvements percentages are found to be 46.32 % compared with the case of using ordinary lubrication system. These results are totally supported by Fig. 7a–d which showing the stereoscopic photographs for three-dimensional views

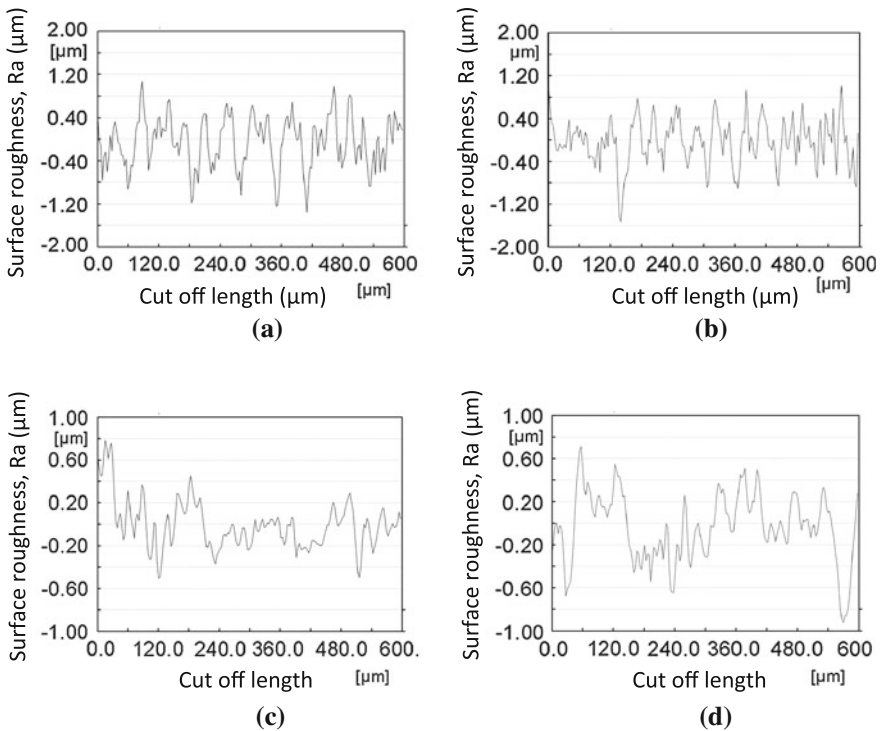


Fig. 5 Measured surface roughness (R_a) at different carbon onion concentration **a** 0 %wt carbon onion concentration **b** 0.5 %wt carbon onion concentration **c** 1.0 %wt carbon onion concentration **d** 1.5 %wt carbon onion concentration

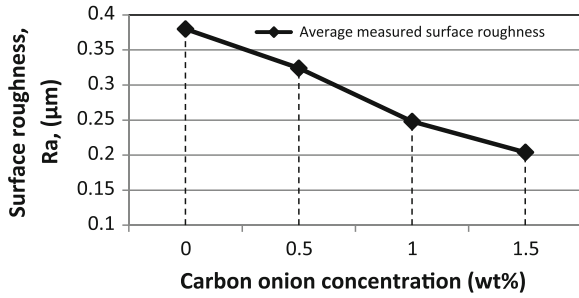


Fig. 6 Average of measured surface roughness (R_a) at different carbon onion concentration

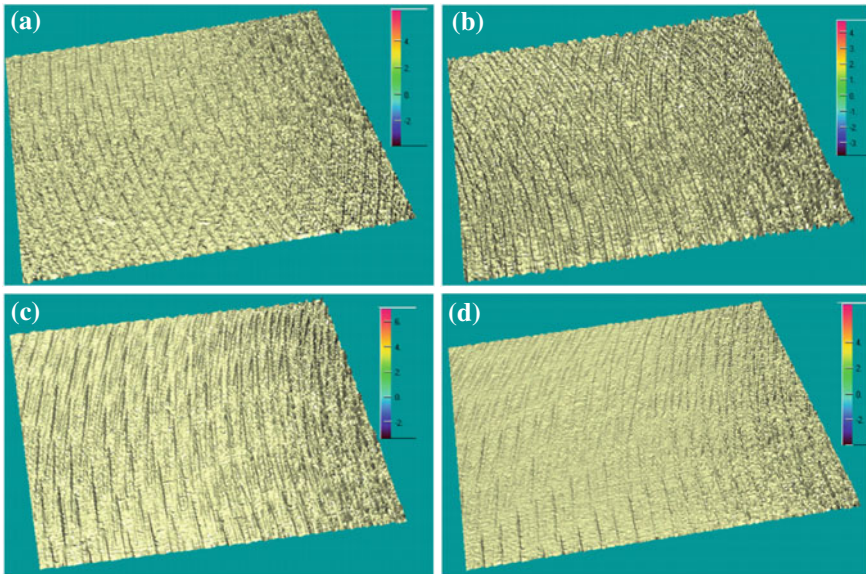


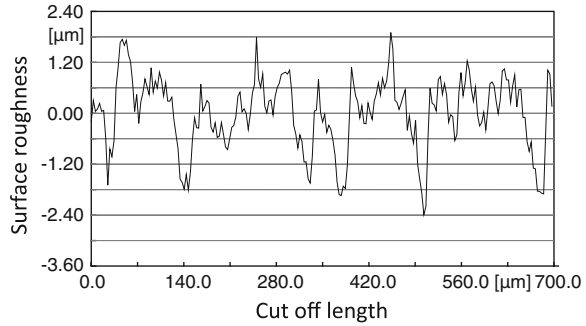
Fig. 7 Stereoscopic photographs for three-dimensional views of machined surface at different concentration **a** 0 %wt carbon onion concentration **b** 0.5 %wt carbon onion concentration **c** 1.0 %wt carbon onion concentration **d** 1.5 %wt carbon onion concentration

of machined surface at different modes of carbon onion concentration. It is clearly shown that the lowest surface roughness is obtained at the highest carbon onion concentration, with a concentration of 1.5 %wt.

3.1.2 Surface Roughness Results Using SiO_2 Nanolubricant

The slot milling test was carried out to investigate machining performance by using the proposed experimental setup. Figure 8 illustrates the samples of measured surface roughness at $5,000 \text{ min}^{-1}$ cutting speed, 100 mm/min feed and 5 mm depth of cut,

Fig. 8 The measured surface roughness at 0.2 wt% SiO₂ nanoparticle concentration



using 0.2 wt% SiO₂ nanoparticle concentration. While Fig. 9 shows surface roughness variations at different SiO₂ nanoparticle concentrations. As can be seen from Fig. 9, the best roughness were obtained at 1.0 wt% of SiO₂ concentration in which 36.82% better compared with the case of using ordinary lubrication system.

The mechanism behind such phenomena could be elaborated as being due to the increment of SiO₂ concentration, which increases the existence of nanoparticles at the tool-chip interface, and these nanoparticles serve as spacers, which eliminate the contact at tool-chip interface. In high speed machining processes, the high heat generated changes elastohydrodynamic lubrication to boundary lubrication. The spherical nanoparticles cause a rolling effect in between the rubbing surfaces, and reduce the coefficient of friction [14, 15]. The low friction behavior of nanoparticles effectively minimizes the frictional effects at the tool-chip interface and thus improves the machined surface. Moreover, when large amounts of nanoparticles exist in cutting oil, it would collide and impeded by the asperities on the work surface and hence generate a better machined surface.

On the other hand, the extensively dispersed SiO₂ nanoparticles in cutting oil are facilitated by a high pressure air stream at the cutting zone, showing good per-

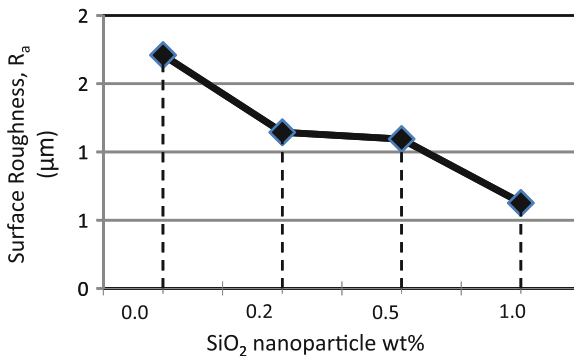


Fig. 9 Average of measured surface roughness at different SiO₂ concentration

formance in improving the machined surface. The atomized mist mixed with SiO_2 nanoparticles suspended lubrication exhibited more efficient feeding at the cutting zone compared with the flood lubrication method. The presence of nanoparticles at the tool-chip interface acts as a polisher, since an interacting force is induced at the interface and hence, improves the surface quality.

3.2 The Morphology of the Machined Surface

For further investigation on machining performance using SiO_2 nanolubrication, the Field Emission Scanning Electron Microscopy (FESEM) was utilized to examine the machined surface morphology. At the beginning, substrate cleaning was required to remove all unwanted surface contamination. Substrate surface finish was achieved by etching in hot solutions of sodium hydroxide to remove minor surface imperfections. To remove surface oxides, which are a combination of inter metallic, metal and metal oxides remaining on the surface after cleaning/etching, an aqueous solution containing an oxidizing inorganic acid, phosphoric and sulfuric acids, simple and complex fluoride ions, organic carboxylic acid, and manganese in its oxidation state was used. The formation and growth of the protective SiO_2 thin film on the machined surface were examined through surface elemental mapping analysis.

3.2.1 The Formation and Growth of the Protective SiO_2 Thin Film on the Machined Surface

Figure 10 shows the FESEM image of machined surfaces produced at four different SiO_2 concentrations of 0, 0.2, 0.5 and 1.0 wt%. Clearly, many protective thin films were produced on the feed marks of the machined surface containing billions of SiO_2 nanoparticles which provide much less friction and thermal deformation, as shown in Fig. 10b–d. These regular thin film formations grew when the SiO_2 concentration was increased from 0.2 to 1.0 %wt. It was also observed in the surface layer that small exfoliations or shedding of the thin film occurred, as is illustrated by Fig. 10d. This could be explained by the fact that the increment of nanoparticle concentration increases the viscosity of cutting oil. In this case, more nanoparticles exist between the tool-chip interface and these nanoparticles will serve as spacers which eliminate the tool-chip contact friction. Moreover, due to the porous nature of spherical SiO_2 nanoparticles, it could impart high elasticity, which augments their resilience in a specific loading range and enhances the gap at the tool-chip interface [16]. Therefore, with the extreme pressure of additives in cutting oil and the existence of a gap at tool-chip interface, there is a high contact resistance which induces the formation of film on the workpiece surface through chemical reaction [17]. In addition, the generation of high heat in the cutting zone will change elasto-hydrodynamic lubrication to boundary lubrication. This results in the formation of thin protective films on the surfaces, as per Fig. 10. The increment of nanolubricant concentration

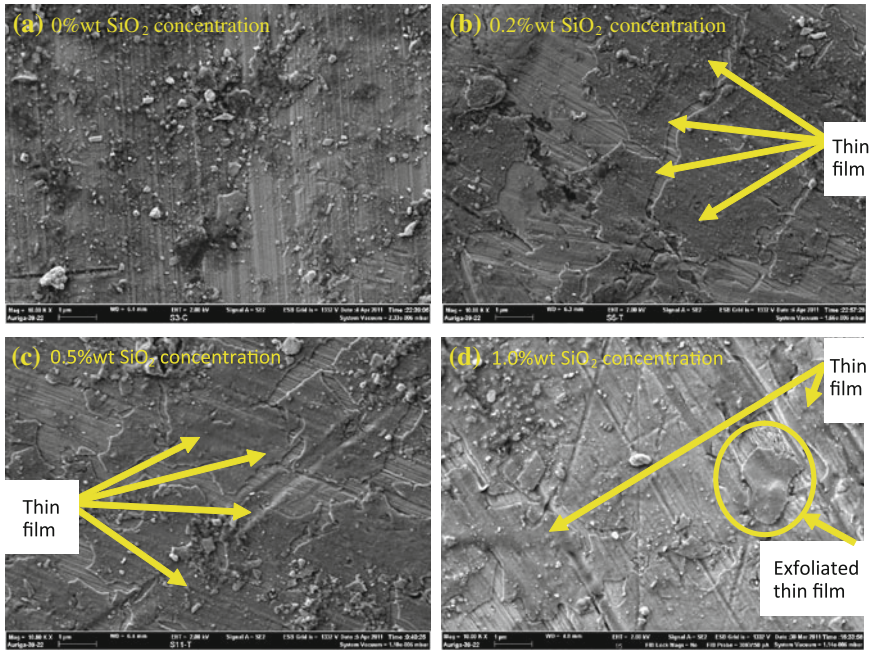


Fig. 10 FESEM on sample (a), (b), (c) and (d) which machined with 0, 0.2, 0.5 and 1.0 wt% of SiO_2 concentration respectively

increases the growth of thin protective film on machined surfaces due to the breaking process. In other words, during machining the high number of nanoparticles rubs with the asperities at the workpiece surface and, thus, the newly created surface is exposed to cutting oil more frequently. In this way, strong chemical interaction is formed between nanolubricant and newly created surface and, therefore, a more intensive protective film is formed. This process definitely increases the quality of the machined surface, successfully enhancing the surface properties and reducing its coefficient of friction [18, 19].

3.2.2 Surface Elemental Mapping

Surface elemental mapping was employed to investigate the relation between the machined surface quality and the orientation and distribution of SiO_2 nanoparticles on the machined surface. Elemental mapping of samples machined with 0.2, 0.5 and 1 wt% are shown in Fig. 11a–c, respectively. Figure 11a shows that at 0.2 wt% SiO_2 nanoparticles, the polishing track orientation matched the SiO_2 nanoparticle distribution on the machined surface, especially at the exfoliated thin film. By using 0.5 wt% SiO_2 nanoparticles, higher amounts of SiO_2 nanoparticles were present compared to the case of using 0.2 wt%, as shown in Fig. 11b. With 0.5 wt%

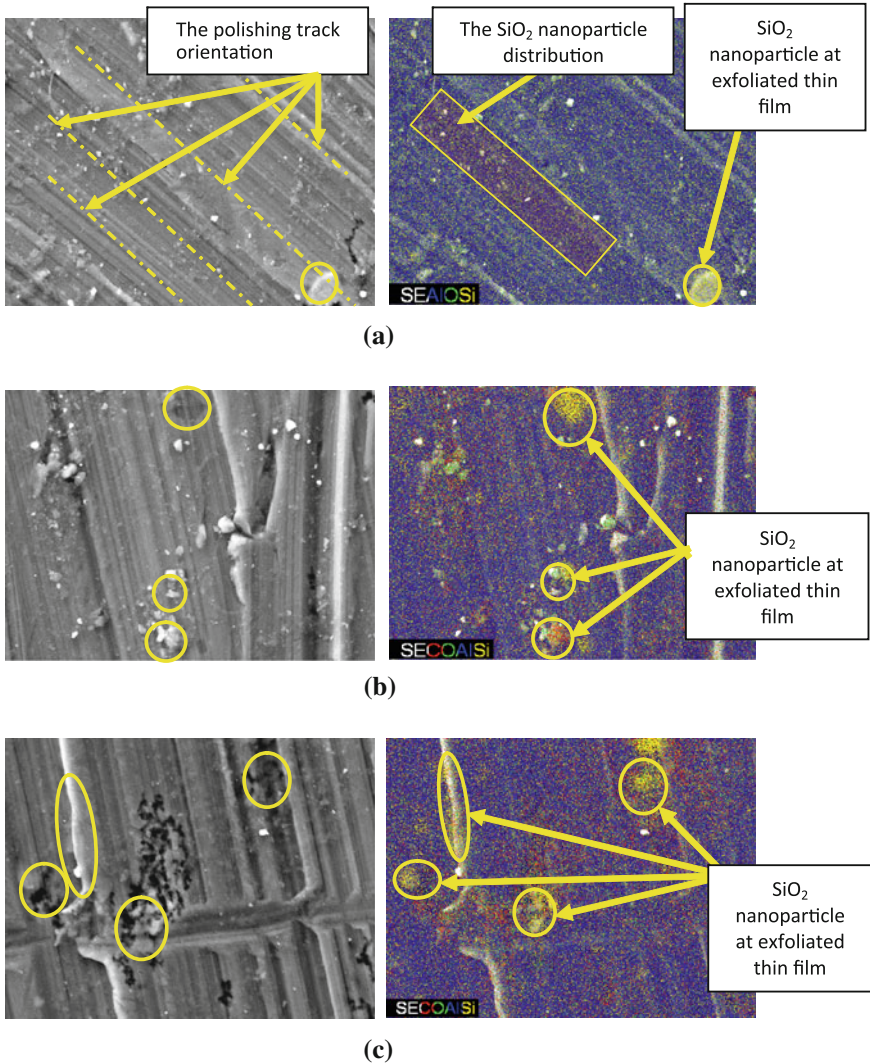


Fig. 11 Surface elemental mapping of sample machined with different SiO₂ nanoparticle concentration **a** Machined surface using 0.2 wt% of SiO₂ nanoparticle concentration **b** Machined surface using 0.5 wt% of SiO₂ nanoparticle concentration **c** Machined surface using 1.0 wt% of SiO₂ nanoparticle concentration

SiO₂ nanoparticles the track of embedded nanoparticles on the machined surface was clearly seen. Eventually the nanoparticles could be ploughed off but left debris nanoparticles behind. When the SiO₂ nanoparticle concentration was increased up to 1.0 wt% as illustrated in Fig. 11c, higher amounts of SiO₂ nanoparticles were embedded on the machined surface compared to the cases of 0.2 and 0.5 wt%. It was

clearly seen that the nanoparticles were burnished on the porous alumina. Several SiO_2 nanoparticles were partially embedded into the surface and some showed the tracks from being ploughed off, and more nanoparticle debris was left on it.

Following the results of elemental mapping shown in Fig. 11a–c, the mechanism of nanoparticles was found to assist the cutting operation and can be categorized into the three levels. At the first level, the particles were partially embedded on the machined surface when they collided with its asperities due to extremely high pressure in the cutting zone; the particles were sheared and changed shape because of compression. The sheared off debris continues to assist the cutting, but not as well as spherical nanoparticles which exercise rolling with a low coefficient of friction. At the second level, with a higher concentration of nanoparticles, the partially embedded particles were ploughed off by new nanoparticles, and both then continued to polish the surface. The ploughed off particles left a thin exfoliated film on the contact area due to the damage from high loading [20–22]. Meanwhile, when nanoparticle concentrations continued to increase, those nanoparticles were impregnated into the pore of the surface and were then sheared by other incoming nanoparticles. The rolling of nanoparticles leads to the formation of an easily sheared lubrication film as well as the asperities on the surface, thus the surface is being polished and enhanced in quality. Therefore, as shown in the results, the 1.0 wt% concentration provided the best machined surface morphology compared to other concentrations.

4 Discussion

In this study, the carbon onion and SiO_2 are used as solid nanoparticle and mixed with ordinary mineral oil at different concentrations in order to investigate the surface quality improvement in CNC end milling machined of Al2017-T4 and Al6061-T6, respectively. For both cases, it is clearly seen that highest carbon onion and SiO_2 concentration are producing the best surface quality. This could be explained as depicted in Fig. 12 with the fact that the deformation of the chip is flowing over the tool leads to localized regions of intense shear occurring due to the friction at the rake face, which is known as secondary shear. At higher plastic deformation chip is welded to the tool face and hence effectively changes tool geometry and rake steepness. This results in poor surface finish since the bits of the welded chip will eventually break off and stick to the workpiece. These bits tend to be problematic due to the work-hardening which they underwent very hard and abrasive.

Applying the lubrication system to the tool-chip interface will reduce the coefficient of friction leading to better surface quality. However, introducing of the carbon onion and SiO_2 nanolubrication system would show much less friction and much better surface quality. This is mainly attributed to the tribological properties of the nanoparticle which reduces the coefficient of friction at the interface during the machining as it is acting as billions of nano-scale quasi-spherical structure rolling elements, as shown in Fig. 13.

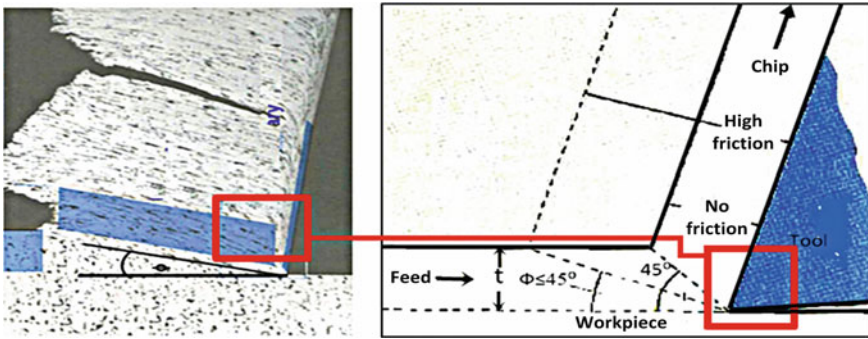


Fig. 12 Shear mechanism in cutting zone [23]

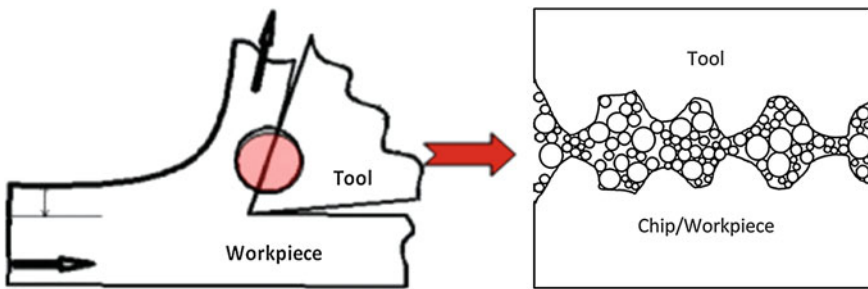


Fig. 13 Rolling element in the tool-chip interface [24]

5 Conclusion

In this study, the surface quality improvement of CNC end milling machined Al2017-T4 and Al 6061-T6 using carbon onion and SiO₂ nanolubrication system has been investigated, respectively. Based on the results obtained, the highest concentration of carbon onion and SiO₂ nanoparticles produces the best surface quality. In addition, surface roughness reduction percentage in case of using carbon onion and SiO₂ nanoparticle are found to be 46.32 and 36.82 %, respectively, compared with those obtained when using ordinary lubrication oil. The results are mainly attributed to the tribological properties of the carbon onion and SiO₂ solid nanoparticle, which act as billions of nano-scale spherical structure rolling elements at the tool-chip interface. Consequently, the coefficient of friction at tool-chip interface is reduced significantly. With such excellent properties of carbon onion and SiO₂ nanoparticles, it might be a new effective way as an alternative to flood lubrication due to the environmental issues involved.

Acknowledgments The authors would like to acknowledge the University of Malaya, Malaysia and Tokyo Institute of Technology, Japan for providing the necessary facilities and resources for

this research. This study was partially funded by HIR Grant no. HIR-MOHE-D000001-16001. The authors gratefully acknowledge the Ministry of Higher Education Malaysia for the financial support.

References

1. Sayuti M, Tanaka T, Sarhan AAD, Saito Y, Hamdi M (2012) Surface quality improvement in CNC milling machined aerospace AL-2017-T4 alloy using carbon onion nanolubrication with DLC coated cutting tool. In: Lecture notes in engineering and computer science : proceeding of the world congress engineering 2012, WCE 2012, London, U.K., pp 1487–1491
2. Lathkar GS, Bas USK (2000) Clean metal cutting process using solid lubricants. In: Proceeding of the 19th AIMTDR conference. Narosa Publishing House, IIT Madras, pp 15–31
3. Suresh Kumar Reddy N, Venkateswara Rao P (2006) Enhancement of machinability of AISI1045 steel using molybdenum disulphide as a solid lubricant. In: 2006 ASME international mechanical engineering congress and exposition, IMECE2006. American Society of Mechanical Engineers, Chicago, IL, United States, 5–10 November 2006
4. Belgasim O, El-Axir MH (2010) Modeling of residual stresses induced in machining aluminum magnesium alloy (Al–3Mg). In: Proceedings of the world congress on engineering 2010, WCE 2010, London, U.K., June 30–July 2 2010
5. Spanoudakis P, Tsourveloudis N, Nikolos I (2008) Optimal selection of tools for rough machining of sculptured surfaces. In: Proceedings of the international multiConference of engineers and computer scientists 2008, IMECS 2008, Hong Kong, 19–21 March 2008
6. Dilbag S, Rao PV (2008) Performance improvement of hard turning with solid lubricants. *Int J Adv Manuf Technol* 38:529–535
7. Yassin IN, Hamdi M, Fadzil M, Norhirmi MZ (2011) Investigation into new development of minimal quantity lubricant (MQL) system in high speed milling of H13. In: UK–Malaysia–Ireland engineering science conference 2011 (UMIES 2011): Faculty of Economics & Administration, University Malaya, Kuala Lumpur
8. Klocke F, Eisenblätter G (1997) Dry cutting. *CIRP Ann Manuf Technol* 46(2):519–526
9. Sarhan AAD, Hassan MA, Matsubara A, Hamdi M (2011) Compensation of machine tool spindle error motions in the radial direction for accurate monitoring of cutting forces utilizing sensitive displacement sensors. In: Proceedings of the world congress on engineering (2011) WCE 2011, London
10. Reddy NSK, Nouari M (2011) The influence of solid lubricant for improving tribological properties in turning process. *Lubr Sci* 23(2):49–59
11. Deshmukh SD, Basu SK (2006) Significance of solid lubricants in metal cutting. In: 22nd AIMTDR
12. Hirata A, Igarashi M, Kaito T (2004) Study on solid lubricant properties of carbon onions produced by heat treatment of diamond clusters or particles. *Tribol Int* 37:899–905
13. Street KW, Marchetti M, Wal RLV, Tomasek AJ (2004) Evaluation of the tribological behavior of nano-onions in Krytox 143AB. *Tribol Lett* 16:143–149
14. Ramana SV, Ramji K, Satyanarayana B (2010) Studies on the behaviour of the green particulate fluid lubricant in its nano regime when machining AISI 1040 steel. *Proc Inst Mech Eng Part B J Eng Manuf* 224(10):1491–1501
15. Shenoy BS, Binu KG, Pai R, Rao DS, Pai RS (2012) Effect of nanoparticles additives on the performance of an externally adjustable fluid film bearing. *Tribol Int* 45(1):38–42
16. Zhang B-S, Xu B-S, Xu Y, Gao F, Shi P-J, Wu Y-X (2011) CU nanoparticles effect on the tribological properties of hydrosilicate powders as lubricant additive for steel–steel contacts. *Tribol Int* 44(7–8):878–886
17. Lin YC, So H (2004) Limitations on use of ZDDP as an antiwear additive in boundary lubrication. *Tribol Int* 37:25–33

18. Lee K, Hwang Y, Cheong S, Choi Y, Kwon L, Lee J, Kim SH (2009) Understanding the role of nanoparticles in nano-oil lubrication. *Tribol Lett* 35:127–131
19. Zalnezhad E, Sarhan AAD, Hamdi M (2013) Optimizing the PVD TiN thin film coating's parameters on aerospace AL7075-T6 alloy for higher coating hardness and adhesion with better tribological properties of the coating surface. *Int J Adv Manuf Technol* 64(1–4):281–290
20. Kragelsky IV, Zolotar AI, Sheiwekhman AO (1985) Theory of material wear by solid particle impact—a review. *Tribol Int* 18(1):3–11
21. Raadnui S (2005) Wear particle analysis—utilization of quantitative computer image analysis: a review. *Tribol Int* 38(10):871–878
22. Williams JA (2005) Wear and wear particles—some fundamentals. *Tribol Int* 38(10):863–870
23. Trent EM, Wright PK (2000) *Metal cutting*, 4 edn. Butterworth-Heinemann, Boston
24. Yan J, Zhang Z, Kriyagawa T (2011) Effect of nano-particle lubrication in diamond turning of reaction-bonded SiC. *Int J Autom Technol* 5(3):307–312

A Price-Based Decision Policy to Mitigate the Tragedy of the Commons and Anti-Commons

M. Sumbwanyambe and A. L. Nel

Abstract In developing countries subsidies play an important role and are used in one way or another to extend the information and communication services to the information “have nots” through subsidized communication services. However, subsidies may have an impact on network resource utilization, quality of service and the amount of revenue generated. For example, subsidies may lead to low Quality of Service (QoS) and high resource utilization while in some instances unsubsidized services may lead to high quality of services and low utilization of resources. This *see-saw effect* may eventually lead to market failure and it may, now and then, destroy market efficiency. This phenomenon calls for a combined study, in which the relationship between subsidy, price, QoS and resource utilization is investigated. In this chapter, the impact of subsidies on quality of service and resource utilization in multitier communities is investigated. We try to find a middle ground between implementation of subsidy policy and its effects on QoS and resource utilization in a network.

Keywords Heterogeneous communities · ICTs · Pricing policy · Quality of service · Subsidy · Tragedy of the anti-commons · Tragedy of the commons

1 Introduction

In one way or another, over-pricing or under-pricing of networks service provision in developing countries has created resource problems associated with under-usage or over-usage of such resources [1–4]. Such under-usage or over-usage of resources in

M. Sumbwanyambe (✉) and A. L. Nel
Department of Mechanical Engineering Science, University of Johannesburg,
P.O. Box 524, Auckland Park 2006, South Africa
e-mail: sumbwam@gmail.com

A. L. Nel
e-mail: andren@uj.ac.za

rural areas or underserved regions of developing countries, arise from incompletely defined and enforced pricing policies and legal framework within such countries. This situation is further compounded by the problems associated with subsidy allocation by governments of developing countries when trying to promote social and economic agendas for its “needy” people [5].

Different opinions have risen on the usage of subsidies to promote social and economic agendas in developing [6]. For example, Sumbwanyambe and Nel, Levin and Hardin [4, 7, 8] argue that the usage of subsidies to enhance social and economical growth in a competitive market is not feasible and may distort market efficiency. Other social-economical proponents advocating for the implementation of subsidies have argued, to the contrary, that subsidies are a necessity to a developing country as the means of promoting, through lowering down of prices, social and economic growth in purely monopolistic market economies.

For instance “a novel idea to take voice and data services to the most rural areas could see Vodacom and MTN paid subsidies to do the job. The operators have to promise high-quality services even the poorest people can afford in return for having up to 80% of infrastructure subsidized. But the cost will not affect taxpayers as the cash will come from the Universal Service Fund, to which the operators themselves contribute” [9].

Another study conducted by Alesina and Rodrik, [10] showed that income disparities had an adverse effect on the country’s economic growth; making it necessary for subsidies to be used in correcting such disparities. They showed that in more imbalanced communities, social and economic growth is lower because the demand for fiscal redistribution financed by distortionary taxation is higher.

Amegashie in [6] points out that implementing a subsidy, “by reducing the price of the commodity, may increase the consumption of the commodity towards the equilibrium (perfectly) competitive quantity, given that output was initially too low” and if chosen properly a subsidy may “move the economy towards the perfectly competitive equilibrium quantity”.

Without doubt, income inequality in developing countries fuels social discontent and increases socio-political instability. The uncertainty in the politico-economic environment reduces investment which in turn reduces growth in underserved areas.

Subsidies, though seen as the means of promoting social and economic agendas in developing countries can create the tragedies associated with public resource usage or something-for-nothing resources [4, 11]. Given a subsidy rate, consumers of developing countries usually anticipate a net social benefit derived from free resources due to subsidy or under pricing of such resources. Anticipation of net social benefits from such resources may generate a damaging rush from consumers to exploit the resource, which may result in the tragedy of the commons. By definition, the tragedy of the commons is a situation when “multiple owners are each endowed with the privilege to use a given resource and no one has the right to exclude another. When too many such people have the privileges to resource usage, the resource is likely to be over used [8]”.

On the contrary, when no subsidy is given, consumers face no differential between the perceived utility and the cost of the resource, as such very few consumers or users

will create a damaging rush towards resource utilization creating a no social and pecuniary benefit to users. Generally, exogenous factors, such as exorbitant pricing or the absence of subsidy rates can decrease network resource usage and can make network usage in developing countries less effective. Heller in [2] defines the anti-commons as a situation where “multiple owners are each endowed with the right to exclude others from a scarce resource, and no one has an effective privilege of use”.

Indeed, the implementation of subsidies may create a situation where the commodity is over utilized, because it is highly subsidized, or underutilized, because it is under subsidized. This may create a problem that affects the market efficiency and social and economic growth of developing countries.

In light of the above problems, correct subsidy driven policies are necessary in order to avoid the tragedy of the commons and the tragedy of the anti-commons. In this chapter we develop a subsidy driven policy to prevent the tragedy of the commons or anti-commons in heterogeneous community. This chapter is organized as follows: Sect. 2 gives the necessity of subsidies in social and economic augmentation of developing countries, Sects. 3 and 4 gives an introduction on subsidies in developing countries. We present the modeling parameters in Sects. 5 and 6 introduces the tyranny of subsidies on network resources. Section 7 presents the model and graphical and numerical analysis of our model. We conclude our study in Sect. 8.

2 Subsidies as a Response to Meeting Social Objectives

Subsidy usage in developing countries has both social and economic objectives, defined in this chapter as reducing the effects of poverty and promoting social and economic development in rural or underserved areas of developing countries. Having provided ICT infrastructure in underserved regions of the country, the cumulative importance of service provision will, however, depend on the adoption and usage of communication services. In underserved regions of developing countries very few people may adopt or use information resources because of the competing basic needs and uncompetitive market prices [12, 13].

Until such a time when the prices of ICTs are low and government intervention through subsidy, it is very unlikely that government efforts to bridging the digital divided may bear fruit. It is therefore necessary, if governments wants to promote social and economic objectives, for governments to “determine priority underserved areas and accelerate the provision of Universal Service and Access through integrated government interventions” [14] such as sector specific subsidy injection.

When objectives for promoting social-economic growth in rural areas fall short of the intended objectives, they generally lack other complimentary government initiatives needed to undertake the economic aspects of such a development that ICT provision is suppose to compliment. Oyedemi [15] points out that “the failure of some of these policies is not solely inherent in the policies alone, but also in forces from other social factors that render many access program unrealistic”. Therefore

social-economic ICT provision directed at the poor rural people will probably fail without substantial complementary interventions by government, education institutions, regulatory institutions, and private institutions.

3 Effect of Price Decision Policies in Heterogeneous Communities

In heterogeneous communities, communities with users who have diverse variations in income distributions, price decision policies are usually difficult to make especially when there are subsidies involved. Governments of developing countries are faced with decisions that require the promotion of social and economic development whilst in the same vein trying to balance the economic dynamics of the overall country. To this effect, pricing of communication services in a heterogeneous society is in itself a multi-faceted matter which requires sustainable, long term pricing solutions.

Currently, there is only one form of pricing (undifferentiated service) which is not adequate to meet customer needs and satisfaction. Accordingly, in order to achieve sustainable development in a heterogeneous environment, pricing must go beyond understanding customer needs to sensitivities around pricing based on the proposed customer's position in social hierarchy. A brief outline of problems associated with undifferentiated pricing is given below:

- If the too much a subsidy is given the tragedy of the commons will occur. The tragedy of the commons is a situation in which this type of pricing will draw in a lot of subscribers, who are driven exclusively by narrow economic interest-acting as rational agents, who will congest the network, resulting in poor QoS, poor system performance, and the resultant revenue loss for ISPs leading to collapse of the entire enterprise.
- However, if no subsidy is given a situation exists when certain subscribers will not subscribe to the services due to lack of subsidized services. Therefore, only a few consumers will be drawn to the resource resulting in under usage of the resource. Considering an economics point of view, we would say that tragedy of the anti-commons would result from the fact that resources are not exploited even for situations in which the marginal productivity is positive.
- If we price decision policies of communication services are done in accordance with a position of consumer in an economic hierarchy (one category model may be a means test based model i.e. in broad categories perhaps called the poor and the rich) care must be taken so that the information there is no piggybacking among customers. This arbitrage based change in usage will appear as long as the two types of consumers or customers have a non-network-only based form of differentiation.

4 Pricing and Subsidies in Developing Countries

Pricing of information and communication services in rural areas is of prime importance if social and economic development is to be realized [16]. Nevertheless, extremely poor households in developing countries, who live on less than ZAR 16 per day, find it difficult to afford telecommunications services which, in most of the rural areas, are overpriced. In pricing of such services, different pricing schemes and models may be used to cater for consumers or users with differing sensitivities towards price. Historically, especially in developed countries, such schemes have evolved from subsidy driven pricing to competition as noted by Hayashi [17]:

“Under state monopoly or regulated private monopoly, the mechanism was called cross subsidization that was built into the tariff structure. Usually long distance callers subsidized local callers, business users subsidized household users, and offices and houses in densely populated subsidized those in sparsely populated area”. Hayashi [17] further notes that:

“As countries moved into the privatization and competition phase, the cross subsidization scheme became increasingly unsustainable. Competition was introduced into the long distance market, as a result of which the source of subsidies dried up”.

Various pricing strategies must be applied in a heterogeneous or multi-tier society if social and economic objectives proposed by government are to be realized. However, in developing countries pricing for information services still remains a challenge to most communication and information services providers particularly in a community which has an unbalanced economic system between the information “haves” and “have nots” [18]. The problem governments should consider then is: how should governments of developing countries design a pricing model that will maximize or encourage social and economical development, and promote an inclusive information society, given an observable income disparity among users? While the more general problem is to determine the total subsidy budget as well as to maximize the net surplus (the benefit minus the cost), however, governments tend to focus only on the sub-problem of the subsidy rate path over a given specific budget. Such pricing mechanisms tend to be catastrophic, do not reflect costs involved and do not meet the social and economic objectives set forth by governments of developing countries.

5 The Parameters

As defined in [11, 19], we consider a multi-tier community whose population profile comprises of heterogeneous type of consumers with an observable disparity in reservation prices; the information “haves” (N_2) and “have-nots” (N_1). In many multi-tier communities, as in case of developing countries, customers have different service requirements. In such developing countries, service providers can offer different prices for each consumer and can choose a suitable price based on consumers’

needs and acceptance of the quoted price and adjust their price accordingly so as to obtain an optimal price at which both types of consumers will be more willing to pay for the service. As Sumbwanyambe, Nel and Clarke [1, 11] have stated, “these parameters can be learned through a market price adaptive research which estimates the number of consumers that accept a given price. In fact, the process of learning price parameters in such a market is a dynamic process with the aim of adjusting the quoted price in line with the consumers’ behavior towards price dynamics”. We will assume that the two types of consumers have differing reservation price.

Definition 1: Reservation price (sometimes called the threshold price), p_{thi} for $i \in (1, 2)$ is the price at which a customer is indifferent between subscribing to the current network or opting out of the current network and subscribing to the other network or dropping them all.

Furthermore we assume that p_1 and p_2 is the price that is payable to the service provider by the consumers N_1 and N_2 , respectively, where $p_1 = \beta p_2$, for $0 < \beta < 1$ and $p_2 \in (0, p\infty)$ and $c_{\{1,2\}}$ is the cost of providing such services to N_1 and N_2 and β is the subsidy factor provided by the government. We use a simple cut-off price rule to determine whether the user would subscribe to the ISP or not. In the current context the cut-off rule is based on users’ reservation price i.e. p_{th1} and p_{th2} , for N_1 and N_2 consumers. For any given price, those customers whose reservation price are greater than or equal to the given price will purchase the service provided by the ISP or the service provider. A price setting service provider in collaboration with government uses the information in the distribution of reservation prices to choose its optimal price.

6 The Tyranny of Subsidy and Price Sensitivity on Network Resources

Subsidies and price sensitivity has an adverse effect on resources utilization in multi-tier communities. From the users’ point of view, subsidies affect their behavior in terms of network resources utilization, which is correlated to their price sensitivities. From the government and service provider’s perspective, choosing the right subsidy is of great importance in maximizing revenue and enhancing the optimal usage of resources. Therefore, correct subsidization of deserving customers or consumers in underserved areas is of great importance if tragedies in such communities are to be avoided. In underserved region of South Africa and Zambia willingness to pay is associated with a number of factors. Primarily, any heterogeneous user is assumed to maximize utility subject to a predetermined level of his or her income.

In a heterogeneous society (especially in developing countries), subsidy and price sensitivity can lead to the tragedy of the (anti) commons [1, 2, 11]. If too large a subsidy β is given and the price sensitivity is close to 1 (i.e. $\alpha = 1$), consumers will

demand more than is available, creating the tragedy of the commons [1, 7]. In order to prevent the tragedy of the commons, service providers and government policy makers have to cut the subsidy level and increase the price which will eventually reduce the number of consumers. The decrease in the number of consumers (especially rural consumers) may then lead to the tragedy of the anti-commons [1, 2], leaving a gap in government policy and objectives of promoting social and economic growth. In such an eventuality, the government will have to increase the subsidy and lower the prices yet again in order fulfill their objectives. This creates what is known as a *see-saw effect*. The following example demonstrates this effect:

Let's take that $0 < p_1 \leq p_{th1}$ and $0 < p_2 \leq p_{th2}$, there exists a NE at which all users i.e. N_1 and N_2 , will attempt to maximize their utilities, $u_{pth1,p1}$ and $u_{pth2,p2}$, given the reservation price p_{th1} and p_{th2} and the price p_1 and p_2 charged by the service provider. Such a development will lead to both users i.e. the information "haves" and "have nots" trying to maximize their utility. This utilization of resources without restraint will, more likely, result in the "tragedy of the commons". Once more, if $0 \leq p_{th1} \leq p_1$ and $0 \leq p_{th2} \leq p_2$ the tragedy of the anti-commons is a more likely outcome. If the utility of one user is positive and the other user is negative, then the free rider problem is likely to occur. The following definitions are applicable in equilibrium:

Definition 2: Given that $0 \leq p_1 \leq p_{th1}$ and $0 \leq p_2 \leq p_{th2}$ the population profile of N_1 and N_2 users are in equilibrium if it does not pay for any member of the group to opt out of any ISP.

Definition 3: Given that $0 \leq p_{th1} \leq p_1$ and $0 \leq p_{th2} \leq p_2$ the population profile of N_1 and N_2 users are in equilibrium if it does not pay for any member of the group to subscribe to any ISP.

Definition 4: Given any price p_1 and p_2 the population profile of N_1 and N_2 users are not in equilibrium if it pays for any member of the group to join another group.

Having put up the above definitions, the following Lemma will show that given the prices, p_1 and p_2 , and the reservation price, and if all users receive equal negative or positive utility, users are unlikely to switch to another group or free ride or opt out.

Lemma 1: Given $0 \leq p_1 \leq p_{th1}$ and $0 \leq p_2 \leq p_{th2}$ and $p_1 < p_2$, for any values of p_1 and p_2 the situation is in equilibrium if $u_{pth2,p2} - p_2 = u_{pth1,p1} - p_1$.

Proof: From Definition 7.3 the situation is in equilibrium if the information "have-nots" and "haves" do not have a unilaterally decision to change from their present strategy to another. If the utility of the information "haves" is less than 0 i.e. $u_{pth2,p2} - p_2 < 0$, no N_2 users will subscribe to the ISP. Similarly if the utility of the information "haves-nots" is less than zero i.e. $u_{pth1,p1} - p_1 < 0$, there will be no N_1 users who will subscribe to the ISP. Otherwise if $u_{pth2,p2} - p_2 < u_{pth1,p1} - p_1$ a free rider problem is likely to occur. From the Lemma above we can draw the following conclusions:

- If there are no N_2 users subscribing to the ISP then the utility of such population group is zero i.e. $u_{pth2,p2} - p_2 < 0$
- If there are no N_1 users subscribing to the ISP then it means that the utility of the N_1 is zero i.e. $u_{pth1,p1} - p_1 < 0$

These two statements will lead us to the following corollary describing the structure of the equilibrium.

Corollary: Define $p_i = [p_1, p_2, p_{th1}, p_{th2}]$ an equilibrium has the following structure:

It is the tragedy of the commons if $u_{pth2,p2} - p_2 \geq 0$ and $u_{pth1,p1} - p_1 \geq 0$ where $p_2 \approx 0$ and $p_1 \approx 0$.

It is the tragedy of the anti-commons if $u_{pth2,p2} - p_2 \leq 0$ and $u_{pth1,p1} - p_1 \leq 0$ for all users in the network.

It is likely to be a free rider problem when $u_{pth2,p2} - p_2 \leq 0$ and $u_{pth1,p1} - p_1 \geq 0$ and $u_{pth2,p2} - p_2 \geq 0$ and $u_{pth1,p1} - p_1 \leq 0$.

7 Subsidy Driven Policy and Social Dilemmas

In this section we provide a numerical and a graphical analysis of results on how the subsidy affects the number of users and ultimately how it affects the revenue of the service provider. We take into consideration the dynamics of a heterogeneous community with an observable difference between the reservation price of the information “haves” and information “have-nots”.

7.1 A Local Decision Procedure for the Tragedy of the Commons and the Tragedy of the Anti-Commons Problem

This section gives a brief overview of a distributed algorithm to solve the problem of the tragedy of the commons and the tragedy of the anti-commons and also discusses the convergence of the algorithm to equilibrium. The algorithm can be used to determine the optimal values of α and β and the maximum revenue at which the utilization of network resources will be optimal. The algorithm is as follows:

Algorithm:

Step 1: Define the initial values $p_2, p_{th1}, p_{th2}, \alpha$ and β and bandwidth.

Step 2: Calculate the number of users by using equation below

$$N_{total} = \left(\frac{\beta p_2}{p_{th1}}\right)^{-\alpha_1} - 1 + \left(\frac{p_2}{p_{th2}}\right)^{-\alpha_2} - 1$$

Step 3: Calculate the amount of packets (T_{total}) sent by users by using

$$T_{total} = N_{total} \lambda$$

Step 4: If T_{total} is greater than bandwidth, add $\beta+0.1$ and p_2+1

Step 5: If T_{total} is less or equal to 0 (bandwidth usage is minimal) start $\beta - 0.1$ and $p_2 - 1$

Step 6: Go to step 3, 4, 5

Step 7: Calculate the final value of p_2 and $\frac{p_1}{\beta}$

Step 8: If final $p_2 = \frac{p_1}{\beta} = p_{th2}$. STOP. Optimal β solution found

Step 9: Calculate maximum revenue Π using equation 1 and the value of p_2 by equating equation 2=0

Step 10: Government and ISP (service provider) maintains values of β and p_2 Repeat step 8 and 9

Convergence to equilibrium: The system reaches equilibrium when all agents reach Step 9 of the algorithm. At this state, the government and the service provider feels that any increment or decrement in the value of β and p_2 will either reduce or increase its utility. Hence the number of packets in the system does not change. After Step 4 the government and the service provider may either decrease the value of subsidy or increase the price p_2 . After Step 3 the service provider calculates the number of packets or bandwidth used. (Note that in this case the number of packets used is proportional to the number of users subscribing to the ISP hence number of packets can be measured in terms of number of users). If the amount of packets sent is less than the total bandwidth, the government increments the subsidy factor and the ISP decrements the price p_2 . A rational government and ISP can reason after Step 3 that to prevent over-utilization of the resource (tragedy of the commons) it should either decrease the subsidy or increase the price by using Step 4. Similarly if the government wants to prevent the tragedy of the anti-commons it should increase the subsidy or decrease the price p_2 by using Step 5.

The government and ISP can use probabilistic searching scheme outlined in Steps 4 through 6 to reach its optimum load. If the total number of packets sent is greater than the bandwidth, the government starts with an initial increment probability of subsidy factor by a factor of 0.1 and the price is increased by a determined increment factor. If the subsidy increment produces a lower utility for the users, the government and the service provider will follow Step 5. Both the government and the ISP keep on probing in this manner until an optimal solution is reached where no users will deviate from his present status. At such a point government and ISP maintains the value of β and p_2 as shown in Fig. 4. Our claim is that when equilibrium is reached, the combined load on the resource is exactly the critical load or capacity of the

resource; i.e. the agents are using the resource optimally. They have reached this optimality through a distributed decision procedure using only local knowledge and with the directive of government.

7.2 Numerical and Graphical Analysis

In a developing country a complex relationship that exists between heterogeneous communities escalates the problem of developing a two tier workable pricing model. Customers have different income levels, sensitivities toward price and different reservation prices for a particular service. This problem is further compounded by the problem of subsidies and how to properly allocate subsidies to the “needy” or the information “haves nots”, without creating the tragedy of commons and the tragedy of the anti-commons. When over-subsidization and over-pricing of network resources creates the undesirable outcomes of the tragedy of commons and tragedy of anti-commons, it is necessary for the service provider and the government to subsidize and price network resources in an optimal way. In the next section we provide a numerical analysis and graphical analysis to make our statement clearer.

Numerical analysis: We provide a numerical analysis by applying an algorithm and the pricing model as proposed by Sumbwanyambe and Nel in [2] as follows:

$$\Pi = \lambda \left(\left(\frac{\beta p_2}{p_{th1}} \right)^{-\alpha} - 1 \right) (p_2 - c) + \lambda \left(\left(\frac{p_2}{p_{th2}} \right)^{-\alpha} - 1 \right) (p_2 - c) \quad (1)$$

To find the optimal prices and an optimal subsidy rate at which no user will unilaterally change his decision, we consider the service provider’s profit maximization problem (see Eq. 1). Often times finding such optimal prices entails checking all the likely values of p_2 and subsidy factor β and ask ourselves; “is this a Nash Equilibrium?” At times it is possible to eliminate actions iteratively to narrow the cases that need to be checked. Since, profit functions are continuously differentiable, concave and the price p_2 is always positive, we can take the first order conditions of Π as follows:

$$\frac{\partial \Pi}{\partial p_2} = \frac{\lambda}{\left(\frac{\beta p_2}{p_{th1}} \right)^{-\alpha} - 1} + \frac{\lambda}{\left(\frac{p_2}{p_{th1}} \right)^{-\alpha} - 1} + \frac{\lambda \alpha (c - p_2)}{p_{th2} \left(\frac{p_2}{p_{th2}} \right)^{\alpha+1}} + \frac{\lambda \alpha \beta (c - p_2)}{p_{th1} \left(\frac{\beta p_2}{p_{th1}} \right)^{\alpha+1}} \quad (2)$$

To find the concavity of the above profit function, we take the partial derivatives of Eq. 2. Clearly, the second order conditions imply that marginal profits should slope downwards with respect to the ISP’s own action. Further differentiation of Eq. 2 provides us with an expression as follows:

$$\frac{\partial^2 \Pi}{\partial p_2^2} = \left(\frac{2\lambda\alpha}{p_{th2} \left(\frac{p_2}{p_{th2}}\right)^\alpha + 1} \right) - \left(\frac{2\lambda\alpha\beta}{p_{th1} \left(\frac{\beta p_2}{p_{th1}}\right)^\alpha + 1} \right) - \left(\frac{\lambda\alpha(c - p_2)(\alpha + 1)}{p_{th2}^2 \left(\frac{p_2}{p_{th2}}\right)^{\alpha+2}} \right) - \left(\frac{\lambda\alpha\beta^2(c - p_2)(\alpha + 1)}{p_{th1}^2 \left(\frac{\beta p_2}{p_{th1}}\right)^{\alpha+2}} \right) \tag{3}$$

To begin with, we assume that the government and the service provider lower the price and the subsidy factor as shown in Table 1. At NE selfish users will use the network resources without restraint. Given that the price is low, and government is heavily subsidizing the information “have nots”, there will be an increase in the number of users trying to maximize their utility which may lead to the tragedy of commons. Consider again, that this time the service provider increases its price and the government heavily subsidizes the information “have nots”. As observed from Table 1, increasing the subsidy and increasing the price p_2 without consideration will result in the information “haves” failing to pay for the network services when the price set reaches the reservation price p_{th2} .

For example when the price of network services is approximately equal or above the reservation price, the information “haves” won’t be able to pay for the network services i.e. $p_{th2} = 100 = p_2$. This scenario may result in the free rider’s problem and may disrupt the once stable telecommunication market or economic system of the country.

Table 2 paints a picture of how the value of subsidy factor β affects the number of information “have nots”. At no subsidy or when subsidy from the government is approximately equal to zero, the numbers of information “have nots” users won’t be able to pay for the service because $p_1 = p_{th1}$. Tables 3 and 4 shows the effect of users’ sensitivity towards price at different values of β .

Graphical analysis: In Figs. 2, 3 and 4, we evaluate the effect of subsidy on revenue maximization, given any price p_2 and the threshold price of consumers i.e. p_{th1} and p_{th2} at a constant sensitivity α Fig. 1, for example, shows that for a subsidy

Table 1 Expected amount of revenue with varying p_2 : when $p_{th1} = 40, p_{th2} = 100$

Users	P_{th1}	P_{th2}	P_2	P_1	β	α	Revenue	Normalized number of users
N_1	40		20	5	0.1	0.3	291.30	1.4565
N_2		100	20	5	0.1	0.3	124.13	0.6207
N_1	40		50	20	0.1	0.3	692.85	0.8661
N_2		100	50	20	0.1	0.3	184.91	0.2311
N_1	40		90	35	0.1	0.3	903.02	0.5644
N_2		100	90	35	0.1	0.3	51.38	0.0321
N_1	40		100	40	0.1	0.3	928.28	0.5157
N_2		100	100	40	0.1	0.3	0	0

Table 2 Expected amount of revenue with varying β : when $p_{th1} = 40, p_{th2} = 100$

Users	P_{th1}	P_{th2}	P_2	P_1	β	α	Revenue	Normalized number of users
N_1	40		50	5	0.1	0.3	692.00	0.8661
N_2		100	50	5	0.1	0.3	184.00	0.2311
N_1	40		50	20	0.4	0.3	184.91	0.2311
N_2		100	50	20	0.4	0.3	184.00	0.2311
N_1	40		50	35	0.7	0.3	32.69	0.0409
N_2		100	50	35	0.7	0.3	184.00	0.2311
N_1	40		50	40	0.8	0.3	0	0
N_2		100	50	40	0.8	0.3	184.00	0.2311

Table 3 Expected amount of revenue with varying α : when $p_{th1} = 40, p_{th2} = 100$

Users	P_{th1}	P_{th2}	P_2	P_1	β	α	Revenue	Normalized number of users
N_1	40		50	5	0.3	0.1	57.41	0.1031
N_2		100	50	5	0.3	0.1	57.41	0.0718
N_1	40		50	20	0.3	0.3	184.91	0.3421
N_2		100	50	20	0.3	0.3	184.91	0.2311
N_1	40		50	35	0.3	0.7	499.60	0.9869
N_2		100	50	35	0.3	0.7	499.60	0.6245
N_1	40		50	40	0.3	1	800.00	1.6667
N_2		100	50	40	0.3	1	800.00	1

Table 4 Optimal subsidy rate, β , with varying p_{th1} and p_{th2} : when $\alpha_1 = 0.4, \alpha_2 = 0.7$

Type of users	P_{th1}	P_{th2}	P_2	P_1	β	α	ISP revenue
N_1	10		50	5.5	0.11	0.4	216.119
N_2		90	50	5.5	0.11	0.7	407.2042
N_1	40		50	16.55	0.331	0.4	338.66
N_2		120	50	16.55	0.331	0.7	676.55
N_1	70		50	27	0.54	0.4	216.119
N_2		130	50	27	0.54	0.7	407.304

factor of $\beta = 1$ (in this case we assume that the government does not subsidize the information “have nots”), the revenue generated from the information “have nots” is, actually, dependent on the threshold value or the maximum reservation price p_{th1} that they are willing to pay. This phenomenon, especially in underserved areas of developing countries, may sometimes result in the tragedy of anti-commons since very few information “have nots” will be more willing to subscribe to the ISP. For all intents and purposes, it is worthwhile to note that the information “haves nots” are contributing very little towards the ISP’s overall revenue since no subsidy is being



Fig. 1 Price discrimination may lead to: **a** underuse in Gautrain (tragedy of anti-commons) or **b** overuse in metro-rail (tragedy of the commons)

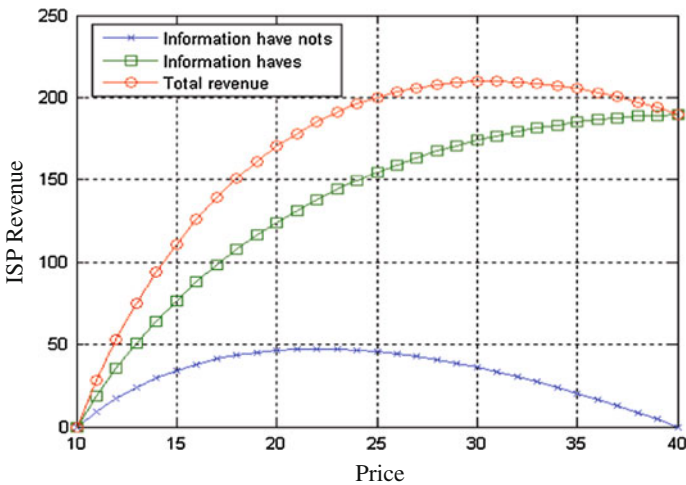


Fig. 2 Revenue versus price at $\alpha = 0.3, p_{th1} = 40, p_{th2} = 100, \beta = 1, c = 10$

given by the government towards the ISP (see Fig. 3). In order to obtain maximum revenue from the users (as shown in Fig. 2), it will be meaningful for the ISP to charge the information “have nots” and “haves” an optimal price of 30 units so as to maximize revenue.

Since the main aim of the government is to promote social and economic growth in underserved regions, Fig. 3 would represent a much more desirable scenario in promoting social and economical development in underserved areas (ICT access for all in developing countries). In actual fact, Fig. 3 represents a desirable outcome that will objectively fulfill government policies, as it allows more information “have nots” to subscribe to the ISP at a subsidized price. However, such a situation would definitely promote the tragedy of the commons and may lead to market failure in once stable markets. A more close analysis of Fig. 3, in reality, reveals that the information “haves” will fail to subscribe to the ISP, at a higher price e.g. at 100 units. Since this is

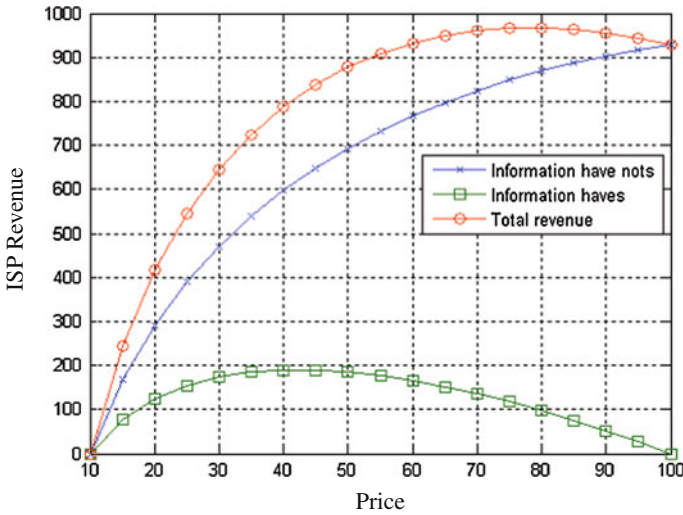


Fig. 3 Revenue versus price at $\alpha_{1,2} = 0.3$ $p_{th1} = 40$ $p_{th2} = 100$ $\beta = 0.1$ $c = 10$

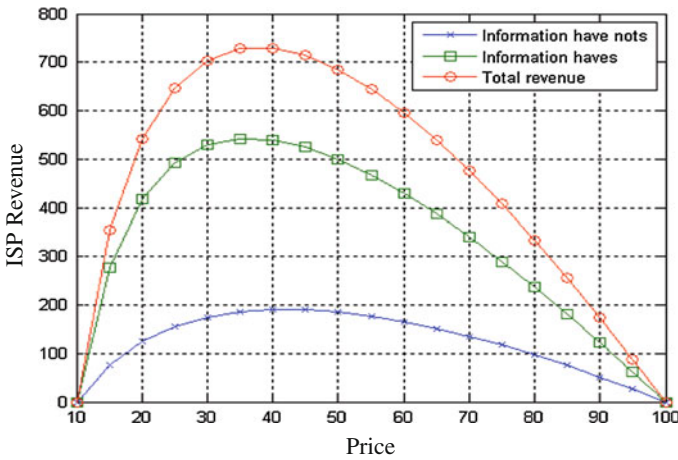


Fig. 4 Revenue versus price at $\alpha = 0.3$ for N_1 , $\alpha = 0.7$ for N_2 and $p_{th1} = 40$ $p_{th2} = 100$ $\beta = 0.4$ $c = 10$

the reservation price of information “haves”. Thus, Fig. 2 highlights the implications of heavy subsidy towards revenue maximization in a heterogeneous society, and shows that at an optimal price of 80 units (partially paid by the government) there will be a skewed information access between the information “haves” and the information “have nots” due to unbalanced subsidy distribution. Figure 4, on the other hand shows, that the revenue of an ISP is maximum at an optimal price of 40 units. In our view,

Fig. 4 represents an equilibrated game between the information “haves” and “have nots” as the price of both groups seem to follow the same trajectory. Actually, Fig. 4 presents a case where both groups of users seem to follow the same trajectory price and revenue path. Both users (as depicted in Fig. 4) will not subscribe to the ISP at the same reservation price or threshold price of p_{th2} . Contrary to Figs. 2, 3, and 4 represents a desirable outcome between the information “haves” and “have nots” as it represents the same price versus revenue curve trajectory for both the information “haves” and “have nots”. Generally, Fig. 4, in point of fact, may somehow prevent the tragedy of the commons and tragedy of the anti-commons in heterogeneous society as no member of a group is disadvantage due to skewed pricing.

8 Conclusion

In this chapter we have analyzed the effect of subsidy driven procedure on the resource usage, revenue maximization and the impact of subsidy policy in multitier communities. We have shown that if implemented properly, subsidies can prevent the tragedy of the (anti) commons and could be the solution to promoting social and economical development in developing countries especially in sub-Saharan Africa. Our proposed subsidy driven pricing model or policy addresses the importance of a balanced subsidy pricing scheme with a view of achieving social and economical balance while enhancing the usage of network resources efficiently. Using this framework, we have shown that a government subsidy in a developing country can result in the increase or decrease of information “have nots” leading to the tragedy of (anti) commons. We have also shown that a correct subsidy, given customer sensitivities, can promote desired revenue and that the revenue is a concave function of price. Thus far, we have derived the optimal revenue given the price, subsidy factor β and α as shown in the Figs. 2, 3 and 4.

Acknowledgments This work was partially supported by the University of Johannesburg commonwealth bursary fund.

References

1. Sumbwanyambe M, Nel AL (2012) A subsidy driven decision procedure to mitigate the tragedy of the commons and anti-commons. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE, UK, London, pp 1651–1656, 4–6 July 2012
2. Heller MA (1997) The tragedy of the anti-commons: property in the transition from Marx to markets. *Harvard Law Rev* 3:621–688
3. Ward MR Rural telecommunications subsidies do not help. <http://www.uta.edu/faculty/mikeward/JRAP.pdf>
4. Levin SL (2010) Universal service and targeted support in a competitive telecommunication environment. *Telecommun Policy J* 34:92–97

5. Ramos B, Saeed K, Pavlov O (2010) The impact of universal service obligations and international cross-subsidies on the dispersion of telephone services in developing countries. *Socio-Econ Plan Sci* 44:57–72
6. Amegashie JA (2006) The economics of subsidies. *Crossroads* 6(2):7–15
7. Sumbwanyambe M, Nel AL, Clarke WA (2011) Challenges and proposed solutions towards telecentre Sustainability: a Southern Africa case study. In: *Proceedings of IST-Africa 2011*, Gaborone, Botswana
8. Harding G (1968) The tragedy of the commons. *J Sci* 168:1246–1248
9. TIBA (2012) South Africa's universal service agency to offer big players subsidy to take voice and data to most rural areas. <http://www.balancingact-africa.com/news/en/issue-no-398/telecoms/south-africa-s-unive/en>. Accessed Feb 2012
10. Alesina A, Rodrik D (1994) Distributive politics and economic growth. *Q J Econ* 109(2):465–490
11. Sumbwanyambe M, Nel AL (2012) Subsidy and revenue maximization in developing countries. In: *Lecture notes in engineering and computer science: international multicongress of engineers and computer scientist*, pp 1568–1573
12. Esselaar S, Gillwald A, Moyo M, Naidoo K (2009/2010) Towards evidence-based ICT Policy and regulation. *S Afr Sector Perform Rev* 2
13. Habeezu S (2009/2010) Towards evidence-based ICT policy and regulation. *Zambian Sector Perform Rev* 2:1–38
14. DOC (2011) Doc documents. www.doc.gov.za
15. Oyedemi TD (2009) Social inequalities and the South Africa ICT access policy agendas. *Int J Commun* 3:151–168
16. Padayachie RR (2011) Budget vote of the department of communications. <http://www.doc.gov.za/>
17. Hayashi T (2011) Fostering globally accessible and affordable ICTs. <http://www.itu.int/osg/spu/visions/papers/accesspaper.pdf>
18. Sumbwanyambe M, Nel AL, Clarke WA (2011) Optimal pricing of telecomm services in a developing country: a Game theoretical approach. In: *Proceedings of the 10th IEEE Africon conference*. Livingstone Zambia, pp 24–26
19. Jagannathan S, Almeroth KC (2002) Prices issues in delivering E-content on-demand. *ACM SIGecom Exch* 3:19–27

Modeling Emergency Department Using a Hybrid Simulation Approach

Norazura Ahmad, Noraida Abdul Ghani, Anton Abdulbasah Kamil
and Razman Mat Tahar

Abstract Within hospital, emergency department is one of the most important unit that involves complex patient movement flow and detailed operational activities. As an integrated system, the efficiency of emergency department depends on its interaction between inter-departmental units and intra-departmental elements. Over the years, with the rapid development of computer technology, there has been a rising trend of using simulation modeling to improve healthcare operations. Discrete-event simulation (DES) has become a popular and effective decision-making tool for modeling the stochastic operational activities in a system. However for a whole system approach, system dynamics (SD) has advantages over DES. SD does not require vast data and is able to capture the interdependency relations between different units in an integrated system. Both approaches have strengths and weaknesses that may support and complement each other. An integrated model of both approaches will provide a realistic view of a complex system. This chapter provides an overview of the hybrid simulation modeling applications to emergency department.

N. Ahmad (✉)

School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara,
Malaysia, 06010 Sintok, Kedah, Malaysia
e-mail: norazura@uum.edu.my

N. A. Ghani · A. A. Kamil

School of Distance Education, Universiti Sains Malaysia, USM,
11800 Georgetown, Pulau Pinang, Malaysia
e-mail: noraida@usm.my

A. A. Kamil

e-mail: anton@usm.my

R. Mat Tahar

Faculty of Technology Management, Universiti Malaysia Pahang, 26300
Gambang, Kuantan, Malaysia
e-mail: razman779@ump.edu.my

Keywords Complex system · Discrete-event simulation · Emergency department · Healthcare system · Hybrid simulation · System dynamics

1 Introduction

An emergency department (ED) is recognized as the front door of hospital, where patients arrive without prior appointment. Also known as accident and emergency department, this unit is a critical health unit where the battle between life and death is sometimes inevitable. With a person's life at stake, the concern of ED administrators is to minimize the waiting time and length of stay while providing a satisfactory treatment care.

In ED care, the key bottleneck is usually the resource or activity in treatment process that requires a patient spending a longer time in the system. The difference in demand and supply often builds up bottlenecks, which in turn will increase the waiting times. However, bottlenecks in ED may originate from factors external to ED [1] because some of the resources are shared resources with other units in the hospital. Due to the complex patient flows and interdependency with other units such as wards and labs, the use of analytical models is definitely not a preferable choice to represent ED. One effective approach that has been proven to be a good and flexible tool for modeling such a complex system is simulation. Through application of simulation modeling, identifying bottlenecks, evaluating the current existing system as well as testing proposed alternatives can be explored without directly affecting patient care or altering the real system. This approach has received a great attention among researchers in healthcare [2].

Brailsford [3] describes the increasingly acceptance of simulation among healthcare researchers is grounded on three main points. Firstly, healthcare systems require a stochastic approach to deal with uncertainties and variability in the system. Secondly, healthcare systems are usually complex systems that require effective modeling technique. Thirdly, human interactions in healthcare systems need an approach that allows interaction between modeler and user. These aspects are all embedded in the simulation approach and it justifies the escalating usage of simulation in healthcare systems.

In healthcare, most of the systems consist of interrelated units and therefore it is difficult to study the system in isolation from other supporting units. With this in mind, administrators are not only interested to track the status of individual entities but are also concerned on the impact of a unit to other units in a healthcare system. In the case of ED operations, the admission process describes the complexities of relations between many units in hospital. If a patient needs to be admitted, a bed is requested in the appropriate nursing unit. The availability of a bed is affected not only by the capacity of the relevant unit, but also by the admission of elective patients who compete for the same beds as emergency patients. Therefore, ED modeling should demonstrate the detail operational level as well as the relationships with other supporting units in the hospital. This is where the discrete-event simulation (DES) and

system dynamics (SD) modeling can fit in. Both approaches have strengths and weaknesses that may support and complement each other. Integration of both approaches will be able to grasp the detailed and dynamic complexity issues simultaneously [4].

This chapter is an extended version of a paper presented in the WCE 2012 conference [5]. We discuss the potential of integrating two familiar simulation approaches, DES and SD in modeling a healthcare system, particularly ED. We begin the chapter by a discussion on previous works of ED modeling. Then, in the following section we provide a discussion on the combination of DES and SD modeling in healthcare. Previous works using hybrid simulation are also presented and the chapter ends with brief conclusions.

2 Simulation Models for Emergency Department

Over the years, healthcare demand is rising as does the population [6]. The aging of our population and lack of places for affordable acute medical treatment have resulted in increasing number of attendances to ED. Many patients experience very lengthy waits before treatment and some may leave without being treated. Besides waiting times, other issues that are related to the ED are resource allocation and overcrowding. To mitigate this issue, a significant amount of research has been conducted in ED and some are briefly reviewed in this section.

Many researchers have been using Operations Research (OR) or Management Sciences (MS) methods to solve problems in healthcare domains [7]. However, simulation outnumbered other OR/MS approaches in studies related to ED [8]. In ED modeling, many researchers have concentrated on resource allocation, improving patient flow and reducing patients' waiting time. For example, [9–11] simulated ED to determine the optimal number of staff to run ED more efficiently. On the other hand, Medeiros et al. [12] introduced Provider Directed Queuing (PDQ) to improve patient flow in ED. The PDQ listens to patient complaints and conducts medical evaluation before sending patients to emergency room. Another study by Komashie et. al. [13] studied ED to find the hidden causes of long waiting time. They developed a DES model to determine the impact of critical resources (beds, doctors and nurses) on key performances (queuing times and length of stay). Cochran and Roche [14] described a different approach to improve the performance of ED. They proposed a multi-class queuing network model to increase the capacity of an ED to treat patients. A methodology was introduced to model across any patient mix, arrival volume and operational performance. Others have also used simulation to estimate future capacity of new ED facilities or expansion of current ED [15].

There have been several excellent studies in ED using SD modeling [16–18]. Royston et al. [16] described several applications of system dynamics modeling to problems in the UK's National Health System (NHS). One of the applications described in detail in the paper uses system dynamics to develop a better understanding of the interactions between the emergency care system and the social care system. Lane et al. [17] however focused on the micro level of the healthcare system by

modeling patients' flow at an ED to investigate the sensitivity of waiting time to hospital bed numbers. The developed model was used to explore scenarios which involved changes in bed capacity as well as in ED demand. On the other hand, Brailsford et al. [18] developed a high-level model of patients' flows in the entire Nottingham health emergency care system. The model was useful to identify the potential bottlenecks in the healthcare system and was readily adopted by policy makers to test various policy scenarios.

From the literature, it has been shown that simulation has been progressively applied in ED modeling [19]. Perhaps one obvious reason is that ED is the front door of hospital for many and therefore issues like waiting time and length of stay quickly become apparent to the public. Since the system experts such as doctors and nursing staff are busy to provide adequate care they do not have much time to study analytical models with complex numerical equations to determine the possible alternative solutions. Animation features embedded in simulation model have advantages over the analytical techniques and thus motivate the adoption of simulation among healthcare researchers in ED modeling. However this does not mean that successful ED simulation models are easy to develop [20, 21].

There are a few references that delineate the fundamental principles for conducting a successful simulation study of a healthcare system. Mahachek [22] introduced structured tutorials on simulation steps that are required in conducting a DES model for healthcare system. Isken et al. [23] present a general framework for simulating outpatient obstetrical clinics that is also applicable to other outpatient clinics. In addition, Eldabi and Young [24] discussed a framework selection tool that enabled healthcare administrators to choose appropriate modeling tools in modeling healthcare system. From literature, the need for healthcare modeling using combination methods [25] are also addressed as does the hybrid simulation applications in healthcare [5, 26, 27]. The following section provides an overview of combination between DES and SD in healthcare.

3 DES and SD: A Hybrid Simulation in Healthcare

It is well known that the use of DES modeling is widely established in manufacturing and business environment [28]. A special property of DES is that the modeling of systems can be represented by series of events at discrete time intervals. It is a stochastic modeling approach grounded on queuing theory where movement of entities in queuing system is governed by probability distributions. Entities in the system flow around a network of services or processes and may have characteristics which determine their trail through the network.

DES permits one to track the status of individual entities in a system and compute performance measures associated with the entities. The tracking is done by focusing on times at which the various events occur. In DES modeling great emphasis is given on the effect of random variation. Such stochastic effects are not considered in analytical models and are even less important in SD models. Another advantage of DES

is the animations and graphics visualization embedded in software packages that has enhanced its visual interactivity. This makes DES an ideal tool to communicate with healthcare administrators. The animations provide justification for factual figures and help in a better understanding of the system. However, healthcare researchers do not favor DES to model the holistic view of healthcare integrated system [2] since the complexity will increase exponentially with the size of the model [27]. For this reason SD has an advantage over DES.

SD modeling begins with the identification of causal factors that affect the performance of a system. Once identified, the causal factors are drawn in diagrams to reveal the relationship between each factor. The diagrams are known as Causal Loop diagrams and it is then translated to Stock-Flow diagrams. A causal diagram consists of variables that are connected by arrows denoting the causal relationship between the variables. Feedback and unanticipated effects of key elements in the system is basis to understanding the dynamic complexity of the system. The graphical description of the system based on Stock-Flow diagrams can be translated into a mathematical description. Only Stock-Flow diagrams are formulated in the simulation model. Further information on the diagrams is discussed in detailed by Sterman [29].

SD approach as introduced by J.W. Forrester in the late of 1950s, gives real insights of how the system behaves in the short and longer period [30]. Besides, unlike other traditional analysis that studies a problem by separating and solving it into smaller parts, SD methodology provides a holistic view of a system by looking at possible interactions among subsystems. From the above discussion, apparently DES and SD have strengths to offer in healthcare as well as drawbacks that need to be overcome. From the marriage of DES and SD the detailed and dynamic complexity of ED operations can be captured.

The application of a hybrid simulation initially begins in manufacturing sectors. In most cases, the developed models were created for software development process as well as supply chain management. For example, Martin and Raffo [31] and Setamanit et al. [32] combined DES and SD to model discrete activities within a continuous environment in software development process. Rabelo et al. [33] and Helal et al. [34] on the other hand, deployed the detailed event approach of DES for modeling operational decisions, whereas SD is used to capture the strategic decisions in manufacturing sectors. Similarly, Reiner [35] and Venkateswaran et al. [36] have integrated DES and SD in supply chain management.

In healthcare sector, Giachetti et al. [37] develop a hybrid model for an outpatient clinic to determine the practicality of having an open access policy. Chahal and Eldabi [27] stressed that healthcare problems have bigger scope when compared to other sectors. Therefore, they proposed and applied a generic framework for hybrid simulation in healthcare. A non-technical approach for integrating DES and SD in healthcare simulation is introduced by Zulkepli [6]. The simulation model is developed for complex patient pathways in integrated healthcare system.

Though not much hybrid simulation models are developed in healthcare, there have been considerable debates in applying both detailed and whole system approach to healthcare systems [4, 21, 24, 27]. Integrating the DES and SD approaches would be challenging but yet beneficial as the hybrid model will provide detailed and

individual patient analysis as well as a whole system approach for the real healthcare systems.

4 The Proposed Model

4.1 System Model Description

The ED understudy is open 24h every day a week and receives an average of 1563 patients weekly. The process flow begins with the patient arrival. Patients arrive by either ambulance or as walk in cases. Regardless of mode of arrival, the arriving patient stops at a registration counter for registration. At the same time a medical assistant will triage the patient. However, critical patients will be sent directly to the critical area and bedside registration will be performed at a later time by the registrar.

Once triaged, the patient moves to the waiting area and wait to be called for treatment. If there is availability of a doctor, the patient moves to the treatment area and sees the doctor. Delay of doctor to patient contact depends on severity of illnesses. Doctors will decide if the patient needs further tests such as clinical lab tests or X-rays. Results obtained will be reviewed by doctors and a decision is made upon the results. Some patients will be observed temporarily in ED's observation ward before release for discharge. Discharged patients can either be released to go home or send to hospital wards for further treatment.

4.2 The Framework

In order to capture the detailed and dynamic complexity of ED, the methodological process for this research consists of a two-stage framework. Interrelations between ED and other units in the hospital are evaluated using SD, while DES is used to capture the operational level inside the ED. Output from DES model will be passed to SD model to study the feedback effects of elements in the system. Stage 1 is the initial stage of the study that requires sufficient comprehension on the problems understudy. Problems are identified by communicating with the system owner. Then, several observations are conducted to understand patient flows in the ED. Interviews with administrators and staffs regarding process flows of the department are also carried out. Next, after obtaining a clear view of the system understudy, a set of objectives are established to alleviate problems in the ED. At this stage, objectives are re-examined to identify whether the objectives can be achieved using a DES model, SD model or by both. Upon deciding to use a single technique or hybrid approach both DES and SD are compared from modeling perspective and system perspective.

The implementation of Stage 2 depends on the outcome from Stage 1. Based on objectives in Stage 1, variables are identified and divided into DES and SD model.

Next, both models are developed based on DES modeling steps and SD modeling steps. At this phase format interactions between both models are identified. The models are run in parallel where both models are run concurrently and information are exchanged while running. Once the models are completed, several tests including qualitative validation will be performed before model implementation. Validation is carried out with participation from ED administrators. This should be done to increase the confidence in the developed model among the system owners.

The hybrid model is based on actual ED system of a government hospital in Penang, Malaysia. The developed model is created to have a better understanding of intra-departmental and inter-departmental interactions of the system and how patient flow in the ED is influenced by capacity availability and process change. AnyLogic 6.6 software package is used to create both DES and SD models.

4.3 The Hybrid Model

There are two types of hybrid interaction, namely parallel and cyclical [38]. Parallel interaction requires models to be run simultaneously while in cyclical interaction, model is run model by model. In this study, models are run in parallel as the variables are linked in space and time. Both models are created using AnyLogic Software that is capable to integrate DES and SD models in a similar interface. The DES model is developed to capture the operational level inside the ED. The model consists of the main model which is subdivided into three sectors representing patient flows according to the triage zones: Green, Yellow and Red. Figure 1 shows a screen shot of the ED model logic.

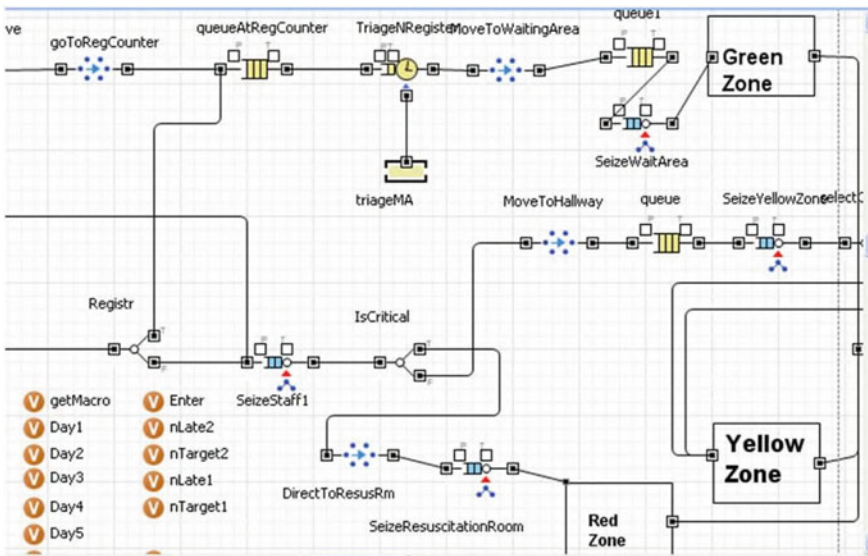


Fig. 1 A part of DES model logic

Table 1 Preliminary result

Performance measure	Output
Mean wait time for triaged Green	15.20 min (95 % CI* = 12.4–18)
Mean treatment time for triaged Green	48.81 min (95 % CI = 40.71–56.92)
Average length of stay for triaged Green	70.24 min (95 % CI = 60.34–80.14)
Mean wait time for triaged Yellow	3.5 min (95 % CI = 3.09–3.91)
Mean treatment time for triaged Yellow	103.44 min (95 % CI = 93.69–113.18)
Average length of stay for triaged Yellow	116.63 min (95 % CI = 107.25–126.01)

The model is run 12 times for seven day replication length with one day warm up period. Result from the 12 replications is compared to actual 12 weeks data that record the number of arrivals and number of patient in each triage zone. Preliminary results from the DES model reveal that an average of 1613 patients visits the ED per week. Of the average total patients, 89.77 % were triaged as Green, 9.3 % were triaged as Yellow and less than 1 % were triaged as Red. Table 1 shows that on resource utilization, medical assistants score the highest utilization with an average of 70 %. This result parallels with the administration's claim that medical assistants usually are the busiest resources because they assist doctors in treatment processes and also serve for ambulance services. For room utilization, waiting area has the highest value, which is 70 %. As for Green triaged patients, results also reveal that they wait on average approximately 15 min to see the doctors and spend around an hour in the system. On the other hand, for Yellow triaged patients, the mean wait time is 3.5 min and the average length of stay is 116.63 min.

From interviews with ED administrators, four main sectors have been identified affecting ED services. They are human resources, lab tests/X-ray, ambulance and wards. The principle objective of the SD model is to determine the impact of each factor to ED operations. The model is used to investigate whether sharing resources between the treatment process and ambulance service affect ED performance, and as well as to identify whether turnaround times between ED and labs increase the length of stay in the system.

The SD model consists of five sectors that show interrelationships between ED operations and other units such as ambulance, wards, labs and human resource. For each of the sector, causal loop diagrams are developed that act as the dynamic hypothesis for computer model construction. Figure 2 shows the causal loop diagram of the operation sector at the case study ED.

Loops [1a] and [1b] act to deplete the number of patients in ED by discharging them either to home or send to wards. Loop [2] shows an increase in the number of patients entering the ED increases the number of patients in ED and thus, increases the request to see the doctors. This will affect doctor availability and patient's length of stay. Besides, an increase in the number of emergency patients will also increase the request for lab tests. Causal loops for other sectors were also developed and then converted into formal computer model using AnyLogic 6.6.

approaches the intra-departmental elements and inter-departmental relations of the system can be captured. The developed model can be used by the administrators to assist the decision making process in improving the hospital emergency system.

References

1. Kolb EMW, Lee T, Peck J (2007) Effect of coupling between emergency department and inpatient unit on the overcrowding in emergency department. In: Proceedings of the 2007 winter simulation conference, pp 1586–1593
2. Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete event simulation in healthcare clinics: a survey. *J Oper Res Soc* 50:109–123
3. Brailsford SC (2007) Tutorial: advances and challenges in healthcare simulation modeling. In: Proceedings of the 2007 winter simulation conference, pp 1436–1448
4. Chahal K, Eldabi T (2008) Applicability of hybrid simulation to different modes of governance in UK healthcare. In: Proceedings of the 2008 winter simulation conference, pp 1469–1477
5. Ahmad N, Ghani NA, Anton AK, Mat Tahar R (2012) Emergency department problems: a call for hybrid simulation, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering 2012, WCE 2012, 4–6 July, 2012, London UK, pp 1470–1474
6. Zulkepli J (2012) A theoretical framework for hybrid simulation in modeling complex patient pathways. PhD thesis 2012, Brunel University, London, UK (unpublished), p 1
7. Garg L, McClean S, Barton M (2008) Is management science doing enough to improve healthcare? *Proc World Acad Sci Eng Technol* 30:76–80
8. Ruohonen T, Neittaanmaki P, Teittinen J (2006) Simulation model for improving the operation of the emergency department of special healthcare. In: Proceeding of the 2006 winter simulation conference, pp 453–458
9. Ahmed MA, AlKhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur J Oper Res* 198(3):936–942
10. Mahapatra S, Koelling CP, Patvivatsiri L, Fraticelli B, Eitel D, Grove L (2003) Pairing emergency severity index5-level triage data with computer aided system design to improve emergency department access and throughput. In: Proceedings of the 2003 conference on winter simulation, pp 1918–1925
11. Takakuwa S, Shiozaki H (2004) Functional analysis for operating emergency department of a general hospital. In: Proceedings of the 2004 winter simulation conference, pp 2003–2011
12. Medeiros DJ, Swenson E, DeFlitch C (2008) Improving patient flow in a hospital emergency department. In: Proceedings of the 2008 winter simulation conference, pp 1526–1531
13. Komashie A, Mousavi A (2005) Modeling emergency departments using discrete event simulation techniques. In: Proceedings of the 2005 winter simulation conference, pp 2681–2685
14. Cochran JK, Roche KT (2009) A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Comput Oper Res* 36:1497–1512
15. Baesler FF, Jahnsen HE, DaCosta M (2003) The use of simulation and design of experiments for estimating maximum capacity in an emergency room. In: Proceedings of 2003 winter simulation conference, pp 1903–1906
16. Royston G, Dost A, Townshend J, Turner H (1999) Using system dynamics to help develop and implement policies and programmes in healthcare in England. *Syst Dyn Rev* 15:293–313
17. Lane DC, Monefeldt C, Rosenhead JV (2000) Looking in the wrong place for healthcare improvements: a system dynamics study of an accident and emergency department. *J Oper Res Soc* 51(5):518–531
18. Brailsford S, Lattimer VA, Tarnaras P, Turnbull JC (2004) Emergency and on-demand healthcare: modelling a large complex system. *J Oper Res Soc* 55(1):34–42

19. Brenner S, Zeng Z, Liu Y, Wang J, Li J, Howard PK (2010) Modeling and analysis of the emergency department at university of Kentucky Chandler hospital using simulations. *J Emerg Nurs* 36(4):303–310
20. Brailsford SC, Churilov L, Liew SK (2003) Treating ailing emergency departments with simulation: an integrated perspective. In: Anderson J (ed) *Proceedings of western multiconference on health sciences simulation*. Florida, Society for Modelling & Simulation International (SCS), San Diego, CA
21. Brailsford SC, Desai SM, Viana J (2010) Towards the holy grail: combining system dynamics and discrete-event simulation in healthcare. In: *Proceedings of the 2010 winter simulation*, pp 2293–2303
22. Mahacheck A (1992) An introduction to patient flow simulation for healthcare managers. *J Soc Health Syst* 3:73–81
23. Isken MW, Ward TJ, McKee TC (1999) Simulating outpatient obstetrical clinics. In: *Proceedings of the 1999 winter simulation conference*, pp 1557–1563
24. Eldabi T, Young T (2007) Towards a framework for healthcare simulation. In: *Proceedings of the 2007 winter simulation conference*, pp 1454–1460
25. Eldabi T, Paul RJ, Young T (2007) Simulation modelling in healthcare: reviewing legacies and investigating futures. *J Oper Res Soc* 58(2):262–270
26. Zulkepli J, Eldabi T (2011) Technique for improving care integration models. In: *Proceedings of the european, Mediterranean & middle eastern conference on information systems 2011*, Athens, Greece, pp 46–58 May 30–31, 2011
27. Chahal K, Eldabi T (2008) Which is more appropriate: a multi-perspective comparison between system dynamics and discrete event simulation. In: *Proceedings of the (2008) European Mediterranean conference on information systems*, Dubai, UAE
28. Kuljis J, Paul RJ, Stergioulas LK (2007) Can healthcare benefit from modeling and simulation methods in the same way as business and manufacturing has? In: *Proceedings of the 2007 winter simulation conference*, pp 1449–1453
29. Serman J (2000) *Business dynamics: system thinking and modeling for a complex world*. McGraw Hill, Boston
30. Ford A (1999) *Modeling the environment: an introduction to system dynamics models of environmental systems*. Island Press, Washington DC
31. Martin RH, Raffo D (2000) A model of the software development process using both continuous and discrete models. *Softw Process Improv Pract* 5(2–3):147–157
32. Setamanit S, Wakeland W, Raffo D (2007) Using simulation to evaluate global software development task allocation strategies. *Softw Process Improv Pract* 12(5):491–503
33. Rabelo L, Helal M, Jones A, Min HS (2005) Enterprise simulation: a hybrid system approach. *Int J Comput Integr Manuf* 18(6):498–508
34. Helal M, Rabelo L, Sepúlveda J, Jones A (2007) A methodology for integrating and synchronizing the system dynamics and discrete event simulation paradigms. In: *Proceedings of the 25th international conference of the system dynamics society*, vol 3, pp 1–24
35. Reiner G (2005) Customer-oriented improvement and evaluation of supply chain processes supported by simulation models. *Int J Prod Econ* 96(3):381–395
36. Venkateswaran J, Son YJ, Jones AT, Min HSJ (2006) A hybrid simulation approach to planning in a VMI supply chain. *Int J Simul Process Model* 2(3–4):133–149
37. Giachetti RE, Centeno EA, Centeno MA, Sundram R (2005) Assessing the viability of an open access policy in an outpatient clinic: a discrete event and continuous simulation modelling approach. In: *Proceeding of the 2005 winter simulation conference*, pp 2246–2255
38. Chahal K, Eldabi T (2010) A generic framework for hybrid simulation in healthcare. In: *Proceedings of the system dynamics conference 2010*, (April 10, 2011). <http://www.systemdynamics.org/conferences/2010/proceed/papers>

The Challenge of Adopting Minimal Quantities of Lubrication for End Milling Aluminium

Brian Boswell and Mohammad Nazrul Islam

Abstract End milling is a very common metal cutting process used for the machining of most types of metal. The process is inherently intermittent causing the tool tip edge to constantly fluctuate between various levels of temperatures, specifically from cold to 300 °C when cutting Al alloy. During dry end milling cutting temperatures need to remain within the design specifications of the tool tip. Even working with Al alloy the tool tip is subjected to thermal cyclic stresses. Conventional wisdom states that it is essential to use flood cooling during end milling, as intermittent cooling increases the effect of thermal shock and build up edge. Al alloy—unlike other materials—needs cutting fluid to avoid smearing the insert edges and to improve the surface finish. Modern machining companies constantly face the challenges of environmental issues that affect the manufacturing costs of machined parts. New environmental manufacturing techniques need to be developed for companies to remain competitive in the future. The research presented in this paper represents the experimentation involved in determining a suitable environmental alternative to using copious amounts of cutting fluid during end milling of Al alloy. Previous experimental evaluation of Minimal Quantities of Lubrication (MQL) when applied to the machining of Al alloy has proved to be inconclusive.

Keywords End milling · Environmental issues · Flood coolant · Intermittent cooling · Minimal quantities of lubrication · Thermal shock

B. Boswell (✉) · M. N. Islam
Curtin University, GPO Box U1987, Perth, WA 6845, Australia
e-mail: b.boswell@curtin.edu.au

M. N. Islam
e-mail: m.n.islam@curtin.edu.au

1 Introduction

The machining process involves removing unwanted material from the workpiece in the form of chips, and is one of the principal methods of manufacturing. According to Childs [1] the wealth of nations can be judged by their investment in machining. Modern manufacturing trends require parts to be produced quickly, and with as small a carbon footprint as possible. This is directing machining processes towards higher cutting speeds, lower waste and improved part quality, making it necessary to use coolant. Having adequate cooling and lubrication during metal cutting is essential to reduce thermal shock and stresses generated during machining. These factors have a substantial impact on the tool life, quality, and the power consumed. The customary end milling process uses copious amounts of liquid coolant [2], with the liquid coolant being used to increase the tool life and to improve the workpiece surface finish.

Even with the recognition of the aforementioned benefits liquid coolant use needs to be reduced as it supports the growth of micro-organisms [3] such as aerobic bacteria, anaerobic bacteria and fungi which are the most notable microbial group [4]. All metal cutting operations will benefit from the development of a more environmentally tool cooling method, as ecological and health costs can be contributed to the cutting fluid. Obviously the health and safety aspects of using cutting fluids adds to the cost of metal cutting as suitable disposal of the cutting fluid is needed [5]. A number of alternative cooling methods have been trialed to help reduce the amount of liquid coolant used, while having no adverse effect on the machining performance [6]. One such method is called Minimal Quantities of Lubrication (MQL) [7]. This is where an extremely small amount of lubricant is blasted by air into the cutting zone.

Previous research has shown that MQL has been effective in prolonging tool life when machining steel workpieces [8, 9]. A research paper by Rahman [10] examined the design of a new cooling system which used liquid mist and air in end milling. Additional research for machining steel conducted by Rahman [11] has shown MQL to be compatible to that of flood coolant for cutting conditions within the following range: cutting speed 75–125 m/min, feed rate 0.01–0.03 mm/tooth, and a depth of cut of 0.35–0.7 mm. Figure 1 shows the flank wear and surface roughness recorded for a feed rate of 0.015 mm/tooth, and depth of cut of 0.35 mm.

When contrasted with the production cutting conditions it was found that the flank wear had increased as expected, and that MQL was not as effective due to the higher cutting temperatures. As yet there has not been the same rigorous research into the machining of Al alloy. One paper has investigated the effect of using MQL on tool wear, chip morphology and surface finish produced during machining A356 Al alloy at high cutting speed. It was found that the use of MQL to replace flood coolant in high speed machining of Al alloy had been demonstrated to be successful [12].

However, there were some technical issues still to be resolved such as tool wear and machine reliability. The machining parameters selected for this research were typically those which are suitable for most CNC milling machines. A hypo-eutectic grade of Al-Si alloy (6061) was selected as it has medium to high strength properties, and has a machinability rating of 1.9 [2]. The machinability of Al alloys primarily

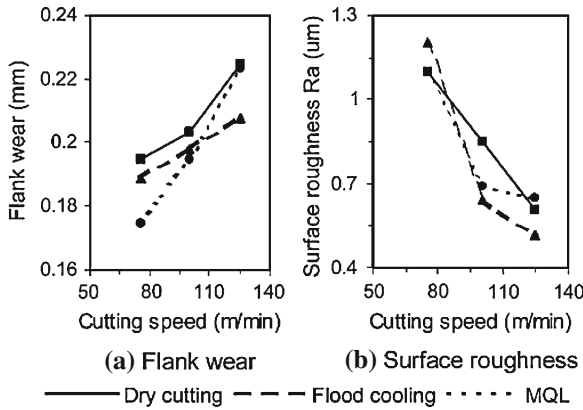


Fig. 1 Effect of cutting speed on tool wear and surface roughness by Rahman [10]

depends on the Si content of Al-Si alloy (Si content between 0.4 and 0.8 %). Unlike most other milling applications, cutting fluid should always be used when machining Al to avoid smears and to improve the surface finish. The dominant wear criteria when machining Al alloy is built-up-edge, burr formation and poor surface finish. Burr formation or surface finish is best to use as tool life criteria as it is difficult to observe wear on the tool tip when machining Al alloy.

This research endeavours to show the effectiveness and suitability of MQL to prolong tool life during the end milling of Al alloy. The power contribution proportioned to the cooling system is a revised and extended version of our previous work [13]. The Taguchi method [14] was used to strategize the experimental procedure and optimized the experimental machining parameters used in the tests.

2 Cutting Tests and Set-Up

The metal cutting tests consisted of a Leadwell vertical machining center (V-30), a Kistler three component dynamometer (Type 9257BA) and a Yokogawa CW140 clamp on power analyser. An Airtx vortex tube (Model 20008) with an inlet pressure of 85 psi supplied chilled air at a temperature of -5°C . The compressed air used was supplied from the workshop airline. The MQL was delivered from a Uni-max cutting tool lubrication system which distributed atomised coolube metalwork lubricant to the cutting zone. This system operates on the same principle as a Serv-O-Spray allowing the lubricant to be sprayed from a single air source, which allows adjustment to the amount of lubricant delivered to the cutting zone. The proper selection of cutting fluid is often neglected in machining practice as many cutting fluids which are suitable for use with ferrous materials are not suitable for machining Al alloys. A traditional emulsified cutting fluid (Cocol ultra cut) was used for the wet test

machining as this is suitable for Al. The cutting tool selected for all of the tests was a Sandvik single tip tool (R390-012A16-11L) with a coated tungsten carbide insert (R390-11 T3 08E-NL H13A). All cooling nozzles used during the tests were kept at approximately 25 mm from the tool during all tests. The combined cold air and MQL nozzle was able to be placed closer to the cutting zone. A single tooth cutter was selected to avoid the influence of tool run out on the flank wear, and to simplify the analysis of the tests. The workpiece was clamped onto the dynamometer that in turn was secured onto the machine table of the vertical milling center. Cutting forces were then recorded onto the computer's hard disk for later analysis. Figure 2 shows the cutting test set-up.

Cutting tests were carried out using five conditions; dry, flood, cooled air, MQL and combine cooled air with MQL. Typical machining practices were used to machine the face ensuring that the tool tip was constantly removing 70 % of the material along the tool path. The cutting forces and power were measured for each face machined. The cutting conditions used were selected to reflect typical working conditions as shown in Table 1.

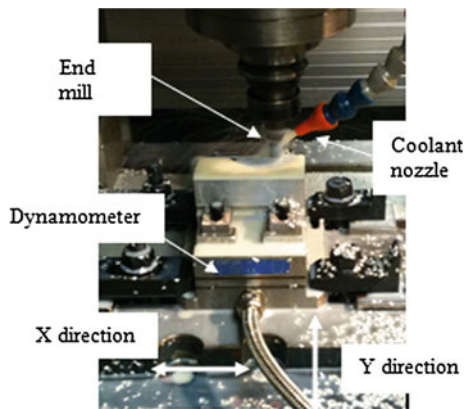


Fig. 2 Force measuring set-up

Table 1 Cutting test settings

Cutting speed (m/min)	DOC (mm)	Feed rate (mm/min)	Cooling parameter
135	3	400	Minimal quantities of lubrication
150	3	400	Cooled air
165	3	400	Cooled air combined with minimal quantities of lubrication

Dry and flood workpieces were also end milled to provide the appropriate comparison of these two extreme conditions. In this research the tool failure criteria applied to the cutting edges were:

1. Part surface quality and burr formation.
2. Flank wear V_B greater than 0.3 or maximum notch wear of 1.0 mm.
3. Dramatic change in tool forces and cutting power.

All tool tips were examined for wear after the machining of each test sample by using a tool maker microscope. The surface roughness of the workpieces was measured by a Mitutoyo portable stylus type surface roughness tester.

3 Results and Discussions

It is necessary during machining to adopt sustainable machining, clearly demonstrated when considering the CO₂ emissions in the atmosphere. Peer-reviewed estimates of the social cost of carbon (net economic cost of damage from climate change aggregated across the globe) for 2005 have an average value of US\$12 per tonne of CO₂. The range of published evidence indicates that the net damage costs of climate change are projected to be significant and to increase over time [15]. It has been established for a number of years that the cost of traditional coolant can contribute up to 17 % of the cost of producing the part due to disposal costs of the coolant. Now parts being produced by machining will need to take into account the cost of carbon. Clearly manufacturers have the additional challenge of being environmentally friendly in their production, while still being cost effective. The only machining parameter that can be reduced is the use of liquid coolant when machining Al alloy. This is of a great concern to the manufacturing industry, as the economy of the machining process largely depends on selecting the best machining conditions for the machine. The best cooling practices are used in an effort to reduce the total amount of greenhouse gas produced during machining. For example the functional parameter of tool life may be defined as the time when the generated part surface no longer meets the quality requirement.

The total production energy will consist of power needed for all auxiliary equipment such as cutting fluid, chip handling tool changers, computers and machine lubrication systems. The actual cutting energy required for manufacturing is at most 14.8 % of the total energy required for machining parts, and the other 85.2 % is continually being used by the auxiliary equipment [16]. For today's standalone machine tools the auxiliary equipment also requires substantial power up to 40 % in many cases. The actual cutting power varies depending on the cutting requirements, whereas the auxiliary power is required for the entire time the machine tool is powered on. This represents non-producing power input. Unfortunately, there is no option other than to use auxiliary equipment if companies want to remain competitive. However, Toyota found that coolant contributed 31.8 % to the auxiliary energy usage of their machining centre system. This identifies that the only sustainable alternative

option is to use a different cooling systems to reduce the energy used. As the other auxiliary energy uses do not have as much promise in reducing the energy requirements at this time. Reducing the reliance on traditional coolant makes the process more sustainable as waste disposal measures are eliminated.

Although Al alloys are some of the most machinable of the common materials used there are still some machining issues to be aware of. The low melting point of the material and having one of the highest coefficients of expansion along with relative softness and elasticity makes it necessary to dissipate the generated heat. Otherwise, it is difficult to maintain tolerances of the workpiece. Al alloys normally have significant amounts of Si causing them to be adhesive, promoting rapid heat generation resulting in chip welding and built-up-edge. Rapid machining of Al alloy allows the development of a hard Al oxide film on new exposed surfaces. This oxide film produces galling and smearing rather than good chip formation causing rapid degeneration of the cutting edge. All tested workpieces generate large amounts of data; a typical output from the dynamometer for dry machining is shown in Fig. 3. The dynamometer was used to measure the forces acting on the tool tip for each new machined face of the workpiece. Examining this output helps identify areas of interest during the cutting process. A more detailed analysis of two revolutions of the cutter is possible as shown in Fig. 4 which examines X direction forces (the milling machine table moving left or right), and the Y direction forces (the milling machine table moving to the back, or moving to the front) as shown in Fig. 2. The cutting force shown in Fig. 5 combines the X and Y forces at the same 400 s point.

The output from the dynamometer allows changes in the cutting conditions to be observed, with a more detailed cutting analysis possible when the sample time is reduced. However, this would restrict the analysis of the cooling process used during machining of the workpiece face. A detailed analysis of the cutting engagement of the tool tip can clearly be seen over two revolutions of the tool in Fig. 4; both the X and Y directional forces are given. Examination of the two cutting cycles for MQL

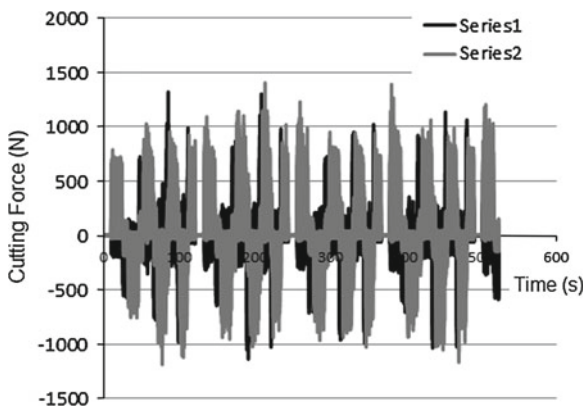


Fig. 3 Cutting forces for dry machining

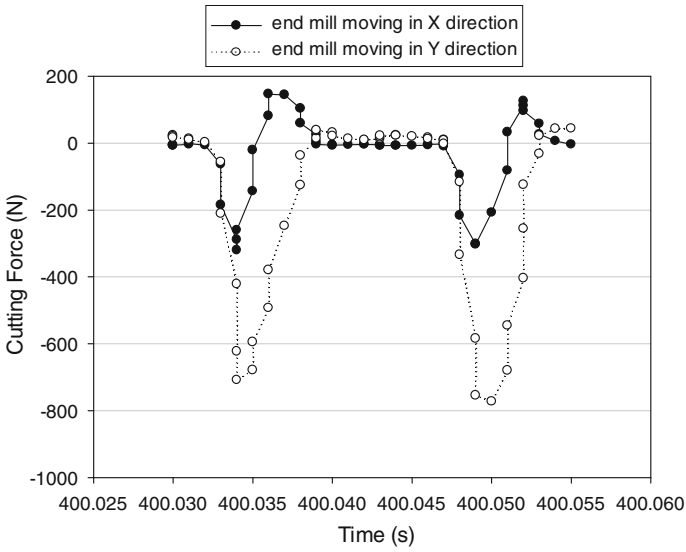


Fig. 4 Dry X and Y forces at 150 m/s

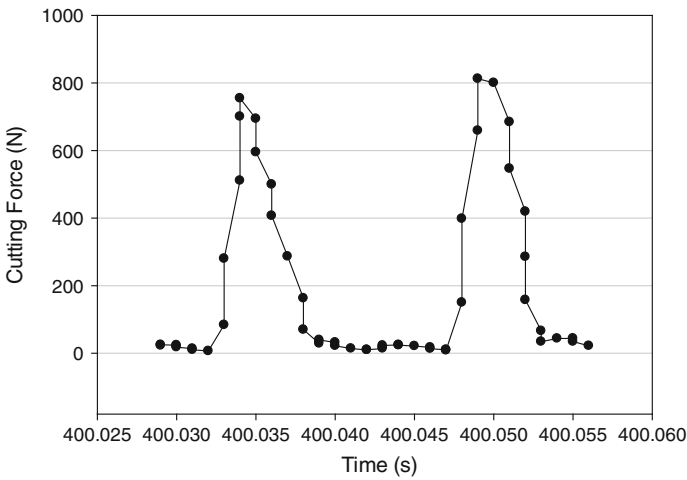


Fig. 5 Dry cutting force at 150 m/s

cooling in Fig. 6, and dry cutting in Fig. 5 shows a large reduction in cutting forces for MQL.

Figure 7 shows the highest cutting force and power for each cooling parameter including dry and flood cutting for reference purposes. Using the cutting power to determine the cutting performance is found to be difficult for Al alloys, as there is only a slight power change with respect to wear on the tool. Also, the power used

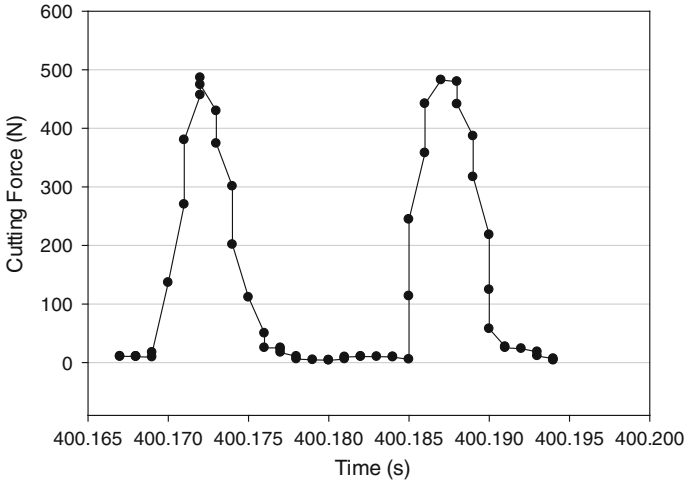


Fig. 6 Maximum cutting force during MQL at 150 m/s

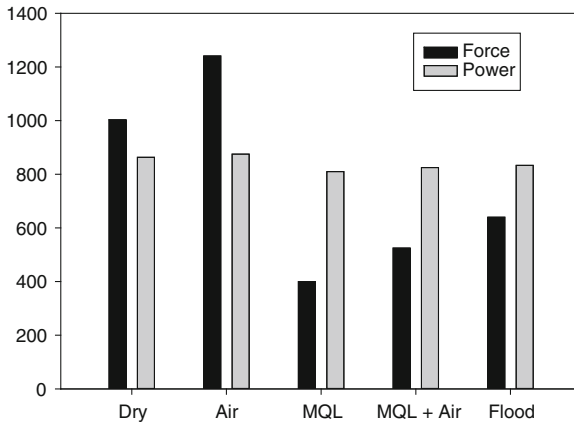


Fig. 7 Cutting forces and power

for the different cooling conditions was relevantly constant making it unsuitable to determine the best condition. The cutting forces were used therefore as one of the conditions in determining the effectiveness of the cooling process.

As expected when the cutting speed increased the cutting force reduced, which was observed for all the cooling methods. Figure 8 shows this for MQL and MQL + air.

The second method used to determine the effectiveness of the cooling process is by examining the surface finish (Fig. 9) as anomalies on the surface are quite apparent.

The surface finish data in Fig. 10 showed that MQL produced comparable surface finish as flood end milling even at higher cutting speed. However, the workpiece retained much of the generated heat.

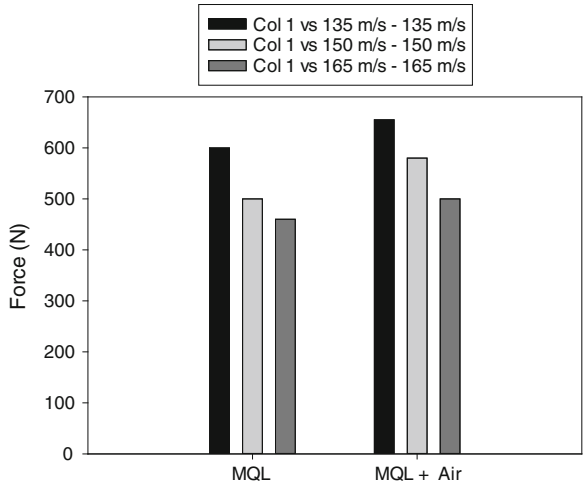


Fig. 8 Cutting forces during MQL and Air

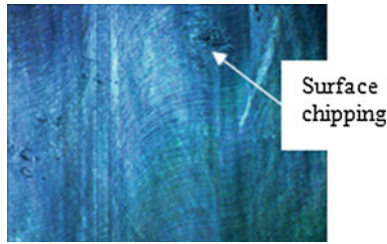


Fig. 9 Surface damage indicating tool tip deterioration

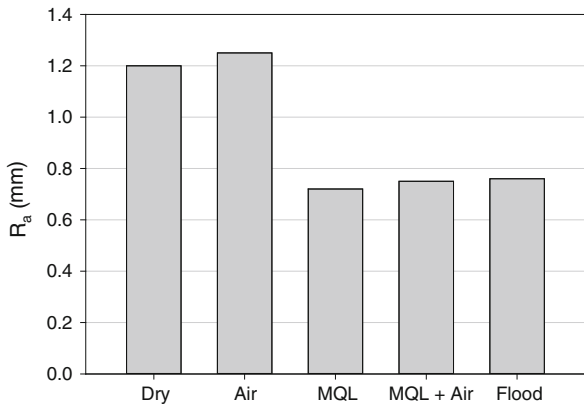


Fig. 10 Surface roughness at 150 m/min

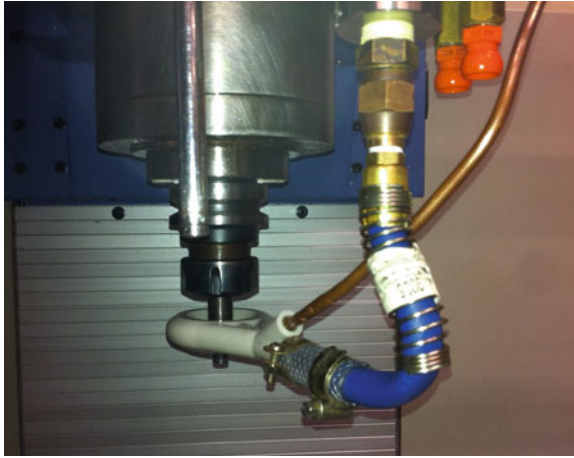
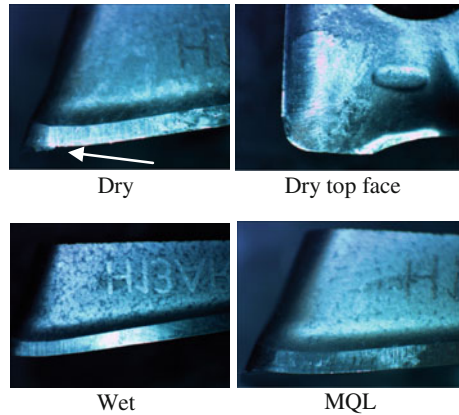


Fig. 11 Combined cold air and MQL nozzle

The challenge for MQL when machining Al alloy is to dissipate the heat from the workpiece to ensure dimensional integrity. Combining air cooling with MQL seemed to be the obvious answer to keep the workpiece cool while achieving good surface finish workpieces. The cold air cooling test for Al alloy did not achieve the same improvements at the tool tip as was obtained when machining steel. The temperature at the tool tip was too low to induce thermal cracking reducing tool life. Previous research conducted by Kelly [17] and Diakodimitris [18] determined that the alignment of the nozzle in relation to the tool can optimise the tool life when MQL is machining Al alloy. To facilitate this new cooling nozzles were designed incorporating cold air and MQL. A number of designs were tested with the most efficient design being used in this research (Fig. 11). To reduce the workpiece temperature the tool tip was constantly surrounded with cold air with the addition of a small quantity of vegetable oil. The mist levels produced by the nozzle are important for two reasons. First, the effectiveness of the machining operations is dependent on both the concentration of mist that reaches the cutting zone, and the oil droplet size of the mist. The design of the nozzle also needs to assist in reducing airborne emissions of the oil mist as it is directed at the cutting zone [19]. This is important for occupation, health and safety purposes.

Generally the most inclusive cooling method is shown by the most reduction in tool wear. However, machining Al alloy wear is difficult to determine, even with the use of a microscope, as depicted by the tool tip pictures in Fig. 12, where only the dry tool tip shows the start of a built up edge at the arrow.

Fig. 12 Tool tip wear



4 Conclusions and Future Work

The demand for environmentally sustainable manufacturing is the primary drive to discover a technology offering a solution that reduces the use of liquid coolant. However, determining the effectiveness of the cooling parameters cannot be judged simply by considering one function only. Metal cutting is a very complex system, and a small change in cutting conditions can have major consequences. To determine the best cooling method it was necessary to consider a number of factors.

- Did the cooling method increase tool life?
- Had the workpiece met the tolerance?
- Was the surface finish appropriate?
- Was the method more sustainable than traditional wet coolant?

It was shown in Fig. 7 that MQL had the lowest cutting force followed by MQL + cooled air, indicating an efficient machining performance. In addition MQL and MQL + cooled air achieved surface finishes as compatible to that of flood coolant. These results confirm that MQL + air cooling met two of the criteria needed to be considered as equivalent as or better than traditional flood cooling. These goals were all achieved by combining air cooling + MQL in a suitable redesigned nozzle. Although air cooling with the use of a small amount of vegetable oil is not a totally dry process it is quite close and therefore is sustainable.

The results have shown that cold air + MQL can be used for end milling with normal production cutting speeds, feed rates and depths of cut. It must be observed that the operational use of the nozzle may be considered cumbersome and for this reason machine operators may not like employing this cooling system in practice. Further work is necessary to examine how the nozzle can be made to be ‘user friendly’ for the machine operator.

References

1. Childs THC, Maekawa K, Obikawa T, Yamane Y (2000) *Metal machining—theory and applications*. Elsevier, Amsterdam
2. Drozda TJ (ed) (1976) *Machining. Tool and manufacturing engineers handbook*. Mc Graw-Hill, Dearborn
3. Soković M, Mijanović K (2001) Ecological aspects of the cutting fluids and its influence on quantifiable parameters of the cutting processes. *J Mater Process Technol* 109:181–189
4. Tant CO, Bennett EO (1958) The growth of aerobic bacteria in metal-cutting fluids. *Appl Microbiol* 6:4
5. Calvert GM, Ward E, Schnorr TM, Fine LJ (1998) Cancer risks among workers exposed to metalworking fluids: a systematic review. *Am J Ind Med* 33:282–292
6. Hwang Y-K, Lee C-M, Park S-H (2009) Evaluation of machinability according to the changes in machine tools and cooling lubrication environments and optimization of cutting conditions using Taguchi method. *Int J Precis Eng Manuf* 10:65–73
7. Weinert K, Inasaki I, Sutherland JW, Wakabayashi T (2004) Dry machining and minimum quantity lubrication. *CIRP Ann Manuf Technol* 53:511–537
8. Boswell B (2010) An experimental approach to determining the effectiveness of minimum liquid cooling for end milling 1040 steel. Presented at the 6th Australasian congress on applied mechanics, ACAM 6, Perth
9. Boubekri N (2010) A technology enabler for green machining: minimum quantity lubrication (MQL). *J Manuf Technol Manag* 21:556
10. Rahman M, Senthil Kumar A, Salam MU (2002) Experimental evaluation on the effect of minimal quantities of lubricant in milling. *Int J Mach Tools Manuf* 42:539–547
11. Rahman M, Senthil Kumar A, Manzoor UIS, Manzoor UIS (2001) Evaluation of minimal of lubricant in end milling. *Int J Adv Manuf Technol* 18:235–241
12. Kishawy HA, Dumitrescu M, Ng EG, Elbestawi MA (2005) Effect of coolant strategy on tool performance, chip morphology and surface quality during high-speed machining of A356 aluminum alloy. *Int J Mach Tools Manuf* 45:219–227
13. Boswell B, Islam MN (2012) Feasibility study of adopting minimum quantities of lubrication for end milling aluminium. In: *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012 (WCE2012)*, London, 4–6 July 2012, pp 1358–1362
14. Roy R (1990) *A primer on the Taguchi method*. Society of Manufacturing Engineers, Dearborn
15. Sutherland JW, Rivera JL, Brown KL, Law M, Hutchins MJ, Jenkins TL, Haapala KR (2008) Challenge for the manufacturing enterprise to achieve sustainable development. In: *The 41st CIRP conference on, manufacturing systems*, pp 15–18
16. Gutowski T, Murphy C, Allen D, Bauer D, Bras B, Piwonka T, Sheng P, Sutherland J, Thurston D, Wolff E (2005) Environmentally benign manufacturing: observations from Japan Europe and the United States. *J Clean Prod* 13:1–17
17. Kelly JF, Cotterell MG (2002) Minimal lubrication machining of aluminum alloys. *J Mater Process Technol* 120:327–334
18. Diakodimitris C, Hendrick P, Iskandar Y (2010, 12 Dec 2011) Study of minimum quality cooling (MQC) on the tool temperature in milling operations. http://msepr.engr.wisc.edu/phocadownload/cirp44_study%20of%20minimum%20quantity%20cooling.pdf
19. Dasch JM, Kurgin SK (2010) A characterisation of mist generated from minimum quantity lubrication (MQL) compared to wet machining. *Int J Mach Mach Mater* 7:14

Fine-Tuning Negotiation Time in Multi-Agent Manufacturing Systems

W. L. Yeung

Abstract Global market competition has put intense pressure on the manufacturing industry to become more agile and responsive to market changes. Multi-agent systems (MAS) provide a decentralised control architecture that can reduce complexity, increase flexibility, and enhance fault tolerance for manufacturing systems. Shop floor control applications can be designed based on the paradigm of agent negotiation. This often involves the contract net protocol (CNP) and previous research has suggested that the timing parameters of CNP can affect significantly the performance of agent negotiation. This chapter discusses the combinatorial variations of these parameters using a discrete-event simulation case study.

Keywords Cycle time · Multi-agent systems · Negotiation protocol · Performance · Simulation · Work-in-progress

1 Introduction

Agent-based software technology plays an important role in the manufacturing industry for achieving agility [21]. At the shop floor control level, multi-agent systems (MAS) support a heterarchical approach to the dynamic scheduling and dispatching of jobs in the presence of variabilities in manufacturing resources, production disturbances, uncertain arrival of parts, etc. [9]. Other manufacturing applications of MAS include design [1], enterprise resource planning [24], scheduling and monitoring [5], supply chain management [14] and robotics [11].

The *contract net protocol* (CNP) [19] was originally proposed for cooperative problem solving in a distributed processing environment based on a negotiation metaphor [7]. In a typical shop floor application scenario of CNP, a part agent

W. L. Yeung (✉)
Lingnan University, Hong Kong, China
e-mail: wlyeung@ln.edu.hk

announces the processing requirements of a part to all machine agents through a task announcement message. Such a message is received and checked by every machine agent against its own capabilities and capacity to decide whether to respond with a bid message. Bid messages contain information such as estimated processing time which are used the part agent to evaluate the received bids. An award message is then sent to notify the selected bidder (machine) agent. Note that the bidding process is *time-bound*: an announced task is open for bidding for only a limited period of time and if no bids at all are received, the part agent would normally re-announce the task for another round of bidding.

While the CNP approach has been applied to agent-based manufacturing control with promising results in many studies, there have been some concerns with the performance of the agent negotiation process. The performance issue of *message congestion* was initially mentioned in [19]. In [13], the issue is highlighted as a potential problem in agent-based manufacturing control systems. In [6], the CNP approach is applied to the cooperative scheduling of production and maintenance activities wherein resource agents always submit bids in response to relevant task announcements, resulting in a time consuming negotiation process. It is pointed out in [8] that distributed agent-based manufacturing control systems could exhibit chaotic behavior, raising doubts about their predictability, reliability and performance.

In our previous case study [23], we applied discrete-event simulation to the performance analysis of CNP-based negotiation processes. Whereas individual timing parameters were considered separately in previous studies, our case study considered the *combinatorial variations* of two major timing parameters, namely open-for-bidding time and commitment duration, of a negotiation process. Furthermore, we measured the impact of these variations on the number of bids received for each task announcement as an important performance indicator: the smaller the number of received bids, the less likely a task would get an optimal bid, hence undermining the overall performance of the system.

In this chapter, we provide further elaboration on the results of our case study which contribute to a deeper understanding of the time-bound behavior of a CNP-based negotiation process. In particular, this helps explain the sometimes “chaotic” behavior in which agents struggle to cope with moderate workload and offers further insights into the fine-tuning of CNP-based control schemes for better performance. We also discuss directions for further work.

2 Multi-Agent Manufacturing Systems

Multi-agent systems (MAS) have become a major paradigm for manufacturing control in a variety of applications [2, 18]. In [9] a heterarchical approach is proposed for scheduling manufacturing resources based on autonomous agents representing machines, parts, and operators. Advantages of this approach include reactivity to disturbances, reduced complexity and fault tolerance; whereas the lack of predictability,

poor ability to define optimal loadings, lack of analytical solutions and possibility of deadlock are the main disadvantages [10].

The *contract net protocol* (CNP) [19] was originally proposed for cooperative problem solving in a distributed processing environment based on a negotiation metaphor [7]. In a typical shop floor application scenario of CNP, a part agent announces the processing requirements of a part to all machine agents through a task announcement message. Such a message is received and checked by every machine agent against its own capabilities and capacity to decide whether to respond with a bid message. Bid messages contain information such as estimated processing time which are used by the part agent to evaluate the received bids. An award message is then sent to notify the selected bidder (machine) agent.

The effectiveness of using multi-agent systems in manufacturing systems and supply chain management has been extensively studied, primarily through simulation experiments. In [14], there are over a hundred studies on the use of multi-agent systems (MAS) in manufacturing systems and supply chain management. Very often, these studies involved the use of prototype systems and/or simulation models in validating the performance of the proposed MAS designs.

In [17], the performance of a CNP-based distributed scheduling scheme is compared with a centralised scheduling scheme in a simulation study with both schemes using the shortest processing time as the main criterion in the scheduling decisions. The simulation results show that the distributed scheme performs significantly better than the centralised scheme in terms of late jobs, waiting time, tardiness and mean flow-time. The CNP approach has also been applied to deal with challenging operating conditions. In [3, 4], the performance of CNP-based control is studied based on different agent internal strategies for reconfigurable manufacturing systems in handling unpredictable exceptions (e.g. machine breakdowns). Machine agents negotiate with each other in the event of breakdowns to determine the transfer of parts from broken to reconfigured machines. Their simulation results suggest that, with intelligent internal strategies based on heuristic rules and fuzzy logic, agents can negotiate dynamically in allocating part transfers to achieve better overall average part flowtime.

3 Message Congestion

Message congestion was a major concern in the original proposal of CNP for task allocation in a sensor network [19]. In [13], message congestion is highlighted as a potential problem in agent-based manufacturing control systems. In [6], CNP is applied to the cooperative scheduling of production and maintenance activities and noted that resource agents always submitted bids in response to relevant task announcements, resulting in a time consuming negotiation process. In [20], a simulation study on a CNP-based multi-agent manufacturing system is reported with a focus on the negotiation process. The results suggest that the *task evaluation time* (TET) incurred by resource (machine) agents is the main influence on the negotiation

process' performance. While the length of TET depends largely on the nature of the task description, an increase in TET leads to longer waiting time for the part agent as well as more task re-announcements. The latter increase directly contributes to message congestion. Furthermore, more time spent on task evaluation means a higher chance for resource agents to miss bidding deadlines, resulting in lower bidding rate. Similar results have been found in another simulation study [22] in which it is concluded that negotiation time affected the performance of their proposed CNP-based integrated process planning and scheduling approach.

4 Negotiation Time

In [20], the task evaluation time (TET)—time required by a resource agent to evaluate a task announcement and formulate a bid—is considered as the main parameter in a simulation study. The results suggest that longer TET demands longer open-for-bidding duration. This also means that resources agents are tied up for longer time in each round of bidding and can only afford to bid less frequently. As a result, the chance of getting bids from *all* resource agents eligible for a particular task is reduced and, the lower the chance gets, the less likely the task would get an optimal bid, hence undermining the overall system performance. At the same time, the chance of not getting any bids at all in a bidding round increases and indeed the simulation results suggest that tasks could be re-announced several times before receiving any bids.

Figure 1a illustrates a scenario in which two tasks were announced at about the same time and only one received bids; the other task did not get any bids within the bidding time limit as the two machine agents had been tied up with the first task. In the original CNP [19], bids are binding and a resource agent is allowed to bid and be awarded multiple tasks that have to be queued for processing. For manufacturing control applications, however, tasks often carry completion deadlines and hence unrestricted queued processing would not be acceptable. It follows that a resource agent has to bid within its own capacity or risk failure in honouring contract awards. In [16], the different levels (stages) of commitment and the use of time-limited commitment in CNP-based negotiation are considered.

The *commitment duration* of a bid is considered in [15] as another important timing parameter affecting the performance of a CNP-based negotiation process. Bidders are required to submit binding bids within their capacities but the validity of a bid is limited to a certain commitment duration. If an award is received before the end of the commitment duration, the bidder is bound to it; if no award is received by the end of the commitment duration, the bidder's obligation to accept an award is relieved. The results of a simulation experiment suggest that the length of bid commitment duration affects the outcomes of the negotiation process [15]. Specifically, extending the commitment duration lowers the risk of "premature" expiry of bids (Fig. 1b illustrates such a risk), thereby increasing the success rate of negotiation.

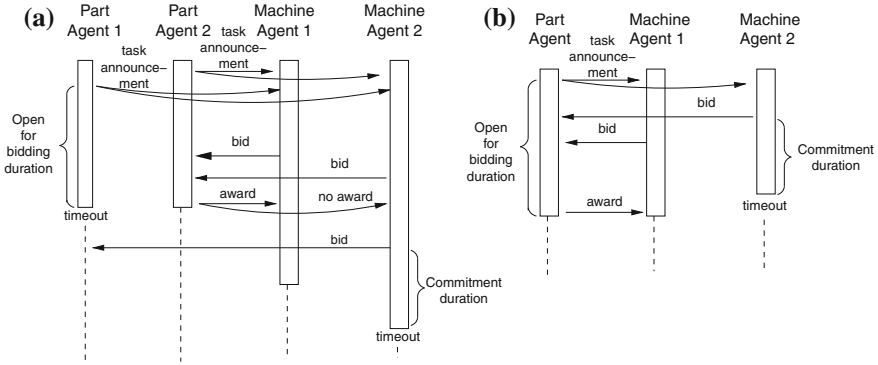


Fig. 1 Negotiation scenarios

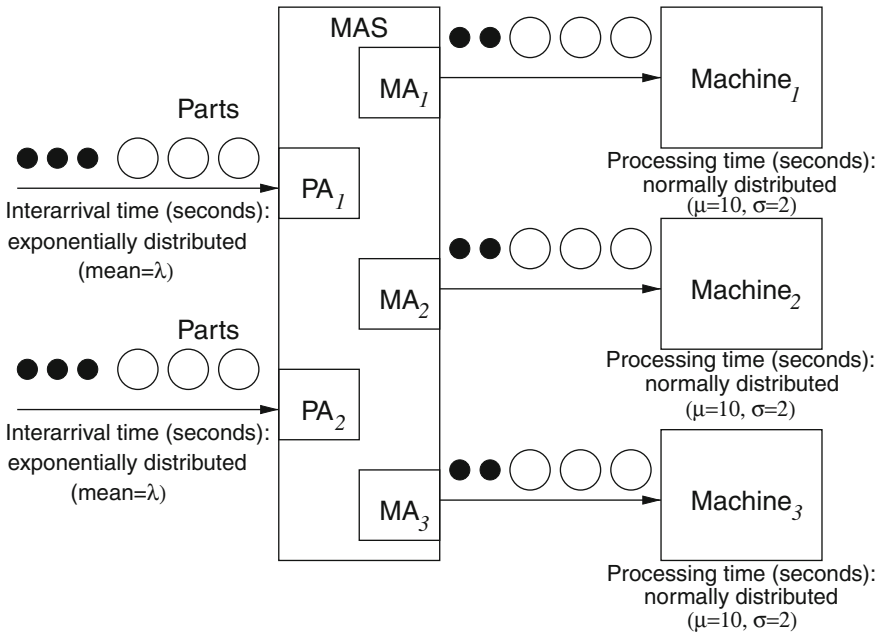


Fig. 2 Configuration of the multi-agent manufacturing system

5 Simulation Case Study

We present in this section a simulation case study of a hypothetical flexible manufacturing system with an aim to obtain a deeper understanding of the performance of timed-bound negotiation as discussed in the preceding section. Figure 2 depicts the configuration of the system. Parts arrive at the system from two separate sources and the arrival rate at each source is exponentially distributed with an mean inter-arrival

time of λ seconds. There are three machines labelled as Machine₁, Machine₂ and Machine₃ which can process any parts from either one of the sources. The amount of processing time (in seconds) for a part is estimated to be normally distributed with a mean and a standard deviation of 10 and 2. A multi-agent system is responsible for allocating parts to the machines dynamically. It involves five software agents: PA₁ and PA₂ represent parts from the two respective sources whereas MA₁, MA₂ and MA₃ represent the three machines, respectively. Details of the negotiation protocol are presented in the following subsection, followed by the methods and results of the simulation experiment.

5.1 Negotiation Protocol

The negotiation protocol in this case study is based on the contract net protocol. Each part agent interacts independently with all three machine agents via messages. When a part arrives, the responsible part agent will send out a task announcement message to all machine agents with details about the type as well as some physical characteristics of the part. If a machine is capable of processing the announced part, the responsible machine agent will respond by sending back a bid message containing the estimated amount of processing time; otherwise, it simply ignores the current announcement.

We assume that it takes a non-negligible amount of time for the machine agent to compute and formulate a bid (i.e. the task evaluation time) and the amount, x in seconds, is estimated to be distributed (triangularly) as follows:

$$f(x) = \begin{cases} \frac{2(x-a)}{(m-a)(b-a)} & \text{for } a \leq x \leq m \\ \frac{2(b-x)}{(b-m)(b-a)} & \text{for } m \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $a = 3$ (minimum), $m = 5$ (mode) and $b = 9$ (maximum).

Part agents are programmed to accept bids (i.e. open for bidding) for a fixed period of time which we call the *open for bidding duration* (OBD) before making an allocation decision based on the received bids. Part agents simply select machines with the *shortest estimated processing time*. Bid selection results are announced immediately to bidders and parts are queued for processing according to the results.

However, if a part agent does not receive any bids during the OBD time, it will repeat the cycle of announcement and bidding for the part until an award is made. Then, and only then, the part agent will begin processing the next part in waiting. On the other hand, task announcement messages are also processed one at a time at each machine agent. Furthermore, we assume that *machine agents will keep at most one task announcement message in their input buffers*; incoming messages beyond the buffer limit are simply ignored and dropped. Finally, if a submitted bid misses

the OBD deadline, the bidder will not receive any bid selection results. Figure 1b illustrates such a case. As a safeguard against indefinite waiting, machine agents are programmed to wait for only a certain duration, i.e. the *commitment duration* (CD).

5.2 Methods

We followed the approach of [3] in developing our simulation test-bed on the Arena discrete event simulation platform by Rockwell Software, Inc. Arena is not only well suited for modelling shop floors of production systems [12], but also suitable for modelling the workflow behaviour of multi-agent systems.

For the experimental design and steady-state statistical analysis of simulation output, we adopted the truncated-replication strategy as discussed in [12]. The simulation model covers the timing of both the agent negotiation process as well as the physical machine processes. However, it does not account for any networking overheads or latencies in the transmission of messages; they are assumed as negligible in the case study.

We conducted simulation runs using different combinations of OBD and CD values (OBD = 5, 7.5, 10.0, 12.5 s and CD = 10, 12.5, 15, 17.5, 20 s). Each simulation run had a run length of 150h and was initially replicated for 10 times. We carried trial runs with various mean part inter-arrival times and settled with a range of mean part inter-arrival times ($\lambda = 100, 150, 200$ s) that maintain the system in a steady state. Then, by plotting the system's WIP against time during each run (in Arena's Output Analyzer), we identified a suitable warm-up period of 30h for all the simulation runs included in our analysis.

The performance of the system was examined by measuring the average work-in-progress (WIP) and average cycle time of parts. Furthermore, we measure the number of parts receiving bids from all three machine agents. Whenever necessary, the number of replications in a simulation run was increased to ensure that the 95 % confidence interval half-width of each performance measure was no more than 2.5 % of its average.

5.3 Results

Figure 3 shows the average work-in-progress (WIP) under various combinations of open-for-bidding (OBD) and commitment durations (CD). An OBD of 7.5 s represents the optimal choice in our simulation experiment. Longer OBDs (10 s & 12.5 s) increase the minimum time required for negotiating a task and hence keep the part a little bit longer in the system. With an OBD = 5 s, machine agents tend to miss bidding deadlines more likely, given that it takes 3–9 s with mode = 5 s for a machine agent to come up with a bid. This increases the chance of having to reannounce a task (upon all machine agents missing the bidding deadline) and contributes to a higher

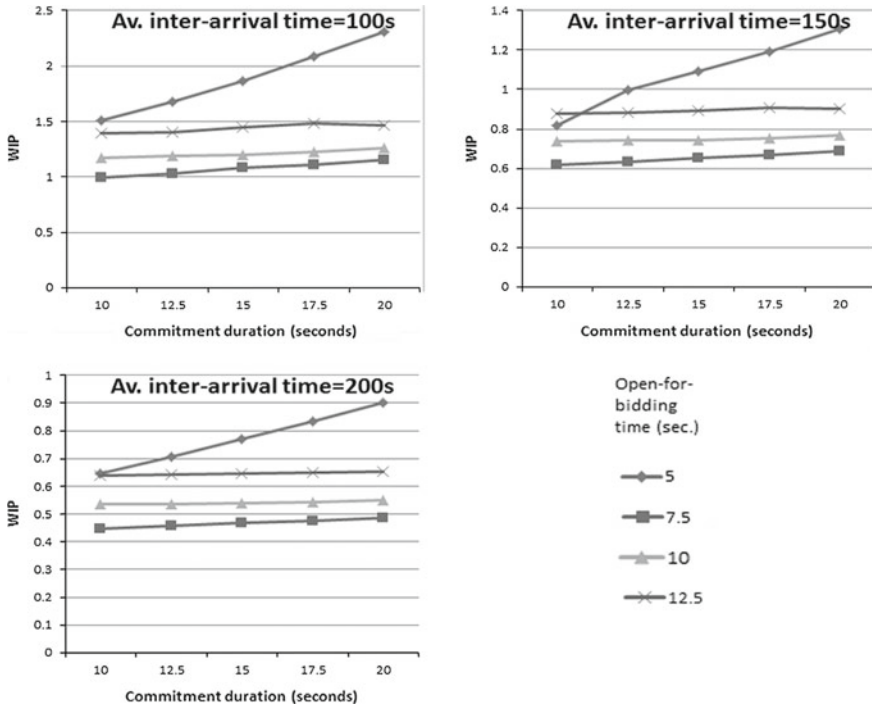


Fig. 3 Simulation results: WIP

average WIP level. Furthermore, when bidding deadlines are missed, machine agents wait for the whole length of CD before giving up and, the longer the waiting time, the less frequently they can bid again, driving up further the chance of re-announcements and hence the level of WIP. This explains the impact of CD on WIP in the case of $OBD = 5$ s in our results.

Some simulation runs, however, failed to complete due to a limit imposed by the (academic version of) Arena software on the number of active entities. When this limit is exceeded, the simulation run stops with an error. Using the animation feature of Arena, we found long queues of parts waiting for the two part agents to handle when the simulation run terminated abnormally, while both physical machines were idle. This suggests the presence of bottlenecks in the multi-agent system.

The above situations tend to occur when OBD exceeds CD by at least 5 s. Since 5 s is the mode time for a machine agent to formulate and submit a bid, the CD times set in the above situations are simply too short, allowing many bids to expire prematurely as in the case shown in Fig. 1b. It follows that part agents can frequently fail to get any bids for their announced tasks and are forced to reannounce them. This effectively increases the loading on part agents. In some cases, they fail to keep up with the building up of their input queues and the simulation runs terminated abnormally; in other cases, parts get re-announced repeatedly so many times before

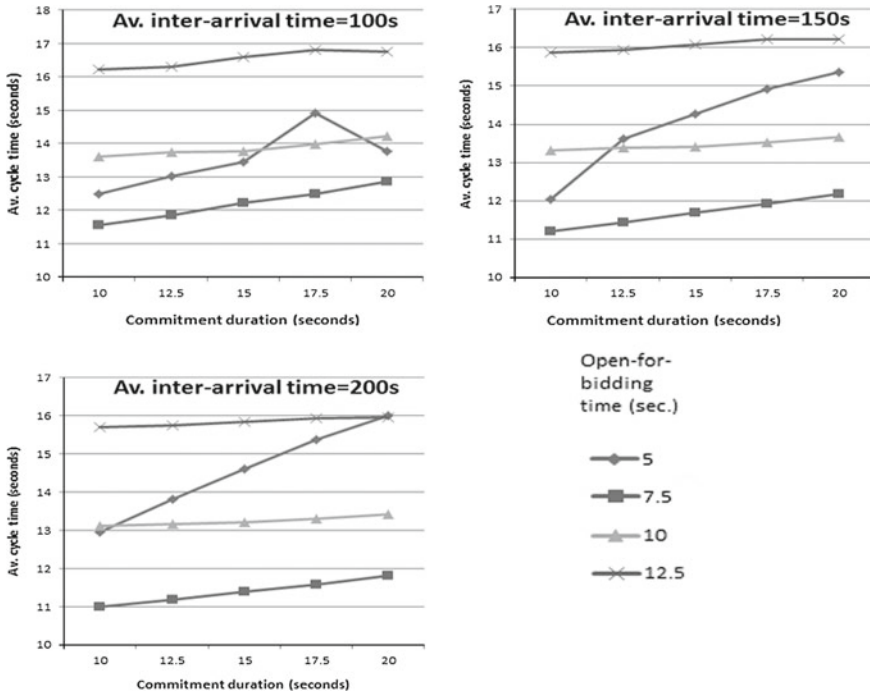


Fig. 4 Simulation results: Average part cycle time (in seconds)

receiving any bids, resulting in very long cycle time for them. Table 1 highlights these long (maximum) cycle time cases.

In terms of average cycle time (see Fig. 4), an OBD of 7.5 s again represents the optimal choice in our simulation experiment. In general, a longer OBD time lengthens the cycle time of a part and hence a OBD of 12.5 s tends to get the highest cycle time averages in the experiment results. On the other hand, when OBD is so short (OBD = 5 s) that machine agents easily miss the bidding deadlines as explained above, average cycle time also suffers from more task re-announcements and longer commitment durations.

The exceptions are that when OBD exceeds CD by 5 s or more, many bids expire prematurely as discussed above, resulting in very long cycle times for certain tasks. Putting aside these exceptions, we can see that CD time affects cycle time significantly only in the cases of OBD = 5 s. In these cases, OBD is so short that machine agents often miss the bidding deadlines, driving up by several times the number of task re-announcements (see Fig. 5) as well as the maximum cycle time (see Table 1); worse still, when deadlines are missed, machine agents wait for the whole length of CD before giving up and, the longer the waiting time, the less frequently they can bid again, driving up further the chance of re-announcements and hence the average cycle time of parts.

Table 1 Simulation results: Maximum part cycle time (in minutes)

Parts inter-arrival time (sec.)	Open-for-bidding duration (OBD) = 5s							Open-for-bidding duration (OBD) = 7.5s						
	Commitment duration (CD) (in sec.) =							Commitment duration (CD) (in sec.) =						
	5	7.5	10	12.5	15	17.5	20	5	7.5	10	12.5	15	17.5	20
100	3.98	1.72	1.82	2.15	2.53	2.83	3.25	1.31	1.43	1.57	1.81	1.90	1.78	2.34
150	6.89	1.34	1.46	1.58	1.91	2.01	2.13	1.23	2.46	1.21	1.33	1.90	1.49	1.74
200	5.08	1.16	1.23	1.37	1.45	1.53	1.79	0.96	0.96	0.96	1.26	1.29	1.24	1.50
Parts inter-arrival time (sec.)	Open-for-bidding duration (OBD) = 10s							Open-for-bidding duration (OBD) = 12.5s						
	Commitment duration (CD) (in sec.) =							Commitment duration (CD) (in sec.) =						
	5	7.5	10	12.5	15	17.5	20	5	7.5	10	12.5	15	17.5	20
100	49.93	1.61	1.69	2.35	2.01	2.28	2.06	N/A	12.30	1.88	2.10	2.58	2.96	2.85
150	95.54	1.41	1.36	1.51	1.70	1.61	2.03	N/A	6.69	1.54	1.55	1.97	1.82	1.88
200	N/A	1.24	1.34	1.39	1.23	1.34	1.63	N/A	3.93	1.29	1.33	1.64	1.62	1.51

Figure 6 compares the average counts of tasks receiving three bids under the different combinations of OBD and CD times. The comparison reveals that longer OBD time (10s & 12.5s) has the advantage of allowing more tasks to receive bids from all machine agents and potentially improving the optimality of bid selection. This may be considered as an offset against the relatively longer (average) cycle time. On the other hand, except when CD times are set too short as discussed above, lengthening the commitment duration does not appear to increase the three-bid counts.

On the whole, OBD plays the more decisive role in our experiment on the performance of the agent negotiation process and this includes WIP, average cycle time and the number of bids received for each task.

6 Summary and Discussion

Our simulation study examines the performance of time-bound negotiation based on CNP under different combinations of open-for-bidding and commitment durations in a manufacturing control application. The results show that there are trade-offs between, on one hand, the efficiency in terms of WIP and cycle time, and on the other hand, the integrity of the negotiation process—longer OBD time improves integrity by allowing more tasks to get the maximum number of bids, whereas cutting the OBD time tends to speed up the negotiation process at the expense of integrity. Furthermore, cutting the OBD time beyond a certain point would actually hinder the negotiation process by disrupting its “rhythm” and, in these cases, extending the commitment duration would further worsen the negotiation performance.

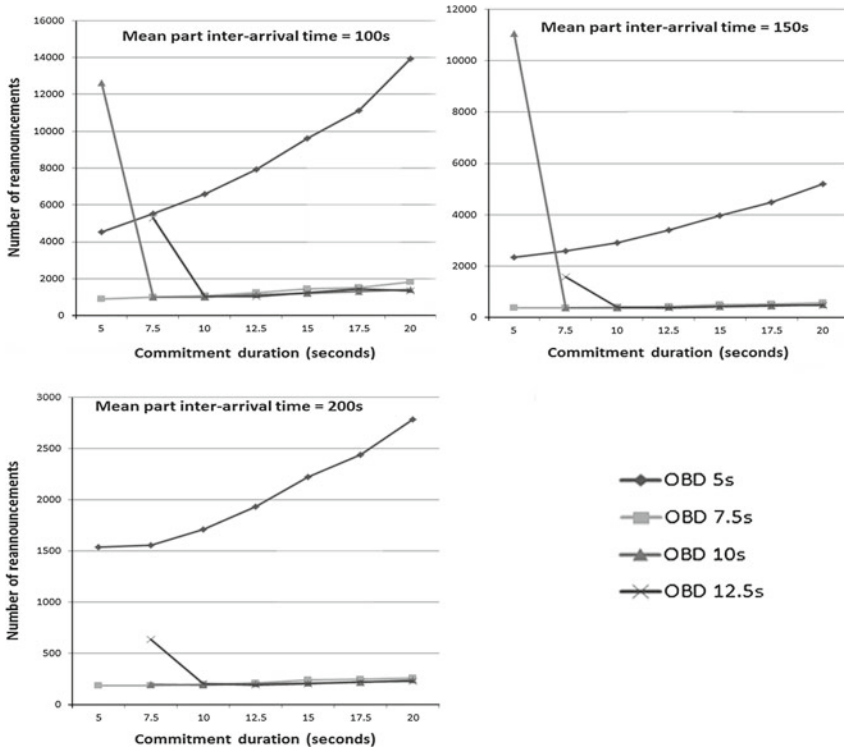


Fig. 5 Simulation results: Number of task re-announcements per simulation run

Our experiment also helps explain the seemingly “chaotic” behavior of the negotiation in some cases where the cycle time of some parts lasts nearly fifty times above the average. This tends to happen when the commitment duration is too short and machine agents tend to de-commit their bids prematurely before the bidding results are due. Our results show that simply lengthening the commitment duration solves the problem at the expense of only slight increase in cycle time and the number of re-announcements. On the other hand, contrary to our intuition, lightening the system load (alone) can actually aggravate the problem. This was the case in the experiment when OBD and CD were set to 10s and 5s, respectively, and the average and maximum cycle times actually went up when the arrival rate of parts decreased (see Fig. 4 and Table 1).

Message congestion is another potential issue affecting negotiation performance. From the point of view of our experiment, it is a by-product of the inappropriate setting of timing parameters. In such cases, tasks are re-announced repeatedly many times until they are allocated. We have assumed that machine agents keep at most one task announcement message in their input buffers and simply drop any incoming messages beyond the buffer limit. The limit is put in place to prevent machine agents from being overwhelmed by task re-announcement messages and, without it, machine

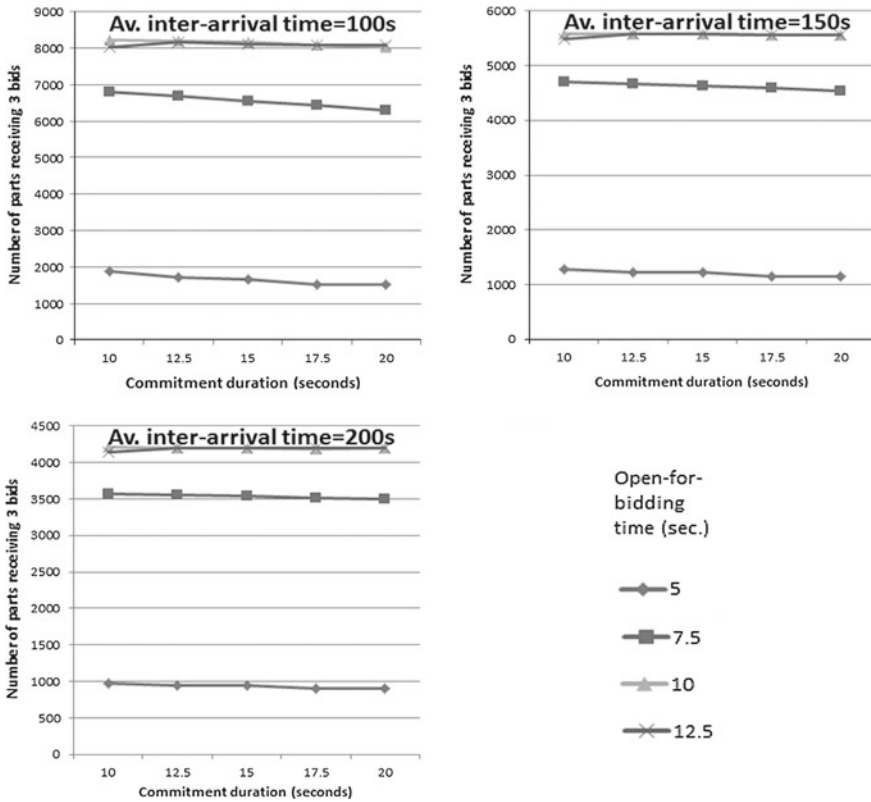


Fig. 6 Simulation results: Average number of parts receiving three bids

agents would find themselves handling such messages long after the relevant tasks have been allocated.

7 Conclusion and Further Work

The results of our simulation case study show that the agent negotiation process requires the judicious setting of its timing parameters in order to run smoothly and efficiently. To extend our work further, variations of other timing parameters (e.g. machine setup time) can be incorporated into the experiment so as to study their performance implications.

Apart from fine-tuning the negotiation timing parameters, restricting the “audience” of task announcement messages can be an effective means of reducing message congestion, although it can also impact on the optimality of task allocation. With

judicious and dynamic audience restriction (selection), agents could, over time, self-organise themselves into clusters that handle certain task types more efficiently.

References

1. Aldea A, Banares-Alcantara R, Jimenez L, Moreno A, Martinez J, Riano D (2004) The scope of application of multi-agent systems in the process industry: three case studies. *Expert Syst Appl* 26(1):39–47
2. Baker AD (1998) A survey of factory control algorithms that can be implemented in a multi-agent heterarchy: dispatching, scheduling, and pull. *J Manuf Syst* 17(4):297–320
3. Bruccoleri M, Renna P, Perrone G (2005) Reconfiguration: a key to handle exceptions and performance deteriorations in manufacturing operations. *Int J Prod Res* 43(19):4125–4145
4. Manfredi Bruccoleri, Michele Amico, Giovanni Perrone (2003) Distributed intelligent control of exceptions in reconfigurable manufacturing systems. *Int J Prod Res* 41(7):1393–1412
5. Caridi M, Cavalieri S (2004) Multi-agent systems in production planning and control: an overview. *Prod Plan Control* 15(2):106–118
6. Coudert T, Grabot B, Archimede B (2002) Production/maintenance cooperative scheduling using multi-agents and fuzzy logic. *Int J Prod Res* 40(18):4611–4632
7. Davis R, Smith RG (1983) Negotiation as a metaphor for distributed problem solving. *Artif Intel* 20:63–109
8. Duffie NA, Prabhu VV, Kaltjob PO (2002) Closed-loop real-time cooperative decision-making dynamics in heterarchical manufacturing systems. *J Manuf Syst* 21(6):409–418
9. Duffie NA, Piper RS (1986) Nonhierarchical control of manufacturing systems. *J Manuf Syst* 5(2):137–139
10. Frayret JM, D'Amours S, Montreuil B (2004) Coordination and control in distributed and agent-based manufacturing systems. *Prod Plan Control* 15(1):42–54
11. Gorbenko A, Mornev M, Popov V (2011) Planning a typical working day for indoor service robots. *IAENG Int J Comput Sci* 38(3):176–182
12. Kelton WD, Sadowski RP, Sturrock DT (2007) *Simulation with Arena*, 4th edn. McGraw Hill, New York
13. Krothapalli NKC, Deshmukh AV (1999) Design of negotiation protocols for multi-agent manufacturing systems. *Int J Prod Res* 37(7):1601–1624
14. Lee J-H, Kim C-O (2008) Multi-agent systems applications in manufacturing systems and supply chain management: a review paper. *Int J Prod Res* 46(1):233–265
15. Lee KJ, Chang YS, Lee JK (2000) Time-bound negotiation framework for electronic commerce agents. *Dec Supp Syst* 28(4):319–331
16. Sandholm T, Lesser V (1995) Issues in automated negotiation and electronic commerce: extending the contract net framework. In: Lesser VR, Conry S, Demazeau Y, Tokoro M (eds) *Proceedings of first international conference on multi-agent systems*. AAAI Press, Menlo Park, California pp 328–335
17. Shaw MJ (1987) A distributed scheduling method for computer integrated manufacturing: the use of local area networks in cellular systems. *Int J Prod Res* 25(9):1285–1303
18. Shen W, Norrie DH (1999) Agent-based systems for intelligent manufacturing: a state-of-the-art survey. *Knowl Inf Syst* 1(2):129–156
19. Smith RG (1980) The contract net protocol: high-level communication and control in a distributed problem solver. *IEEE Trans Comput* 29:1104–1113
20. Tilley KJ (1996) Machining task allocation in discrete manufacturing systems. In: Clearwater SH (ed) *Market-based control: a paradigm for distributed resource allocation*, Chap 9. World Scientific Publishing, River Edge, pp 224–252
21. Van Dyke Parunak H (1996) Applications of distributed artificial intelligence in industry. In: O'Hare GMP, Jennings NR (eds) *Foundations of distributed artificial intelligence*, Chap 4. Wiley, New York, pp 139–164

22. Wong TN, Leung CW, Mak KL, Fung RYK (2006) An agent-based negotiation approach to integrate process planning and scheduling. *Int J Prod Res* 44(7):1331–1351
23. Yeung WL (2012) Performance of time-bound negotiation in agent-based manufacturing control. In: *Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, London, UK, 4–6 July 2012*, pp 1524–1529.
24. Zhou N, Xing K, Nagalingam S (2010) An agent-based cross-enterprise resource planning for small and medium enterprises. *IAENG Int J Comput Sci* 37(3):252–258

Adhesive Bonding of Attachments in Automotive Final Assembly

Loucas Papadakis, Vassos Vassiliou, Michalis Menicou, Manuel Schiel
and Klaus Dilger

Abstract In modern societies there is an increasing concern regarding the environmental impact of automobiles driving automotive manufacturers to develop lighter and, thus, less fuel consuming vehicles. Customers' protection during crash is a major demand which motivates automotive manufacturers to improve production processes which can satisfy the highly demanding market. Simultaneously, the introduction of new manufacturing techniques is strongly correlated with additional costs, which should be analyzed and quantified, in order to prove the sustainability of such processes for automotive production. This chapter considers adhesive bonding for joining attachments (i.e. roof components) on painted automotive shell surfaces as a potential technique in volume production. In order to introduce such type of adhesive joining process in current production lines, different process chain scenarios are proposed depending on the paint type in order to achieve the required strength of connection, especially during crash loads. Production costs are gathered and a proposed cost analysis is presented for evaluating the suggested scenarios aiming to identify cost intensive procedures.

L. Papadakis (✉) · M. Menicou
Department of Mechanical Engineering, Frederick University,
Y. Frederickou Str.7, 1036 Nicosia, Cyprus
e-mail: l.papadakis@frederick.ac.cy

M. Menicou
e-mail: eng.mm@frederick.ac.cy

V. Vassiliou
Frederick Research Center, Frederickou Str.7, 1036 Nicosia, Cyprus
e-mail: eng.vv@frederick.ac.cy

M. Schiel · K. Dilger
Institute for Joining and Welding (ifs), Technische Universitaet Braunschweig,
Langer Kamp 8, 38106 Braunschweig, Germany
e-mail: m.schiel@tu-bs.de

K. Dilger
e-mail: k.dilger@tu-bs.de

Keywords Adhesive bonding · Car body · Cost analysis · Final assembly · Laser ablation · Painting

1 Introduction

The principle adoption of functionally integrated components and modules in automobile construction, without which light automobile construction today would be greatly limited, is necessary in the completion of the exterior of automotive body shells [1, 2]. The most appropriate way to attach components and modules to painted car shells is through an appropriate low-temperature joining process with a similar finishing paint as the auto body. An assembly of the external components before the application of a finishing paint is, in the majority of cases, not possible because the affiliated oven-curing process cannot be withstood by the components and inhibits the functional integration of the painting process [3].

As a result of exacting specifications for the above application area, the best suitable joining technology is adhesive bonding. The industry standard adhesive bonding on finish painted surfaces nowadays provides the required strength only partially with respect to automobile structure strength and crash safety. An exception is the bonded wind screen of some automobiles. Here extensive test to prove the structural characteristics of the bonded parts with used colors have to be carried out. In order to find solutions to this problem, it is relevant to know the properties of the finish paint from its composition, which depends on the specific compounds and the “process history” of the paint. It is known that through various oven-times and temperature settings, the paint can fluctuate in adhesive strength from “structural rigidity” to “not-adhesive”, but these properties can only be measurable retroactively [4]. Additionally, an important issue which impacts the adhesive joint is the strength of the paint itself depending on its compositions, i.e. metallic or non-metallic.

A measurement technique for the adhesive bond system, and consequently for the adoption of functionally integrated components and essential paint characteristics, is not currently available. Furthermore, the interrelationship between process history and composition of the paint and its adhesive capability were only recently understandable [5].

The goal of this chapter is to provide an understanding of how the total adhesive bond system including substrates, electro-coating, primer, paint layers and 2-component polyurethane adhesive (2C-PU) influences the adhesive strength on attachments. A deeper look into how accessories can be attached with an adhesive agent so as to provide the necessary joint strength and crash safety will also be undertaken. Furthermore, the results provided from this study will enable the discovery of optimized high-demand bonds.

Finally, in order to introduce such type of adhesive joining process in current production lines, different process chain scenarios are proposed depending on the paint type in order to achieve the required strength of connection, especially during crash loads. Appropriate process sequences are suggested, with which the preparation

of coated surfaces, the treatment of painted surfaces and the application of an adhesive coat can all be incorporated in the production. Hereby, production costs are gathered and a proposed cost analysis is presented and explained for the suggested process chain scenarios in order to identify cost intensive procedures and secure their long-term sustainability in automotive production.

2 State of the Art

2.1 Organic Coated Sheets

In the field of vehicle manufacturing today, there are various methods used in the painting process, all of which not only vary among the various manufacturers, but also vary internally among factories of individual manufacturers [6]. A cathode electrophoretic coating is used first as a primer coat. In the following step of the process, a water-based filler, or a functional coat, is applied to the surface. Before a water-based color and/or effect lacquer is used, the paint from the underlying layer is hardened in the same manner as the electrophoretic coat. This next coat is then electrostatically applied in a uniform color evenly over the surface. After the electrostatic paint coat, an “effect lacquer” may be applied using a wet-on-wet, pneumatic coating process, with which the effect of the metallic pigments will be more visible and can be easily repaired in the future if necessary. After an intermediate drying process, in which water is removed, a wet-on-wet application of the 2C clear lacquer and its hardening in the finishing coat dryer follows. Figure 1 illustrates the layer configuration of the system substrate-coating-paint as utilized by automobile manufactures.

The release of information about other painting procedures (i.e. the hardening of material through ultra violet-radiation) allows a greater bandwidth of the various

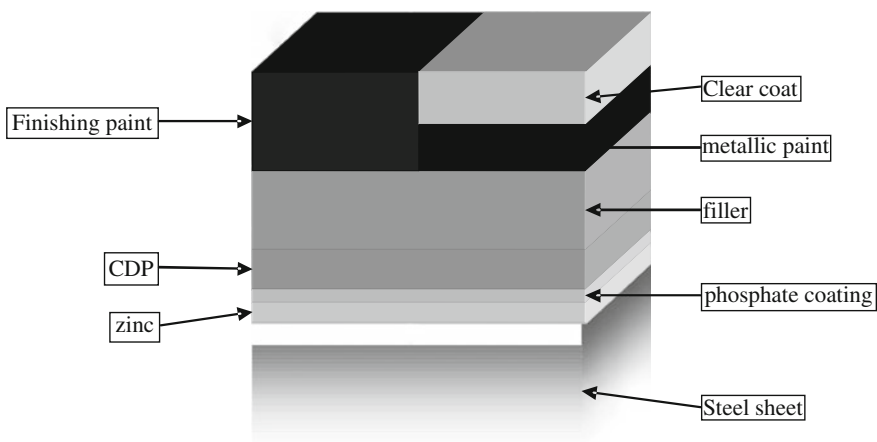


Fig. 1 Layer configuration of automotive substrate-coating-paint system

methods to be made available [7]. Current research of paint layering procedures after the stresses involved in an automobile crash are only relevant to respective stone chipping resistance of the paint [8]. There are numerous additional investigations in the area of the improvement of adhesive strength between metal and polymer surfaces [9–11].

The performance of paint layering under the stresses of high speed travel was, until now, only investigated from the perspective of stone impact resistance. These investigations allowed carrying out a direct interrelation between stone impact resistance and the bond strength and elasticity of the paint [12]. Furthermore, the large influence of interlacing temperatures and interlacing times on the bond strength and elasticity of paint coats were investigated. Other authors investigated the stone impact resistance of paint on plastic, specifically how aesthetic paints in automobile construction are being used [8]. These investigations have discovered that paint damage in contrast to metallic substrates is almost exclusively dependent upon plastic. For higher impact resistance of paint, the existence of a glass transition made a difference in low temperature tests.

2.2 Joining of Coats of Sheet Steel

The joining of painted metal through the use of an adhesive agent is widely used in industrial fabrication. In 1999, the adhesion of completely painted aluminum panels on the side quadrants of Light Rail Vehicles (LRV) was made possible through developed process technology [13].

Different possibilities in the area of joining surface coated steel materials were demonstrated in the literature [14]. The fact that the adhesive medium applied to the coated steelwork adhered to the surface but did not compromise the strength or adhere to the substrate displays the research worthiness of this work.

Formation and adhesion through hybrid bonding techniques were used for the integration of new materials in a transportation structure [15]. On this topic of synergy, the completion and the properties of the emerging connections under quasi-static, dynamic and impulsive strain are admonished. Through this work, the possibilities that the use of dissimilar materials, the isolation of the joining partners and the avoidance of seam corrosion are elucidated. Finally, the “patchwork blank” technology offers certain advantages as an alternative to the currently-used “tailored banks,” in which the cutting process could be reduced and a higher flexibility could be obtained through the adhesion of various sheets of material [16].

2.3 Crash Behavior of Steel Sheeting

There are numerous research studies on the topic of the behavior of steel sheeting during a crash. Various steel components were investigated in terms of their dif-

ferent tensile strengths [17]. Moreover, the dependency between tensile strengths and temperature in regards to the steel sheets the automobile industry adopted (dual-phase steel, trip-steel and bake-hardening steel) was also investigated [18]. The tensile strength differences between the various steel types above are quite pronounced.

A practical investigation method for determining the crash resistance offered by adhered steel sheeting connections was also achieved [19]. Research using different shapes of crash devices was undertaken, with which aptitude for the accuracy of crash-test values and the predictions of the construction components were accurately ascertained.

3 Mechanical Analysis of Bond System Properties

In order to identify the potential of the application of adhesive boning on lacquered automotive shells various bond system configurations were analyzed in terms of their mechanical failure. Following bond systems joined with a 2C-polyurethan adhesive were investigated in terms of their mechanical properties by means of (a) lap-joint tensile-shear tests and (b) butt-joint tension tests:

- (i) plain steel sheets
- (ii) electro-coated steel sheets
- (iii) primed sheets
- (iv) lacquered sheets with
 - white non-metallic color and
 - silver metallic color.

3.1 Lap-Joint Tensile-Shear Tests

The lap-joint tensile-shear tests were performed after DIN EN 1465 for the above five bond system configurations stated above. The location and type of failure for each adhesive bond configuration in case of tensile-shear overlap tests with a sheet thickness of 0.8 mm as well as the respective lap-shear strengths are illustrated in Fig. 2.

The experiments show that the connection fails in the first four bond system configurations through all the involved layers with a clear reduction of the bond strength up to 34 % [20]. In case (iv) of the lacquered surface with white non-metallic color the mechanical failure is observed within the whole adhesive bond connection including the electro-coat, the primer and the white paint, the varnish and the adhesive. In case (i) of blasted metal sheet probes the connection between adhesive and metal surface proves the highest strength representing the lap-shear strength of 2C-PU of 17 MPa. Blasted steel probes were used for the performance of the experiments instead of metal sheet probes with a smooth surface, which would cause a very early failure of

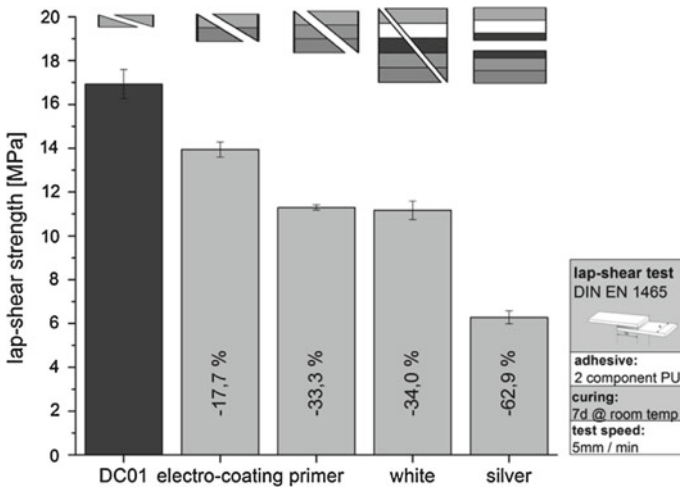


Fig. 2 Failure mechanisms of the adhesive bond system joined by 2C-PU on (i) plain steel sheets, (ii) electro-coated steel sheets, (iii) primed sheets, and (iv) white and (v) silver painted sheets after tensile-shear tests

the adhesion zone. Finally in case of the adhesive bonding on painted silver metallic surface (v) the weakest link in the bond system proved to be the silver paint layer, since the failure occurs solely here. The bond system with silver paint indicates an even higher reduction of strength in the paint cohesive zone reaching even almost the half of the lap-shear strength of the bond on white non-metallic paint [20].

3.2 Butt-Joint Tension Tests

The butt-joint tests were performed after ISO 11003-2 again for the three bond system configurations: adhesive bond with a 2C-PU on (i) plain steel sheets and on lacquered sheets with (ii) white non-metallic color and (iii) silver metallic color. The location and type of failure for each adhesive bond configuration in case of but-joint tension tests with cylindrical probes as well as the respective maximum tensile stresses are illustrated in Fig. 3.

The performed tests prove that the adhesive bonds fail within the adhesive-varnish-paint system in case (ii) of white paint, which indicates an almost equal tensile strength of the white non-metallic paint compared to the 2C-polyurethan adhesive in case (i). Evidence of this is the only slight decrease of the total bond tensile strength of 8%. In the contrary in case of butt-joint tension with silver metallic paint the bond failure is located entirely within the silver paint layer. This signifies that the silver paint is the weakest link in such kind of bond systems. This conclusion is also supported by the considerable decrease of 40% of the bond tensile strength in case of the silver paint bond [20].

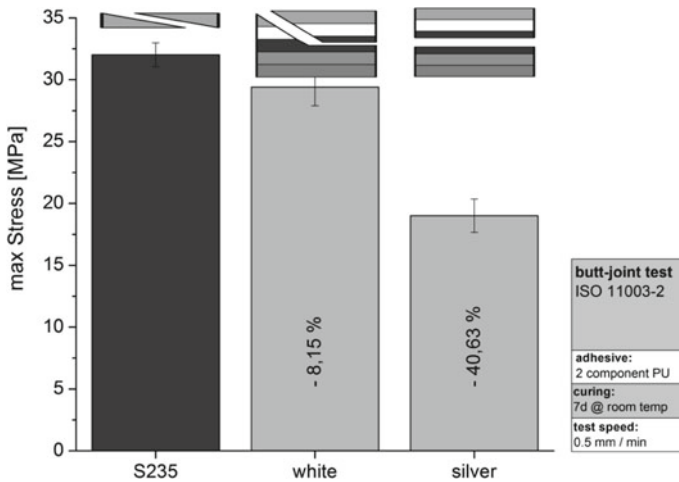


Fig. 3 Failure mechanisms of the adhesive bond system joined by 2C-PU on (i) *plain* steel sheets and (ii) *white* and (iii) *silver* painted sheets after butt-joint tension tests

4 Process Alternative Scenarios

4.1 Case-Study

The investigation for adhesive bonding alternatives during the automotive body production focused on the case of roof attachment to the automobile frame. This is a typical example in the automobile industry, especially in case components of different material are to be adequately attached on the car body. Thus, the case of a composite roof attached to the steel car frame, a feature most manufacturers increasingly integrate to their models portfolio, is considered. Based on this case-study the baseline and alternative proposed scenarios were identified. According to these the cost analysis was carried out in order to determine the cost increment to the baseline scenario.

4.2 Current Situation

The classic process chain in automotive production includes three major areas: the body shell work or body-in-white, the painting process and the final assembly. The body shell work is subdivided into the component manufacture, e.g. through deep forming, extrude molding, casting etc., the mechanical or thermal joining, i.e. laser or spot welding, and the application of structural adhesive bonding for integrating single components and assemblies into an integrated body shell [21]. In turn, the painting

process encompasses various applications as electro-coating (EC), hardening of EC and structural adhesives, primer and top color coating. Finally, during the final assembly mainly mechanical joining of attachments and bonding of non-structural parts occur [22].

In terms of the case-study, the current situation was set to be the adhesive bonding on painted automotive surfaces. As discussed in the aforementioned Sects. 3.1 and 3.2, adhesive joining on silver paint yielded high reduction of strength a result clearly illustrated in Figs. 2 and 3. That said, the identified alternative production scenarios for tackling the inadequate strength results were decomposed into their activities and their impact on costs was examined.

4.3 Process Chain Alternatives

First attempts in the automotive industry in Europe to increase the adhesive bond quality and strength during the final assembly phase prerequisites a masking of the surfaces to be adhered with a corresponding unmasking operation prior and after the color painting process respectively. The masking will be placed on the EC-layer. A further alternative proposal involves the pre-treatment of already painted body surfaces by means of laser process for the paint removal prior to adhesive bonding. This alternative solution aims to increase the strength of the bond system since the adhesion will occur solely on the electro-coat (cf. Fig. 2). During such process application it is important to ensure the EC-layer consistency so as to avoid corrosion.

For the purposes of this chapter two alternative scenarios are identified as shown below. Nevertheless in the context of this contribution the cost analysis results presented here refers only to Scenario 1.

- Scenario 1: laser ablation, i.e. paint removal, prior to final assembly
- Scenario 2: masking and unmasking process steps during the painting process after EC and color painting respectively.

These proposed alternatives are visualized in Fig. 4. Here the high-level processes are listed with the option to integrate the suggested alternatives in order to achieve the required bond quality and strength for different product variants, i.e. metallic or non metallic paint, panoramic roof etc.

5 Alternatives Cost Analysis

A costs analysis model was developed for the economic investigation of integrating the identified alternative scenarios to the existing production line. The scenarios described in Sect. 4 are essential for raising the standards of production in respect to the adhesive joining. The methodological framework followed is that of engineering economic analysis.

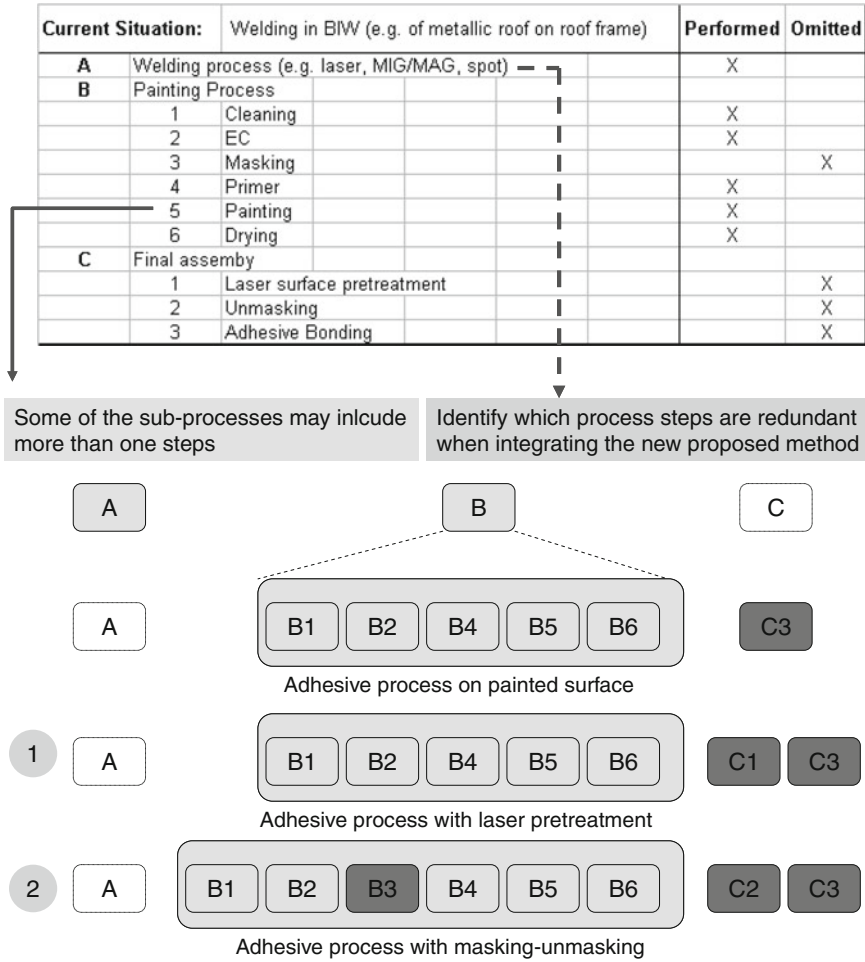


Fig. 4 Current situation and definition of alternative process chain scenarios

5.1 Engineering Economic Analysis

Engineering economics or alternatively the engineering economy, as its name states, is an area in which economics and engineering are combined. Essentially, one could say that it is the use of fundamental economic techniques applied to engineering projects and subsequently to engineering investment decisions. Sullivan et al. [23] report that engineering economy involves the systematic evaluation of the economic merits of proposed solutions to engineering problems.

5.2 Parameters and Model

The analysis for Scenario 1, started by identifying the cost groups which will be affected. For this to be effective a clear understanding of which process would be added was needed as well as how the baseline scenario's process sequence would be altered. Figure 4, although simplistic, was used as a roadmap.

Next, based again on Scenario 1, it was necessary to identify which cost groups are influenced and how. In particular, (a) direct labor; (b) utilities; (c) maintenance; (d) parts of manufacturing overheads; (e) R&D expenses; and finally (f) depreciation, were in some degree affected by the proposed Scenario 1. Note that the cost groups (a) direct material; (b) building associated costs—part of manufacturing overheads—and (c) tooling, were assumed not to be affected by the scenario.

For each cost group the inherent variables, which some authors call cost elements [23], which determined the level of influence were set. An example is the cost per kWh and the electricity consumption for the case of utilities costs. In addition, various key operational variables have also been quantified. These include the following variables: (i) overall production rate; (ii) new process' cycle time (broken down into setup-processing-post time); (iii) working days; and (iv) shift(s) duration. Finally, for the cost analysis to be carried out successfully financial variables were quantified. Those of paramount importance were (a) investment, (b) possible salvage value of machinery; (c) tax rate; (d) interest rate, (e) inflation rate; (f) investment horizon; and finally (g) capital structure i.e. 100% loan financing. Table 1 depicts all recognized variables.

Note that since the alternative Scenario 1 suggested supplementary activities within the production process and not replacements of existing ones, the impact was clearly negative.

Subsequently major operationally and facility constraints were identified. In brief, some examples of these constraints were (a) the space requirements for the new processes which were examined; and (b) the number of conveyor belts at the specific cell.

The final step of the cost analysis was to construct the respective model to evaluate Scenario 1 and 2 based on a quantitative criterion. The model was constructed in MS Excel, an intuitive and powerful tool. The criterion chosen was the cost per unit where unit referred to car frames. This criterion defined the additional economic burden the company must carry in order to adopt a specific production scenario; hence, it was denoted as ΔC . Its magnitude encloses all operational changes each scenario was requiring and worked as simple but yet effective comparison criterion for the scenarios. Scenario 1 model started by quantifying how many car frames could be processed within the two workday shifts based on the ablation speed, as given by the manufacturer, and the required ablation area, as set by specific automobile model. This number was then divided by the current production rate and the result was rounded upwards to the nearest integer. With this manner the number of required machinery was quantified considering current setting production rate as a hard constraint that needed to be fulfilled. Following, the remaining cost variables

Table 1 Cost groups and variables

Cost groups	Variables
Direct labor	# of employees Annual salary (incl. benefits, allowances, training, etc)
Utilities	Electricity consumption Cost of electricity as € per kWh Water consumption Cost of water as € per lt _{H2O}
Maintenance	Cost (in-house) Period
Man. Overheads ^a	As % of Investment Capital cost (cost of machinery)
Depreciation	Horizon (Years) ^b Straight-line method
R&D	3-year average % calculated on total running cost
<i>Financial variables</i>	
Capital cost (investment)	Cost of primary machinery Cost of secondary machinery
Salvage value	Value
Installation costs	As % of total cost
Tax rate	As %
Capital Structure	100% Loan Duration (Years)
Interest rate	As %
Inflation rate	As %
Investment horizon	Years
<i>Operational variables</i>	
Production rate ^b	Cars per day
Cycle time	Setup, processing and post processing time
Processing area per car	Area in cm ²
Factory working days	Days per year
Workers working days	Days per year
Shift	Number of shift(s) Duration

^aSome elements of this cost group were unaffected

^bBased on technology experts these may vary according to how fast the technology matures

^cThis was kept constant

and subsequent cost groups were quantified on the calculated number of ablation lasers required.

5.3 Results

The cost-per-unit criterion was projected over the selected investment horizon which was set to 15 years. It portrayed, in a reliable manner, how much the respective

company would have had to pay for integrating Scenario 1. Firstly, the analysis yielded that 2 new equipment machineries (ablation lasers and supporting equipment) were needed and an additional time Δt of 75.12 s was required per car frame. As a result the first 5 years the criterion's order of magnitude was to the level of hundreds of euros, from just above 105 to close to 96 euros. Next, from year 7 up to the end, the order of magnitude dropped dramatically to just a few euros. Exemptions of the declining pattern were years 6 and 12 at which cost spikes were present due to the scheduled maintenance and associated costs that were assigned. Three main reasons contributed to the creation of the pattern. These were (a) the depreciation period; (b) the loan duration; and (c) the maintenance interval. The first point is subjective since this is a technology intensive process and depends highly on its maturity rate. In this model it was set to 5 years. The second reason is an operational issue thus the laser manufacturer was consulted. Last but not least, the loan duration was arbitrarily set to 5 years.

Validation of the aforementioned results is highly correlated to the validity of the analysis methodology and the reliability and validity of the model's assumptions. The assumptions were set based on extensive literature review of similar works [23, 24] and comments of industry experts extracted through interviews conducted within the context of the collaborative research project CrabLacs. Furthermore, the model was constructed and its parameters were addressed based on the well documented principles of engineering economic analysis [23]. Consequently, the presented analysis is valid and the extracted results provide the best estimation of the criterion. However, it is likely that results can potentially deviate from real-life results since every model is a representation of reality but never reality itself.

6 Conclusions and Outlook

In this chapter the state-of-the-art of adhesive bonding mechanisms on different coating configurations is introduced. Hereby, the bond strength is investigated in respect of the coat layer composition and the weaknesses of bond systems on metallic paint are identified. Based on the experimental conclusions and the available means, alternative process chain scenarios are proposed with the aim to overcome the disadvantages identified during adhesive bonding of attachments on painted surfaces in automotive final assembly lines. In order to quantify the costs of the suggested scenarios a proposed cost analysis and an appropriate model are presented including all relevant cost groups and variables. The proposed cost analysis contributes to the justification of the viability of the modern manufacturing process in respect of their additional costs versus the improved product quality. Hereby, the operational conditions of the entire process chain taking into consideration plant parameters and, most importantly, constraints of preceding processes that may affect the proposed, subsequent ones will affect the cost analysis results and will have to be considered for further investigations. Moreover, the total process chain will be necessary to be

examined using methods based on business process simulation in order to establish key indexes not addressed here such as buffer levels and queuing times.

Acknowledgments The presented investigations and results were accomplished in the context of the collaborative project CrabLacs, which was financed in the framework of the 8th Joint CORNET Call for Trans-national Collective Research Project Proposals by the German Federation of Industrial Research Associations (AiF) and the Research Promotion foundation (RPF) of Cyprus.

References

1. Daniels J (1984) Design implications of adhesive bonding in car body construction. *Int J Adhesion Adhesives* 4(1):5–8
2. Barnes TA, Pashby IR (2000) Joining techniques for aluminium spaceframes used in automobiles, Part II—adhesive bonding and mechanical fasteners. *J Mater Process Technol* 99:72–79
3. Fathil MFBM (2008) Painting process improvement for automotive industry, University Malaysia Pahang, Nov 2008
4. Kessel A, Dilger K (2006) Kleben auf Lack. Einfluss der Lacktrocknung auf die Hafteigenschaften. *J für Oberflächentechnik* 46(9):56–59
5. Kessel A (2009) Lackautohäsion in kraftübertragenden Klebverbindungen im Automobilbau. Dissertation Technical University Braunschweig, Shaker, Aachen
6. Woelfel G (2005) BMW Werk Leipzig setzt neue Maßstäbe in der Automobillackierung. *Die Oberflächen-Zeitung*, Germany
7. Hilt M (2000) Perspektiven der UV-Strahlungstrocknung von Klarlacken. *Metalloberfläche* 54(11):43–45
8. Nothhelfer-Richer R, Kunz G, Eisenbach CD (2005) Stone chip resistance of coatings on plastic substrates. *Stuttgarter Kunststoff Kolloquium* 19:1–6
9. Scholz W (2003) Improving substrate wetting with coating additives. *Adhesive and finishing symposium*, vol 28, Munich
10. Kleber W (2000) Elektrostatische Oberflächenbeschichtung, Teil 4: Technologische Anforderungen an die Anlagentechnik. *Metalloberfläche* 54:48–51
11. Mechtel M, Melchior M (2003) 2-K-Wasserklarlack für die Automobil-Lackierung. *J für Oberflächentechnik* 43:42–46
12. Somborn R (2007) Lackwunden durch Steinschläge. Unzureichende Stein schlagbeständigkeit, Farbe und Lack 113(1):50–52
13. Schmatz G (1999) Verkleben von fertiglackierten Aluminiumpaneelen im Seitenwandbereich von LRV (light rail vehicle) aus prozess-technischer Sicht. *Praxis-Forum, Arbeitskreis Automobil*, Vienna
14. Fluegge W (2004) Fügen von oberflächenbeschichteten Stahlwerkstoffen. Conference report: DVS-Report, Salzgitter
15. Meschut G, Friedrich H (2003) Zukünftige Fügekonzepte für Automobil-strukturen in Mischbauweise, Conference report: Dresdner Leichtbausymposium
16. Lamprecht K, Geiger M (2005) Experimental and numerical investigation of the formability of laser welded patchwork blanks. Conference report: advanced materials research
17. Borsutzki M, Matthissen D, Schaumann TW, Sieg HJ (2005) Modern methods of material characterization during the development of new steels for automotive application. Conference report: SCT, international conference on steels in cars and trucks
18. Bleck W, Frehn A, Larour P, Steinbeck G (2004) Untersuchungen zur Ermittlung der Dehnratenabhängigkeit von modernen Karosseriestählen. *Materialwissenschaft und Werkstofftechnik* 35(8):505–513

19. Hahn O, Kurzok JR, Oeter M, Brede M, Hesebeck O, Dilge K, Schmid G (2002) Untersuchungen zum Crashverhalten geklebter und hybridgefügter Stahlblechverbindungen. Forschung für die Praxis, Paderborn
20. Schiel M, Hanssen E, Fraunhofer M, Dilger K (2011) Determination of mechanical properties of adhesive joints on painted substrates. In: International conference on structural adhesive bonding, Porto, 7–8 July 2011
21. Papadakis L (2008) Simulation of the structural effects of welded frame assemblies in manufacturing process chains. Dissertation, Technische Universitaet Muenchen
22. Papadakis L, Vassiliou V, Menicou M, Schiel M, Dilger K (2012) Adhesive bonding on painted car bodies in automotive production lines: alternatives and cost analysis. Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE 2012, 4–6 July 2012 U.K , London, pp 1382–1387
23. Sullivan GW, Wicks ME, Kowlling CP (2009) Engineering economy, 14th edn. Pearson Education, New Jersey
24. Roy R, Souchoroukov P, Shehab E (2011) Detailed cost estimating in the automotive industry: data and information requirements. *Int J Prod Econ* 133:694–707

Uncertainty Components in Performance Measures

Sérgio Dinis Teixeira de Sousa, Eusébio Manuel Pinto Nunes
and Isabel da Silva Lopes

Abstract Data quality is a multi-dimensional concept and this research will explore its impact in performance measurement systems (PMSs). Despite the large numbers of publications on the design of PMSs and the definition of critical success factors to develop Performance Measures (PMs), from the data user perspective there are possibilities of finding data quality problems, that may have a negative impact in decision making. This work identifies and classifies uncertainty components of PMSs, and proposes a qualitative method for PMs' quality assessment. Fuzzy PMs are used to represent uncertainty that is present in any physical system. A method is also proposed to calculate an indicator of the compliance between a fuzzy PM and its target value, that can serve as a risk indicator for the decision-maker.

Keywords Data quality · Fuzzy sets · Performance measurement · Performance measures · Quality management · Risk management · Uncertainty

1 Introduction

Performance measurement systems (PMSs) are receiving increasing attention from academics and practitioners particularly after the development of the Balanced Scorecard (BSC) [1], and many PMSs are available nowadays [2, 3]. Nevertheless, this subject is not new and, for example, quality gurus such as Crosby, Feigenbaum, or Deming recognized the importance of performance measurement as an activity

S. D. T. de Sousa(✉) · E. M. P. Nunes · I. da Silva Lopes
Centro Algoritmi, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: sds@dps.uminho.pt

E. M. P. Nunes
e-mail: enunes@dps.uminho.pt

I. da Silva Lopes
e-mail: ilopes@dps.uminho.pt

within quality management. Recently, many publications on the design of PMSs and about its implementation and use have been published. There is, however, a lack of investigation on the uncertainty associated with such performance measures (PMs).

The uncertainty is a quantitative indication of the quality of the result. It is an unavoidable part of any measurement and it starts to matter when results are close to critical decision limits. When uncertainty is evaluated and reported in a specified way it indicates the level of confidence that the value actually lies within the range defined by the uncertainty interval.

The PMS's purpose is to contribute to both the goals and the sustainability of the organisation [3], through the decision-maker that acts on the organization based on PMs. However, if uncertainty is present in physical systems it should be reflected in the PMS. Furthermore, there are many measurement capability studies of "hard" variables, but there are few attempts to deal with attribute data and "soft" PMs (based on subjective assessment), such as customer satisfaction. Failure to deal with such uncertainty will result in simplified models of reality.

The information (or data) quality field of research has established that information quality is a multi-dimensional concept [4]. Many works [5–7] have identified different sets/categories of quality dimensions: Intrinsic, contextual and reputational [6]; internal, data related, external, system-related [8]; objective and subjective [7]; syntactic, semantic, pragmatic and physical [9]. These categories, applicable to data or information in general, could also be applied to specific areas [10] such as the PMS.

In the traditional formulation of a PMS, most PMs are affected by imprecision and vagueness but they are represented using a number that is not able to represent uncertainty. A good decision-making model needs to tolerate vagueness and imprecision because these types of the non-probabilistic uncertainty are common in decision-making problems [11].

The hypothesis is that organisations need to reflect the uncertainty of its physical systems and contextual factors in their PMs to improve their models. This identification of uncertainty in PMS is the first step to reduce such uncertainty. The problem is how to overcome this situation or how to deal with data uncertainty. Several ways can be used to represent the uncertainty [12] such as: standard deviation, probability distribution, fuzzy numbers, scenarios and quartiles.

Fuzzy Set Theory have proved to be a successful in handling imprecise and vague knowledge that characterize this kind of problems, and it has been applied in a variety of fields in the last decades.

The second hypothesis of this work is that fuzzy sets are sufficient to represent the uncertainty in PMs.

The research methodology to characterise PMs' uncertainty will comprise both deductive and inductive stages. This chapter starts with a literature review on the field of performance measurement and uncertainty to develop through deductive logic a conceptual and theoretical structure about the classification of uncertainty in PMs. This chapter presents the findings of this deductive research which will later be tested through case studies, to allow another step of inductive research to support, change or refute the proposed characteristics of the PMs.

The work described in this chapter was presented in the World Congress on Engineering 2012 [13].

2 Literature Review

2.1 Performance Measures

Juran and Godfrey [14] argue that “the choice of what to measure and the analysis, synthesis, and presentation of the information are just as important as the act of measurement itself” and emphasise the system to which the measurement process belongs. The measurement process consists of steps needed to collect data and present results.

A thorough understanding of the existing measurement systems, formal and informal, spoken and unspoken, as they are perceived [15] must be achieved, i.e. the overall framework in which the PMS operates should be understood [14].

According to Macpherson [16] there are two approaches to identifying PMs: top-down and bottom-up. Using the first approach, the search for PMs is based on the mission and vision of the organisation. The latter, on the other hand, is determined by what data is currently available and has the advantage of being cost effective by only focusing on visible data [16]. A third approach [17] is outside (or customer)—inside (or internal processes), endorsing the argument about the importance of looking at the organisation from the customer’s viewpoint [18].

Regardless of the approach used, there are two basic types of PM in any organisation—those related to results, and those that focus on the determinants of the results [15]. This suggests that it should be possible to build a performance measurement framework (PMF) around the concepts of results and determinants.

Perhaps the best know PMF is Kaplan and Norton’s Balanced Scorecard (BSC) [19]. It seems to be the most influential and dominant concept in the field. The authors of the BSC suggested the definition of strategy maps to describe the cause-and-effect relationships between the identified measures, but according to Wilcox and Bourne [20] these relationships are outdated, because the organisation and its context are dynamic.

Kanji and Sá [21] integrated BSC with TQM principles and its critical success factors (CSFs) resulting in a model which focussed on measuring how an organisation is performing from an outside perspective. Bititci et al. [22] developed a model for an integrated and dynamic PMS. As the previous framework it should have: an external and internal monitoring system. Basu [23] also argued that the PMs should be more externally focused for the total network and a formal senior management review process with two-way communication to all partners was essential to success.

The Performance Prism’ authors [24] refer to the importance of identifying stakeholders’ contributions, as they are part of a reciprocal relationship with the organi-

sation. They also argue that it is necessary to start to think about measurement as the process of gathering management intelligence.

2.2 *PMs Quality*

To contribute to the planning phase of the PMS, CSFs about data quality are identified in the literature. PMs should be [16, 25, 26]:

- relevant (C1);
- credible (C2);
- precise (C3);
- valid (C4);
- reliable (C5); and
- frequent (C6).

Other CSFs discussed in the performance measurement literature are:

- data collection and methods for calculating the PMs must be clearly defined [27] (C7);
- presentation of PMs must be simple [18] (C8);
- PMs must be flexible [26], including being tied to desired results [28] (C9);
- more extensive use should be made of subjective data [25] (C10); and
- ratio-based performance criteria are preferred to absolute numbers [27] (C11).

However, the designing of a PMS may not comply with all of these requirements and thus its quality should be assessed using the commonly accepted dimensions of data quality. Data quality dimensions commonly referred in literature are [29]:

- accuracy;
- completeness;
- timeliness; and
- consistency.

Many other classifications of Information Quality exist, for example, [5] refer:

- *intrinsic*—implies that information has quality in its own right;
- *contextual*—must be considered within the context of the task at hand; it must be relevant, timely, complete, and appropriate in terms of amount, so as to add value; and
- *representational and accessibility*—emphasize the importance of computer systems that store and provide access to information; that is, the system must present information in such a way that it is interpretable, easy to understand, easy to manipulate, and is represented concisely and consistently; also, the system must be accessible but secure.

Galway and Hanks [30] classify data quality problems as operational, conceptual and organizational. Associated with operational data problems there is an implied presumption that, were the data correct, the user could directly utilize them in making the necessary decision(s).

2.3 Characterising Uncertainty

Any measurement is subject to imperfections; some of these are due to random effects. Repeated measurements will show variation because of such random effects. When uncertainty is evaluated and reported in a specified way it indicates the level of confidence that the value actually lies within the range defined by the uncertainty interval.

Uncertainty of measurement “is a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand” [31]. Thus the uncertainty, in metrology, is a quantitative indication of the quality of the result. It gives an answer to the question, how well does the result represent the value of the quantity being measured? It allows users of the result to assess its reliability, for example for the purposes of comparing results from different sources or with reference values.

Uncertainty can be expressed as a quantity, i.e., an interval about the result. “Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a specification or standard. It is therefore necessary that there be a readily implemented, easily understood, and generally accepted procedure for characterizing the quality of a result of a measurement, that is, for evaluating and expressing its uncertainty” [31]. This is common knowledge in metrology but, apparently, it is not being applied in ordinary PMs.

There is a wide variety of reasons why uncertainty is present in PMSs. Particularly, to reliability studies [32] presents three main reasons: (i) dependence on subjective information in the form of expert judgments; (ii) the relaxation of dependence on precise statistical models justified by physical arguments; (iii) the exact system structure and dependence relations between components are known, which may well be unrealistic. These relationships are conditioned by the system’s environment and may generate contradictory information, vagueness, ambiguity data, randomness, etc. In reliability studies, the vagueness of the data have many different sources: it might be caused by subjective and imprecise perceptions of failures by a user, by imprecise records of reliability data, by imprecise records of the appropriate tools for modelling vague data, and unsuitable statistical methodology to handle these data as well [33].

According to ISO 1012 [34], Sect. 7.3, the measurement uncertainty shall be estimated for each measurement process covered by the measurement management system and all known sources of measurement variability shall be documented. If these requirements are to be applied in all PMs of the organization there would be the need to identify all sources of variability. However, few works [35–38] report the inclusion of such variability in their studies.

Both [35] and [36] considered uncertainty in manufacturing systems and argue that reducing it, is a means to improve the system. Other studies have included uncertainty in project scheduling [38], inventory control [37], or supply chain management [39]. However, specific components of PMS's uncertainty and its classification, to facilitate systematic studies, are not known.

2.4 Methods to Deal with Uncertainty of PMs

Traditionally, uncertain parameters in inventory control and supply chain management problems have been treated as stochastic processes and described by probability distributions [39]. A probability distribution is usually derived from evidence recorded in the past [37]. This requires a valid hypothesis that evidence collected are complete and unbiased, and that the stochastic mechanism generating the data recorded continues in force on an unchanged basis [39]. However, there are situations where all these requirements are not satisfied and, therefore, the conventional probabilistic reasoning methods are not appropriate [39]. In this case, uncertain parameters can be specified based on the experience and managerial subjective judgment. Often, an expert may feel that a given parameter is within a certain range and may even have an intuitive feel for the best value within that range [37].

It may be convenient to express these uncertainties using various imprecise linguistic expressions [39]. Fuzzy sets are found to be useful in representing these approximate qualifiers, due to their conceptual and computational simplicity. The typical membership functions can represent fuzzy customer demand, fuzzy external supplier reliability or fuzzy lead time [39], just to name a few.

To deal with uncertainty in scheduling environment other approaches (apart from stochastic and fuzzy) are found in literature: reactive, proactive and sensitivity analysis [38], while [36] argue that for complex processes, methodologies based on artificial intelligence and simulation should be used.

On production planning, [36] identified a need for further research in incorporating all uncertainty in an integrated manner and in the development of empirical works that compare the different modelling approaches with real case studies.

Several authors [40, 41] used a fuzzy AHP approach to assign weights to certain PMs, while [42] used it to obtain weights in multicriteria multifacility location problems.

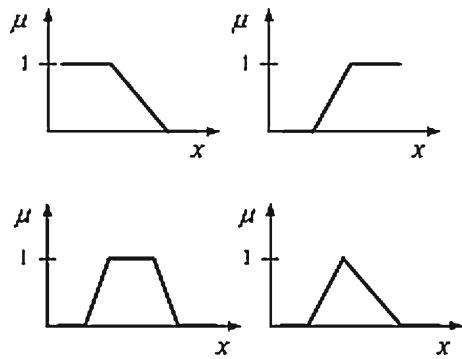
The costs incurred by organisations to manage uncertainty should not be ignored [29], and different methods to deal with uncertainty have different requirements and associated costs. In risk management a parallel situation can be established because identified risks are not all subject to the same detailed subsequent treatment, for example qualitative methods for risk assessment, may be enough for lower level risks, while quantitative techniques would be economically reasonable for higher level risks. Similarly, methods to deal with uncertainty have associated costs, and if some components of uncertainty, are small compared to others, it could be unjustifiable to

make a detailed determination of all its components. This is also expressed in ISO 10012 (Sect. 7.3.1).

2.5 Fuzzy Sets

The numerical assessment of fuzzy parameter/data and linguistic variables such as some PMs on customer satisfaction is done by using adequate membership function which determines the degree of membership in each input fuzzy set. The design of a fuzzy model is not trivial and several approaches [43, 44] have been proposed to identify the shape of elementary PM. The most usual solution is to use triangular and/or trapezoidal membership functions (see examples in Fig. 1) because of the advantages in terms of their manipulation.

Fig. 1 Commonly used membership functions



Having reviewed PMSs and uncertainty, the next section will address the classification of PMs' uncertainty.

3 Uncertainty Components in Performance Measures

Given the multitude of information about quality dimensions and problems this work will identify uncertainty components or data quality problems of PMs. A basic quality tool, the Cause and effect diagram which is usually used to group the main causes of a problem by controllable factors of a company was developed in order to identify the uncertainty component (Fig. 2).

The uncertainty components were classified in three main groups: Intrinsic (mainly related with its design), Data Collection (refers to real time data quality problems introduced by the collection method), and, finally, the PM definition (on the customer perspective use of PMs).

Examples of the uncertainty components are provided in [45]. A brief description of the uncertainty components is provided in Table 1.

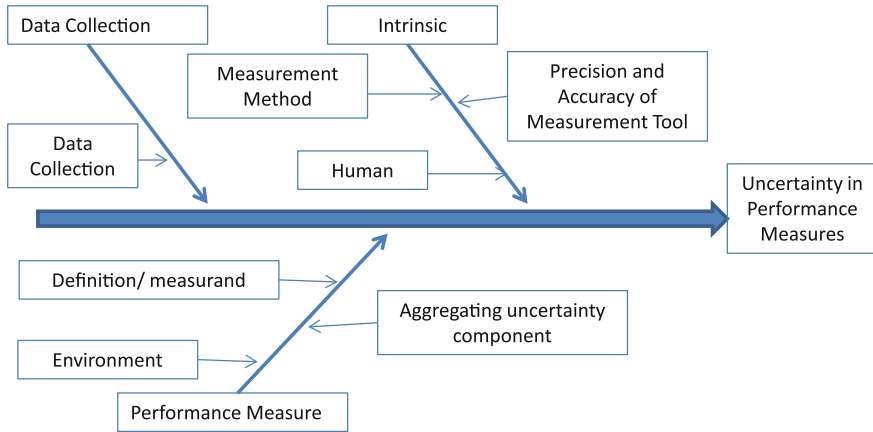


Fig. 2 Cause-and-effect diagram of uncertainty components in PMs

4 Modelling Uncertainty in Performance Measures

4.1 Uncertainty Qualitative Assessment

Quantitative methods for uncertainty modelling usually require more resources and data than qualitative ones. It is proposed a less demanding assessment method, a qualitative assessment, based on the analysis of uncertainty components, which are represented through fuzzy membership functions. The first step to characterise uncertainty would be to identify what Uncertainty Components are associated with each PM.

The second step would be to classify the uncertainty level of each Uncertainty Component. In structured systems (such as automotive manufacturing plants) risk assessment and FMEAs, typically, use a 10 item Likert scale. As PMs apply to less structured systems, the authors propose a scale with three levels but similar solutions with other levels are also feasible. For example, a scale for UC_A component could be:

No Uncertainty—There is a recognized formula that derives from theory and is not scientifically questioned.

Some uncertainty—An agreed formula is, apparently, recognised as adequate by all stakeholders.

High uncertainty—The formula was defined without consensus and may be changed.

After building similar scales to each uncertainty component, a matrix could relate each PM with each uncertainty component. This matrix would be a tool to decide which uncertainty components would be further studied, and could provide evidence to change existing PMs. The uncertainty reduction of the PMS would provide less risk in decision making.

Table 1 Uncertainty components of performance measures

Intrinsic	Data collection	Performance measure
<i>Measurement method uncertainty component (UC_MM):</i> Error on the method used to perform the measurement.	<i>Data collection (equipment/operator) uncertainty component (UC_DC):</i> Errors in the introduction and acquisition of data.	<i>Definition/Measurand uncertainty component (UC_D):</i> Difference between what is intended to be measured and what is really measured with the chosen PM
<i>Precision and accuracy of measurement tool (UC_PA):</i> The precision of the measurement tool can be determined by the repeatability and reproducibility (R&R) study.		<i>Environmental uncertainty component (UC_E):</i> Difference between what is intended to be measured and what is really measure motivated by changes in the system that occurred after the introduction of the PM
<i>Human assessment component (UC_H):</i> Uncertainty introduced by a subjective judgment when the measurement system relies on human.		<i>Aggregating uncertainty component (UC_A):</i> Uncertainty present in other PMs used to obtain a new PM

The next step to model uncertainty would be to represent it by fuzzy membership functions.

4.2 Modelling Uncertainty

Let us assume that PMs fuzzy numbers will represent the uncertainty components identified in previous section. Let $M = f(x_1, x_2, x_3, \dots, x_n)$ be the analytical model for a given crisp PM, M. This model maps the n inputs $(x_1, x_2, x_3, \dots, x_n)$ into the output space. It is now intended to extend such a mapping to fuzzy sets $M = f(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots, \tilde{x}_n)$.

The arithmetic by intervals and the Zadeh’s extension principle [46] constitute the two fundamental methods to handle fuzzy numbers. These methods have shown to be suitable for simple mathematical operations.

Let \tilde{P}_i to be a PM represented by a fuzzy number and $\tilde{\psi}_{ij}$ to be the fuzzy weight to calculate an aggregate PM, \tilde{M}_j . This PM is the result of an aggregator operator Θ of PM, \tilde{P}_i each weighted by $\tilde{\psi}_{ij}$.

$$\tilde{M}_j = \Theta \left(\tilde{\psi}_{ij}, \tilde{P}_{ij} \right) \quad (i = 1, 2, \dots, n) \forall j \tag{1}$$

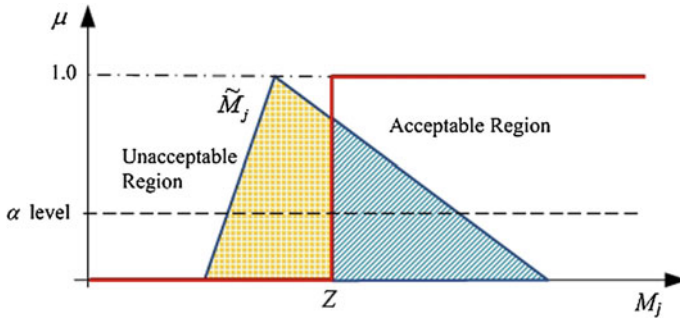


Fig. 3 Fuzzy performance measure and target (or acceptable region)

The membership function of the PM, \tilde{M}_j , given by Eq. (1), is a surface with the possible values of M_j . Under these circumstances arises the difficulty of interpreting the result. Frequently, a defuzzification is performed of the membership function of the performance measurement to obtain a crisp number. In this operation a lot of information is lost that could be relevant to the decision process. Thus the fuzzy result is richer than the crisp number.

Let Z be the target value for the PM, \tilde{M}_j . This value represents the acceptance/rejection region of the decision maker. To capture the uncertainty of the assessment system, the PM, M_j , is represented by a fuzzy number \tilde{M}_j , and the target (or acceptable region) is a crisp value Z (Fig. 3).

In [47], the *Compliance* of two membership functions \tilde{M} and Z was calculated as a fuzzy measure of compatibility (Eq. 2). There are several measures to quantify the compatibility of two fuzzy numbers [48]. The overlapping area between two membership functions (i.e. a fraction of the total area of the PM) represents the concept of *Compliance* better than other compatibility measures, such as possibility and necessity. Thus,

$$Compliance = \frac{\text{overlapping of membership functions of PM and target } Z}{\text{total area of membership function of PM}} \tag{2}$$

The following assumptions are considered when using Compliance (C):

- The maximum of C is equal to 1, and that happens for any level of α -cut such as, $\tilde{M}(\alpha) \geq Z \forall \alpha \in (0, 1]$
- The minimum value of C is equal to 0, and that happens for any level of α -cut, such as $\tilde{M}(\alpha) < Z \forall \alpha \in (0, 1]$
- C provides a consistent ranking to assess the degree to which a fuzzy number complies with target. It is a monotonic increasing function.

5 Conclusion and Future Research

This work provides a classification of uncertainty components that affect the quality of PMs. Can each uncertainty component be decomposed into a systematic and random part? This decomposition will allow the identification of causes that, if changed, could reduce uncertainty.

The first contribution of this work is to provide a general classification of sources of uncertainty that could affect PMs, based on dimensions of information quality. This would allow the establishment of a common theoretical framework to classify uncertainty in the field of Performance Measurement. Secondly, it would provide a basis for practitioners to provide evidence about the uncertainty of existing PMSs.

The second contribution of this work is the creation of a risk indicator associated with a decision that represents the uncertainty the decision maker faces given a PM with uncertainty and a decision criterion.

The development of methods to propagate the uncertainty of the PMs throughout the PMS and through different hierarchic levels is being pursued by the authors in another research project.

This work will be extended to deal not only with the uncertainty in the PMS but also with the uncertainty of the decision criteria.

The representation of uncertainty in performance measurement systems is at the centre of this on-going research. This part of the research presents the findings of this deductive research which will later be tested through case studies, to allow another step of inductive research to support, change or refute the proposed elements for the PMS.

Acknowledgments This work was financed with FEDER Funds by Programa Operacional Fatores de Competitividade—COMPETE and by National Funds by FCT—Fundação para a Ciência e Tecnologia, Project: FCOMP-01-0124-FEDER.

References

1. Kaplan R, Norton D (1992) The balanced scorecard—measures that drive performance. *Harvard Bus Rev* 69: 71–79
2. Bourne M (2004) *Handbook of performance measurement*. Gee Publishing, London
3. Verweire K, Van den Berghe L (2003) Integrated performance management: adding a new dimension. *Manag Decis* 41(8):782–790
4. Pipino LL, Lee YW, Wang R (2002) Data quality assessment. *J Commun ACM* 45(4):211–218
5. Lee YW, Strong DM (2002) AIMQ: a methodology for information quality assessment. *Inf Manag* 40(2):133–146
6. Stvilia B, Gasser L (2007) A framework for information quality assessment. *J Am Soc Inf Sci Technol* 58(12):1720–1733
7. Ge M, Helfert M, Abramowicz W, Fensel D (2008) *Data and information quality assessment in information manufacturing systems*. Springer, Austria
8. Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. *Commun ACM* 39(11):86–95

9. Dai Y, Su Y (2009) Assuring the information quality of production planning and control in Tobacco Industries. Fourth international conference on cooperation and promotion of information resources in science and technology, COINFO '09
10. Madnick SE, Wang RY (2009) Overview and framework for data and information quality research. *J Data Inf Qual* 1(1):1–22
11. Yu C-S (2002) A GP-AHP method for solving group decision-making fuzzy AHP problems. *Comput Oper Res* 29(14):1969–2001
12. Durbach IN, Stewart TJ (2011) An experimental study of the effect of uncertainty representation on decision making. *Eur J Oper Res* 214(2):380–392
13. Sousa SD, Nunes EP, Lopes IS (2012) Data quality assessment in performance measurement. *Lecture notes in engineering and computer science: proceedings of the world congress on engineering, WCE 2012, London, UK, 4–6 July 2012*, pp 1530–1535
14. Juran JM, Godfrey AB (1999) *Juran's quality handbook*. McGraw-Hill, New York
15. Neely A, Gregory M, Platts K (1995) Performance measurement system design. *Int J Oper Prod Manag* 15(4):80–116
16. Macpherson M (2001) Performance measurement in not-for-profit and public-sector organisations. *Meas Bus Excell* 5(2):13–17
17. Seddon J (2002) Changing management thinking. Q2002—a world quality congress—46th EOQ congress, UK
18. Tenner A, DeToro I (1997) *Process redesign*. Addison-Wesley, Harlow
19. Kaplan RS, Norton DP (2001) *The strategy-focused organization*. Harvard Business School Press, Massachusetts, Boston
20. Wilcox M, Bourne M (2003) Predicting performance. *Manag Decis* 41(8):806–816
21. Kanji G, Sá P (2002) Kanji's business scorecard. *Total Qual Manag* 13(1):13–27
22. Bititci U, Turner T, Begemann C (2000) Dynamics of performance measurement systems. *Int J Oper Prod Manag* 20(6):692–704
23. Basu R (2001) New criteria of performance management. *Meas Bus Excell* 5(4):7–12
24. Neely A, Adams C, Kennerley M (2002) *The performance prism: the scorecard for measuring and managing business success*. Financial Times Prentice Hall, London
25. Schalkwyk J (1998) Total quality management and the performance measurement barrier. *TQM Mag* 10(2):124–131
26. Ghalayini A, Noble J, Crowe T (1997) An integrated dynamic performance measurement system for improving manufacturing competitiveness. *Int J Prod Econ* 48:207–225
27. Globerson S (1985) Issues in developing a performance criteria system for an organisation. *Int J Prod Res* 23(4):639–646
28. Franco M, Bourne M (2003) Factors that play a role in "managing through measures". *Manag Decis* 41(8):698–710
29. Batini C, Cappiello C (2009) Methodologies for data quality assessment and improvement. *J ACM Comput Surv* 41(3):1–52
30. Galway LA, Hanks CH (2011) Classifying data quality problems. *IAIDQ's Inf Data Qual Newsl* 7(4):1–3
31. JCGM/WG 1 (2008). *JCGM 100:2008—GUM 1995 with minor corrections—evaluation of measurement data—guide to the expression of uncertainty in measurement*, JCGM
32. Coolen FPA (2004) On the use of imprecise probabilities in reliability. *Qual Reliab Eng Int* 20(3):193–202
33. Nunes E, Faria F, Matos M (2006) Using fuzzy sets to evaluate the performance of complex systems when parameters are uncertain. *Proceedings of safety and reliability for managing risk*, vol 3. Lisbon, pp 2351–2359
34. ISO-1012 (2003) *ISO 10012 measurement management systems—requirements for measurement processes and measuring equipment*, ISO
35. Wazed MA, Ahmed S (2009) Uncertainty factors in real manufacturing environment. *Aust J Basic Appl Sci* 3(2):342–351
36. Mula J, Poler R, García-Sabater J, Lario FC (2006) Models for production planning under uncertainty: a review. *Int J Prod Econ* 103(1):271–285

37. Petrovic D, Roy R, Petrovic R (1999) Supply chain modelling using fuzzy sets. *Int J Prod Econ* 59(103):443–453
38. Herroelen W, Leus R (2005) Project scheduling under uncertainty: survey and research potentials. *Eur J Oper Res* 165(2):289–306
39. Petrovic D (2001) Simulation of supply chain behaviour and performance in an uncertain environment. *Int J Prod Econ* 71(103):429–438
40. Lam K-C, Lam MC-K, Wang D (2008) MBNQA-oriented self-assessment quality management system for contractors: fuzzy AHP approach. *Constr Manag Econ* 26(5):447–461
41. Hu AH, Hsu C-W, Kuo T-C, Wu W-C (2009) Risk evaluation of green components to hazardous substance using FMEA and FAHP. *Expert Syst Appl* 36(3, Part 2): 7142–7147
42. Bashiri M, Hosseini-zhad SJ (2009) A fuzzy group decision support system for multifacility location problems. *Int J Adv Manuf Technol* 42(5–6):533–543
43. Ross T (1995) *Fuzzy logic with engineering applications*. McGraw-Hill, New York
44. Klir GJ (1995) *Fuzzy sets and fuzzy logic: theory and applications* Upper Saddle River. Prentice Hall, N.J
45. Sousa SD, Nunes EP, Lopes I (2011) On the characterisation of uncertainty in performance measurement systems. In: Putnik GD, Ávila P (eds) *Business sustainability 2.0*. Guimarães: School of Engineering, University of Minho; Porto: ISEP, School of Engineering, Polytechnic of Porto, pp 82–88
46. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1(1):3–28
47. Nunes E, Sousa SD (2009) Fuzzy performance measures in a high uncertainty context proceedings of the IX Congreso Galego de Estatística e Investigación de Operacións, Ourense
48. El-Baroudy I, Simonovic P (2003) New fuzzy performance indices for reliability analysis of water supply systems. *Water Resources Research Report*, The University of Western Ontario, Department of Civil and Environmental Engineering

Decision Making of Industrialized Building System: A Supply Chain Perspective on the Influence of Behavioral Economic Factors

Sharifah Akmam Syed Zakaria, Graham Brewer
and Thayaparan Gajendran

Abstract Decision making skills with a wide range of decision scenario are relevant to the members of construction supply chain as decision making processes are essential to be understood in ensuring the success of a building project. Currently, there are likely that few decisional issues in the application of Industrialized Building Systems (IBS) require the invention of a completely new dimension and outlook. There is an increasing use of IBS to substitute conventional construction methods in Malaysia, however, the use of IBS technology is often resisted, apparently on grounds other than simply technological. The aim of this chapter is to present the influence of behavioral economic factors on IBS decision making as perceived by IBS supply chain members in Malaysia. Conclusions are drawn and recommendations are made with respect to the perception of behavioral aspects, socio-economic and IBS technology associated with building projects.

Keywords Behavioral · Building · Construction · Decision · Economic · Technology

1 Introduction

In this study, IBS technology decision is considered as human responses to the direct and indirect effects of new building technology and unpredictability for the purpose of lessening negative consequences or enhancing beneficial consequences and

S.A.S. Zakaria (✉) · G. Brewer · T. Gajendran
School of Architecture and Built Environment, Faculty of Engineering and Built Environment,
University of Newcastle, Callaghan NSW 2308, Australia
e-mail: sharifahakmam.syedzakaria@uon.edu.au

G. Brewer
e-mail: Graham.Brewer@newcastle.edu.au

T. Gajendran
e-mail: Thayaparan.Gajendran@newcastle.edu.au

it presents a series of unique challenges for decision makers in construction industry. Thus, in deciding on Industrialized Building System (IBS) technology adoption, it is also important to understand economic and behavioral factors in construction environment. The focus of this study was on the decision making of IBS technology adoption in construction industry which should be considered as part of the broader topic of construction management. IBS is a construction technique in which components are manufactured in a controlled environment, on or off site, transported, positioned and assembled into a structure with minimal additional site work [1]. IBS or off-site production is the focus of many government and private initiatives to increase the productivity of the building projects and construction industry. Accordingly, IBS technology policy over these years has focused on better understanding the practice behind the scientific and technical aspects of IBS technology itself [2]. Therefore, based on this argument, this chapter is materialized to evaluate the perception of IBS supply chain members based on their knowledge and understanding in relation to the behavioral and economic factors of decision making in the IBS technology adoption of building projects in Malaysia. Moreover, behavioral economics can be applied in technical research on human and social cognitive to better understand technology adoption decisions, particularly on IBS technology [3]. This chapter provides the result of recent works in behavioral economics and technology adoption from the qualitative perspective of IBS decision making. First, the study will be discussing a series of key behavioral economic concepts. Next, the study is intended to determine whether IBS decision makers are concerned and influenced by the behavioral economic factors of IBS decision making.

2 IBS Technology Adoption and Decision Making Process

The current thinking on IBS is that the contractors prefer to choose conventional building system rather than proposing IBS system since the shifting of building system from conventional to IBS is not motivated by cost factors and furthermore, most contractors have been exposed and trained in conventional building system for decades and there is an abundance of cheap foreign workers in Malaysia [4]. Thus, IBS decision should provide for a response to new information as the construction technology unfolds. The unfolding will often occur over a long time scale, a requirement that particularly demands creativity in dealing with the economical and technological context. In making economic choice, Camerer [5] discovered that decision makers are influenced by market and social factors to rationalize and predict a new phenomenon requires one to understand the environment adaptation in addition to the evolutionary of psychology in decision making in terms of variances in economic behavior such as socialization, cultural adaptations and individual differences. There has also been a change in housing construction technology from the conventional system to a wider application of an industrialized building system as the concept of industrialization of the construction industry in Malaysia has been strongly supported by the federal and state governments [6]. Thus, understanding

the nature of decision making has been indirectly recognized as a vital component of IBS technology adoption [7], in which construction stakeholders and supply chain members should decide on building technology matters. In order for these changes to come about, Gomez [8] suggested that less emphasis should be placed on managing isolated facts and concepts, but more emphasis should be placed on extensive and overarching themes including contextual matters and their nature, besides the technology concerns themselves. Although the preference and demand for IBS technology adoption are perceived as ‘promising and upcoming’ but the decisions and actions of IBS are indecisive because there are also huge considerations on contextual factors that might influence IBS decisions, actions, and implementations [9].

3 Economic Factors and Decision Making

The decision making of using modularization technology for construction projects is based on economic aspect as an important factor besides five other influencing factors such as plant location, labor related problems, environmental, and organizational considerations, plant characteristics besides project risks [10]. Moreover, a person’s situational awareness is also critical to the success of a decision process in any dynamic real world [11]. According to Fredholm [12], there are four points of view which sum up important aspects of the problem to make decisions that are: (a) the quality of the decision problems concerning on the type and degree of complexity, the span of time and space besides the type and degree of dynamics (b) decision making as cognitive processes concerning the ways in which the decision maker mentally cope with the problems of decisions (c) the experiences and knowledge of the decision maker based on their experiences and knowledge of the current problem situation, and (d) the complexity of the coordinating contextual concerns which consists of decision making command, control and processes. The whole decision making context in a real situation seems to be a function of these four aspects and each aspect depends on the others. Furthermore many real world decisions are of a dynamic nature in dynamic environments [13].

4 Behavioral Factors and Decision Making

The term behavior refers to anything a person does, typically in response to internal and external events which can be overt actions that are open to view or covert activities that are not openly shown [14]. According to Svenson [15], human decision making involves different stages in decision process and psychological theory reveals that decisions are first based on a problem or post-decision as problems appear and solved at different levels withstand the roughness of the future. In general, decision makers’ evaluation on an item’s value on some dimension is derived from binary comparisons between a small number of choice alternatives available in the

working memory [16]. Besides that, emotional states have powerful effects on how individuals process chance events or evaluate outcome [17] and feeling states are also important modulators of decision making behavior [18]. Therefore, the effect of an internal state change of an individual who subjectively transform probabilities into decision weight and outcomes into values, also depends on the individual aspiration, expectation, and situation [19]. Based on these developments, behavioral economics has now generated a large number of studies showing how descriptive and procedural variables that are psychologically important and actually influence behavior in many settings [20].

5 Construction Supply Chain

In ensuring profitability and competitiveness, supply chain management should be linked to business process to capture synergy of inter and intra company management with the integration of managerial components and behavioral components such as management methods, power, and leadership structure, risk and reward structure, and culture and attitude that are less tangible and difficult to access [21]. According to Faizul [22], in construction industry, IBS supply chain is the management of the current IBS delivery process with the transformation from on-site to off-site activities, so that each process of project execution and implementation must be strategized to reduce risks and bring maximum value. Although construction stakeholders appreciate the introduction of IBS in the construction industry but are not as eager to adopt it, thus creating a need for further understanding of stakeholders in the use of IBS that relates to their environment [23]. Since IBS supply chain members are also the entity of construction stakeholders, the consideration of their perspectives from decision making point of view, is the key element at this level. This is the context that establishes the nature of competitive landscape that indicates to IBS supply chain entities the possible influences of IBS technology adoptions [24]. Therefore, there is a need for IBS supply chain members to look beyond IBS decision per se [25] and to consider the conceptual decision frame and the behavioral economic context of IBS technology adoption.

6 Conceptual Framework of Behavioral Economic Factors in IBS Decision Making

Much work within construction economics is concerned with understanding and modeling the processes and consequences of decision making among construction firms. This is not a straightforward issue of applying behavioral elements in technological and technical construction practices. Behavioral economic is needed to clarify the monetary and social processes underlying IBS technology adoption, to deter-

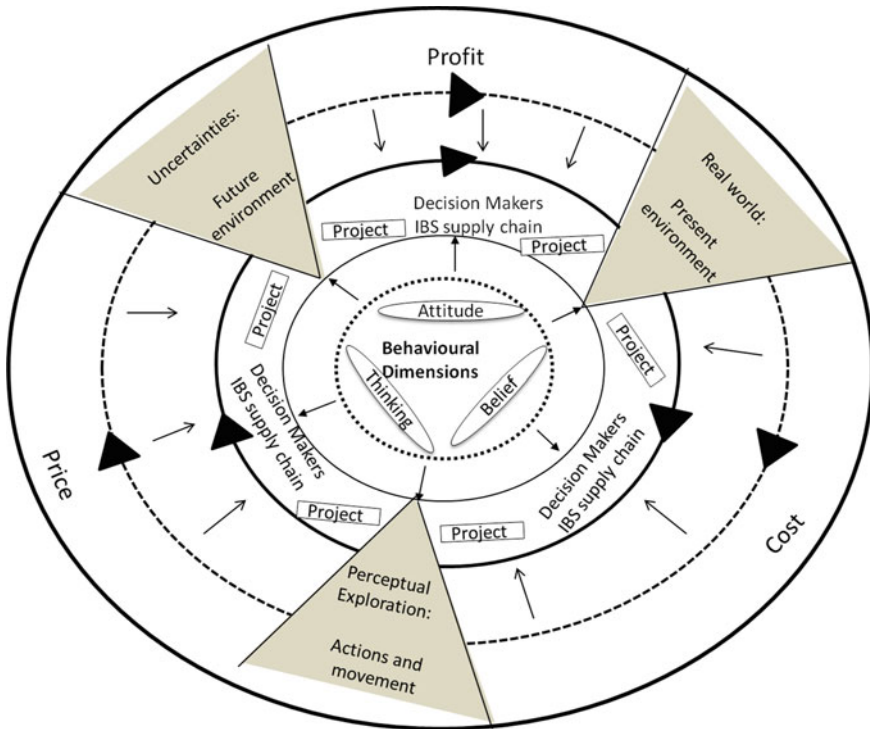


Fig. 1 Contextual framework of behavioral factors in IBS decision making (Adapted from Zakaria [26])

mine the effects of economic influences on IBS decision making, to specify effective behavioral strategies for shifting from the conventional construction methods and to assist policy-makers in formulating and improving IBS technology adoption. Thus, it requires the transformation and integration of behavioral aspects, strategic management, economics, technology management and construction practices knowledge.

A crucial decision making issue in IBS technology adoption can be adequately addressed by incorporating behavioral economic aspects and how best to reveal technology adoption and responses to the policy of construction practices. The usefulness of predicting the behavior of decision makers in construction industry with respect to IBS policy and technology assessment is not well known and models of decision making in this area should be developed for a variety of specific situations. As the science of decision making emerged, it is necessary to understand the aspect of human behavior in order to handle the decision making of technology adoption. As illustrated by Fig. 1, behavioral economic factors in IBS decision making can be represented in a contextual framework with the relationship between the way of thinking and behaving of individuals as a decision makers and their economic environment.

Behavioral economic factors in decision making explore individuals and groups decision making in their social and economic context, by giving major focus to IBS

supply chain members and IBS technology adoption based on their attitudes, belief and thinking, besides other economic aspects such as costs, price and profit. This framework also presents the process of IBS decision making based on the perceptual exploration of decision action and movement, the uncertainties of future environment and the reality of present environment.

7 Research Objectives and Methodology

In this study, it is aimed to fill an existing gap in decision practices in IBS technology and to provide a descriptive viewpoint of how decision makers actually deal with IBS technology adoption. This study intends to:

- (a) Determine the influence of behavioral economic factors in IBS decision based on the perception of decision makers.
- (b) Discover the important elements of behavioral economic factors to be considered into the decision making of IBS technology adoption and
- (c) Synergize the behavioral economic factors as perceived IBS supply chain members in IBS decision making.

Based on the review of published literature and a developed conceptual framework, three key dimensions for each economic and behavioral factor were recognized as being important for this study. As for economic factor, it involves the dimensions of price, cost and profit considerations which influence the decision making of IBS technology adoption. Meanwhile, for behavioral factor, the dimensions are attitude, belief and thinking. The exploratory interviews were semi-structured in nature with two major guiding themes and lasted around an hour. The target interview participants were focused on IBS supply chain members within the Malaysian construction industry based on a performing IBS project, particularly a building project. Nine respondents were identified in this study namely the IBS supply chain members consisting of an architects, a quantity surveyor, a contractor, a civil engineer, a consultant, a developer, a project manager and an IBS manufacturer. The interviews were recorded and transcribed verbatim. The results of the interview were analyzed with a computerized approach using QRS Nvivo, Version 10 and qualitative research methods for data analysis were employed.

8 Results and Discussions

In the discussion of the results, they are presented using two major themes representing behavior and economic factors as set in the interview question with three dimensions respectively.

8.1 Behavioral Factors

The results show that there was consensus among IBS supply chain members on the influence of behavioral factors such as attitude, belief and thinking on IBS decision making. These factors were perceived as contributing to the decision making of IBS technology adoption in building projects. Further results and discussion on behavioral factors are as follows:

8.1.1 Attitude

The dimension of attitude obtained a very high perceived influence on IBS decision making among IBS supply chain members. This situation corresponds to participants who attached high absolute importance to attitude element in IBS technology decisions. The consideration of attitude was certain and the participants suggested that:

- *“When comes to the general public, it is not that easy to convince them about IBS. Their attitudes have been changing and are becoming important in the support of IBS use.”* (Design Architect)
- *“What is important here is the combination of people’s attitudes and mentalities. Based on my experience, all these attitudes and mindset have made IBS implementation slow as there are a lot of minor things to consider.”* (Developer)

This raises the issue as what type of attitude that should be considered in IBS decision making and which one the priority in this consideration. In relation to the type of people attitude, there was agreement among IBS supply chain members that attitude should reflect the people’ commitment, initiative, optimistic and positive which should be changed accordingly as reflected by this statement:

- *“...and they are committed and things really will be in place..... that people are still and we are all living in their own comfort zone..... if there is a will, there is a road for that. That’s the way; it can be implemented in a way. It is the matter of the decision maker to commit themselves to that and then they pass the same commitment.”* (Civil Engineer)

The participants were the most vocal about the importance of attitude in terms of people’s values and opinion in the decision making of IBS technology adoption. These elements appeared to be significant for IBS supply chain members as attitude is an important internal factor that shapes technology choice.

- *“...we can make some suggestions by giving our opinions.”* (Design Architect)
- *“So, of course their opinions are useful in IBS adoption.”* (Project Manager)

In terms of perspectives outside the project or the supply chain, it seemed that other construction players have their own thought on IBS technology adoption. People’ attitude on non-monetary aspect in general, and more specifically in IBS technology

adoption also became a major issue of interest, not only among direct interest group involved in the IBS supply chain, as notified by the participants.

8.1.2 Belief

In terms of beliefs, with regard to principle and trust on IBS technology adoption in building projects, the expectations of IBS supply chain members were that construction players with experience in IBS projects have expressed more concern about their own and other beliefs, since they have to contribute effort and energy toward their own IBS projects as stated by these participants:

- *“...about the significance of IBS in buildings projects. So, again is essential for all stakeholders and IBS supply chain to maintain a belief in itselfthat the project owner is not the only one who believes in the project.....”* (Design Architect)
- *“...more success stories for contractors and consultants to believe and trust IBS.... We believe that the high cost is due to lack of economy of scale in IBS...”* (Contractor)

With regard to IBS technology decisions, beliefs were also explored to determine the concern about feelings, senses and perceptions among IBS supply chain members. It is discovered that all participants in this study evaluated the state of their belief more positively in the decision making of IBS technology adoption.

- *“I always believe that people’s perception is influencing IBS especially when they look at IBS.”* (Quantity Surveyor)
- *“...to start with something, perhaps they will take the term even longer. This is what I strongly feel.in the long run, so we have to make sense that the promotional activity....”* (Civil Engineer)

In summary, all beliefs attribute received a high perceived importance level among these IBS supply chain members. As a consequence, their beliefs are rather unbiased and well-justified, thus what people belief is considered moderately influencing IBS decision making compared to people’s attitudes.

8.1.3 Thinking

Nonetheless, the attribute of thinking received a moderate importance than the attribute of attitudes and beliefs. With respect to thinking attribute, IBS supply chain members perceived the element of thinking to be fundamental in certain areas of IBS decision making. As one design architect stated:

- *“In our situation, the architect would automatically think about IBS construction without thinking much about it during design.”* (Design Architect)

Participants also stated that the outlook or viewpoints of other construction stakeholders were also helpful in making the real IBS decision. The decision makers had to look at each project on a case by case basis with the consideration of other stakeholders' views on IBS technology adoption in project developments as mentioned by the participants:

- *“...developers are more into IBS because for them they view IBS as a cost saving manner.”* (Quantity Surveyor)
- *“.....the client forming a view on the relative importance of matters such as the relationships between key activities,....”* (Design Architect)

In relation to the dimension of thinking, the element of mindset was also perceived by IBS supply chain members as influencing the decision making of IBS technology adoption.

- *“We need to widen people’s mindset, widen their view on what is IBS because they only see IBS is concrete, steel and plastic because of these are not sustainable.”* (Project Manager)
- *“...when people are already exposed to IBS, it is easier to change people’s mindset. Those mindsets that previously prefer the conventional method or the old way have now been changed to IBS.... all these attitudes and mindset have made IBS implementation slow as there are a lot of minor things to consider...”* (Developer)

In summary, the participants also highlighted the significance of mindset in IBS decision making. People’s mindsets have to be revolutionized in changing their attitude towards IBS technology adoption before any key IBS decisions will be made. Although there are variations amongst the IBS supply chain members, as the result of the interviews, it appeared that understanding human nature and its related factors were influential in IBS decision making.

8.2 Economic Factors

One of the main benefits of obtaining information from the IBS supply members who were interviewed was that the most of the participants had been involving in the construction industry for a considerable time with a certain length of experience. They had discovered the importance of economic consideration in the decision making of IBS project. IBS supply chain members in this selected project considered financial aspect, as a part of an economic factor to be important and this was stated in three major dimensions namely cost, price and profit.

8.2.1 Cost

It was interesting to note that in deciding on IBS technology adoption in a building project, cost factor was often a most contentious issue and involved significant

considerations among IBS supply chain members. One member felt that additional cost had been a kind of investment in IBS technology adoption. One civil engineer stated that:

- “...there will be some kind of investment in addition to the cost.... there is a cost factor that I have to consider that part as well the potential return in the long run,.... in a way they can help to share the cost with their clients” (Civil Engineer)

Increasingly, there was seen to be a need for IBS supply chain members who are directly or indirectly involved in IBS decisions to determine and verify cost information, particularly to compare project costs between IBS technology adoption and conventional building method. The need for this has grown rapidly in recent times, with many participants realized that many construction stakeholders needing to justify cost-effective aspect and requiring more valid and reliable IBS cost information, as mentioned by this participant:

- “I estimate the cost and then make the comparison between the conventional and these types of IBS... we have to be based on the amount and form of cost data and quantitative techniques to be adopted before we can propose the estimations for any decision making... construction is cost a factual process designed to give a reliable estimation or prediction... on budget through cost management.” (Quantity Surveyor)

Cost factor was seen to be the most important means for IBS supply chain members to adopt IBS technology to gain competitive and comparative advantages in building projects. The aspects of cost control, cost savings, cost estimation, cost comparison, cost effectiveness and cost management were perceived as crucial elements of any IBS decisions.

8.2.2 Price

While the cost was considered as most important factor from the economic standpoint of IBS decision making, price factor was also regarded as significant by IBS supply chain members. The price factor was also verified because it is perceived as being a project priority and a competitive source as stated by two participants:

- “...cost and price are always the priority,... the high cost is due to lack of economy of scale in IBS projects..” (Contractor)
- “...consider offering a competitive price in order for them to be sustainable in the construction industry.” (Project Manager)

Price factor was particularly important for the supply chain members of an IBS project that had a diverse industry exposure often having exceptional project requirements. Besides the price factor itself, IBS decision making was also influenced by the value aspect of pricing.

- *“...IBS benefits should be viewed from the perspectives of price, cost and value.”* (Design Architect)
- *“...the commissioning of option appraisals, analysis of outcomes and choice of the best option to ensure best value for money is obtained.”* (Client)

Thus, the overall price of a building project was a key consideration. However, equally important was the price of IBS components or building materials. In summary, the price element of economic factors was profoundly affected IBS decision making as revealed by one developer:

- *“In terms of economic factors, they are also helpful in our decisions. For instance, if the price of product is too high, we are not able to adopt IBS because we have to consider steel, to make moulds and so on.”* (Developer)

8.2.3 Profit

Profit was claimed as another important influencing factor in IBS decision making as perceived by IBS supply chain members. In the increasingly competitive and uncertain business environment today, project developments require a profit oriented approach in decision making as revealed by one developer.

- *“Usually, for government projects they go for IBS because of many factors but for private projects, their decisions are mainly due to cost factors as they are more profit oriented.”* (Developer)

In terms of monetary return from the IBS technology adoption, IBS supply chain members also perceived profit margin as crucial in project developments. Generally, the cost of IBS supplies may increase at any given time, leading to a lower profit margin, vice versa. The participants stated that:

- *“...extra cost is being added up for main contractor’s profit margin.”* (Contractor)
- *“...if the construction implements IBS and can have real profit...”* (Design Architect)

In summary, profit is one of the primary indicators used by IBS decision makers to ascertain a project or a company’s performance. It can be concluded that economic factors such as cost, price and profit in this study have influenced the decision making of IBS technology adoption among IBS supply chain members.

From the results, it can be presumed that the concerns of behavioral and economic factors are likely in practice when deciding on IBS technology decisions. However, the influence of economic factors on IBS decision making were slightly significant compared to behavioral factors. This is due to the fact that technology decisions are more influenced by economic factors due to the nature and dynamics of building project developments in the construction industry.

Although IBS technology or off-site manufacturing can contribute to a change in the industry, it depends on behavioral changes in order to be widely adopted.

The outcome of this study suggests that there are few areas that could be explored for future research and development. The method of survey conducted in this research was specific and prepared for certain targeted participants. Thus, for future research, a survey combining both personal interviews and questionnaires could be extended to a larger sized target respondent that will give more comprehensive outlook for the immediate and future course of action in IBS decision making.

9 Conclusions

It can be concluded that the influence of behavioral factors in IBS decision making can contribute towards explaining the drivers or barriers of IBS technology adoption that are manifesting in IBS decision makers and also how it can assist in identifying operational IBS requirements and guidelines measures to facilitate the best or good practice in IBS decision making. By considering behavioral economic factors in IBS decision making, this kind of decision principle tool can help clarify non-technical matters and provide a more practical framework for IBS decision making. Yet, despite its limitations, IBS decision based on behavioral economic factors can still perform as a useful function in the progress and growth of IBS technology adoption. The consideration of behavioral economic factors in IBS decision suggests that decision makers in construction industry should be aware of a number of non-technical concerns when considering IBS adoption decisions.

Acknowledgments S. A. S. Zakaria: Author thanks Universiti Sains Malaysia for the scholarship of this study and University of Newcastle, Australia for the support.

References

1. CIDB (2007) IBS digest at Malbex in IBS digest, Special issues on 24th Malaysian international building: construction industry development board (CIDB)
2. CIDB (2010) Industrialized building system (IBS) roadmap 2011–2015. Construction Industry Development Board (CIDB), Kuala Lumpur
3. Zakaria SAS, Brewer G, Gajendran T (2012) Behavioral economics perspective in exploring the decision making of industrialized building systems in Malaysia. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012, UK, London, pp 1421–1426, 4–6 July 2012
4. Kadir MRA, Lee WP, Jaafar MS, Sapuan SM, Ali AAA (2006) Construction performance comparison between conventional and industrialized building systems. *Malays Struct Surv* 24(5):412–424
5. Camerer C (1988) Bounded rationality in individual decision making. *Exp Econ* 1:163–183
6. Agus MA (2002) The role of state and market in the Malaysian housing sector. *J Hous Built Environ* 17:49–67
7. Abdullah MR, Egbu C (2010) Selection criteria framework for choosing industrialized building systems for housing projects. In: Proceedings of the 26th annual ARCOM conferenc., Leeds, UK, pp 1131–1139

8. Gomez CP (2006) Reconceptualizing the management of technology in construction: a Malaysian perspective. In: Proceedings of the 22th annual ARCOM conference, Leeds, UK, pp 781–788
9. Zakaria SAS, Brewer G, Gajendran T (2012) Contextual factors in the decision making of industrialized building system technology. *World Acad Sci Eng Technol* 67:489–497
10. Gupta VK, Fisher DJ, Murtaza MB (1996) A consortium sponsored knowledge-based system for managerial decision making in industrial construction. *Interfaces* 26(6):9–23
11. Singh H, Petersen LA, Thomas EJ (2010) Understanding diagnostic errors in medicine: a lesson from aviation. *Qual Saf Health Care* 15:159–164
12. Fredholm L (1997) Decision making in major fire fighting and rescue operation. In: Fin R, Salas E, Strub M, Martin L (eds) *Decision making under stress*. Ashgate Publishing, England
13. Kerstholt JH (1997) Dynamic decision making in situations. In: Fin R, Salas E, Strub M, Martin L (eds) *Decision making under stress*. Ashgate Publishing, England
14. Sarafino EP (1996) *Principles of behaviour change*. Wiley, New York
15. Svenson O (1992) Differentiation and consolidation theory of human decision making: a frame of reference for the study of pre- and post-decision processes. *Acta Psychol* 80:143–168
16. Steward N, Charter N, Baron GDA (2006) Decision by sampling. *Cogn Psychol* 53:1–26
17. Loewenstein GF, Bazerman MH, Thompson L (1989) Social utility and decision making in interpersonal contexts. *J Pers Soc Psychol* 57(3):426–441
18. Montague PR, King-Kasas B, Cohen JD (2006) Imaging valuation models in human choice. *Ann Rev Neurosci* 29:417–448
19. Kahneman D, Tversky A (1984) Choices, values and frames. *Am Psychol* 39:341–350
20. DellaVigna S (2009) Psychology and economics: evidence from the field. *J Econ Lit* 47:315–372
21. Lambert DM, Coopers MC (2000) Issues in supply chain management. *Ind Mark Manag* 29(1):65–83
22. Faizul NA (2006) Supply chain management in construction industry, Malaysian international IBS exhibition, MIIE 2006, Kuala Lumpur, Malaysia, 21–24 Nov 2006
23. Onyeizu EN, Bakar AHA (2011) Assessing key factors in design in the industrial building systems (IBS) approach: stakeholders' opinion in Malaysia. *Int J Acad Res* 3(4):168–175
24. Kamar KAM, Hamid ZA (2011) Supply chain strategy for contractor in adopting industrialized building system (IBS). *Aust J Basic Appl Sci* 5(12):2552–2557
25. Shukor ASA, Mohammad MF, Mahbub R, Ismail F (2011) Supply chain integration in industrialised Building System in the Malaysian construction industry. *Built Hum Environ Rev* 4(1):108–121
26. Zakaria SAS, Brewer G, Gajendran T (2011) Psychology in the decision making of industrialized building systems (IBS): a field of application. In: Proceedings of the international academic forum of Asian conference on psychology and behavioral science, ACP, Osaka, Japan, pp 68–78, 20–22 March 2011