Advances in Intelligent Systems and Computing 662

Radek Silhavy
Petr Silhavy
Zdenka Prokopova *Editors*

# Applied Computational Intelligence and Mathematical Methods

## Computational Methods in Systems and Software 2017, vol. 2

Springer

**Advances in Intelligent Systems and Computing**

Volume 662

## *About this Series*

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Radek Silhavy · Petr Silhavy
Zdenka Prokopova
Editors

# Applied Computational Intelligence and Mathematical Methods

## Computational Methods in Systems and Software 2017, vol. 2

Springer

*Editors*

Radek Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlín
Czech Republic

Zdenka Prokopova
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlín
Czech Republic

Petr Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlín
Czech Republic

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-67620-3 ISBN 978-3-319-67621-0 (eBook)
DOI 10.1007/978-3-319-67621-0

Library of Congress Control Number: 2017952853

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This book constitutes the refereed proceedings of the Computational Methods in Systems and Software 2017 (CoMeSySo 2017), which will be held on September 2017.

CoMeSySo 2017 conference intends to provide an international forum for the discussion of the latest high-quality research results in all areas related to intelligent systems. The addressed topics are the theoretical aspects and applications of Software Engineering in Intelligent Systems, Cybernetics and Automation Control Theory, Econometrics, Mathematical Statistics in Applied Sciences and Computational Intelligence.

CoMeSySo 2017 has received (all sections) 129 submissions, 70 of them were accepted for publication. More than 40% of accepted submissions were received from Europe, 30% from Asia, 17% from Africa, and 12% from America. Researches from 27 countries participated in CoMeSySo conference.

The volume Applied Computational Intelligence and Mathematical Methods brings new approaches and methods to real-world problems and exploratory research that describes novel approaches in the mathematical methods, computational intelligence methods, statistics, and software engineering in the scope of the intelligent systems.

The editors believe that readers will find following proceedings interesting and useful for their own research work.

July 2017

Radek Silhavy
Petr Silhavy
Zdenka Prokopova

# Organization

## Program Committee

## Program Committee Chairs

| | |
|---|---|
| Petr Silhavy | Department of Computers and Communication Systems, Faculty of Applied Informatics, Tomas Bata University in Zlin, Czech Republic |
| Radek Silhavy | Department of Computers and Communication Systems, Faculty of Applied Informatics, Tomas Bata University in Zlin, Czech Republic |
| Zdenka Prokopova | Associate Professor at Department of Computers and Communication Systems, Tomas Bata University in Zlin, Czech Republic |
| Krzysztof Okarma | Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland |
| Roman Prokop | Department of Mathematics, Tomas Bata University in Zlin, Czech Republic |
| Viacheslav Zelentsov | Doctor of Engineering Sciences, Chief Researcher of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), Russian Federation |
| Lipo Wang | School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore |
| Silvie Belaskova | Head of Biostatistics, St. Anne's University Hospital Brno, International Clinical Research Center, Czech Republic |

## International Program Committee Members

| | |
|---|---|
| Pasi Luukka | North European Society for Adaptive and Intelligent Systems & School of Business and School of Engineering Sciences Lappeenranta University of Technology, Finland |
| Ondrej Blaha | Louisiana State University Health Sciences Center New Orleans, New Orleans, USA |
| Izabela Jonek-Kowalska | Faculty of Organization and Management, The Silesian University of Technology, Poland |
| Maciej Majewski | Department of Engineering of Technical and Informatic Systems, Koszalin University of Technology, Koszalin, Poland |
| Alena Vagaska | Department of Mathematics, Informatics and Cybernetics, Faculty of Manufacturing Technologies, Technical University of Kosice, Slovak Republic |
| Boguslaw Cyganek | Department of Computer Science, University of Science and Technology, Krakow, Poland |
| Piotr Lech | Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland |
| Monika Bakosova | Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology, Bratislava, Slovak Republic |
| Pavel Vaclavek | Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech Republic |
| Miroslaw Ochodek | Faculty of Computing, Poznan University of Technology, Poznan, Poland |
| Olga Brovkina | Global Change Research Centre Academy of Science of the Czech Republic, Brno, Czech Republic |
| Elarbi Badidi | College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates |
| Gopal Sakarkar | Shri. Ramdeobaba College of Engineering and Management, Republic of India |
| V.V. Krishna Maddinala | GD Rungta College of Engineering and Technology, Republic of India |
| Anand N. Khobragade | Maharashtra Remote Sensing Applications Centre, Republic of India |

Abdallah Handoura — Computer and Communication Laboratory, Telecom Bretagne – France

## Organizing Committee Chair

Radek Silhavy — Tomas Bata University in Zlin, Faculty of Applied Informatics

## Conference Organizer (Production)

OpenPublish.eu s.r.o.

Web: http://comesyso.openpublish.eu
Email: comesyso@openpublish.eu

## Conference Website, Call for Papers

http://comesyso.openpublish.eu

# Contents

**Spatially Augmented Analysis of Macroeconomic Convergence with Application to the Czech Republic and Its Neighbors** . . . . . . . . . . . 1
Tomáš Formánek

**Trend-Cycle Decomposition of Economic Activity in the Czech Republic** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
Ondřej Čížek

**Parallel Matrix Multiplication for Business Applications** . . . . . . . . . . . . . 24
Mais Haj Qasem and Mohammad Qatawneh

**Content Generation for Massively Multiplayer Online Games with Genetic Algorithms** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
Tiago Alves, Jorge Coelho, and Luís Nogueira

**An Autonomous Architecture for Managing Vertical Elasticity in the IaaS Cloud Using Memory Over-Subscription** . . . . . . . . . . . . . . . 50
Bouaita Riad, Zitouni Abdelhafid, and Maamri Ramdane

**A Security Framework for Cloud Data Storage(CDS) Based on Agent** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
Oussama Arki and Abdelhafid Zitouni

**Proposal for the Design of a New Technological Infrastructure for the Efficient Management of Network Services and Applications in a High Complexity Clinic in Colombia** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
Leonel Hernandez, Humberto Villanueva, and Sandra Estrada

**Initial Centroid Selection Optimization for K-Means with Genetic Algorithm to Enhance Clustering of Transcribed Arabic Broadcast News Documents** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
Ahmed Mohamed Maghawry, Yasser Omar, and Amr Badr

**An Imperialist Competitive Algorithm to Solve the Manufacturing Cell Design Problem** . . . . . . . . 102
Ricardo Soto, Broderick Crawford, Rodrigo Olivares,
Héctor Ortega, and Boris Almonacid

**Optical Character Recognition System for Czech Language Using Hierarchical Deep Learning Networks** . . . . . . . . 114
Arindam Chaudhuri and Soumya K. Ghosh

**A Percentile Transition Ranking Algorithm Applied to Knapsack Problem** . . . . . . . . 126
José García, Broderick Crawford, Ricardo Soto, and Gino Astorga

**SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments** . . . . . . . . 139
Hichem Rahab, Abdelhafid Zitouni, and Mahieddine Djoudi

**Enrichment Ontology Instance by Using Data Mining Techniques** . . . . . 150
Kunyanuth Kularbphettong

**Solving the Set Covering Problem Using Cat Swarm Optimization Algorithm with a Variable Mixture Rate and Population Restart** . . . . . . 156
Broderick Crawford, Ricardo Soto, and Hugo Caballero

**Analysis of Students' Behavior Based on Educational Data Mining** . . . . 167
Kunyanuth Kularbphettong

**Inference Algorithms in Latent Dirichlet Allocation for Semantic Classification** . . . . . . . . 173
Wan Mohammad Aflah Mohammad Zubir, Izzatdin Abdul Aziz,
Jafreezal Jaafar, and Mohd Hilmi Hasan

**Between Data Mining and Predictive Analytics Techniques to Cybersecurity Protection on eLearning Environments** . . . . . . . . . . . . 185
José Manuel, Raul Cordeiro, and Carla Silva

**A Performance Evaluation of Chi-Square Pruning Techniques in Class Association Rules Optimization** . . . . . . . . 195
Han Chern-Tong and Izzatdin Abdul Aziz

**A Hybrid Method Based on Intelligent Water Drop Algorithm and Simulated Annealing for Solving Multi-depot Vehicle Routing Problem** . . . . . . . . 204
Absalom E. Ezugwu, Micheal O. Olusanya, and Aderemi O. Adewumi

**Information Retrieval Based on the Extracted Social Network** . . . . . . . . 220
Mahyuddin K.M. Nasution, Rahmad Syah, and Maria Elfida

**Strategic Decision Method Structured in SWOT Analysis and Postures Based in the MAGIQ Multicriteria Analysis** . . . . . . . . . . . 227
Sergio Alexandre Barreira Forte,
Sergio Henrique Arruda Cavalcante Forte, and Placido Rogério Pinheiro

**Automatic Structuring of Arabic Normative Texts** . . . . . . . . . . . . . . . . . . 238
Ines Berrazega and Rim Faiz

**Computer Aided Analysis of the Mobile Crane Handling System Using Computational Intelligence Methods** . . . . . . . . . . . . . . . . . . . . . . . . 250
Wojciech Kacalak, Zbigniew Budniak, and Maciej Majewski

**The Study of the Impact of Technological Innovation Network on Dual Innovation** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 262
Cheng Song, Longying Hu, and Haiyan Yuan

**Concept of Econometric Intelligence System: OLAP Applications in the Ambient Intelligence Environment** . . . . . . . . . . . . . . . . . . . . . . . . . 279
Jan Tyrychtr, Martin Pelikán, Hana Štiková, and Ivan Vrana

**Earthquake Ground Motion Attenuation Modeling Using Levenberg-Marquardt and Brute-Force Method** . . . . . . . . . . . . . . . . . . . 287
Edy Irwansyah, Bayu Kanigoro, Priscilia Budiman, and Rokhana D. Bekti

**Advanced Approach for Observability of Distributed Systems Using Internal Pointwise Sensor** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 298
Amine Bouaine and Mostafa Rachik

**Using the Method of System Dynamics to Forecast Additional Manpower Needs in Murmansk Region** . . . . . . . . . . . . . . . . . . . . . . . . . 310
Vitaliy Bystrov, Svetlana Malygina, and Darya Khaliullina

**Ordinary Kriging and Spatial Autocorrelation Identification to Predict Peak Ground Acceleration in Banda Aceh City, Indonesia**. . . . . 318
Rokhana D. Bekti, Edy Irwansyah, Bayu Kanigoro, and Theodorick

**Multicriteria Problem Structuring for the Prioritization of Information Technology Infrastructure Problems** . . . . . . . . . . . . . . . . . 326
Carolina Ferreira Gomes Silva, Plácido Rogério Pinheiro,
and Odecília Barreira da Silva Benigno

**Agent Based Modelling Approach of Migration Dynamics** . . . . . . . . . . . 338
Samira Boulahbel-Bachari, Nadjia El Saadi, and Alassane Bah

**Applying of the Classifications Trees Method in Forecasting of Risk Groups of Intolerant Behavior** . . . . . . . . . . . . . . . . . . . . . . . . . 350
I.V. Vicentiy, S.M. Eliseev, and A.V. Vicentiy

**Prediction of Attacks Against Honeynet Based on Time Series Modeling** . . . . . 360
Pavol Sokol and Andrej Gajdoš

**The Cognitive Approach to the Coverage-Directed Test Generation** . . . . 372
Anna Klimenko, Galina Gorelova, Vladimir Korobkin, and Petr Bibilo

**An Agent Based Model to Study the Impact of Intra-annual Season's Variability on the Dynamics of Aedes Vexans and Culex Poicilipes Mosquito Populations in North Senegal (Ferlo)** . . . . . 381
Python Ndekou Tandong Paul, Alassane Bah,
Papa Ibrahima Ndiaye, and Jacques André Ndione

**Author Index** . . . . . 393

# Spatially Augmented Analysis of Macroeconomic Convergence with Application to the Czech Republic and Its Neighbors

Tomáš Formánek(✉)

University of Economics, Prague, Czech Republic
formanek@vse.cz

**Abstract.** This paper deals with macroeconomic convergence at the NUTS2 level for the following six countries: Czechia, Slovakia, Poland, Hungary, Germany and Austria. Prominent spatial dependencies are identified and compared to the Solow-Swan type convergence. The estimation and testing is performed using spatial panel data methodology. At the theoretical and empirical level, properties and performance of spatial panel models are compared with classical cross-sectional and panel (non-spatial) approaches. Given the variety of available frameworks of modeling and estimation of spatial dependencies, significant proportion of this paper is devoted to model specification and robustness analysis issues. Also, topics relevant for appropriate interpretation of the estimated spatio-temporal models are included.

**Keywords:** Macroeconomic convergence · Spatial dynamics · Panel data

## 1 Introduction

Macroeconomic convergence is often studied in terms of GDP per capita dynamics. Such approach is based on the neoclassical Solow-Swan model of long run growth and the corresponding analysis framework provided by Mankiw et al. [8]. Their approach leads to a convenient and empirically testable "$\beta$-convergence" model that estimates and evaluates the presumed inverse relationship between the growth rate of per capita output over a finite time period and the output level at the beginning of the period. The underlying hypothesis for $\beta$-convergence is quite simple: we assume that poorer economies take advantage of their potential and grow faster than the richer ones. In the long-run, this leads to wealth equalization among originally heterogeneous economies. Although the intuition behind $\beta$-convergence may be simple, empirically we are dealing with complex processes, prone to diverse types of shocks and influences. Such models can be approached and studied from many perspectives, e.g. focusing on different types of assumptions relevant for the growth convergence models, as discussed e.g. in [8]. In this contribution, the focus is on spatio-temporal aspects of macroeconomic convergence.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_1

In recent growth literature, there is a prominent turn from cross-country analyses towards the sub-national scale: see Piras and Arbia [12] for examples and an exhaustive list of references. At the regional scale, the closed-economy paradigm as in [8] is no longer appropriate - regional economies typically operate as prominently open and interconnected. For the EU regions analyzed in this article, three main drivers of convergence through regional interactions may be pointed out: Unification is institutionalized and incorporated in most EU policies. Also, factor mobility (labor, capital) and trade relations play an increasingly important role. Finally, technology & knowledge diffusion processes provide a positive push to poorer regions.

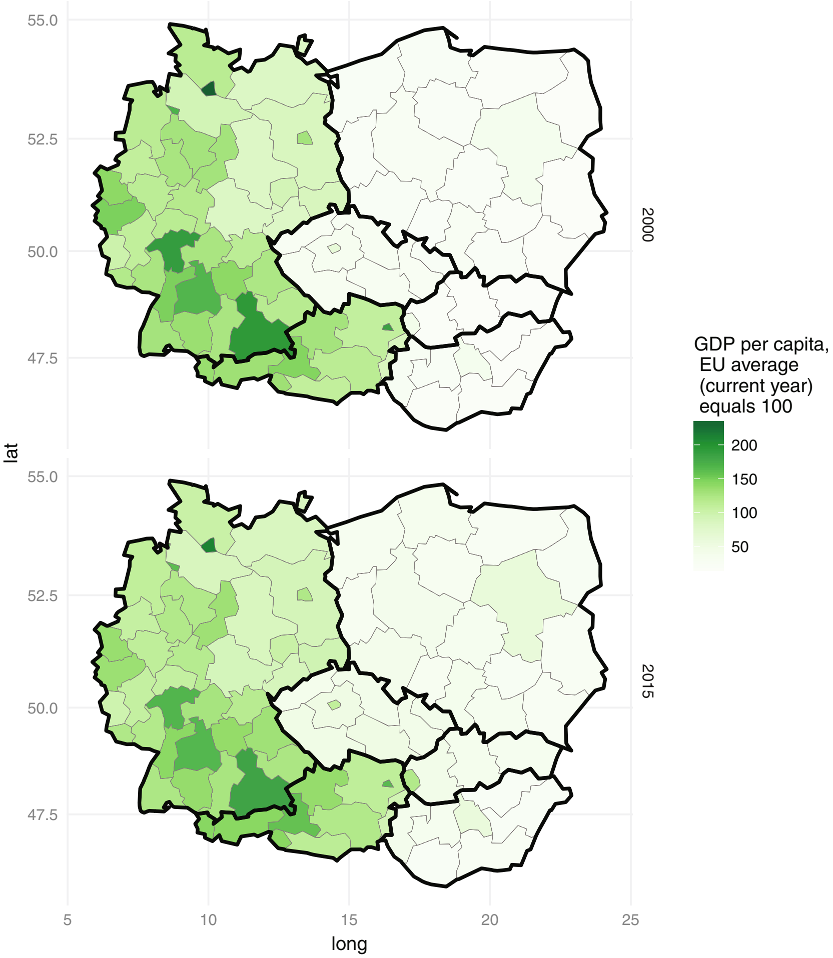


**Fig. 1.** Choropleths with 2000 & 2015 relative GDP per capita - NUTS2 level

Theoretically, the best way to control for such regional interactions would be to directly include labor, capital and goods movements, etc. into the growth models. In practical terms, such approach is impossible due to data availability issues, especially with variables such as capital flows and technology diffusion. Here, spatial panel data methods may provide an indirect, yet feasible and reliable framework to regional growth and convergence analyses.

Spatial panel data methods - if properly applied - can correct for the inherent bias in classical cross-sectional growth models (see [12] for discussion) as follows: the bias generated by regional differences is controlled for by the explicit inclusion of regional effects (individual heterogeneities within the panel data paradigm as in [14]) to the model. Spatial interdependencies are also explicitly accounted for (see [4] or [7]). Moreover, spatial panel approach allows us to accurately differentiate between the two effects (panel and spatial).

This paper provides a thorough integration of spatial modeling to the panel data-based analysis of regional convergence dynamics in terms of GDP per capita. In contrast with previous attempts in this field of research (see [12]), proper interpretation of the ceteris-paribus effects is used here (see discussion provided in the next section). The theoretical part of spatial panel analysis is accompanied by a $\beta$-convergence model describing regional growth dynamics at the NUTS2 level for the following six countries: Czechia, Slovakia, Poland, Hungary, Germany and Austria. For illustration of the convergence process, Fig. 1 compares relative GDP per capita levels as of 2000 and 2015: prominent & stable spatial patterns are apparent, while the presumed time convergence is not quite visually identifiable.

The remainder of this paper is structured as follows: Section two covers key methodological topics of the spatial panel approach along with references to fundamental literature. Section three provides an application to the convergence topics outlined. Section four and the list of references conclude.

## 2 Spatial Econometrics and Spatial Panel Data Methods

Spatial econometric models address the presence of effects such as economic spill-overs between neighboring regions. In spatial econometrics, data need to be geo-coded by means of the latitude/longitude geographic coordinates system, as distances (and/or common borders) are used to estimate spatial dependencies. The variety of available approaches towards modeling and estimation of spatial dependencies implies that researchers usually have to consider several spatial structure settings to evaluate model efficiency and robustness; see e.g. [4] or [13] for theoretical and empirical contributions to the field.

Due to different research scopes (say, ecology vs. housing prices), spatial dependency definitions may differ. However, spatial interactions play an important role and we need to adjust our methodologies to incorporate spatial autocorrelation: the extent to which a value of a given attribute in one location depends on values observed in nearby locations. Moran [11] and Geary [6] are often cited as the founding fathers of spatial econometrics, yet the actual framework for contemporary spatial econometrics was established by Cliff and Ord [3], by introducing a flexible spatial weights specification. Spatial weights are usually calculated in a two-step approach: First, a square spatial matrix $\boldsymbol{S}_N$ is established for a given set of $N$ spatial (geo-coded) units. Next, a corresponding spatial weights matrix $\boldsymbol{W}_N$ is constructed for use in spatial models such as (3).

Spatial matrices and neighbor definitions are based on dummy variables: the $s_{ij}$ elements of $\boldsymbol{S}_N$ equal 1 if the two spatial units $i$ and $j$ are neighbors and 0 otherwise. Usually, two units are considered neighbors if they are located sufficiently near each other. Diagonal elements of $\boldsymbol{S}_N$ are set to zero by definition (units are not neighbors to themselves). Formally, individual elements of the symmetrical spatial matrix $\boldsymbol{S}_N$ may be defined as follows:

$$s_{ij} = s_{ji} = \begin{cases} 0 & \text{if} \quad i = j, \\ 0 & \text{if} \quad d_{ij} > t, \\ 1 & \text{if} \quad d_{ij} \leq t, \end{cases} \tag{1}$$

where $d_{ij}$ is some adequate measure of distance between units' representative location points (centroids) and $t$ is an ad-hoc defined maximum neighbor distance threshold. Usually, pure geographical distances are used in (1), yet technological or other convenient dimensions (highway/railway infrastructure, work-commuting intensities, etc.) are often applied according to the particular research focus. The elements of $\boldsymbol{S}_N$ are determined by the ordering of the data (spatial units), which can be arbitrary.

Elhorst [4] provides two formal stability conditions for spatial models that may be restated as follows: (a) The row and column sums of any $\boldsymbol{S}_N$ matrix should be uniformly bound in absolute value as $N \to \infty$. (b) The row and column sums of $\boldsymbol{S}_N$ should not diverge to infinity at a rate equal to or faster than the rate of sample size growth. Both conditions reflect the "spatial weak dependency" assumption that correlation between two spatial units should converge to zero as their distance increases to infinity. Finally, it should be noted that $\boldsymbol{S}_N$ is generally assumed to be time-invariant. Also, many possible alternative approaches to $\boldsymbol{S}_N$ construction exist (contiguity, $k$NN, etc.) - see e.g. [1] or [2].

Once the $\boldsymbol{S}_N$ matrix is established, its corresponding spatial weights matrix $\boldsymbol{W}_N$ is often constructed by simple row-standardizing, so that the row weights sum up to 1. Before estimating spatial econometric models, we need to apply preliminary tests for spatial autocorrelation in the observed variables. Many spatial autocorrelation test statistics are available from [1], yet Moran's $I$ seems to be the most widely used:

$$I_{y_t} = \frac{N}{S} \ddot{\boldsymbol{y}}_t' \boldsymbol{W}_N \ddot{\boldsymbol{y}}_t (\ddot{\boldsymbol{y}}_t' \ddot{\boldsymbol{y}}_t)^{-1}, \tag{2}$$

where $\ddot{\boldsymbol{y}}_t$ is the centered vector of the $N$ spatial observations of some variable $y$ under scrutiny at time $t$. $S$ is the standardization factor, corresponding to the sum of all elements of the spatial weights matrix $\boldsymbol{W}_N$. The expected value of Moran's $I$ under the null hypothesis of no spatial autocorrelation is: $-1/(N-1)$. Following [1], we can use $var(I_{y_t})$ to calculate a $z$-score and test for statistical significance: whether neighbor units are more similar to one another than they would be under spatial randomness. The sign of Moran's $I$ discriminates between positive and negative spatial autocorrelation.

Once significant spatial dependence in observed data is identified, spatial regression may be used to account for such situation. Specifically, individual observations of some variable $y_{it}$ (i.e. variable $y$ observed at $i$-th geo-unit at time $t$) may be expressed in terms of weighted averages of their neighbors' values: Given the row-standardized $\boldsymbol{W}_N$, we may write $\mathit{SpatialLag}(y_{it}) = \boldsymbol{w}_i \boldsymbol{y}_t$, i.e. the spatially defined expected value of $y_{it}$ is a product of the $i$-th row of $\boldsymbol{W}_N$ and the observed neighboring $\boldsymbol{y}_t$ values. Additional discussion on spatial autocorrelation is provided in [1] and [4].

As the observed geo-coded data are often replicated in time, spatial panel models can be used to depict interactions among variables across spatial units as well as over time. The application part of this paper is based on a general form of a static panel model that includes both the spatial effects (spatial lags) for the dependent variable and the spatially autocorrelated error terms. Such model may be outlined as

$$\boldsymbol{y} = \lambda \left( \boldsymbol{I}_T \otimes \boldsymbol{W}_N \right) \boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{3}$$

where $\boldsymbol{y}$ is a $NT \times 1$ column vector of dependent variable observations ($i = 1, 2, \ldots, N$ denotes cross-sectional units and $t = 1, 2, \ldots, T$ relates to the time dimension). $\boldsymbol{X}$ is a $NT \times k$ matrix of $k$ exogenous regressors, it has full column rank and its elements are uniformly bounded in their absolute values. $\boldsymbol{I}_T$ is an identity matrix and the elements of vector $\boldsymbol{\beta}$ as well as $\lambda$ are parameters of the model. Given the space limitations for this contribution, only random effects (RE) model/estimation approach is briefly discussed. Spatial pooling and fixed effects (FE) model estimation approaches are outlined e.g. in [4] and [9]. With spatial RE models, we implicitly assume that the unobserved individual effects are not correlated with other regressors. Although this is a very strong assumption, one can simply adopt the Mundlak-Chamberlain method of dealing with correlated random effects (CRE) - see [10] for detailed discussion of the CRE estimation applied to spatial panels. Hence, using the notation in [9], the error term $\boldsymbol{u}$ from (3) and its variance may be described as

$$\boldsymbol{u} = \left( \boldsymbol{\iota}_T \otimes \boldsymbol{I}_N \right) \boldsymbol{\mu} + \left( \boldsymbol{I}_T \otimes \boldsymbol{B}_N^{-1} \right) \boldsymbol{v}, \tag{4}$$

$$\boldsymbol{\Omega}_u = var\left( \boldsymbol{u} \right) = \sigma_\mu^2 \left( \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \otimes \boldsymbol{I}_N \right) + \sigma_v^2 \left[ \boldsymbol{I}_T \otimes \left( \boldsymbol{B}_N' \boldsymbol{B}_N \right)^{-1} \right], \tag{5}$$

where $\boldsymbol{\iota}_T$ is a unit vector ($T$×1) and $\boldsymbol{I}_N$ is an identity matrix. Vector $\boldsymbol{\mu}$ holds the time-invariant and spatially uncorrelated individual effects with $\mu_i \sim \mathit{IID}\left(0, \sigma_\mu^2\right)$. $\boldsymbol{B}_N = \left( \boldsymbol{I}_N - \rho \boldsymbol{W}_N \right)$ is assumed non-singular and features a spatial error autoregressive parameter $\rho$ where $|\rho| < 1$. $\boldsymbol{v}' = \left( \boldsymbol{v}_1', \ldots, \boldsymbol{v}_T' \right)$ is a vector of innovations that vary both over cross-sectional units and across time with $v_{it} \sim \mathit{IID}\left(0, \sigma_v^2\right)$.

Maximum likelihood (ML) or generalized moments (GM) procedures allow us to estimate the $\boldsymbol{\beta}$, $\lambda$ and $\rho$ parameters, along with $\sigma_v^2$ and $\sigma_\mu^2$. The testing of random effects assumptions with respect to an estimated spatial panel model (RE vs. FE tests) is an essential part of RE model-estimation and verification. Although derivations of the estimators and corresponding inference tests lie beyond the scope of this paper, they are readily available from [9].

Besides model estimation and testing, it is crucial to keep in mind some basic interpretation issues related to models featuring a spatial lag in the dependent variable. The estimated $\beta$ parameters of the model (3) do not form a proper basis for description of model dynamics. A partial derivative approach to interpretation of the impacts from changes to the regressors constitutes a more valid basis for model interpretation: As we simulate a change in $x_{rit}$ - the $r$-th explanatory variable for spatial unit $i$ at time $t$ - we expect the dependent variable in the $i$-th unit to change (direct effect) and also, for $\lambda \neq 0$, we expect some non-zero effects on the dependent variables in neighboring units (indirect effects). For proper direct and indirect effect quantification, we have to use partial derivatives. Even with static spatial panel models, this type of dynamics is relatively complex to describe. However, we may use the notation as in LeSage and Pace [7] to provide a relatively simple overview, starting with reduced form of a cross-sectional spatial model:

$$(\boldsymbol{I}_N - \lambda \boldsymbol{W}_N)\, \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \alpha \boldsymbol{\iota}_N + \boldsymbol{u}, \tag{6}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $(N \times 1)$, $\boldsymbol{X}$ is $(N \times k)$ and $\alpha$ is the intercept. Equation (6) can be conveniently rewritten for subsequent interpretation as

$$\boldsymbol{y} = \sum_{r=1}^{k} \boldsymbol{S}_r(\boldsymbol{W}_N)\, \boldsymbol{x}_r + \boldsymbol{V}_N(\boldsymbol{W}_N)\, \boldsymbol{\iota}_N \alpha + \boldsymbol{V}_N(\boldsymbol{W}_N)\, \boldsymbol{u}, \tag{7}$$

where

$$\begin{aligned} \boldsymbol{S}_r(\boldsymbol{W}_N) &= \boldsymbol{V}_N(\boldsymbol{W}_N)\, \boldsymbol{I}_N \beta_r = (\boldsymbol{I}_N - \lambda \boldsymbol{W}_N)^{-1}\, \boldsymbol{I}_N \beta_r, \\ \boldsymbol{V}_N(\boldsymbol{W}_N) &= (\boldsymbol{I}_N - \lambda \boldsymbol{W}_N)^{-1} = \boldsymbol{I}_N + \lambda \boldsymbol{W}_N + \lambda^2 \boldsymbol{W}_N^2 + \lambda^3 \boldsymbol{W}_N^3 + \dots \end{aligned}$$

The direct impacts and spillover effects (indirect impacts) for a cross-sectional spatial model (7) are given by

$$\frac{\partial y_i}{\partial x_{ir}} = \boldsymbol{S}_r(\boldsymbol{W}_N)_{ii} \quad ; \quad \frac{\partial y_i}{\partial x_{jr}} = \boldsymbol{S}_r(\boldsymbol{W}_N)_{ij} \tag{8}$$

where $\boldsymbol{S}_r(\boldsymbol{W}_N)_{ij}$ is a scalar term, element of the matrix $\boldsymbol{S}_r(\boldsymbol{W}_N)$.

Now, we may conclude our derivation of impacts for the spatial panel model (3): for spatial panel models, $\boldsymbol{S}_r(\boldsymbol{W}_N)$ may be generalized to

$$\boldsymbol{S}_r(\boldsymbol{\mathcal{W}}) = (\boldsymbol{I}_{NT} - \lambda \boldsymbol{\mathcal{W}})^{-1}\, \boldsymbol{I}_{NT} \beta_r. \tag{9}$$

Here, $\boldsymbol{\mathcal{W}} = (\boldsymbol{I}_T \otimes \boldsymbol{W}_N)$ is a block-diagonal matrix, with $T$ blocks of spatial matrices $\boldsymbol{W}_N$. To express the impacts from an estimated spatial panel model, we only need to substitute $\boldsymbol{S}_r(\boldsymbol{\mathcal{W}})$ for $\boldsymbol{S}_r(\boldsymbol{W}_N)$ in expression (8). For additional discussion of impacts' variances and statistical significance tests, see [7].

A potential weakness of the spatial econometric approach arises from the fact that $\boldsymbol{W}_N$ matrices cannot be estimated along with model parameters. Rather, $\boldsymbol{W}_N$ needs to be specified prior to model estimation. There is little theoretical background for choosing the "right" $\boldsymbol{W}_N$ specification. The consequences of possible $\boldsymbol{W}_N$ misspecification are discussed in [5] within a context of cross-sectional spatial models - along with a convenient likelihood-based algorithm for choosing an optimal $\boldsymbol{W}_N$ from a given group of alternative matrices. This rather intuitive approach is also used for robustness verification of the estimated empirical spatial panel model featured in the next section.

Finally, it may be argued that much of the spatial effects and dependencies are attributable to omitted variable factors. However, spatial autocorrelation can be conveniently interpreted as a proxy for multiple real and theoretically sound, yet practically unobservable spatial effects - many spatial interactions and their dynamic features are difficult to accurately describe and structure in a way that would facilitate informative and consistent model estimation (e.g. the technological and capital transfers between NUTS2 regions as discussed above). Hence, spatial models may provide a useful, interpretable and functional approach towards regional (macroeconomic) data analysis.

## 3 Macroeconomic Convergence - Empirical Results

A relatively simple yet efficient spatial panel $\beta$-convergence model may be outlined as

$$\log\left(\frac{y_{it}}{y_{i,t-1}}\right) = \lambda\left[\sum_{j=1}^{N} w_{ij}\,\log\left(\frac{y_{it}}{y_{i,t-1}}\right)\right] + \beta\,\log\left(y_{i,t-1}\right) + a_i + u_{it} \qquad (10)$$

where $y_{it}$ is the GDP per capita observed in the NUTS2 region $i$ at time $t$. Eurostat's "nama_10r_2_gdp" dataset is used, with annual 2000–2015 GDP observations recorded in 2010-constant prices. All the observed data exhibit strong positive spatial autocorrelation when tested using the Moran's $I$ statistic (2).

Logarithmic transformation provides the desired growth-rate interpretation: the LHS of Eq. (10) is the annual growth of real per capita income. The first element on the RHS is the spatial lag and it follows from (3). Besides observed variables, coefficient $\lambda$ and $\beta$ along with the time-invariant & region specific effects $a_i$ constitute the functional form of the spatial panel model. Finally, $u_{it}$ is the error term with properties described in (4). Although Eq. (10) contains the first time-lag of $y_{it}$, the model doesn't have an actual dynamic specification. In fact, Eq. (10) is an empirical implementation of the general Eq. (3).

Equation (10) is estimated using a balanced panel of 82 NUTS2 regions (depicted in Fig. 1) across 16 years. Hence, a total of 1.312 individual observations of GDP per capita are collected from the following EU members: Austria (9 NUTS2 regions), Czechia (8 regions), Germany (38 regions, from those 8 - plus Berlin - are from the former East-Germany), Hungary (7 regions), Poland

(16 regions) and Slovakia (4 regions). The applied modifications and restrictions to model (10) are described below and summarized in Table 1.

There is an ongoing discussion related to the estimation of $\beta$-convergence models using datasets covering relatively short time periods (see [12]). Convergence is a long-term phenomenon, therefore wider time spans would increase the accuracy of tackling true convergence dynamics (instead of adjustments towards some trend after random shocks). Although data availability restrictions may not be circumvented, we may take advantage of some rather non-restrictive assumptions: Our dataset features regions at diverse development stages - compare the GDP per capita among South-German regions, spatial units in the former East-Germany, Czechia and the Eastern parts of Poland. When properly addressed (using spatial panel models), regional heterogeneity can add confidence to our estimates, thus somewhat compensating for the limited time-span currently available from Eurostat.

**Table 1.** Estimated alternative specifications of the $\beta$-convergence model

| Model specification (classical approach) | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| $\hat{\beta}$ | −0.00418 | −0.00393 | −0.01070 | −0.01070 |
| s.e.($\hat{\beta}$) | (0.00026) | (0.00036) | (0.00128) | (0.00128) |
| $t$ value | [−16.13869] | [−11.02666] | [−8.33446] | [−8.33078] |
| Pr(> $\|t\|$) | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Model specification (spatially augmented) | (e) | (f) | (g) | (h) |
| Direct impacts | −0.00216 | −0.00149 | −0.00116 | −0.00084 |
| simulated s.e | (0.00027) | (0.00031) | (0.00023) | (0.00030) |
| $z$ score | [−8.04931] | [−4.83524] | [−5.07915] | [−2.78725] |
| Pr(> $\|z\|$) | 0.00000 | 0.00000 | 0.00000 | 0.00532 |
| $\hat{\lambda}$ | −0.13135 | −0.07829 | 0.89381 | 0.87027 |
| s.e.($\hat{\lambda}$) | (0.06732) | (0.07547) | (0.01121) | (0.01320) |
| $t$ value | [−1.95118] | [−1.03744] | [79.70808] | [65.90509] |
| Pr(> $\|t\|$) | 0.05104 | 0.29953 | 0.00000 | 0.00000 |

Table 1 illustrates the importance of controlling for both region specific effects and spatio-temporal dynamics in $\beta$-convergence models. Eight restricted and augmented forms of the Eq. (10) are used for estimation and comparison of the $\beta$ coefficients: (a) comes from a basic pooled regression estimate, with both individual and spatial effects ignored - $\lambda$ is set to zero and the $a_i$ intercept is identical for every $i$-th unit. In (b), model (a) is augmented by two dummy variables: one controls for the 2009 drop in output due to the global crisis, while the other dummy variable distinguishes "old EU" regions - NUTS2 regions in Austria and in the former West Germany - from their post-communist counterparts.

Although statistically significant, the two dummies are not reported in Table 1, as they only serve to filter out some prominent inconsistencies in the data generating process (DGP) under scrutiny, i.e. to obtain accurate convergence indicators with proper ceteris-paribus validity. All estimates omitted from this paper are available from the author upon request, along with the raw data and R source code. (c) is a panel model with spatial effects omitted ($\lambda = 0$), estimated using the FE "twoways" method as in [14], thus controlling for the unobserved individual and time effects. Specification (d) augments (c) by using the same two auxiliary dummies as introduced in (b) - again, the reason is to add controls for the two distinct influences affecting GDP growth and thus filtering them out from the the pursued $\beta$-convergence dynamics estimation.

Models (a) to (d) lack spatial dependence features, yet they serve for direct comparison with their spatially augmented counterparts (e) to (h). Specification (e) amends the pooled version of (10) by introducing a cross-sectional spatial lag as in (6). Any individual or time effects are ignored here. (f) differs from (e) by featuring dummy variables as in (b) and (d). Specification (g) is the spatial panel model Eq. (10): both region-specific effects and spatial dependencies are accounted for. Again, model (h) is obtained by incorporating our two dummies into (g). The reason for using FE estimation for (c) and (d) models rather than the RE method used in (g) and (h) is motivated by the need to incorporate the "two-ways" (individual and time) effects, which are not implemented into the RE estimator for non-spatial panel models in R. This does not hamper with our analysis: given the RE assumptions from [14] are satisfied, consistency of estimators in the last two columns of Table 1 is not affected.

All models presented in Table 1 are statistically significant at the 5% significance level and were subjected to the usual model testing and verification procedures as proposed in [9] and [14]. For example, the RE assumptions for (g) and (h) were tested using the $\chi^2$ (Spatial Hausman) test as in [9] as well as evaluated by generalizing both models into a CRE specification using the Mundlak-Chamberlain approach from [10] and [14]. The residual elements of models (g) and (h) exhibit no spatial autocorrelation at $\alpha = 0.05$. The dummy variables (omitted from Table 1) are statistically significant in all model specifications where they are used, pointing out the important differences in the DGP of growth in GDP per capita. However, statistically speaking, dummies have only a limited effect on the $\beta$ convergence parameters.

Consistently negative estimates of the $\beta$-convergence parameters provide some confidence in the stability and robustness of the underlying convergence processes. Yet, the signs of estimated $\beta$ coefficients need to be put into perspective: On one hand, the results provide evidence in favor of the $\beta$-convergence mechanism considered. On the other hand, when the spatio-temporal dynamics of the DGP is fully and properly accounted for, $\hat{\beta}$ values "fall" (precisely, get attenuated towards zero thus reflecting a slower convergence speed) by an order of magnitude: As we compare the estimated $\beta$-convergence parameter in (c) against (g) and (d) against (h) we can see that convergence dynamics changes from relatively weak to practically negligible. Actually, this situation is also reflected in the formatting of Table 1 that features 5 decimal points - the usual 3 material points would complicate models' comparison.

The estimated spatial autocorrelation parameter $\hat{\lambda}$ in Table 1 describes the effect of spatial interactions - it quantifies the systematic pattern in spatial distribution of the GDP growth rates. As models (e) and (f) do not control for individual and time effects, we can see that the spatial dependence coefficient estimates are severely biased: the negative/insignificant $\hat{\lambda}$ estimates contradict to prior theoretical beliefs, to preliminary Moran $I$ test results as well as to evidence from other published works (e.g. [4,5] and [12]). In contrast, with both spatio-temporal and individual effects properly accounted for, $\hat{\lambda}$ coefficients in (g) and (h) confirm the presence of strong regional spillovers and provide evidence supporting the convergence mechanisms based on spatial lags (presumably through factor mobility, trade and technological relationships, etc.).

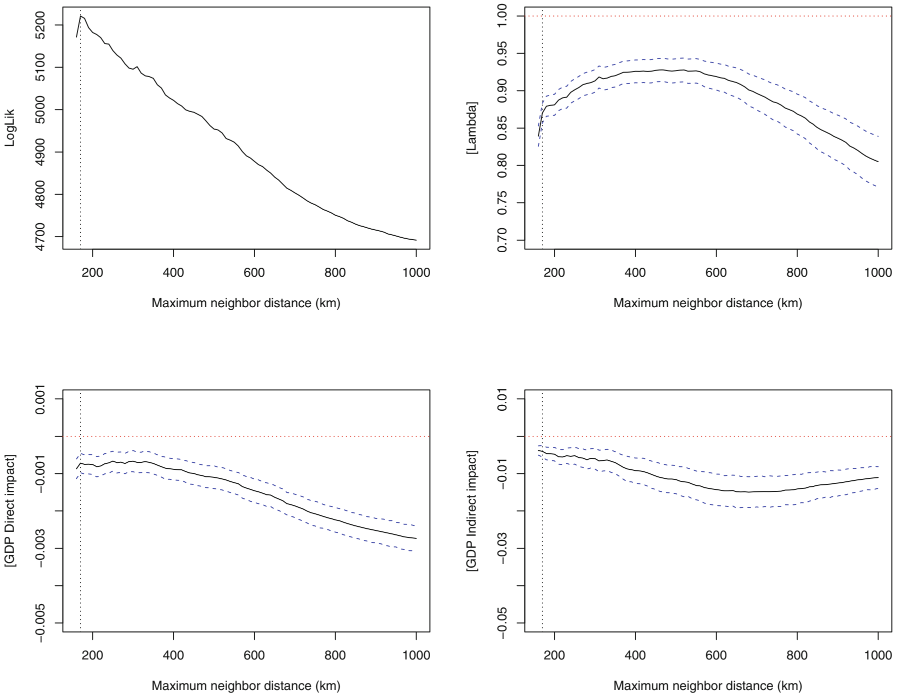


**Fig. 2.** Model stability evaluation: different $\boldsymbol{W}_N$ matrices considered

Overall, the estimated models in Table 1 are built and ordered from the simplest specification (a) towards more realistic setups. Given all theoretical assumptions and the data-based evidence discussed above, we may conclude that models neglecting any (or all) of the unobservable effects (regional specificities and spatio-temporal effects) result in a severely biased $\beta$-convergence parameters: roughly 5 to 10 times stronger as compared to the properly specified spatial panel models (g) or (h). The observed spatial correlations are much more prominent and influential when compared to the Solow-Swan type $\beta$-convergence processes.

Given the need for pre-specification of the $\boldsymbol{W}_N$ matrix (its $w_{ij}$ elements) in Eq. (10), as discussed in Sect. 2, it is advisable to evaluate model stability against changes in the ad-hoc generated neighborhood definitions. A simple yet effective approach is adopted and summarized in Fig. 2: model specification (h) as in Table 1 is estimated using alternative $\boldsymbol{W}_N$ matrices and results from different model setups are compared. The evaluation process starts with a relatively sparse spatial matrix constructed using a maximum neighbor distance threshold set to 160 km (lower thresholds generate disconnected units that are incompatible with the ML estimation of spatial models). Next, neighbor threshold distances are increased and new weight matrices are generated by iterations of 10 km, up to a rather generous maximum neighbor distance of 1.000 km - beyond this threshold, the spatial properties of the model fall apart as the variance of spatial lag elements in (10) quickly falls to zero and spatial weak dependency assumptions are violated). At each iteration, the $\beta$-convergence model is estimated and recorded to Fig. 2: model log-likelihood values are shown, along with $\hat{\lambda}$, direct and indirect (spillover) effects and their asymptotic ± 1 s.e. bands. Finally, the maximized log-likelihoods from Fig. 2 are used to select the "best" $\boldsymbol{W}_N$ matrix from the 85 possibilities considered: the maximum neighbor threshold distance as used in in Table 1 is set to 170 km. Also, Fig. 2 provides enough confidence in overall model robustness.

## 4 Conclusions

This paper pioneers the estimation and interpretation of impacts for spatial panel models in the context of macroeconomic $\beta$-convergence analysis. Compared to previous publications (e.g. [12]), this paper properly addresses the ceteris paribus effects in spatio-temporal models by focusing on the interpretation of impacts and spatial lag parameters instead of the $\beta$ parameters.

Considerable improvement is provided in comparison to the $\beta$ coefficients-based interpretation, which does not describe model dynamics properly under the spatial lag setup. Whenever spatial panel data are available, the framework presented here can extend the classical approach to $\beta$-convergence by controlling for both individual differences and spatial interactions.

The empirical part of this paper does not rule out the Solow-Swan type of macroeconomic convergence ($\beta$-convergence). However, it seems that this type of growth dynamics is more suitable for closed (large) economies. Using the appropriate spatio-temporal methodology, we can see that the regions analyzed exhibit prominent spatial convergence tendencies. The spatial part (spatial clustering) effects are much stronger than the Solow-Swan type $\beta$-convergence.

**Acknowledgement.** Supported by the grant No. IGA F4/58/2017, Faculty of Informatics and Statistics, University of Economics, Prague. Geo-data source: GISCO - Eurostat (European Commission), Administrative boundaries: © EuroGeographics.

## References

1. Anselin, L., Rey, S.J. (eds.): Perspectives on Spatial Data Analysis. Advances in Spatial Science, Springer. Berlin (2010). doi:10.1007/978-3-642-01976-0
2. Bivand, R.S., Pebesma, E., Gómez-Rubio, V.: Applied Spatial Data Analysis with R. Springer, New York (2008). doi:10.1007/978-1-4614-7618-4
3. Cliff, A.D., Ord, J.K.: Spatial Processess: Models and Applications. Pion, London (1981)
4. Elhorst, J.P.: Spatial econometrics: from cross-sectional data to spatial panels. Springer-Briefs in Regional Science. Springer, Berlin (2014). doi:10.1007/978-3-642-40340-8
5. Formánek, T., Hušek, R.: On the stability of spatial econometric models: application to the Czech Republic and its neighbors. In: Kocourek, A., Vavroušek, M. (eds.) Mathematical Methods in Economics, pp. 213–218. TU Liberec, Liberec (2016)
6. Geary, R.C.: The contiguity ratio and statistical mapping. Incorporated Stat. **5**, 115–145 (1954)
7. LeSage, J.P., Pace, R.K.: Introduction to Spatial Econometrics. CRC Press, Taylor & Francis Group, Boca Raton (2009)
8. Mankiw, N.G., Romer, D., Weil, D.N.: A contribution to the empirics of economic growth. Q. J. Econ. **107**(2), 407–437 (1992). doi:10.2307/2118477
9. Millo, G., Piras, G.: SPLM: spatial panel data models in R. J. Stat. Softw. **47**(1), 1–38 (2012). doi:10.18637/jss.v047.i01
10. Miranda, K., Martínez-Ibanez, O., Manjón-Antolín, M.: Estimating individual effects and their spatial spillovers in linear panel data models. Working paper (2015). http://www.reunionesdeestudiosregionales.org/Reus2015/htdocs/pdf/p1294.pdf
11. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrica **37**, 17–23 (1950)
12. Piras, G., Arbia, G.: Convergence in per-capita GDP across EU-NUTS2 regions using panel data models extended to spatial autocorrelation effects. Statistica **67**(2), 157–172 (2007). doi:10.6092/issn.1973-2201/3513
13. Tepperová, J., Zouhar, J., Wilksch, F.: Intra-EU migration: legal and economic view on jobseekers' welfare rights. J. Int. Migr. Integr. **17**, 1–20 (2016). doi:10.1007/s12134-016-0509-6
14. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge (2010)

# Trend-Cycle Decomposition of Economic Activity in the Czech Republic

Ondřej Čížek(✉)

University of Economics, Prague, Czech Republic
cizeko@vse.cz

**Abstract.** The aim of the paper is to decompose two important economic variables (GDP and unemployment rate in the Czech Republic) into a cyclical and a trend component by applying a state space methodology. An unobserved component model is econometrically estimated by the method of maximum likelihood. The likelihood function is constructed using the square root version of the Kalman filter. The results are economically interpreted and it is found that (1) the cyclical component of output and unemployment rate has already recovered from an initial shock at the beginning of the economic crisis in 2008, (2) there has been a persistently decreased growth of the trend component of output after the outbreak of the economic crisis, (3) the trend component of unemployment rate has been constant during the current crisis which suggests that possible hysteresis effects have not played an important role yet, (4) the growth of the GDP trend component is highly volatile in the Czech Republic.

**Keywords:** State space approach · Unobserved components model · Trend-cycle decomposition · Economic crisis

## 1 Introduction

There was a considerable decline in economic activity at the beginning of the current economic crisis in 2008 in the Czech Republic. The goal of the presented paper is to shed some light on the following questions: (1) has economic activity already recovered from the initial decline and to what extent, (2) is there evidence of permanently increased unemployment rate or not, (3) is long-run growth of GDP permanently decreased? Bivariate unobserved component model of GDP and unemployment rate is formulated in this paper in order to answer these questions. The model will be written in state space form which is a methodology commonly applied not only in technical sciences but also in economics (Zeng, Wu [17]). The model will be estimated by the method of maximum likelihood. The likelihood function is constructed using the square root version of the Kalman filter which has better numerical properties compared to the basic form of the filter (Anderson, Moore [1], Chui, Chen [5]).

Trend-cycle decomposition methodology has a long tradition in macroeconometrics (Nelson [14]) and is described in detail in Dagum and Bianconcini [7]. Empirical papers discussing trend-cycle decomposition of economic activity include Cerra and Saxena [4] who discuss this issue using regime switching methodology. Ball [2] estimates trend component of GDP for OECD countries by applying the concept of

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_2

potential output and a production function approach. Ball [2] and Barro [3] (and many others cited in these papers) found evidence that deep recessions have permanent effects on output. Similar results are found in the presented paper for the case of the Czech Republic.

The paper is organized as follows. Econometric methodology is described in Sect. 2. The subsequent Sect. 3 is the empirical part of the paper presenting results and economic discussion. The final Sect. 4 concludes.

## 2 Econometric Methodology

### 2.1 Model

The model decomposing real GDP and unemployment rate into trend and a cycle component is presented in this chapter. The applied unobserved components model was formulated by Clark [6] and was also summarized in a textbook treatment by Kim and Nelson [12]. The model equations are given as follows:

$$y_t = n_t + x_t \tag{1}$$

$$n_t = g_{t-1} + n_{t-1} + v_t,\ v_t \sim i.i.d.N\left(0, \sigma_v^2\right), \tag{2}$$

$$g_t = g_{t-1} + w_t,\ w_t \sim i.i.d.N\left(0, \sigma_w^2\right), \tag{3}$$

$$x_t = \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + e_t,\ e_t \sim i.i.d.N\left(0, \sigma_e^2\right), \tag{4}$$

where $y_t$ is log of real GDP, $n_t$ is a stochastic trend component, $x_t$ represents a stationary cyclical component and $v_t$, $w_t$, $e_t$ are independent white noise processes.

The autoregressive process of order two was chosen in the Eq. (4). This is the most common assumption used in empirical literature when modelling cyclical variables as an autoregressive process of the second order is a parsimonious way to model cyclical dynamics.

This standard univariate model is extended into a bivariate model of real GDP and unemployment. The unemployment rate is decomposed into trend and a cycle as follows:

$$U_t = L_t + C_t, \tag{5}$$

$$L_t = L_{t-1} + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N\left(0, \sigma_\varepsilon^2\right), \tag{6}$$

$$C_t = \alpha_0 \cdot x_t + \alpha_1 \cdot x_{t-1} + \alpha_2 \cdot x_{t-2} + \eta_t,\ \eta_t \sim i.i.d.N\left(0, \sigma_\eta^2\right), \tag{7}$$

where $L_t$ is a trend component of unemployment rate, $C_t$ is a stationary component of unemployment rate and $\varepsilon_t$, $\eta_t$ are independent white noise processes.

The cyclical component $C_t$ is assumed to be a function of current and past transitory components of real output which represents a version of Okun's law. The number of lags used in the Eq. (7) was chosen rather arbitrarily. This choice, however, is quite common in the empirical literature (Kim, Nelson [12]).

The transition and measurement equation of the state space representation are written as follows

$$\mathbf{x_t} = \mathbf{A} \cdot \mathbf{x_{t-1}} + \mathbf{u_t}, \tag{8}$$

$$\mathbf{z_t} = \mathbf{D} \cdot \mathbf{x_t} + \mathbf{v_t}, \tag{9}$$

where $\mathbf{x_t} = \begin{bmatrix} n_t & x_t & x_{t-1} & x_{t-2} & g_t & L_t \end{bmatrix}'$, $\mathbf{z_t} = \begin{bmatrix} y_t & U_t \end{bmatrix}'$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{e_1} + \mathbf{e_5} \\ \phi_1 \cdot \mathbf{e_2} + \phi_2 \cdot \mathbf{e_3} \\ \mathbf{e_2} \\ \mathbf{e_3} \\ \mathbf{e_5} \\ \mathbf{e_6} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{e_1} + \mathbf{e_2} \\ \alpha_0 \cdot \mathbf{e_2} + \alpha_1 \cdot \mathbf{e_3} + \alpha_2 \cdot \mathbf{e_4} \end{bmatrix},$$

$$\mathbf{u_t} = \begin{bmatrix} v_t & e_t & 0 & 0 & w_t & \varepsilon_t \end{bmatrix}, \mathbf{v_t} = \begin{bmatrix} 0 & \eta_t \end{bmatrix},$$

$\mathbf{e_j}, j = 1, \ldots, 6$, denotes a 1 × 6 row vector with element $j$ equal to unity and all other elements equal to zero.

Random vectors $\mathbf{u_t}$, $\mathbf{v_t}$ are normally distributed and satisfy the following assumptions commonly assumed in the standard state space model:

$$E\left(\mathbf{v_t} \cdot \mathbf{v_s'}\right) = \begin{cases} \mathbf{\Sigma_{vv}} & \text{for}\, t = s, \\ \mathbf{0} & \text{for}\, t \neq s, \end{cases} \quad E\left(\mathbf{u_t} \cdot \mathbf{u_s'}\right) = \begin{cases} \mathbf{\Sigma_{uu}} & \text{for}\, t = s, \\ \mathbf{0} & \text{for}\, t \neq s, \end{cases} \tag{10}$$

$$E\left(\mathbf{v_t} \cdot \mathbf{u_s'}\right) = \mathbf{0} \quad \text{for all}\, t,\, s, \tag{11}$$

$$E\left(\mathbf{x_0} \cdot \mathbf{u_t'}\right) = E\left(\mathbf{x_0} \cdot \mathbf{v_t'}\right) = \mathbf{0} \quad \text{for all}\, t, \tag{12}$$

$$E(\mathbf{v_t}) = E(\mathbf{u_t}) = \mathbf{0} \quad \text{for all}\, t. \tag{13}$$

Quarterly seasonally adjusted data for unemployment rate $U_t$ and GDP $y_t$ in the Czech Republic from 1996 Q1 to 2015 Q4 were used for the observable variables in the vector $\mathbf{z_t}$. The relevant age structure for unemployed was chosen to be ‘from 24 to 74 years’. The GDP was also calendar adjusted and measured in chain linked volumes (2010) in millions euro. All such data is available at the Eurostat database.

## 2.2 Kalman Filter – Basic Version

The basic version of the Kalman filter algorithm summarized here for convenience is described e.g. in Harvey [11] or Hamilton [10]. The mean of the state vector $\mathbf{x_t}$ conditional on the information known in time $t-1$ is given by

$$\mathbf{x_{t|t-1}} = \mathbf{A} \cdot \mathbf{x_{t-1|t-1}}, \tag{14}$$

where $\mathbf{x_{t|t-1}} \equiv E(\mathbf{x_t}|\mathbf{\Omega_{t-1}})$, $\mathbf{x_{t-1|t-1}} \equiv E(\mathbf{x_{t-1}} \mid \mathbf{\Omega_{t-1}})$ and the information available in time $t-1$ is $\mathbf{\Omega_{t-1}} \equiv (\mathbf{z_1}, \ldots, \mathbf{z_{t-1}}, \mathbf{A}, \mathbf{D}, \mathbf{\Sigma_{uu}}, \mathbf{\Sigma_{vv}})$. The matrices $\mathbf{A}$, $\mathbf{D}$, $\mathbf{\Sigma_{uu}}$ and $\mathbf{\Sigma_{vv}}$ are assumed to be known in this chapter describing the Kalman filter algorithm despite the fact that they depend on unknown parameters $\boldsymbol{\theta} = \left(\phi_1, \phi_2, \alpha_0, \alpha_1, \alpha_2, \sigma_v, \sigma_w, \sigma_e, \sigma_\varepsilon, \sigma_\eta\right)$ in the model presented in Sect. 2.1. The estimation procedure of the parameter vector $\boldsymbol{\theta}$ will be described later in Sect. 2.4.

The prediction error covariance matrix is calculated as follows

$$\mathbf{P_{t|t-1}} \equiv E\left[\left(\mathbf{x_t} - \mathbf{x_{t|t-1}}\right) \cdot \left(\mathbf{x_t} - \mathbf{x_{t|t-1}}\right)' | \mathbf{\Omega_{t-1}}\right],$$

$$\mathbf{P_{t|t-1}} = \mathbf{A} \cdot E\left[\left(\mathbf{x_{t-1}} - \mathbf{x_{t-1|t-1}}\right) \cdot \left(\mathbf{x_{t-1}} - \mathbf{x_{t-1|t-1}}\right)' | \mathbf{\Omega_{t-1}}\right] \cdot \mathbf{A}' + E\left(\mathbf{u_t u_t'}\right),$$

$$\mathbf{P_{t|t-1}} = \mathbf{A P_{t-1} A'} + \mathbf{\Sigma_{uu}}. \tag{15}$$

Distribution of the vector $(\, \mathbf{x_t'} \quad \mathbf{z_t'} \,)'$ conditional on $\mathbf{\Omega_{t-1}}$ is multivariate normal with mean $\left( \mathbf{x'_{t|t-1}} \quad (\mathbf{D} \cdot \mathbf{x_{t|t-1}})' \right)'$ and a covariance given by

$$E\left[\begin{pmatrix} \mathbf{x_t} - \mathbf{x_{t|t-1}} \\ \mathbf{D} \cdot \left(\mathbf{x_t} - \mathbf{x_{t|t-1}}\right) + \mathbf{v_t} \end{pmatrix} \begin{pmatrix} \mathbf{x_t} - \mathbf{x_{t|t-1}} \\ \mathbf{D} \cdot \left(\mathbf{x_t} - \mathbf{x_{t|t-1}}\right) + \mathbf{v_t} \end{pmatrix}' | \mathbf{\Omega_{t-1}}\right]$$
$$= \begin{pmatrix} \mathbf{P_{t|t-1}} & \mathbf{P_{t|t-1} D'} \\ \mathbf{D P_{t|t-1}} & \mathbf{D P_{t|t-1} D'} + \mathbf{\Sigma_{vv}} \end{pmatrix}.$$

Recall a generally known result that a conditional distribution of normally distributed vector $(\mathbf{x'y'})'$ is also normal with mean and covariance given by

$$\boldsymbol{\mu}_{\mathbf{x|y}} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{\Sigma_{xy} \Sigma_{yy}^{-1}} \left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right),$$
$$\mathbf{\Sigma_{xx|y}} = \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}},$$

where $\boldsymbol{\mu}_{\mathbf{x|y}} = E(\mathbf{x}|\mathbf{y})$, $\boldsymbol{\mu}_{\mathbf{x}} = E(\mathbf{x})$, $\boldsymbol{\mu}_{\mathbf{y}} = E(\mathbf{y})$,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}} \\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}} \end{pmatrix} = E\left\{ \begin{bmatrix} \mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} \\ \mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix} \cdot \left[\, (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})' \quad (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})' \,\right] \right\}.$$

Direct application of this result yields that $\mathbf{x_t}$ conditional on $\mathbf{z_t}$ is multivariate normal with mean

$$\mathbf{x_{t|t}} = \mathbf{x_{t|t-1}} + \mathbf{K_t}\left(\mathbf{z_t} - \mathbf{z_{t|t-1}}\right), \tag{16}$$

where

$$\mathbf{K_t} = \mathbf{P_{t|t-1} D'}\left(\mathbf{D P_{t|t-1} D'} + \mathbf{\Sigma_{vv}}\right)^{-1}. \tag{17}$$

The covariance matrix is given by

$$\mathbf{P_t} = \mathbf{P_{t|t-1}} - \mathbf{P_{t|t-1}D'}\left(\mathbf{DP_{t|t-1}D'} + \boldsymbol{\Sigma}_{\mathbf{vv}}\right)^{-1}\mathbf{DP_{t|t-1}}. \tag{18}$$

Equations (14)–(18) together form the recursions of the Kalman filter algorithm.

## 2.3 Kalman Filter – Square Root Version

The basic version of the Kalman filter has disappointing numerical properties. For this reason, square root version of this algorithm is applied in this paper and briefly summarized here for convenience. More details can be found in Anderson and Moore [1] or Grewal and Andrews [9]. The square root version of the algorithm recursively calculates ‘square roots’ of the matrices $\mathbf{P_t}$ a $\mathbf{P_{t|t-1}}$, where the square root of a matrix is a lower triangular matrix $\mathbf{S_t}$ ($\mathbf{S_{t|t-1}}$) satisfying $\mathbf{P_t} = \mathbf{S_t} \cdot \mathbf{S'_t}$ ($\mathbf{P_{t|t-1}} = \mathbf{S_{t|t-1}} \cdot \mathbf{S'_{t|t-1}}$). The matrix $\mathbf{S_{t|t-1}}$ is calculated on the basis of the following equation

$$\begin{bmatrix} \mathbf{S'_{t|t-1}} \\ \mathbf{0} \end{bmatrix} = \mathbf{T} \cdot \begin{bmatrix} \mathbf{S'_{t-1}A'} \\ \boldsymbol{\Sigma}^{\mathbf{1/2'}}_{\mathbf{uu}} \end{bmatrix}, \tag{19}$$

where $\mathbf{T}$ is a square orthogonal matrix ($\mathbf{T'} \cdot \mathbf{T} = \mathbf{I}$) ensuring that $\mathbf{S'_{t|t-1}}$ is lower triangular.[1]

The matrix $\mathbf{S_{t|t-1}}$ is indeed a square root of $\mathbf{P_{t|t-1}}$ which can be easily verified as follows

$$\begin{aligned} \mathbf{S_{t|t-1}} \cdot \mathbf{S'_{t|t-1}} &= \begin{bmatrix} \mathbf{S_{t|t-1}} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{S'_{t|t-1}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{AS_{t-1}} & \boldsymbol{\Sigma}^{\mathbf{1/2}}_{\mathbf{uu}} \end{bmatrix} \cdot \mathbf{T'T} \cdot \begin{bmatrix} \mathbf{S'_{t-1}A'} \\ \boldsymbol{\Sigma}^{\mathbf{1/2'}}_{\mathbf{uu}} \end{bmatrix} \\ &= \mathbf{AS_{t-1}S'_{t-1}A'} + \boldsymbol{\Sigma}^{\mathbf{1/2}}_{\mathbf{uu}}\boldsymbol{\Sigma}^{\mathbf{1/2'}}_{\mathbf{uu}} = \mathbf{AP_{t-1}A'} + \boldsymbol{\Sigma}_{\mathbf{uu}} = \mathbf{P_{t|t-1}}. \end{aligned}$$

The matrix $\mathbf{S_t}$ is calculated according to

$$\begin{bmatrix} \mathbf{F}^{1/2'}_{\mathbf{t|t-1}} & \tilde{\mathbf{K}}'_{\mathbf{t}} \\ \mathbf{0} & \mathbf{S'_t} \end{bmatrix} = \bar{\mathbf{T}} \cdot \begin{bmatrix} \boldsymbol{\Sigma}^{1/2'}_{\mathbf{vv}} & \mathbf{0} \\ \mathbf{S'_{t|t-1}D'} & \mathbf{S'_{t|t-1}} \end{bmatrix}, \tag{20}$$

where the matrix $\bar{\mathbf{T}}$ should have the same properties as the matrix $\mathbf{T}$ and the interpretation of the symbols $\mathbf{F}^{1/2'}_{\mathbf{t|t-1}}$, $\tilde{\mathbf{K}}'_{\mathbf{t}}$ will become clear from the following verification that the matrix $\mathbf{S_t}$ is a square root of the $\mathbf{P_t}$

[1] Specifically, the Matlab function *qr* was used for this purpose. $[\mathbf{QR}] = qr(\mathbf{A})$ calculates an upper triangular matrix $\mathbf{R}$ (with same dimensions as $\mathbf{A}$) and an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{Q} \cdot \mathbf{R}$, or $\mathbf{R} = \mathbf{Q'} \cdot \mathbf{A}$.

$$\begin{bmatrix} \mathbf{F}^{1/2}_{\mathbf{t|t-1}} & \mathbf{0} \\ \tilde{\mathbf{K}}_{\mathbf{t}} & \mathbf{S}_{\mathbf{t}} \end{bmatrix} \begin{bmatrix} \mathbf{F}^{1/2'}_{\mathbf{t|t-1}} & \tilde{\mathbf{K}}'_{\mathbf{t}} \\ \mathbf{0} & \mathbf{S}'_{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{1/2}_{\mathbf{vv}} & \mathbf{DS}_{\mathbf{t|t-1}} \\ \mathbf{0} & \mathbf{S}_{\mathbf{t|t-1}} \end{bmatrix} \cdot \bar{\mathbf{T}}'\bar{\mathbf{T}} \cdot \begin{bmatrix} \boldsymbol{\Sigma}^{1/2'}_{\mathbf{vv}} & \mathbf{0} \\ \mathbf{S}'_{\mathbf{t|t-1}}\mathbf{D}' & \mathbf{S}'_{\mathbf{t|t-1}} \end{bmatrix},$$

or

$$\begin{bmatrix} \mathbf{F}_{\mathbf{t|t-1}} & \mathbf{F}^{1/2}_{\mathbf{t|t-1}}\tilde{\mathbf{K}}'_{\mathbf{t}} \\ \tilde{\mathbf{K}}_{\mathbf{t}}\mathbf{F}^{1/2'}_{\mathbf{t|t-1}} & \tilde{\mathbf{K}}_{\mathbf{t}}\tilde{\mathbf{K}}'_{\mathbf{t}} + \mathbf{S}_{\mathbf{t}}\mathbf{S}'_{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{vv}} + \mathbf{DS}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}}\mathbf{D}' & \mathbf{D}'\mathbf{S}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}} \\ \mathbf{S}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}}\mathbf{D} & \mathbf{S}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}} \end{bmatrix}.$$

Comparison of the corresponding submatrices in the left yields

$$\mathbf{F}_{\mathbf{t|t-1}} = \left(\mathbf{DP}_{\mathbf{t|t-1}}\mathbf{D}' + \boldsymbol{\Sigma}_{\mathbf{vv}}\right),$$

$$\tilde{\mathbf{K}}_{\mathbf{t}} = \mathbf{P}_{\mathbf{t|t-1}}\mathbf{D}\left(\mathbf{F}^{1/2'}_{\mathbf{t|t-1}}\right)^{-1}.$$

Comparing the matrices in the lower right block reveals that

$$\begin{aligned} \mathbf{S}_{\mathbf{t}}\mathbf{S}'_{\mathbf{t}} &= \mathbf{S}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}} - \tilde{\mathbf{K}}_{\mathbf{t}}\tilde{\mathbf{K}}'_{\mathbf{t}} \\ &= \mathbf{S}_{\mathbf{t|t-1}}\mathbf{S}'_{\mathbf{t|t-1}} - \mathbf{P}_{\mathbf{t|t-1}}\mathbf{D}\left(\mathbf{F}^{1/2'}_{\mathbf{t|t-1}}\right)^{-1}\left(\mathbf{F}^{1/2}_{\mathbf{t|t-1}}\right)^{-1}\mathbf{D}'\mathbf{P}_{\mathbf{t|t-1}} \\ &= \mathbf{P}_{\mathbf{t|t-1}} - \mathbf{P}_{\mathbf{t|t-1}}\mathbf{DF}^{-1}_{\mathbf{t|t-1}}\mathbf{D}'\mathbf{P}_{\mathbf{t|t-1}} \\ &= \mathbf{P}_{\mathbf{t}}. \end{aligned}$$

Vectors $\mathbf{x}_{\mathbf{t|t-1}}$ and $\mathbf{x}_{\mathbf{t|t}}$ are calculated in the same way as in the basic form of the Kalman filter (see Eqs. 14 and 16), but the matrix $\mathbf{K}_{\mathbf{t}}$ is calculated as $\mathbf{K}_{\mathbf{t}} = \tilde{\mathbf{K}}_{\mathbf{t}} \cdot \left(\mathbf{F}^{1/2}_{\mathbf{t|t-1}}\right)^{-1}$. This indeed corresponds to the way it is calculated in the basic version of the Kalman fiter which can be seen easily as follows

$$\begin{aligned} \mathbf{K}_{\mathbf{t}} &= \tilde{\mathbf{K}}_{\mathbf{t}} \cdot \left(\mathbf{F}^{1/2}_{\mathbf{t|t-1}}\right)^{-1} \\ &= \mathbf{P}_{\mathbf{t|t-1}}\mathbf{D}\left(\mathbf{F}^{1/2'}_{\mathbf{t|t-1}}\right)^{-1}\left(\mathbf{F}^{1/2}_{\mathbf{t|t-1}}\right)^{-1} \\ &= \mathbf{P}_{\mathbf{t|t-1}}\mathbf{D}\left(\mathbf{F}_{\mathbf{t|t-1}}\right)^{-1}. \end{aligned}$$

## 2.4 Likelihood Function

The likelihood function was computed by the method described in Harvey [11] or Hamilton [10] which is briefly summarized here for convenience. Let us denote $\mathbf{Z}_{\mathbf{T}} = \left(\mathbf{z}'_{\mathbf{1}}, \ldots, \mathbf{z}'_{\mathbf{T}}\right)'$ the column vector with $n$ rows, where $n$ is a multiple of number of observations $T$ and a number of observed variables. The multivariate density of observed data $\mathbf{z}_{\mathbf{t}}$, $t = 1, \ldots, T$ will be denoted by $f_0(\mathbf{Z}_{\mathbf{T}})$ and is a member of

$\{f(\mathbf{Z_T}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Hence, $f_0(\mathbf{Z_T}) = f(\mathbf{Z_T}|\boldsymbol{\theta_0})$, where $\boldsymbol{\theta_0}$ are true (unknown) parameter values. For the model described in Sect. 2.1, the vector $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta} = \left(\phi_1, \phi_2, \alpha_0, \alpha_1, \alpha_2, \sigma_v, \sigma_w, \sigma_e, \sigma_\varepsilon, \sigma_\eta\right)$.

The density $f(\mathbf{Z_T} \mid \boldsymbol{\theta})$ can be viewed as a function of $\boldsymbol{\theta}$ for given data $\mathbf{Z_T}$, i.e. $L(\boldsymbol{\theta} \mid \mathbf{Z_T}) \equiv f(\mathbf{Z_T} \mid \boldsymbol{\theta})$, where the function $L(\boldsymbol{\theta} \mid \mathbf{Z_T})$ is called a likelihood function. The method of maximum likelihood estimates the unknown parameter vector $\boldsymbol{\theta_0}$ by maximizing the likelihood function

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}[L(\theta|\mathbf{Z_T})] \tag{21}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\theta_0}$.

It can be easily shown that the likelihood function can be written as follows

$$L(\boldsymbol{\theta}|\mathbf{Z_T}) = \prod_{t=1}^{T} f(\mathbf{z_t}|\mathbf{Z_{t-1}}, \boldsymbol{\theta}).$$

The vector $\mathbf{z_t}$ is normally distributed in our case. Hence, the distribution of $\mathbf{z_t}$ conditional on $\mathbf{Z_{t-1}}$ is also normal. The mean and covariance of this conditional distribution are obtained from the Kalman filter as follows

$$\mathbf{z_{t|t-1}} \equiv E(\mathbf{z_t}|\boldsymbol{\Omega_{t-1}}) = \mathbf{D} \cdot \mathbf{x_{t|t-1}}, \tag{22}$$

$$\mathbf{F_{t|t-1}} \equiv E\left[\left(\mathbf{z_t} - \mathbf{z_{t|t-1}}\right)\left(\mathbf{z_t} - \mathbf{z_{t|t-1}}\right)'|\boldsymbol{\Omega_{t-1}}\right] = \mathbf{DP_{t|t-1}D'} + \boldsymbol{\Sigma_{vv}}. \tag{23}$$

The likelihood function in this case is given by

$$L(\boldsymbol{\theta}|\mathbf{Z_T}) = \prod_{t=1}^{T}\left[\frac{1}{(2\pi)^{\frac{k}{2}}\left|\mathbf{F_{t|t-1}}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\tilde{\mathbf{z}}_\mathbf{t}'\mathbf{F_{t|t-1}^{-1}}\tilde{\mathbf{z}}_\mathbf{t}\right)\right],$$

where $k$ is the number of observed variables and $\tilde{\mathbf{z}}_\mathbf{t} = \mathbf{z_t} - \mathbf{z_{t|t-1}}$ is so called innovation.

From a numerical point of view, it is easier to maximize the log-likelihood function which is calculated as follows

$$l(\boldsymbol{\theta}|\mathbf{Z_T}) \equiv \ln L(\boldsymbol{\theta}|\mathbf{Z_T}) = -\frac{T \cdot k}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left[\ln\left|\mathbf{F_{t|t-1}}\right| + \left(\tilde{\mathbf{z}}_\mathbf{t}'\mathbf{F_{t|t-1}^{-1}}\tilde{\mathbf{z}}_\mathbf{t}\right)\right]. \tag{24}$$

This function was maximized numerically using standard numerical optimization procedures implemented in Matlab in order to find maximum likelihood estimate $\hat{\boldsymbol{\theta}}$.

# 3 Empirical Application

## 3.1 Results of Econometric Estimation

Econometric estimation was performed in Matlab by maximization of the likelihood function (24). The obtained results are summarized in the following (Tables 1 and 2).

**Table 1.** Econometric estimates—GDP decomposition

| | $\phi_1$ | $\phi_2$ | $\sigma_v$ | $\sigma_e$ | $\sigma_w$ |
|---|---|---|---|---|---|
| Estimate | 1.5396 | −0.6345 | 0.0026 | 0.0055 | 0.0022 |
| Standard error | 0.0672 | 0.0661 | 0.0016 | 0.0010 | 0.0006 |

**Table 2.** Econometric estimates—unemployment decomposition

| | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\sigma_\varepsilon$ | $\sigma_\eta$ |
|---|---|---|---|---|---|
| Estimate | −0.1170 | −0.1649 | −0.1117 | 0.0015 | 0.0002 |
| Standard error | 0.0425 | 0.0568 | 0.0364 | 0.0005 | 0.0011 |

The estimated value $\sigma_e = 0.0055$ indicates that a significant portion of quarter-to-quarter innovations in real GDP is cyclical. Similar results were obtained for the U.S. economy (Clark [6], Kim, Nelson [12]). Nonetheless, the estimated standard error $\sigma_w = 0.0022$ representing the variability of the GDP (long-run) growth is approximately 10 times higher than that reported by Clark or Kim and Nelson.

One possible explanation for why the U.S. economic long-run growth is less volatile than the long-run growth in the Czech Republic is the process of economic transformation. The Czech economy had been opening to the rest of the world. Many international trade barriers had been removed by the entrance to the European Union. Lots of economic reforms had been realized. For these reasons, volatility of the long-run growth of the economy in transition is higher than the volatility of a stable economy.

Negative values of the parameters $\alpha_i$, $i = 0, 1, 2$ confirm an inverse relationship between output and unemployment referred to as Okun's law.

## 3.2 Economic Discussion

The following Fig. 1 illustrates the decomposition of the log of real GDP to its trend and a cyclical component (output gap). There was a sharp decline in the output gap in 2008 and 2009. Nonetheless, it has recovered after this initial shock and has been even positive since 2014 Q1. This finding is interesting as it is in contrast with earlier business-cycle studies for the U.S. economy. Watson [16], Kim and Nelson [12] and Clark [6] attribute most of the variation of U.S. output to the cyclical component.

The graph in the left suggests that there is a change in the GDP trend due to the current crisis. Similar results are found by Perron and Wada [15] who emphasized the importance of changes in the slope of the trend. This suggests that the huge output loss

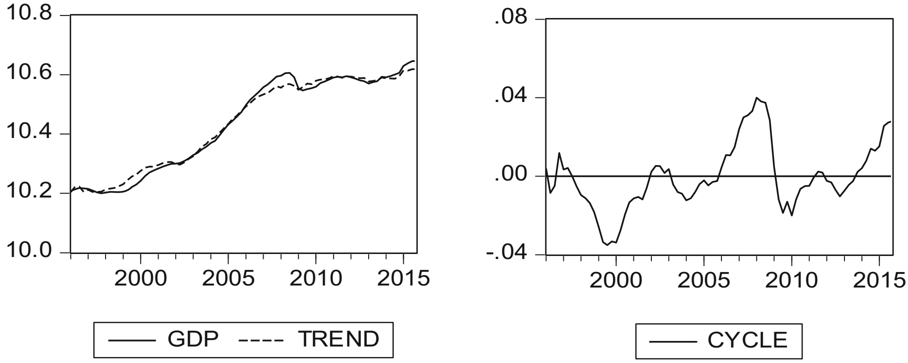


**Fig. 1.** Log of real GDP $y_t$ and its trend $n_t$ and a cycle $x_t$ component

induced by the crisis is permanent and not only transitory. While the trend was upward-sloping from 1996 to 2008, it is practically constant from 2009 to 2015. This result can also be seen in the Fig. 2 which shows that the quarter-to-quarter growth of the GDP trend component $g_t$ has been low since 2009. The mean of the variable $g_t$ in the pre-crisis time period from 1996 to 2008 is 0.0081 which is quite high when compared to the corresponding value 0.0018 for the post-crisis period from 2009 to 2015. These results are in line with other empirical studies analyzing the impact of the current global economic crisis (Barro [3], Ball [2]).

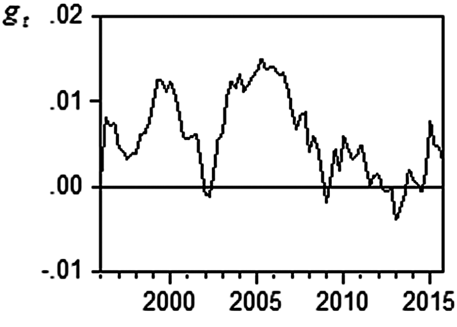


**Fig. 2.** Long-run growth of the GDP trend component $g_t$

These findings suggest an adverse permanent influence of the crisis on the long-run economic growth. Nonetheless, this is only a suggestion. The rigorous evaluation of the influence of the crisis on the long-run growth is left for future research. Such a research would apply the growth theory according to which the growth rate depends on its determinants. Changing these determinants leads to a change in the long-run growth rate. The current economic crisis can be considered to be just one of many determinants of the long-run growth rate. Nonetheless, it is probably the case that the current economic crisis is indeed the most important factor which caused the decreased values of the long-run growth $g_t$ after 2008.

The Fig. 3 shows the trend-cycle decomposition of the unemployment rate.

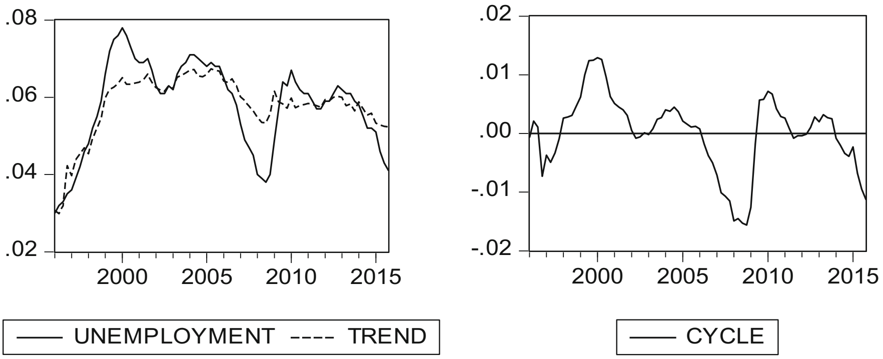


**Fig. 3.** Unemployment rate $U_t$ and its trend $L_t$ and a cycle $C_t$ component

The cyclical component increased substantially in the beginning of the crisis in 2009. Since then, however, the cyclical component of the unemployment rate has been decreasing steadily. The trend component has been practically constant since the beginning of the crisis which suggests that possible hysteresis effects haven't played important role yet.

# 4 Conclusion

The paper applies state space methodology in order to perform trend-cycle decomposition of the GDP and unemployment rate in the Czech Republic. The important finding is that the long-run growth in the Czech Republic has been highly volatile and that there has been a dramatic decrease of this long-run growth since 2008. Similar results of a huge long-term damage to output were found by Ball [2] as well as by many others cited in Ball's influential paper. The trend of unemployment rate has not changed much since the beginning of the crisis while the cyclical component increased sharply at the beginning of the crisis and has been decreasing steadily since then.

The model could be expanded for example by relaxing the common presumption of no correlation between the shocks to the trend and the cycle (Morley et al. [13]). Possible parameter instability due to the economic crisis could also be taken into account by applying regime-switching methodology as by Cerra and Saxena [4]. Detailed analysis along the lines of the growth theory as in Barro [3], or Durlauf, Helliwell, Raj [8] could be applied in order to confirm the suggested hypothesis that the observed decline in the growth of the long-run trend is caused by the current economic crisis.

**Acknowledgements.** Financial support of VŠE IGA IG403036 is gratefully acknowledged by the author. Paper was processed with contribution of long term support of scientific work on Faculty of Informatics and Statistics, University of Economics, Prague (IP 400040).

## References

1. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice-Hall, Englewood Cliffs (1979)
2. Ball, L.M.: Long-term damage from the great recession in OECD countries. In: NBER Working Paper 20185 (2014)
3. Barro, R.J.: Economic growth in East Asia before and after the financial crisis. In: NBER Working Paper No. W8330. National Bureau of Economic Research, Cambridge (2001)
4. Cerra, V., Saxena, S.C.: Did output recover from the Asian Crisis? In: IMF Staff Paper 52(1). International Monetary Fund (2005)
5. Chui, C.K., Chen, G.: Kalman Filtering with Real-Time Applications. Springer, New York (2008)
6. Clark, P.K.: The cyclical component of U.S. economic activity. Q. J. Econ. **102**, 797–814 (1987)
7. Dagum, E.B., Bianconcini, S.: Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation. Springer, Basel (2016)
8. Durlauf, S., Helliwell, J.F., Raj, B.: Long-run economic growth. In: Durlauf, S., Helliwell, J. F., Raj, B. (eds.) Long-Run Economic Growth. Studies in Empirical Economics. Physica-Verlag HD, Basel (1996)
9. Grewal, M.S., Andrews, A.P.: Kalman Filtering-Theory and Practice Using MATLAB. Wiley-Interscience Publication, New York (2008)
10. Hamilton, J.D.: Time Series Analysis. Princeton University Press, Princeton (1994)
11. Harvey, A.C.: Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, London (1989)
12. Kim, C.J., Nelson, C.R.: State Space Models with Regime Switching. MIT Press, Cambridge (1999)
13. Morley, J.C., Nelson, C.R., Zivot, E.: Why are the Beveridge-Nelson and unobserved-components decompositions of GDP So different? Rev. Econ. Stat. **85**, 235–243 (2003)
14. Nelson, C.R.: Trend/cycle decomposition. In: Durlauf, S.N., Blume, L.E. (eds.) Macroeconometrics and Time Series Analysis, pp. 343–346. Palgrave Macmillan, Basingstoke (2010)
15. Perron, P., Wada, T.: Let's take a break: trends and cycles in US real GDP. J. Monet. Econ. **56**(6), 749–765 (2009)
16. Watson, M.: Univariate detrending methods with stochastic trends. J. Monet. Econ. **18**(1), 49–75 (1986)
17. Zeng, Y., Wu, S. (eds.): State-Space Models: Applications in Economics and Finance. Springer, New York (2013)

# Parallel Matrix Multiplication for Business Applications

Mais Haj Qasem(✉) and Mohammad Qatawneh

Computer Science Department, University of Jordan, Amman, Jordan
mais_hajqasem@hotmail.com, mohd.qat@ju.edu.jo

**Abstract.** Business applications, such as market shops, use matrix multiplication to calculate yearly, monthly, or even daily profits based on price and quantity matrices. Matrices comprise large data in computer applications and other fields, which make the efficiency of matrix multiplication a popular research topic. Although the task of computing matrix products is a central operation in many numerical algorithms, it is potentially time consuming, making it one of the most well-studied problems in this field. In this paper, Message Passing Interface (MPI), MapReduce, and Multithreaded methods have been implemented to demonstrate their effectiveness in expediting matrix multiplication in a multi-core system. Simulation results show that the efficiency rates of MPI and MapReduce are 90.11% and 47.94%, respectively, with a multi-core processor on the Market Shop application, indicating better performances compared with those of the multithreaded and sequential methods.

**Keywords:** Business application · Hadoop · MPI · MapReduce · Matrix multiplication

## 1 Introduction

Matrix multiplication is a fundamental operation in linear algebra with related real-life applications. Recently, mathematicians and research scientists have found many applications of matrices due to the advent of personal and large-scale computers, which increased the use of matrices in a wide variety of applications, such as economics, engineering, statistics, and other sciences [14].

Market Shop is a business application that uses spreadsheets for budgeting, sales projections, and cost estimation, making the matrix multiplication useful in these applications. Therefore, using matrix multiplication can dramatically reduce the labor involved in modeling processes that deal with multiple categories of employees, customers, districts, products, or supplies [22].

In addition to these naturally related applications, the focus in using matrix multiplication is computational problems that should be investigated thoroughly to enhance the efficiency of the implemented algorithms for matrix multiplication. Hence, over the years, several parallel and distributed systems for matrix multiplication methods have been proposed to reduce the cost and time of matrix multiplication over multiple processors [5, 15].

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_3

Parallel and distributed computing systems are high-performance computing systems that spread out a single application over many multi-core and multi-processor computers in order to rapidly complete the task. Parallel and distributed computing systems divide large problems into smaller sub-problems and assign each of them to different processors in a typically distributed system running concurrently in parallel. MapReduce [17] and Message Passing Interface (MPI) are among these computing systems [10].

MapReduce is an algorithm design and processing paradigm proposed by Dean and Ghemawat in 2004 [7]. MapReduce enables efficient parallel and distributed computing and consists of two serial tasks, namely, map and reduce. Each serial task is implemented with several parallel subtasks. Specific MapReduce paradigms include MapReduce with expectation maximization for text filtering [31], MapReduce with K-means for remote-sensing image clustering [16], and MapReduce with decision tree for classification [19]. MapReduce has also been used for job scheduling [20] and real-time systems [15].

MPI is a standardized means of exchanging messages among multiple computers running a parallel program across a distributed memory. MPI is generally considered to be the industry standard, and forms the basis of most communication interfaces adopted by parallel computing programmers. MPI is used to improve scalability, performance, multi-core and cluster support, and interoperation with other applications [26].

In the current study, we applied efficient MapReduce matrix multiplication with an optimized mapper set produced by MPI library for real-life business applications. We used this method to demonstrate the performance of business applications by using parallel and distributed computing matrix multiplication compared with those of the multithreaded and sequential methods.

The rest of the paper is organized as follows. Section 2 reviews works that are closely related to using the matrix multiplication in many applications. Section 3 presents the business applications using the matrix multiplication. Section 4 presents all matrix multiplication methods used. Section 5 presents experimental results, and Sect. 6 summarizes and concludes the paper.

## 2 Related Work

Mathematicians and research scientists have found many applications of matrices due to the advent of personal and large-scale computers that increased the use of matrices in a wide variety of applications, such as economics, engineering, statistics, and other sciences.

Traditional sequential algorithms for matrix multiplication consume considerable space and time. To enhance the efficiency of matrix multiplication, Fox [10], Cannon [4], and DNS [9] algorithms have been proposed for parallelizing matrix multiplication. To maximize efficiency, these approaches balance inter-process communication, dependencies, and parallelism level. Parallel matrix multiplication relies on the independence of multiplication, which includes multiple independent element-to-element multiplications and multiple aggregations of independent multiplication results.

Zhang et al. [21] presented an outsourcing computation schema in an amortized model for matrix multiplication of two arbitrary matrices that meet the requirements for both security and efficiency. They compared their scheme functionalities with existing works, such as Fiore's schema [6], Li's schema [11], and Jia's schema [12]. Zhang et al. [21] proved that their schema is more efficient in terms of functionality as well as computation, storage and communication overhead.

Kumar et al. [14] proposed a privacy-preserving, verifiable, and efficient algorithm for matrix multiplication in outsourcing paradigm to solve the lack of computing resources, where the client, having a large dataset, can perform matrix multiplication using cloud server. Kumar et al. [14] evaluated their algorithm on security, efficiency, and variability parameters. With high efficiency and practical usability, their algorithm can mostly replace costly cryptographic operations and securely solve matrix multiplication algorithm.

Ann et al. [2] proposed an approach for calculating the output equation of the hybrid multilayered perceptron (HMLP). They employed a matrix multiplication method simulated and compared with looping method to calculate HMLP output using loops. Their results confirmed that the output of the HMLP calculated using the proposed matrix multiplication method is the same as that calculated using the looping method. The difference is that the processing time of the former is faster than that of the latter for HMLP with more nodes, although the looping method calculated the output faster for HMLP with less nodes.

Afroz et al. [1] proposed a new approach that focused on the time analysis of different matrix multiplication algorithms by combining Karatsuba and Strassen methods for reducing time complexity and analyzing matrix multiplication constructions. They remarked that if the methods are perfect, then the approach can be used to pick between the different algorithms, thus creating a hybrid algorithm.

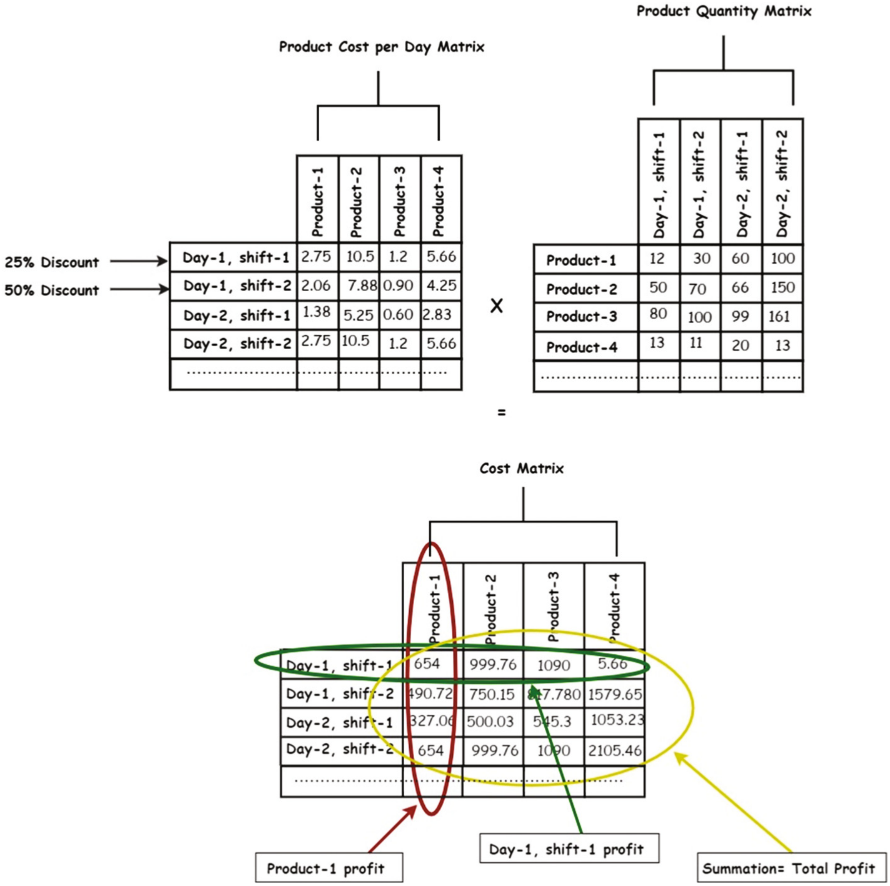
## 3 Business Applications

Market shop is a business application that uses matrix multiplication to calculate yearly, monthly, or even daily profit. Through matrix multiplication, market shop also determines the most purchased product in the market or the effect of purchase quantity due to discount day based on the amount of purchase. Different matrix multiplication methods have been implemented on market shop that uses matrices for the store cost of each product and the quantity purchased for each product.

In the proposed market shop schema, the first matrix collects the cost for each product in the market, and the second matrix stores the quantity size of each product purchased during discount day. Figure 1 illustrates the matrix architecture of the market shop.

- Product cost per day matrix row represents the cost of each product for different days, whereas the column represents the product price for each day before and after discount day.
- Product quantity matrix row represents the quantity of one product purchased during different shifts of discount day, whereas the column represents the quantity of all products purchased during different shifts of the day.

- Cost matrix row represents the total amount cost for all products during different shifts of discount day, whereas the column represents the total amount cost of one product during different shifts of the day. The summation of this matrix equals the total profit of the market.



| | Product-1 | Product-2 | Product-3 | Product-4 |
|---|---|---|---|---|
| Day-1, shift-1 | 2.75 | 10.5 | 1.2 | 5.66 |
| Day-1, shift-2 | 2.06 | 7.88 | 0.90 | 4.25 |
| Day-2, shift-1 | 1.38 | 5.25 | 0.60 | 2.83 |
| Day-2, shift-2 | 2.75 | 10.5 | 1.2 | 5.66 |
| ……… | ……… | ……… | ……… | ……… |

| | Day-1, shift-1 | Day-1, shift-2 | Day-2, shift-1 | Day-2, shift-2 |
|---|---|---|---|---|
| Product-1 | 12 | 30 | 60 | 100 |
| Product-2 | 50 | 70 | 66 | 150 |
| Product-3 | 80 | 100 | 99 | 161 |
| Product-4 | 13 | 11 | 20 | 13 |
| ……… | ……… | ……… | ……… | ……… |

| | Product-1 | Product-2 | Product-3 | Product-4 |
|---|---|---|---|---|
| Day-1, shift-1 | 654 | 999.76 | 1090 | 5.66 |
| Day-1, shift-2 | 490.72 | 750.15 | 847.780 | 1579.65 |
| Day-2, shift-1 | 327.06 | 500.03 | 545.3 | 1053.23 |
| Day-2, shift-2 | 654 | 999.76 | 1090 | 2105.46 |
| ……… | ……… | ……… | ……… | ……… |

**Fig. 1.** Matrix architecture of the market shop

# 4 Matrix Multiplication Methods

## 4.1 MPI

MPI is a library of routines that can be used to create parallel programs in C or Fortran77. It is a library that runs with standard C or Fortran programs, using commonly-available operating system services to create parallel processes and exchange information among these processes.

MPI is designed to allow users to create programs that can run efficiently on most parallel architectures. The design process included vendors (such as IBM, Intel, TMC, Cray, Convex, etc.), parallel library authors (involved in the development of PVM, Linda, etc.), and applications specialists [27].

MPI can also support distributed program execution on heterogeneous hardware. That is, you may run a program that starts processes on multiple computer systems to work on the same problem. This is useful with a workstation farm. These programs cannot communicate with each other by exchanging information in memory variables. Instead they may use any of many MPI communication routines. The two basic routines are MPI_Send, to send a message to another process, and MPI_Recv, to receive a message from another process. The MPI code has been run in IMAN1 which is the Jordan's first and fastest High-Performance Computing resource, funded by JAEC and SESAME [28].

## 4.2 MapReduce

Traditional parallel-based matrix multiplication has been recently replaced with MapReduce, a parallel and distributed framework for large-scale data [17]. Typical MapReduce-based matrix multiplication requires two MapReduce jobs. The first job creates a pair of elements for multiplication by combining input arrays together during map task. The reduce task of this job is inactive at this point. In the second job, the map task independently implements the multiplication operations on each pair of elements. The reduce job aggregates the results corresponding to each output element.

Hadoop is a Java open-source platform used for developing MapReduce applications. Google developed this platform [17]. Figure 2 illustrates the Hadoop architecture.

In this paper, we used the MapReduce-based matrix multiplication proposed by [13], which reduces both time and memory utilization compared with existing schemas [13]. In the proposed technique, matrix multiplication is implemented as an element-to-block schema, as illustrated in Fig. 3. In the first schema; the first array is decomposed into individual elements, whereas the second array is decomposed into sub-row-based blocks. In the second schema, the first array is decomposed into sub-row-based blocks, and the second array is decomposed into sub-column-based blocks. The number of mappers is

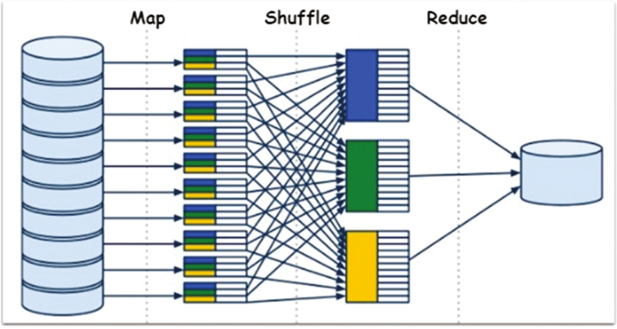


**Fig. 2.** Hadoop MapReduce architecture

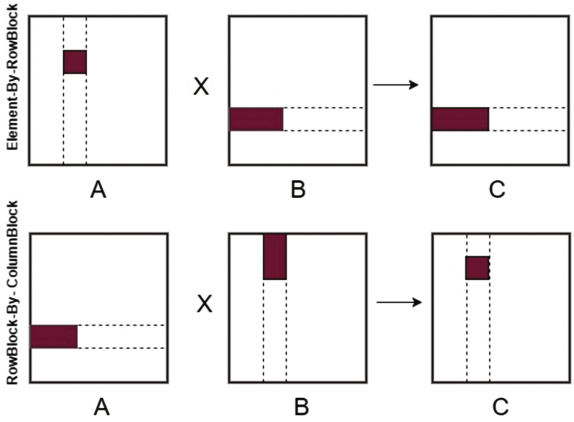


**Fig. 3.** Efficient MapReduce matrix multiplication techniques

determined by the size of the block generated for the second array and selected with the capability of the underlying mapper as basis. Subsequently, a small block size increases the number of blocks, thus requiring additional mappers, and vice versa.

The map task (see Table 1) is responsible for the multiplication operations, whereas the reduce task is responsible for the sum operations. The pre-processing step reads an element from the first array and a block from the second array and then merges them into one file. In matrix multiplication, the entire row in the first array must be multiplied with the entire column in the second array to calculate the results of an element in the output. Thus, the results of each mapper in the proposed schemas are aggregated with other multiplication results in the reduce task.

**Table 1.** Efficient MapReduce matrix multiplication operations

| Scheme | | Input | Output |
|---|---|---|---|
| Element- By- Row-Block | Pre-Process | Files | ‹$a_{ij}$, $b_{kj}$, $b_{kj}$ ........› |
| | Map | ‹$a_{ij}$, $b_{kj}$, $b_{kj}$........› | ‹key,<br>[$a_{ij} * b_{kj}$] ......... [$a_{ij} * b_{kj}$]› |
| | Reduce | ‹key, $c_{ij}$, $c_{ij}$ ......› | ‹key, $c_{ij} + c_{ij}$ ......› |
| Row-Block- By- Column-Block | Pre-Process | Files | ‹$a_{ij}$, $a_{ij}$ ...... $b_{kj}$, $b_{kj}$› |
| | Map | ‹$a_{ij}$, $a_{ij}$ ... $b_{kj}$, $b_{kj}$› | ‹key,<br>[$a_{ij} * b_{kj}$] ......... [$a_{ij} * b_{kj}$]› |
| | Reduce | ‹key, $c_{ij}$, $c_{ij}$ ......› | ‹key, $c_{ij} + c_{ij}$ ......› |

## 5 Experiments and Result

Different matrix multiplication methods have been implemented to compare their efficiencies and time performances in a large matrix. All methods were tested in sparse and dense matrices of different sizes. Results of the tested methods are discussed below.

**MPI:** The MPI code was run in IMAN1, Jordan's first and fastest high-performance computing resource, funded by JAEC and SESAME. We also worked in a Zaina server, an Intel Xeon-based computing cluster with 1G Ethernet interconnection. The cluster is mainly used for code development, code porting, and synchrotron radiation application purposes. In addition, this cluster is composed of two Dell PowerEdge R710 and five HP ProLiant DL140 G3 servers (Table 2).

**Table 2.** Zaina technical details

| | |
|---|---|
| Server | 7 Servers (Two Dell PowerEdge R710 and five HP ProLiant DL140 G3) |
| CPU per server | Dell (2 × 8 cores Intel Xeon) HP (2 × 4 cores Intel Xeon) |
| RAM per server | Dell (16 GB) HP (6 GB) |
| Total storage (TB) | 1 TB NFS Share |
| OS | Scientific Linux 6.4 |

Technical details are given as follows. The results for different numbers of core are shown in Table 3. As shown in the table of results, increasing the core number up to 8 in the matrix, which is less than 500, takes more time due to the small size of problem that does not need a large number of cores, whereas the matrix within the range of 500 does not need more than 8 cores for its problem because it is inefficient and takes more time. In comparison, the matrix size of up to 1000 is more effective and efficient when increasing the number of cores to 32 due to the large size of problems that need a higher degree of parallelism.

**Table 3.** MPI run time result

| Matrix | Core | | | | |
|---|---|---|---|---|---|
| | 2 | 4 | 8 | 16 | 32 |
| *Dense matrix* | | | | | |
| 250 * 250 | 1.245 s | 1.189 s | 1.214 s | 1.284 s | 1.529 s |
| 500 * 500 | 1.992 s | 1.456 s | 1.390 s | 1.501 s | 1.681 s |
| 1000 * 1000 | 7.491 s | 3.435 s | 2.968 s | 2.768 s | 2.655 s |
| 2000 * 2000 | 62.207 s | 22.473 s | 17.569 s | 14.390 s | 10.955 s |
| 4000 * 4000 | 540.819 s | 185.790 s | 135.654 s | 90.827 s | 88.593 s |
| *Sparse matrix* | | | | | |
| 250 * 250 | 1.243 s | 1.190 s | 1.213 s | 1.316 s | 1.552 s |
| 500 * 500 | 1.971 s | 1.453 s | 1.424 s | 1.376 s | 1.691 s |
| 1000 * 1000 | 7.586 s | 3.469 s | 2.786 s | 2.488 s | 2.731 s |
| 2000 * 2000 | 62.498 s | 22.398 s | 17.363 s | 12.719 s | 11.933 s |
| 4000 * 4000 | 537.697 s | 201.295 s | 137.665 s | 107.900 s | 101.656 s |

The speedup is the ratio between the sequential time and the parallel time. The speedup for different numbers of core on sparse and dense matrices of different sizes

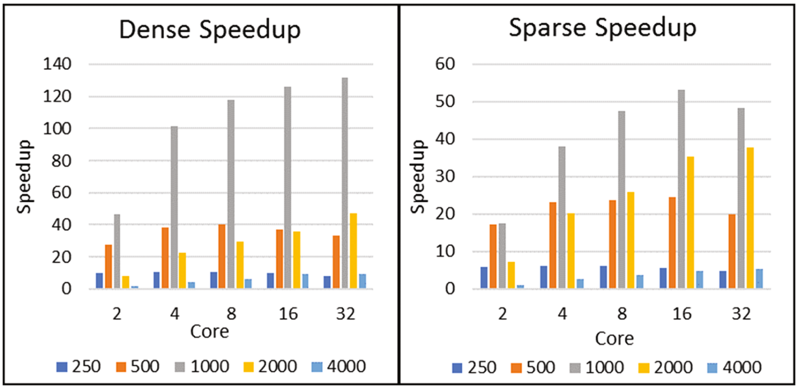


**Fig. 4.** MPI speedup plotting

are illustrated in Fig. 4. The results show that MPI achieves the best speedups values, especially on large number of processors in dense matrices.

**MapReduce:** The MapReduce results of the matrix multiplication using Hadoop for inputs with various sizes are presented. Simple matrix multiplication process on the platform with various block sizes has been run to determine the optimal length to be given to the mapper before running the actual job and the optimal length of block size with a minimum running time was 20. The running time was cut down in the proposed schemes, as the sorting process in the shuffling process was reduced. As the matrix size grows, the stability of the proposed scheme is almost linear. Results are given in Table 4.

The speedup for sparse and dense matrices of different sizes are illustrated in Fig. 5. The results show that MapReduce achieves speedups values less than MPI, especially on large number of processors in dense matrices.

**Multithreaded:** Matrix multiplication using different sizes of thread were tested in various sizes of dense and sparse matrices. Results are given in Table 5. As shown in the table of results, when we increased the thread number up to 4 in the matrix, which is less than 500, more time is required due to the small size of problem that does not need a large number of threads, whereas the matrix within the range of 500 does not need more than 16 cores for its problem. Moreover, the matrix size up to 1000 is more

**Table 4.** MapReduce run time result

| Dense matrix | | Sparse matrix | |
|---|---|---|---|
| Matrix | Time | Matrix | Time |
| 250 * 250 | 6.255 s | 250 * 250 | 4.021 s |
| 500 * 500 | 12.996 s | 500 * 500 | 11.581 s |
| 1000 * 1000 | 98.324 s | 1000 * 1000 | 95.452 s |
| 2000 * 2000 | 116.014 s | 2000 * 2000 | 112.450 s |
| 4000 * 4000 | 321.457 s | 4000 * 4000 | 185.478 s |

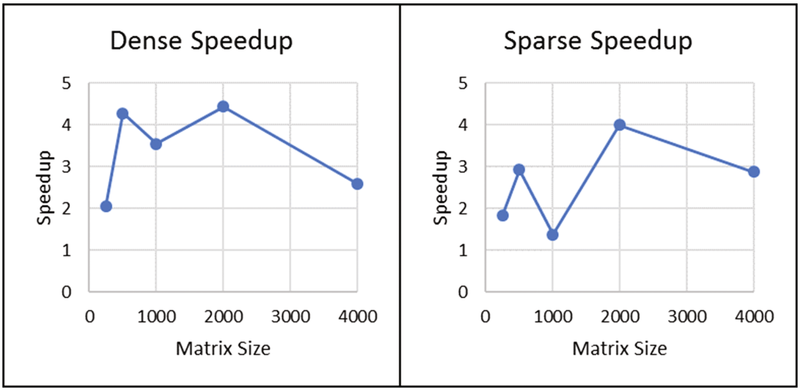


**Fig. 5.** MapReduce speedup plotting

**Table 5.** Multithreaded run time result

| Matrix | Core time | | | | |
|---|---|---|---|---|---|
| | 2 | 4 | 8 | 16 | 32 |
| *Dense matrix* | | | | | |
| 250 * 250 | 11.408 s | 11.527 s | 10.049 s | 9.038 s | 8.255 s |
| 500 * 500 | 45.485 s | 41.675 s | 39.271 s | 36.734 s | 36.996 s |
| 1000 * 1000 | 299.734 s | 234.888 s | 225.802 s | 215.119 s | 204.367 s |
| 2000 * 2000 | 444.012 s | 380.215 s | 315.241 s | 248.562 s | 220.145 s |
| 4000 * 4000 | 750.125 s | 664.021 s | 601.241 s | 521.045 s | 412.547 s |
| *Sparse matrix* | | | | | |
| 250 * 250 | 6.534 s | 6.625 s | 6.447 s | 6.193 s | 6.021 s |
| 500 * 500 | 29.079 s | 27.810 s | 26.829 s | 25.369 s | 23.981 s |
| 1000 * 1000 | 117.060 s | 111.784 s | 106.850 s | 103.137 s | 99.352 s |
| 2000 * 2000 | 287.179 s | 235.122 s | 201.252 s | 162.547 s | 155.842 s |
| 4000 * 4000 | 349.346 s | 299.734 s | 234.888 s | 225.802 s | 215.347 s |

effective and efficient when the number of threads is increased to 32 due to the large size of the problem that needs a higher degree of parallelism.

The speedup for sparse and dense matrices of different sizes are illustrated in Fig. 6. The results show that Multithreaded achieves speedups values less than MPI and MapReduce.

**Sequential:** The sequential results of matrix multiplication were tested in dense and sparse matrices of various sizes of. Results are given in Table 6.

The comparison between MPI and MapReduce results are always faster and more efficient for the different-sized dense and sparse matrices than those of the multi-threaded and sequential methods, as shown in the efficiency table below. The MPI outperformed the MapReduce; thus, the research goal is achieved. Comparison results are given in Table 7 and illustrated in Figs. 7 and 8.

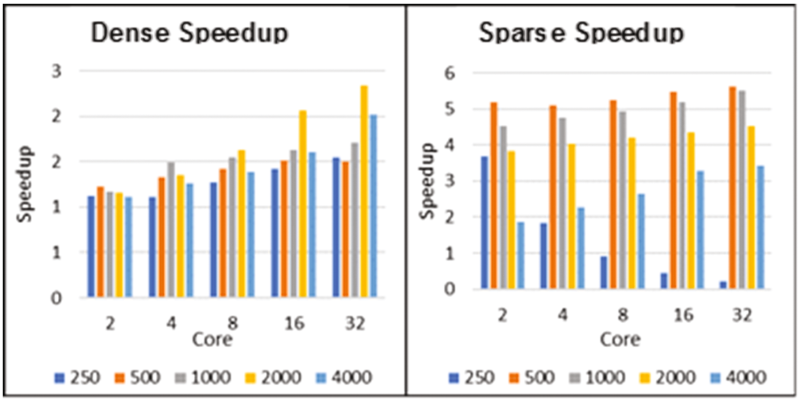


**Fig. 6.** Multithreaded speedup plotting

**Table 6.** Sequential run time results

| Dense matrix | | Sparse matrix | |
|---|---|---|---|
| Matrix | Time | Matrix | Time |
| 250 * 250 | 12.826 s | 250 * 250 | 7.359 s |
| 500 * 500 | 55.656 s | 500 * 500 | 33.875 s |
| 1000 * 1000 | 349.346 s | 1000 * 1000 | 132.083 s |
| 2000 * 2000 | 514.501 s | 2000 * 2000 | 450.485 s |
| 4000 * 4000 | 834.501 s | 4000 * 4000 | 533.269 s |

**Table 7.** Time efficiency result

| Matrix | Sequential | | Multithreaded | |
|---|---|---|---|---|
| | MapReduce | MPI | MPI | MapReduce |
| *Dense matrix time efficiency* | | | | |
| 250 * 250 | 51.23% | 80.24% | 85.60% | 24.23% |
| 500 * 500 | 76.65% | 94.26% | 96.24% | 64.87% |
| 1000 * 1000 | 71.85% | 97.50% | 98.70% | 51.89% |
| 2000 * 2000 | 77.45% | 92.34% | 95.02% | 47.30% |
| 4000 * 4000 | 61.48% | 52.79% | 78.53% | 22.08% |
| *Sparse matrix time efficiency* | | | | |
| 250 * 250 | 45.36% | 83.83% | 80.24% | 33.22% |
| 500 * 500 | 65.81% | 95.94% | 94.26% | 51.71% |
| 1000 * 1000 | 27.73% | 98.12% | 97.50% | 3.93% |
| 2000 * 2000 | 75.04% | 97.35% | 92.34% | 27.84% |
| 4000 * 4000 | 65.22% | 80.94% | 52.79% | 13.87% |

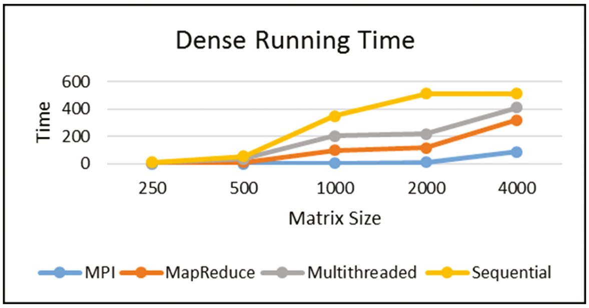


**Fig. 7.** Dense compassion plotting

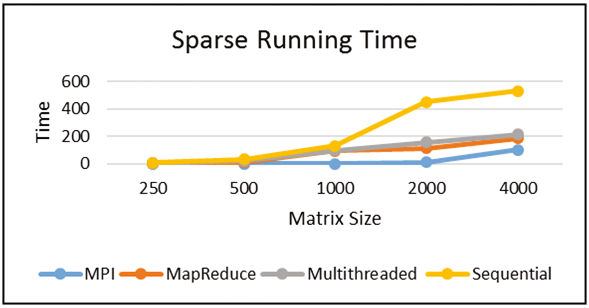


**Fig. 8.** Sparse compassion plotting

## 6 Conclusion

Using the conducted experimental study as basis, MPI and MapReduce matrix multiplication are always faster than the multithreaded and sequential methods, with 90.11% and 47.94% efficiency, respectively. Hence, parallel and distributed computing for matrix multiplication methods have been proposed to reduce the cost and time of matrix multiplication over multiple processors. MPI matrix multiplication is also more efficient, because its matrix size growth outperforms those of the multithreaded and sequential methods.

## References

1. Afroz, S., Tahaseen, M., Ahmed, F., Farshee, K.S., Huda, M.N.: Survey on matrix multiplication algorithms. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 151–155. IEEE, May 2016
2. Ann, L.Y., Ehkan, P., Mashor, M.Y., Sharun, S.M.: Calculation of hybrid multi-layered perceptron neural network output using matrix multiplication. In: 2016 3rd International Conference on Electronic Design (ICED), pp. 497–500. IEEE, August 2016

3. Catalyurek, U.V., Aykanat, C.: Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication. IEEE Trans. Parallel Distrib. Syst. **10**(7), 673–693 (1999)
4. Cannon, L.E.: A Cellular Computer to Implement the Kalman Filter Algorithm. No. 603-Tl-0769. Montana State Univ Bozeman Engineering Research Labs (1969)
5. Coppersmith, D., Winograd, S.: Matrix multiplication via arithmetic progressions. In: Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing, pp. 1–6. ACM, January 1987
6. Fiore, D., Gennaro, R.: Publicly verifiable delegation of large polynomials and matrix computations, with applications. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, pp. 501–512. ACM (2012)
7. Dean, G., Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: OSDI, p. 10. USENIX (2004)
8. Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Commun. ACM **53**(1), 72–77 (2010)
9. Dekel, E., Nassimi, D., Sahni, S.: Parallel matrix and graph algorithms. SIAM J. Comput. **10** (4), 657–675 (1981)
10. Fox, G.C., Otto, S.W., Hey, A.J.G.: Matrix algorithms on a hypercube I: matrix multiplication. Parallel Comput. **4**(1), 17–31 (1987)
11. Li, H., Zhang, S., Luan, T.H., Ren, H., Dai, Y., Zhou, L.: Enabling efficient publicly verifiable outsourcing computation for matrix multiplication. In: 2015 International Telecommunication Networks and Applications Conference (ITNAC), pp. 44–50. IEEE (2015)
12. Jia, K., Li, H., Liu, D., Yu, S.: Enabling efficient and secure outsourcing of large matrix multiplications. In: 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2015)
13. Kadhum, M., Qasem, M.H., Sleit, A., Sharieh, A.: Efficient MapReduce matrix multiplication with optimized mapper set. In: Computer Science On-line Conference, pp. 186–196. Springer, Cham, April 2017
14. Kumar, M., Meena, J., Vardhan, M.: Privacy preserving, verifiable and efficient outsourcing algorithm for matrix multiplication to a malicious cloud server. Cogent Eng. (just-accepted) 1295783 (2017)
15. Liu, X., Iftikhar, N., Xie, X.: Survey of real-time processing systems for big data. In: Proceedings of the 18th International Database Engineering and Applications Symposium. ACM (2014)
16. Lv, Z., et al.: Parallel K-means clustering of remote sensing images based on MapReduce
17. Norstad, J.: A mapreduce algorithm for matrix multiplication (2009). http://www.norstad.org/matrix-multiply/index.html. 19 Feb 2013
18. Thabet, K., Al-Ghuribi, S.: Matrix multiplication algorithms. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **12**(2), 74 (2012)
19. Wu, G., et al.: MReC4. 5: C4. 5 ensemble classification with MapReduce. In: 2009 Fourth ChinaGrid Annual Conference. IEEE (2009)
20. Zaharia, M., et al.: Job scheduling for multi-user mapreduce clusters. EECS Department, University of California, Berkeley, Technical report UCB/EECS-2009-55 (2009)
21. Zhang, S., Li, H., Jia, K., Dai, Y., Zhao, L.: Efficient secure outsourcing computation of matrix multiplication in cloud computing. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE, December 2016
22. Saadeh, M., Saadeh, H., Qatawneh, M.: Performance evaluation of parallel sorting algorithms on IMAN1 supercomputer. Int. J. Adv. Sci. Technol. **95**, 57–72 (2016)

23. Mohammed, Q.: Embedding linear array network into the tree-hypercube network. Eur. J. Sci. Res. **10**(2), 72–76 (2005)
24. Qatawneh, M., Alamoush, A., Alqatawna, J.: Section based hex-cell routing algorithm (SBHCR). Int. J. Comput. Netw. Commun. (IJCNC) **7**(1) (2015)
25. Qatawneh, M.: Multilayer hex-cells: a new class of hex-cell interconnection networks for massively parallel systems. Int. J. Commun. Netw. Syst. Sci. **4**(11), 704–708 (2011)
26. Qatawneh, M.: Embedding binary tree and bus into hex-cell interconnection network. J. Am. Sci. **7**(12) (2011)
27. Mohammad, Q., Khattab, H.: New routing algorithm for hex-cell network. Int. J. Future Gener. Commun. Netw. **8**(2) (2015)
28. Qatawneh, M.: New efficient algorithm for mapping linear array into hex-cell network. Int. J. Adv. Sci. Technol. **90** (2016)
29. Qasem, M.H., Al Assaf, M.M., Rodan, A.: Data mining approach for commercial data classification and migration in hybrid storage systems. World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. **10**(3), 481–484 (2016)
30. Qasem, M.H., Faris, H., Rodan, A., Sheta, A.: Empirical evaluation of the cycle reservoir with regular jumps for time series forecasting: a comparison study. In: Computer Science On-line Conference, pp. 115–124. Springer, Cham, April 2017
31. Lin, J., Dyer, C.: Data-intensive text processing with MapReduce. Synth. Lect. Hum. Lang. Technol. **3**(1), 1–177 (2010)

# Content Generation for Massively Multiplayer Online Games with Genetic Algorithms

Tiago Alves[1], Jorge Coelho[2](✉), and Luís Nogueira[3]

[1] Blip.pt, Porto, Portugal
tiago.alves@blip.pt
[2] ISEP/LIACC, Porto, Portugal
jmn@isep.ipp.pt
[3] ISEP/CISTER/INESC-TEC, Porto, Portugal
lmn@isep.ipp.pt

**Abstract.** Procedural content generation can be defined as the algorithmical creation of game content with limited or indirect user input. In this paper we present a procedural content generation genetic algorithm for massively multiplayer online games. The incremental generation of content by choosing the most appropriate selection of added blocks allows an efficient progress in the game with a small impact on performance and the consequent ability to deploy such type of game in low performance mobile devices.

## 1 Introduction

Massive multi-player online (MMO) games typically support a large number of players, simultaneously interacting with each other in extensive game worlds built with different types of blocks connected together. Normally, MMO games encourage players to explore the unknown in order to evolve. To be addictive, the distribution of points of interest to the players around the game world must be done in an intelligent manner. For example, if it takes too long for a player to reach some point of interest, he may feel a lack of reward for the time he spent, feel bored and leave the game. The same may happen if he finds all the needed points of interest near him without the need to explore the world. At the same time, he must feel that each new place he reaches is somehow unique. If he does not, his desire for exploration will decrease. Therefore, game content is an important factor in keeping players engaged in gaming worlds. In fact, procedural game content generation (PCG) is evolving rapidly, driven by the increasing demand from game development companies [8]. Reducing the game designer's work using PCG methods allows games to be produced faster and cheaper, while preserving quality. They can also drastically reduce game packages size, allowing them to reach a wider audience, including mobile environments. Without reservation we can say that the introduction of PCG methods is an inflection point for mobile gaming. The processing capabilities delivered by current mobile processors will significantly accelerate development of console and PC class games for mobile devices. Games running on current mobile processors will deliver higher level of

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_4

graphics quality and realism that sets a new standard for mobile gaming experience. In a previous work [1], we presented a framework capable of generating worlds with structures that obey a previously configured set of rules, comprising several needs found in current MMO games. To the best of our knowledge there is no other similar framework publicly available. As it is our goal to provide programmers the maximum degree of flexibility, the framework allows the addition of types of content along with new rules that constraint the placement of that content in the game. A key term here is "content". In our definition, content is most of what is contained in a game: levels, maps, game rules, textures, stories, items, quests, music, weapons, vehicles, characters, etc. One consequence of this approach is that it allows the introduction of an arbitrary number of new rules that may impose a performance bottleneck. Additionally, it may happen that, according to the imposed rules, there is no candidate block of content for a given new position in the map. By experience, we found this approach to be too restrictive. It is more appropriate to allow rules to have some degree of flexibility as well. For example, if we have 3 blocks that could be placed in some new position but all of them violate a rule of proximity with other blocks, we may decide to have that position empty or optionally choose the block which has the less negative impact. The contribution of this paper is the solution to these problems with a genetic-based block selection algorithm. The remaining of this paper is organised as follows. In Sect. 2, we present procedural generation techniques and its application in the context of MMO games. In Sect. 3, we present the content generation concepts and related definitions and in Sect. 4 we briefly describe the implementation and detail the block selection process. Finally, in Sect. 5 we conclude the paper.

The terms world and map usually describe the same area in a MMO game. Thus, for the rest of this paper these two terms – world and map – will be used without distinction. The same applies to the terms rule and constraint which describe restrictions on the content generation.

## 2 Procedural Content Generation in MMO Games

In Procedural content generation (PCG) the generated content may coexist with authored content since the two are not mutually exclusive. This concept has been used in gaming context for a long time. For example, in the early eighties, game developers used PCG to deal with the limited storage capabilities that home computers had at that time. Also, PCG is used as a way to provide infinite new experiences to the players as it is the case with recent games like Minecraft [5]. PCG has two main advantages: reducing the effort needed for the generation of content by programmers and provide a solution for the creation of content in devices with limited resources. This last advantage is of great importance since it can make the game available to a broader audience, namely the one that relies on mobile devices. A deep analysis of PCG in game development can be found in [14].

In the development of our framework we studied several games that use PCG techniques: Don't Starve [9] is a single-player survival game with a randomly generated open world. Only the area immediately around the player is shown and more is generated as the player explores the world. Don't Starve distributes resources in an intelligent manner in order to challenge players to find those resources [15]. Rust [17] is a survival multiplayer game played in an open world. In order to survive, players must gather resources from the environment, such as wood and rocks, and craft tools by mixing those resources. Players interact with each other if they are playing in the same server. Each server has its own world that was procedurally generated based on a seed value given when the instance of the game is started. Rust's worlds are divided into large, geographically distinct areas: desert in the South, forest in the center and a snowy one in the North. Each area has its own animals and resources. These animals and resources disappear when consumed by a player and have a timespan to respawn. Since resources are consumed and take time to reappear, players have the need to explore the world to search for more resources. Civilization V [11] is another game that uses procedural generation. Here the players are able to pick the world shape by selecting which generation algorithm to use and setting some generation variables. Every world is different, however the structure is the same every time depending on which algorithm was chosen. In Civilization V, worlds are always structured with hexagonal grids where the world structure is highlighted. More about hexagonal grids in gaming context can be found in [7]. In game, the player leads a civilization searching for different achievements in the scope of research, exploration, and expansion. Pioneers [6] is a turn-based exploration RPG. A turn-based RPG is a type of role playing game where the players face battles that consist of turns. In these turns, the player can command their characters to perform various actions to defeat the opponents. In Pioneers, the player leads a group of travelers in an adventure in a procedurally generated world. In this world, players search for temples and tribes solving puzzles along the way and gathering resources to survive through the year's seasons. The player can only see the world near him. This encourages the player to explore new parts of the world. In [8] a six-layered taxonomy for procedural generation of game content is introduced: *bits*, *space*, *systems*, *scenarios*, *design* and *derived*. The authors of the survey represent this layers as a pyramid in which layers closer to the top may be built with elements from the layers at the bottom. Our area of application is in the *scenarios* layer and more precisely in the *levels* section. *Levels* consist of the playable game space and are of extreme importance in the game world design. The games presented in the previous section use techniques such as pseudo-random number generators (PRNG) for the creation of replicable sequences of random numbers based on a shared seed which can be useful to dynamically create the same map several times. More techniques often used are midpoint displacement algorithms [10] and Perlin Noise [13]. Both techniques are used to generate height maps in order to create realistic looking terrains and landscapes. Other common techniques are Simulation of Complex Systems techniques such as Cellular Automata [2] and Agent-based Simulation [3], Image Filtering (IF) and Spatial algorithms (SA). Genetic algorithms are surveyed in [16] and are studied in the context of the generation of content for games in [14]. Further details about these subjects can be found in [8,14].

## 3 Content Generation Concepts

We propose the decomposition of each world in a set of blocks that can be connected. We define these notions formally in this section.

**Definition 1 (World Template).** *A World Template* $\mathscr{W}_T$ *is a set* $\{\mathscr{I}_d, \mathscr{N}_{sides}, \mathscr{S}_b, \mathscr{S}_c, \mathscr{C}_w\}$ *where* $\mathscr{N}_{sides} \in \{4, 6\}$ *is the number of sides of all block units in the world,* $\mathscr{S}_b$ *is the set of blocks,* $\mathscr{S}_c$ *is the set of connectors and* $\mathscr{C}_w$ *are the world based constraints.*

**Definition 2 (Block).** *A block* $\mathscr{B}$ *is a world unit with the set of properties* $\{\mathscr{I}_d, \mathscr{L}, \mathscr{S}, \mathscr{C}_b\}$ *where* $\mathscr{I}_d$ *is a unique identifier,* $\mathscr{L}$ *is a list of classes,* $\mathscr{S}$ *is a set of pairs* $(s_i, c_i)$ *where side* $s_i$ *has connector* $c_i$ *and* $\mathscr{C}_b$ *are the block based constraints.*

**Definition 3 (Connector).** *A connector* $\mathscr{C}$ *defines the compatibility between blocks and is defined by* $\{\mathscr{I}_d, \mathscr{T}, \mathscr{B}_i, \mathscr{B}_c\}$ *where* $\mathscr{I}_d$ *is the connector unique identifier,* $\mathscr{T}$ *is its type which is one of bl (blacklist) or wl (whitelist),* $\mathscr{B}_i$ *is a set of block identifiers and* $\mathscr{B}_c$ *is a set of block classes.*

*Example 1.* A connector $\mathscr{C}_1 = \{c_1, wl, \{\mathscr{B}_1, \mathscr{B}_2\}, \{\}\}$ means that the side of the connected block can only be connected with a block $\mathscr{B}_1$ or $\mathscr{B}_2$. A connector $\mathscr{C}_2 = \{c_2, bl, \{\mathscr{B}_3\}, \{\}\}$ means that the side of the connected block can be connected with any block except for $\mathscr{B}_3$.

### 3.1 World Instance Generation

The generation of a world instance is based on an initial configuration where the programmer defines a *World template*, its *Blocks* and associated *Connectors* and a set of *Constraints*. These constraints may be of one of the following 2 categories: world based and block based.

World based constraints are the ones related with the map as a whole and defines, for example, the initial map state and how much it can expand. We propose the following world based constraints: *Initial map size*, *Horizontal and vertical boundaries* and *Map center*. On the other hand, as the name denotes, block based constraints are the ones related with the blocks and have effect on the selection process. For example, block based constraints may forbid the placement of a block in a given world position if certain conditions apply. We propose the following block based constraints: *Blacklist connectors*, *Whitelist connectors*, *Maximum occupation*, *Maximum occupation by percentage* and *Minimum distance to other blocks*. As mentioned before the framework allows the insertion of new rules by the programmer.

When a player moves to a new area, the generation process resumes. Also the world instance is evaluated on a periodic base and blocks with an associated lifespan may expire and be unavailable after a given evaluation. World generation stops when a given number of block units, previously defined by the game designer, is available to all the game players.

We divide the world generation in 3 different phases. The first one runs only once at the very start of the world generation to create the initially visible part of the world. The second one runs multiple times to generate more world for a given world position and is triggered, for example, when a player moves to unexplored locations. The third phase runs to invalidate blocks due to timeouts or other constraints. The implementation of the first, third, a previous approach to the second and the communication process is described in detail in [1]. We will focus on a new approach for the second phase.

*Example 2 (World instance generation).* Given a World Template $\mathscr{W}_T = \{s_W, 4, \mathscr{S}_b, \mathscr{S}_c, \mathscr{C}_w\}$, where the set of connectors are:

- $\{c_1, wl, \{\mathscr{B}_1\}\}$
- $\{c_2, wl, \{\mathscr{B}_2\}\}$
- $\{c_3, wl, \{\mathscr{B}_3\}\}$
- $\{c_4, wl, \{\mathscr{B}_1, \mathscr{B}_2, \mathscr{B}_3\}\}$

and the set of blocks are:

- $\{\mathscr{B}_1, d_1, L_1, \{(top, c_2), (right, c_3), (bottom, c_2), (left, c_3)\}\}$
- $\{\mathscr{B}_2, d_2, L_2, \{(top, c_1), (right, c_3), (bottom, c_1), (left, c_3)\}\}$
- $\{\mathscr{B}_3, d_3, L_3, \{(top, c_4), (right, c_4), (bottom, c_4), (left, c_4)\}\}$
- $\{\mathscr{B}_4, d_4, L_4, \{(top, c_2), (right, c_3), (bottom, c_2), (left, c_3)\}\}$

For the sake of clarity, the example above only defines 3 blocks and their compatibility using 4 whitelist connectors. No world or block based constraints were taken in account for the output and only the first $3 \times 3$ map units are represented. We can see that left and right sides of $\mathscr{B}_1$ are only compatible with $\mathscr{B}_3$ blocks and that top and bottom sides are only compatible with $\mathscr{B}_2$. Left and right sides of $\mathscr{B}_2$ are only compatible with $\mathscr{B}_3$ blocks and that top and bottom sides are only compatible with $\mathscr{B}_1$. Finally, all $\mathscr{B}_3$ sides have the same connector ($\mathscr{C}_4$) and so they are compatible with any of the 3 blocks, $\mathscr{B}_1$, $\mathscr{B}_2$ and $\mathscr{B}_3$.

### 3.2 Square Grids

Square grids are the most common grids used in games, primarily because they are easy to use. To reference a certain square of the grid it is enough a pair of cartesian coordinates (x, y) and to reference a certain edge of a square, a unique identifier is needed some (e.g.: letter). Each square has 4 direct neighbours, one per side. The movement directions available for the player in a game is not directly related to the number of sides of each map tile. For example, a map composed by tiles with four sides may allow a player to move diagonally from tile to tile, besides the typical vertical and horizontal movements. If a certain game that uses a square grid allows the player to move in diagonal directions, the player will take longer to reach the next block when moving diagonally than when moving vertically or horizontally. This happens because the square centers are not all at the same distance. There are other grid representations that do not have this problem. For example, hexagonal grids because the centers of all hexagons are at the same distance.

# 4 Implementation

Our implementation consists of 3 different standalone modules: *rule- based- map-generator*, *blocker* and *random-matrix*. Although each module can be used as a unique dependency, *blocker* has *rule-based-map-generator* as a dependency as well as *rule-based-map-generator* has *random-matrix* as one of its dependencies.

The idea of making *rule-based-map-generator* logic completely independent (and therefore an isolated module that can be used by other projects) of the server logic provided by *blocker* aimed to fulfill different needs described by the following case scenarios:

1. the world is generated on the client side with no need of network communications
2. the world is generated on the server side but the game developer will choose how to transmit the world to the clients
3. the world is generated on the server side and sent to clients through socket communication

For case 1 and 2, the game developer should use *rule-based-map-generator* module. For case 3, the game developer should use the *blocker* module. The rule based map generator module includes all the map generation logic and exposes all necessary methods to create world instances by means of an API. Since there are several map structures, being square and hexagonal grids the most common ones, a strategy pattern was used for the generation process. This way each world instance holds the family of algorithms with the behavior needed for its map structure. We decided to center our efforts on the development of square grids because it is one of the most common map structures and its map representations are easier to understand when compared to hexagonal grid maps. The *Blocker* is a wrapper around *rule-based-map-generator* that handles real time communication between web clients and the server using *WebSocket*s [12]. We wanted our framework to be able to open communication sessions between the clients and the server using *WebSocket*s, so both can trade messages without having to poll each other, and therefore reducing the network payload and decreasing request times. In a multiplayer game context, we define *room* as a virtual channel where players can join and interact with each other in the same game world. Sometimes these *rooms* are simply called *servers* by the players but in fact, a single server may provide different *rooms* where players can connect to. *Blocker* allows one to have multiple rooms simultaneously. Each room is running on a different port and has its own world instance. Each player can choose what room to connect to and share the world with other players in the same room.

A central part of the world generation is the block selection process. This process occurs several times during the world generation and its purpose is to choose which set of blocks best fits a given position in the world being explored. The decision is based on the block based constraints and the world state. The framework allows the request of a bounded set of blocks to add to a particular map area and distribute that data among several different clients. It is in this part that the genetic algorithm is used in order to select the most suitable set of blocks minimizing the violation of constraints. We start by explaining how individuals are selected to create a population and then how they are evaluated by means of a fitness function and show new generations are created until a result is selected. We end with some considerations on the algorithm performance and usability.

### 4.1 Individuals Creation

This phase corresponds to the selection of individuals which are candidates to fit the requested map part. Selected blocks obey to the blacklist/whitelist constraints with respect to the adjacent ones (in the set of blocks and in the world map) and can occur more than once. The genetic algorithm is then concerned with all the remaining constraints.

**Definition 4.** *Given blocks $\mathscr{B}_i$, an individual is a set of randomly generated blocks $\mathscr{I} = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_n\}$ from the list of available blocks which are blacklist/white list compatible with adjacent ones and may have repeated occurrences.*

The individual creation process is triggered when a part of the map is requested for the first time. The creation process is composed by 4 phases, one for each quadrant. Each phase may or may not occur depending whether or not the requested part of the map occupies the quadrant corresponding to the phase. The ordering of the phases is the following:

1. Quadrant 1 (the upper right quadrant)
2. Quadrant 2 (the upper left quadrant)
3. Quadrant 3 (the bottom left quadrant)
4. Quadrant 4 (the bottom right quadrant)

The generation process stops when all map positions from the requested part of the map went through the block selection process.

All phases go through each row and column, selecting a suitable block for each position. All phases start with the rows/columns closer to the map center and consequently end with the rows/columns farther from the map center.

To better understand this process we will use an example that occupies the first and fourth quadrants.

*Example 3.* For a given part of the map that goes from (2, −3) to (4, 2), a short description of each phase is the following:

**Phase 1** generates map from all positions in its quadrant, this means from (2, 0) to (4, 2). It starts with the row (y = 0), going horizontally from left to right, selecting a block for each position from (2, 0) to (4, 0). Then it goes through the column (x = 2), going vertically and up, selecting a block for each position from (2, 1) to (2, 2). Then this sequence continues until all map positions have gone through the selection process.

**Phase 2** does no work since the requested part of the map does not occupy its quadrant (2).

**Phase 3** does no work since the requested part of the map does not occupy its quadrant (3).

**Phase 4** generates map from all positions in its quadrant, this means from (2, −3) to (4, −1). It starts with the row (y = −1), going horizontally from left to right, selecting a block for each position from (2, −1) to (4, −1). Then it goes through the column (x = 2), going vertically and up, selecting a block for each position from (2, −2) to (2, −3). Then this sequence continues until all map positions have gone through the selection process.

The application of the rules may allow several possible candidate blocks for each position. Thus, several combinations of blocks are acceptable solutions.

### 4.2 Fitness Function

We start by defining some auxiliary functions:

**Definition 5.** *Given a block $\mathscr{B}_k$, a constraint $c_k \in \mathscr{C}_k$ and a position $\mathscr{P}_k$ of coordinates we define the Constraint Violation Value as:*

$$\mathscr{C}_{vv}(\mathscr{B}_k, \mathscr{P}_k, c_k) \begin{cases} n \text{ } \textit{if violates constraint} \\ 0 \qquad \textit{otherwise} \end{cases} \tag{1}$$

*The n in the function's result is 1 (constraint violated) or $n > 1$ (number of times the constraint is violated) and depends on the chosen approach in the constraint implementation.*

Note that it may be important not only know that the constraint is violated, but also to measure for how much (for example one individual including blocks which already exceed a given maximum limit by 10 elements can be understood as fitter than one that includes blocks which exceed a given maximum limit by 100 elements).

**Definition 6.** *Given a block $\mathscr{B}_k$, a set of constraints $\mathscr{C}_k = \{c_1, \ldots, c_n\}$ and a position $\mathscr{P}_k$ we define Constraint Violation Sum as:*

$$\mathscr{C}_{vs}(\mathscr{B}_k, \mathscr{C}_k, \mathscr{P}_k) = \sum_{i=0}^{n} \mathscr{C}_{vv}(\mathscr{B}_k, \mathscr{P}_k, c_i) \tag{2}$$

The fitness function is defined next.

**Definition 7.** *Given an individual $\mathscr{I} = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_n\}$, a set of block constraints $\mathscr{S}_c = \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_n\}$ where each $\mathscr{C}_k$ is one or more constraints related to block $\mathscr{B}_k$ and a set of positions $\mathscr{P}_b = \{\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_n\}$ where $\mathscr{P}_k$ is the position in the map of block $\mathscr{B}_k$, then the fitness function $\mathscr{F}$ is defined as:*

$$\mathscr{F}(\{\mathscr{B}_1, \ldots, \mathscr{B}_n\}, \{\mathscr{P}_1, \ldots, \mathscr{P}_n\}) = \sum_{i=0}^{n} \mathscr{C}_{vs}(\mathscr{B}_i, \mathscr{C}_i, \mathscr{P}_i) \tag{3}$$

The goal is to minimize this value.

### 4.3 Crossing and Mutating

Crossing and mutation is relevant in order to produce fitter individuals at each iteration. The crossing operator is defined next:

**Definition 8.** *Given an individual* $\mathscr{I}_1 = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_n\}$ *and an individual* $\mathscr{I}_2 = \{\mathscr{B}'_1, \mathscr{B}'_2, \ldots, \mathscr{B}'_n\}$ *select a random element* $k$ *such that* $\mathscr{B}_k$ *and* $\mathscr{B}'_{k+1}$ *are blacklist/whitelist compatible. The new individual is* $\mathscr{I}_{new} = \{\mathscr{B}_1, \ldots, \mathscr{B}_k, \mathscr{B}'_{k+1}, \ldots, \mathscr{B}'_n\}$.

The mutation operator is defined next.

**Definition 9.** *Given an individual* $\mathscr{I}_1 = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_n\}$*, select a random element* $k$ *and* $\mathscr{B}_k$ *can be replaced by a random selected block* $\mathscr{B}'_k$ *such that* $\mathscr{B}'_k$ *is whitelist/blacklist compatible with the* $\mathscr{B}_{k-1}$ *and* $\mathscr{B}_{k+1}$.

### 4.4 Generation of Results

The genetic algorithm is responsible to get the best solution over the set of possible computable solutions and is a rather standard one which is explained in Algorithm 1. For the sake of space and clarity we present a highly simplified example of the block selection process with the genetic algorithm. We assume that all blocks are connector compatible and focus the definition of each block only in its id and the set of associated constraints.

**Algorithm 1.** Genetic Algorithm

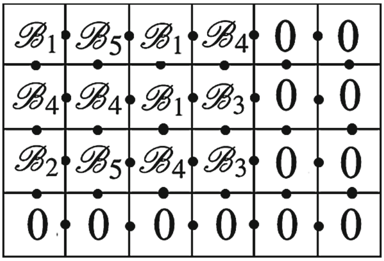
```
Let the number of blocks requested be $n$.
Let an individual be $\mathscr{I} = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_n\}$
Let the total of individuals be $total_{\mathscr{I}}$
Let Pop be $total_{\mathscr{I}}$ randomly generated individuals with or without block repetitions
Let the maximum number of generations be $max_G$
Let the total of produced generations be $total_G = 1$
Let $F_S = \emptyset$

  procedure GENERATE SOLUTION
      while ($\nexists \mathscr{I} : \mathscr{F}(\mathscr{I}) = 0 \wedge total_G < max_G$) do
          $Pop^I$ ← Apply $\mathscr{F}$ to order population by fittest elements
          $Pop^{II}$ ← Apply crossing to the top half of the population (elite)
          $Pop^{III}$ ← Apply mutation with 10% probability
          $Pop^{IV}$ ← Select survivors as top half of the population and add offspring
          $F_S$ ← $F_S \cup$ Fittest element from population
          $total_G$ ← $total_G + 1$
      end while
      return Fittest element from $F_S$
  end procedure
```

*Example 4.* Given a set of 5 types of blocks $\mathscr{T}_B = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_5\}$ that can be used in a game. Given the set of constraints $\mathscr{S}_c = \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_5\}$ where

- $\mathscr{C}_1 = \{max(3), min\_dist(2, \mathscr{B}_2)\}$,
- $\mathscr{C}_2 = \{max(2),\ min\_dist(2, \mathscr{B}_1),\ min\_dist(1, \mathscr{B}_2),\ min\_dist(2, \mathscr{B}_3)\}$,
- $\mathscr{C}_3 = \{max(20),\ min\_dist(2, \mathscr{B}_2),\ min\_dist(2, \mathscr{B}_3)\}$,
- $\mathscr{C}_4 = \{\}$ and $\mathscr{C}_5 = \{max(2)\}$

where $max$ is the maximum number of occurrences of that block type and $min\_dist$ is the minimum distance in number of blocks from that block type to other. The game's grid is described in Fig. 1.



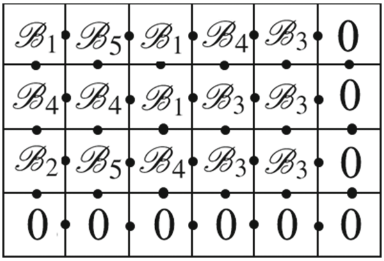
| | | | | | |
|---|---|---|---|---|---|
| $\mathscr{B}_1$ | $\mathscr{B}_5$ | $\mathscr{B}_1$ | $\mathscr{B}_4$ | 0 | 0 |
| $\mathscr{B}_4$ | $\mathscr{B}_4$ | $\mathscr{B}_1$ | $\mathscr{B}_3$ | 0 | 0 |
| $\mathscr{B}_2$ | $\mathscr{B}_5$ | $\mathscr{B}_4$ | $\mathscr{B}_3$ | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

**Fig. 1.** Initial grid

The framework requests 3 new blocks to fit a column of empty positions $(1,5)$, $(2,5)$, $(3,5)$ in a square grid where other positions were already set by previous iterations with the block selection algorithm. The genetic algorithm starts by creating 4 randomly selected individuals from the set of admissible blocks $\mathscr{T}_B$. It results in a population $\mathscr{P} = \{\{\mathscr{B}_5, \mathscr{B}_4, \mathscr{B}_4\}, \{\mathscr{B}_1, \mathscr{B}_1, \mathscr{B}_5\}, \{\mathscr{B}_2, \mathscr{B}_5, \mathscr{B}_3\}, \{\mathscr{B}_2, \mathscr{B}_3, \mathscr{B}_4\}\}$. The application of the fitness function returns these values:

- $\mathscr{F}(\{\mathscr{B}_5, \mathscr{B}_4, \mathscr{B}_4\}, \{(1,5), (2,5), (3,5)\}) = 1$ - because $\mathscr{B}_5$ is on its third occurrence when the maximum is 1.
- $\mathscr{F}(\{\mathscr{B}_1, \mathscr{B}_1, \mathscr{B}_5\}, \{(1,5), (2,5), (3,5)\}) = 3$ - because $\mathscr{B}_5$ already exceeds its maximum occupation by 1 and there are 2 occurrences of $\mathscr{B}_1$ in excess.
- $\mathscr{F}(\{\mathscr{B}_2, \mathscr{B}_5, \mathscr{B}_3\}, \{(1,5), (2,5), (3,5)\}) = 3$ - because $\mathscr{B}_5$ is on its third occurrence when the maximum is 1 and $\mathscr{B}_3$ has 2 neighbors of type $\mathscr{B}_3$ while the minimum distance to that type of block is 2 positions.
- $\mathscr{F}(\{\mathscr{B}_2, \mathscr{B}_3, \mathscr{B}_4\}, \{(1,5), (2,5), (3,5)\}) = 4$ - because $\mathscr{B}_2$ has 2 neighbors of type $\mathscr{B}_3$ and $\mathscr{B}_3$ has 2 neighbors of type $\mathscr{B}_3$.

Following the algorithm, it stops at the forth generation where the population is $\mathscr{P} = \{\{\mathscr{B}_4, \mathscr{B}_4, \mathscr{B}_4\},\ \{\mathscr{B}_4, \mathscr{B}_5, \mathscr{B}_5\},\ \{\mathscr{B}_4, \mathscr{B}_1, \mathscr{B}_2\},\ \{\mathscr{B}_1, \mathscr{B}_5, \mathscr{B}_1\}\}$ because it finds a fittest element: $\{\mathscr{B}_4, \mathscr{B}_4, \mathscr{B}_4\}$ which has 0 conflicts with neighbors and is used to obtain the final grid is described in Fig. 2.



|  |  |  |  |  |  |
|---|---|---|---|---|---|
| $\mathscr{B}_1$ | $\mathscr{B}_5$ | $\mathscr{B}_1$ | $\mathscr{B}_4$ | $\mathscr{B}_3$ | 0 |
| $\mathscr{B}_4$ | $\mathscr{B}_4$ | $\mathscr{B}_1$ | $\mathscr{B}_3$ | $\mathscr{B}_3$ | 0 |
| $\mathscr{B}_2$ | $\mathscr{B}_5$ | $\mathscr{B}_4$ | $\mathscr{B}_3$ | $\mathscr{B}_3$ | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

**Fig. 2.** Final grid

Running several examples showed us that the algorithm works well for our application domain. With a relative short number of generations (10) we get reasonably fit elements which make sense in terms of the game logics and with, a limited impact in performance which was our original purpose. More precisely we ran intensive tests with a grid of 100 × 100 blocks, with 7 types of blocs and 2 types of constraints. Each population is made of 6 individuals with 5 blocks each. The results for requesting a total of 1000 blocks being processed are the following:

- Average fittest element of a request in the beginning of the generation process (randomly generated): 1,212
- Percentage of optimal elements at the beginning of the generation process: 43,2%
- Average fittest element of a request in the end of the generation process (which is going to be delivered to the clients): 0,548
- Percentage of optimal elements at the end of the generation process: 67,6%

### 4.5 Case Study: H1Z1

As a proof of concept we created a world in real time for an open-world survival MMO game called H1Z1 [4]. In H1Z1 players are dropped in the world with nothing but the clothes they wear and a flashlight, and must explore, search and collect food, water and weapons to protect themselves. Having weapons and ammunition is crucial to survive from other players attacks and also from zombies that wander in the world. The world is mainly made of forests and has several points of interest spread over where players can find the means needed to survive. The way these points of interest are spread over the world is what makes the gameplay enjoyable and the players eager to explore new areas, immersing in the game experience.

## 5 Conclusions and Future Work

We have developed a framework for automatic content generation for massively multiplayer online games. The framework makes it easy for game developers to add or replace world generation strategies, so it can support different map structures and the introduction of new rules. This paper extends that framework with a genetic-based block selection algorithm, introducing some flexibility in the application of those rules. Running several syntectic examples showed us that the algorithm works well for our application domain and has a very limited impact in performance, which was our original purpose. There are still some details for future work. We found that some types of blocks appear much more often than others. The random selection of elements does not facilitate this and, although the problem is minimized by the crossing and mutation operators, giving the same probability of occurrence to blocks that should have a very short number of occurrences than other blocks that can occur without restrictions makes it harder to get a very good solution. We foresee that giving weights to types of blocks could minimize this problem.

**Acknowledgments.** This work was partially supported by LIACC through Programa de Financiamento Plurianual of FCT (Portuguese Foundation for Science and Technology).

## References

1. Alves, T., Coelho, J.: A framework for massively multiplayer online game content generation. In: 30th IEEE International Conference on Advanced Information Networking and Applications, AINA 2016. IEEE Computer Society (2016)
2. Chopard, B., Droz, M.: Cellular Automata Modeling of Physical Systems. Cambridge University Press, Cambridge (1998)
3. Davidsson, P.: Multi agent based simulation: beyond social simulation. In: Proceedings of the Second International Workshop on Multi-Agent-Based Simulation-Revised and Additional Papers, MABS 2000, London, UK, pp. 97–107. Springer-Verlag (2001)
4. Daybreak. What is h1z1. World Wide Web (2015). https://www.h1z1.com/what-is-h1z1
5. Duncan, S.C.: Minecraft, beyond construction and survival. Well Played **1**(1), 1–22 (2011)
6. EIGENLENK. Pioneers. World Wide Web (2013). http://www.pioneersgame.com/
7. Red Blob Games. Hexagonal grids. World Wide Web (2013). http://www.redblobgames.com/grids/hexagons/
8. Hendrikx, M., Meijer, S., Van Der Velden, J., Iosup, A.: Procedural content generation for games: a survey. ACM Trans. Multimedia Comput. Commun. Appl. **9**(1), 1:1–1:22 (2013)
9. Klei Entertainment Inc. Don't starve. World Wide Web (2013). http://www.dontstarvegame.com/
10. Jilesen, J., Kuo, J., Lien, F.-S.: Three-dimensional midpoint displacement algorithm for the generation of fractal porous media. Comput. Geosci. **46**, 164–173 (2012)
11. Meier, S.: Civilization v. World Wide Web (2010). http://playrust.com/
12. Mozilla. Websockets. World Wide Web (2015). https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API/

13. Perlin, K.: An image synthesizer. SIGGRAPH Comput. Graph. **19**(3), 287–296 (1985)
14. Shaker, N., Togelius, J., Nelson, M.J.: Procedural Content Generation in Games: A Textbook and an Overview of Current Research. Springer (2015)
15. Sliva, M.: Don't starve review. IGN Entertainment, Inc., May 2013
16. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. Computer **27**(6), 17–26 (1994)
17. Facepunch Studios. Rust. World Wide Web (2013). http://playrust.com/

# An Autonomous Architecture for Managing Vertical Elasticity in the IaaS Cloud Using Memory Over-Subscription

Bouaita Riad[1](✉), Zitouni Abdelhafid[2], and Maamri Ramdane[2]

[1] Lire Labs, Higher School of Technological Education, 21000 Azzaba, Skikda, Algeria
r.bouaita@enset-skikda.dz

[2] Lire Labs, Abdelhamid Mehri Constantine 2 University, 25000 Ali Mendjli, Constanine, Algeria
{abdelhafid.zitouni,ramdane.maamri}@univ-constantine2.dz

**Abstract.** Elasticity is one of the essential properties in Cloud Computing that meets changeable needs of customers, and improves resource utilization for providers. In this context, oversubscription is a very powerful technique to increase resource utilization level as much as possible, which leads to a maximization of the profit of cloud providers. In this paper, we propose an autonomous architecture based on the MAPE-K control loop and using memory oversubscription to improve operating performance in a cloud infrastructure. The overload caused by oversubscription is mitigated by the live migration technique of VMs as well as the use of the network memory of the various physical machines of the cluster through the network. This latter technique is usually used as a replacement technique for the swapping disc as it has more performance.

**Keywords:** Cloud Computing · Oversubscription · Live migration · Network memory · Vertical elasticity

## 1 Introduction

With the development of information technology, modern applications require too much physical resources. These applications are more memory intensive than other resources such as CPU, bandwidth or disk storage. For this purpose, the memory presents a very critical resource in an environment with a very wide access. In this paper, we present an autonomous architecture based on the MAPE-K (Monitor-Analyze-Plan-Execute-Knowledge) control loop introduced by IBM [1] and used in the field of automatics as a self-adaptation technique.

This architecture manages a vertical elasticity in a cluster through the technique of memory oversubscription. This technique is used in several fields. Taking the example of air transport, airlines sell more tickets than the actual number of seats in order to maximize the filling rate in case of discontinuance of certain customers. If the number of customers appears more than expected, they will be moved to another flight or transmitted to another company.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_5

In cloud computing, it is called oversubscription, when providers deploy more resources in the hope that users use less resource than initially requested. This situation certainly leads to an overload situation when the requests exceed the actual physical available capacity [2]. To this end, our architecture mitigates this overload via live migration and network memory.

The global memory in a cluster is actually unusable. For this, our approach combines the live migration of VMs with the use of network memory instead of the local disk swapping given its performance more: (*nano second* vs. *micro second*) [3], asserting that this technique is only useful in very fast networks (Giga Ethernet). In our paper, the oversubscription ratio is dynamically managed by reporting the amount of oversubscribed memory in real time. On the other hand, Live migration makes it possible to share the load on resources other than memory (CPU, Bandwidth,…) between the different physical machines of the Cluster. Live migration therefore contributes proportionately to load balancing.

The remainder of this paper is organized as follows: In Sect. 2, a brief description of oversubscription is introduced in the context of Cloud Computing while presenting the potential risks produced by this technique. Section 3 describes some related work in the field of oversubscription. In Sect. 4 we propose an autonomous architecture inspired by the autonomous control loop for the management of the vertical elasticity in a cluster using memory oversubscription. Finally, in Sect. 5 we conclude our paper by discussing our future work.

## 2 An Overview on Oversubscription in Cloud Computing

As mentioned above, oversubscription is a technique used in different fields. In Cloud Computing, oversubscription (Also called overcommitment or overbooking) is a resource management technique where the sum of user requests for such a resource

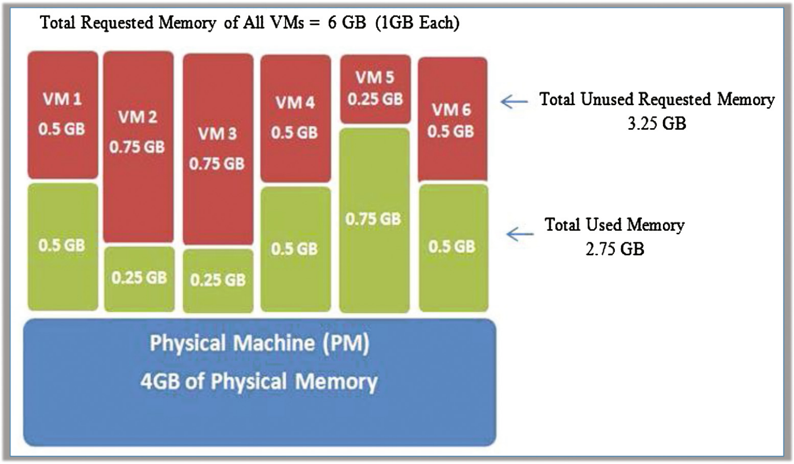


**Fig. 1.** Memory allocation in an oversubscription environment [2]

exceeds the current available capacity in the physical machine (PM). Figure 1 presents an oversubscription environment where the actual physical capacity is 04 GB and the total request size is 06 GB [2]. This technique devotes more resources (CPU, memory, and bandwidth) than is actually available on physical machines hosting a set of applications [4]. Oversubscription can be both on the provider side and on the customer side. In this paper, we focus on provider's oversubscription, where they are efficiently sharing limited resources aiming to maximize their profits and reduce costs, with a Service Level Agreement (SLA) guarantee [5].

## 2.1 Opportunity for Oversubscription

Recent studies show that Cloud customers tend to over-estimate resource requirements for their applications, and indeed use only a portion of these resources. An analysis of the Google data centers for 29 days [6] shows that in a one (01) hour window, overall usage is less than 50% for memory and less than 60% for CPU (Fig. 2).

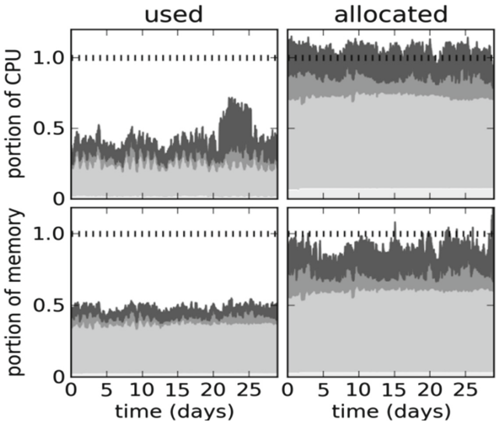


**Fig. 2.** Use vs. allocation (memory/CPU)

Another study performed by Ghosh and Naik [7] on a set of 2193 VMs for a period of one (01) months shows that 84% of the running VMs reach their maximum CPU utilization only for a peak of 20% of the time and that less than 0.7% of VMs only reach their maximum CPU utilization of almost 100%.

These studies show indeed, that the actual use of physical resources is very small compared to the initially allocated amounts of these resources because of an over-estimation of users' needs. Therefore, this over-estimation leads to a waste of resources, from the cloud provider perspective, and higher costs, from the user perspective [8], which gives an opportunity to oversubscribe resources.

While most hypervisors well known such as Xen, KVM, and VMware offer oversubscription ability through techniques such as memory ballooning and disk swapping, these systems do not support dynamic and adaptive oversubscription ratios.

### 2.2 Risks of Cloud Oversubscription

As stated in the previous section, the oversubscription technique is appropriate only with an implicit assumption: Cloud providers assume that the sum of resources used at any time does not exceed actual physical capacity. They also assume that all customers will not show up to use their resources at the same time [2]. As it is difficult to predict the workload, such an oversubscription approach leads to a situation of overload which produces such problems [7]:

- Performance problem: the degradation of the response time of client requests.
- Availability problem: Blocking the VMs and possibly the entire system.

Since these problems occur, a violation of the Service Level Agreement (SLA) promised by providers of the IaaS Cloud affects their credibility [7].

Resource overload can be devastating because it has the potential to prevent any application progress [8]. The oversubscription strategy must be therefore conducted with an overload mitigation technique. In this context, several techniques are detailed by Baset et al. [3].

## 3 Related Work

Our work focuses on resource oversubscription and the associated overload mitigation techniques. In this section, we cite some previous works in this field.

Baset et al. [3] present two solutions to solve the problem of resource overload caused by oversubscription in an IaaS: live migration and quiescing VMs. The selection of the VMs to shut down or to migrate is done according to a well defined policy while respecting the constraints of placement. In that work a comparison between the different strategies to mitigate the overload is carried out: migration, quiescing, or combination between the two. The authors consider here a static oversubscription ratio fixed at two (2) which does not reflect such a state of the system, in addition, the network memory is not exploited.

In their approach *Overdriver*, Williams et al. [9] mitigate memory overload due to oversubscription by two different strategies, where the choice of the appropriate strategy is based on the duration of overload. Unpredictable or transient overloads are handled reactively by a new technique called *Cooperative Swap*: swapping pages across the network. Durable or sustained overloads are handled by live migration. In order to decide which strategy to use, *Overdriver* uses initially fixed thresholds, and then these thresholds are reduced through a learning process available in the *workload profiler* component. This approach is close to ours, while it does not take into account the amount of free memory either locally or across the network to generate a dynamic oversubscription ratio.

Zhang et al. [10] propose an algorithm called *Scattered* which aim to minimize the number of migrations of VMs in an oversubscribed environment. This algorithm mainly avoids an overload cascade due to migration by the appropriate choice of the target physical machine to migrate to. In addition, it minimizes the cost of migration by analyzing the network topology. *Scattered* makes it possible, on the one hand, to maintain a minimum network traffic rate by minimizing the number of migrations, and on the other hand it avoids future risks of overload by migrating VMs having a high correlation in their workloads. This approach is not interested in the ratio of oversubscription and the network memory is still unusable.

Ying [15] proposed an interference aware oversubscription strategy called Sponge. This strategy aims to handle the interference between co-hosting VMs. The oversubscription ratio makes a tradeoff between resource allocation and the performance requirements of VMs hosted in the same PM. In this architecture, the oversubscription module is used to allocate to the VMs a minimum amount of resources based on their performance requirements. The oversubscribed resource entitlements are calculated based on the placement of VMs.

Moltó et al. [11] propose a framework called *CloudVAMP* integrated with cloud management platforms (CMP) in an on-promised cloud infrastructure. This Framework automatically manages vertical elasticity by using memory oversubscription. To handle (predict) memory overload due to oversubscription, the ballooning technique as well as the live migration of VMs are applied. In this approach the oversubscription ratio is dynamically managed and depends on the amount of free memory of the VMs of the local host only, and thus, global memory is not addressed. In our paper, the adopted ratio is an extension of this one, where both local and global memory are considered.

Tomas and Tordsson [8] proposed a framework for resource oversubscription aiming to not exceed the physical capacity of the underlying hosts. They use two essential mechanisms: admission control and scheduling. The first copes with horizontal elasticity and decides whether a user query can be deployed or not using a profiling tool that accurately measures and classifies applications' behaviors. This ranking provides more flexibility by treating the various applications differently. The latter copes with vertical elasticity. It allows - after the deployment decision - to allocate the VM in the appropriate location, taking into account physical capacities in order to avoid performance degradation.

Our work is related to these previous ones. However, it differs to them in significant points. Concerning the oversubscription ratio, unlike the majority of studies that uses a static oversubscription ratio, in our approach, a dynamic ratio is adopted that is adaptable according to the state of the cluster. This ratio is reported dynamically and in real time taking into account the local and global memory. In addition, to mitigate the overload due to oversubscription, conventional approaches using only live migration, have been improved by exploiting the global memory.

## 4 Proposed Approach

The objective of our work is to develop an autonomous architecture managing the vertical elasticity in an IaaS Cloud with the technique of memory oversubscription. The oversubscription technique allows maximizing the resource utilization; however, if the oversubscription ratio is not managed wisely, this leads to a degradation of performance and increasing the risk of SLA violation. In our paper, this ratio depends to the amount of free memory available in the local host as well as the sum of the free memory amounts throughout the network cluster. This technique allows maintaining a dynamic oversubscription ratio instead of using a static ratio depending solely on physical memory. Thanks to the dynamic management of the oversubscription ratio and the real time reporting of the state of the memory, the threshold triggering an elasticity operation is so dynamic. The overload caused by this oversubscription is mitigated by two strategies: live migration and network memory.

## 4.1 Overall Architecture

Our system is based on the MAPE-K self-configuration control loop. This loop includes 04 functions: *Monitor*, *Analyze*, *Plan* and *Execute*. The overall architecture mentioned in Fig. 3 shows the following components:

- *Workload Profiler*: This component covers the *Monitor* function. It is the component responsible for monitoring the workload in real time. It perceives the state of the system and in particular the creation/destruction of VMs, turning on/off VMs, changing the state in a VM. The information collected is sent to the Oversubscription Controller.
- *Oversubscription Controller*: This component covers the *Analyze* function and allows exploiting the information about the amount of memory from the local host and the available in the network and continuously generates (with some granularity) a dynamic oversubscription ratio.
- *Elasticity Manager:* This component is the core of the system. It covers the *Analyze*, *Plan* and *Execute* functions. It retrieves information from other system components to decide which elasticity operation to launch: Live Migration/Network Memory (assuming that other strategies such as Memory Ballooning are not able to mitigate overload). Overload detection and mitigation are discussed in the following sections.

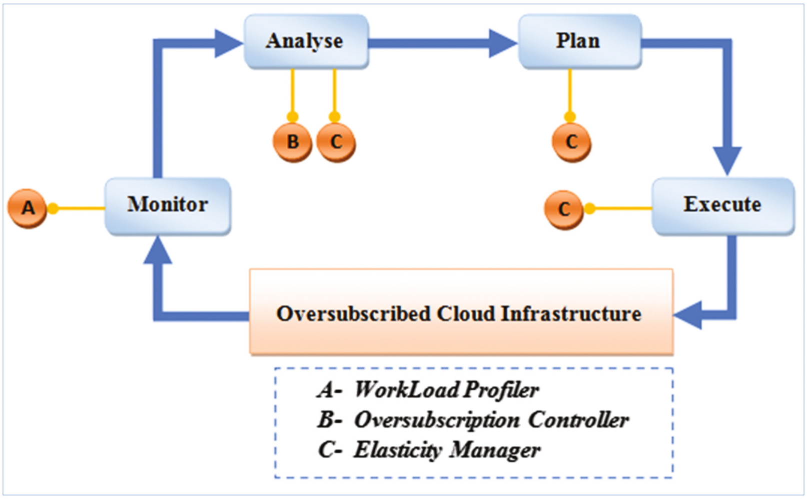


**Fig. 3.** The autonomic control of the cloud infrastructure

## 4.2 Memory Oversubscription

As said previously the oversubscription controller determines the amount of memory oversubscribed at any time. Unlike the works mentioned above and in order to be realistic, this quantity must be calculated taking into account the amount of free memory on the local host as well as that available on the network, instead of using a fixed coefficient. Considering N the number of VMs hosted in a host i, and M the total number of PMs in the cluster, the amount of free memory reported for each VM hosted in a host $PM_i$ is calculated as follows:

$$RM_i = Free_i + R * [\underbrace{(\Sigma_{j=1}^{N} AM_j - \Sigma_{j=1}^{N} UM_j)}_{\substack{\text{Memory Stolen} \\ \text{from the local host}}} + \underbrace{(\Sigma_{j=1}^{M} Free_j / j \neq i)}_{\substack{\text{Memory retrieved through} \\ \text{the network}}}] \tag{1}$$

where:

**RM$_i$**: represents the amount of free memory reported for VMs hosted in a host $PM_i$.
**AM$_j$**: represents the amount of memory allocated to $VM_j$.
**UM$_j$**: The amount of memory actually used in $VM_j$.
**R:** is an oversubscription coefficient (not the oversubscription ratio):
R = 0: means that there is no oversubscription.
R = 1: means that all free memory in the cluster is used for oversubscription
R > 1: means that oversubscription is greater than the actual amount of free memory. If such a solution is not managed carefully, this may increase the risk of overload.

For any host $PM_i$, the amount of free memory is calculated by Eq. (2):

$$Free_i = C_i - \sum_{j=1}^{N} UM_j \tag{2}$$

where $C_i$ represents the real memory capacity of the host $PM_i$.

On each arrival of a request (arrival of a new VM), the quantity of memory requested is compared with that expressed in Eq. (1) to decide whether the request can be served or not.

Equation (1) can be summarized as follows: *the amount of free memory reported for such a VM hosted in any host is equal to the amount of free memory in that host plus the amount of memory oversubscribed. The amount of oversubscribed memory is equal to the sum of the stolen memory quantities of the local host plus the amount of memory retrieved across the network multiplied by a coefficient R.*

Theoretically this value is measured in real time. However, in order to avoid unnecessary oscillations, the system marks only the significant variations: a threshold is used to trigger such update.

### 4.3 Overload Detection

First, an overload can be detected either at the VM level or at the physical host level. Since the first approach requires modifications on applications running in such a VM [3], our approach maintains detection at the physical host level. Memory overload is usually detected by two indicators: paging rate and paging scan rate [3, 9]. For example, for the first indicator an overload is detected if the pagination rate reaches a certain threshold. In our paper, overload is detected intuitively - for a host - if the amount of free memory is negative ($Free_i < 0$). In order to avoid falling into a sudden overload, an overload threshold OT (*Overload Threshold*) is used, and therefore an overload is declared if $Free_i < OT$. OT is calculated as the total amount of memory used in the host plus a percentage of that amount [11], as mentioned in Eq. (3):

$$OT = \sum_{i=1}^{N} UM_i * (P + 1) \tag{3}$$

where $P$ is a percentage.

The overload threshold is used to predict relatively overload. If this threshold is reached, an elasticity operation (Scale UP) is initiated by the *Elasticity Manager* component. If the amount of memory decreases by a certain portion, another elasticity operation (Scale Down) is started. The latter consists of decreasing the allocated memory to the underlying VM.

Since our paper is interested in overload mitigating due to oversubscription, we have focused on the first operation. This operation is performed by the *Elasticity Manager* which can decide the appropriate strategy to use.

### 4.4 Decision on the Strategy to be Used to Mitigate the Overload

If such an overload is indicated, the mitigation procedure of this overload is triggered. In order to decide between live migration and network memory, we proposed the use of a specific indicator: the CPU load for the overloaded PM. To do this:

1. *If a memory overload is detected and the CPU load is low, it is not necessary to migrate the virtual machines, and the network memory is used as a remote paging.*
2. *If memory overload is detected and CPU load is high, live migration of one or more VMs is applied.*

Live migration therefore contributes proportionally to the load balancing between physical hosts in the cluster. This technique is widely recognized in literature as the CPU load balancing caused by jobs (process) and used in the cluster as a stable load balancing strategy [12, 13].

Formally, the decision of the strategy to use is based on a processor utilization threshold (*or CPU length*), and noted: *CpuT* (CPU Threshold).

If live migration strategy is selected, the number of migrations is minimized by two means:

*Migrate first the VM with the largest size and therefore a large memory space is released, which helps to avoid future migrations.*

*Migrate first to the PM with the lowest CPU load. The target PM therefore has a low CPU load, and in the case of a future overload of it, the network memory is preferred while avoiding a migration cascade.*

- Since the CPU load is oscillated, i.e. it may be low at one instant and high at the next instant, it is convenient to use an average load over a certain time interval. It is thus, necessary to keep statistics concerning this load during the last interval of time.
- If such a live migration is performed for a VM from a source host to a destination one, the new CPU load - that is certainly decreased - in the source host is considered, which gives the opportunity to choose the network memory in case of future overload.

### 4.5 Overload Mitigation

As shown in Fig. 4, the virtual disks in the cluster are located on a SAN (*Shared Area Network*) device, which requires – in case of migration – a migration of the memory footprint of the VM only. We notice:

– *$CpuL_i$*: the current CPU load in the host $PM_i$.
– *$RP_i$* (Remote Paging): remote memory paging in the host $PM_i$.

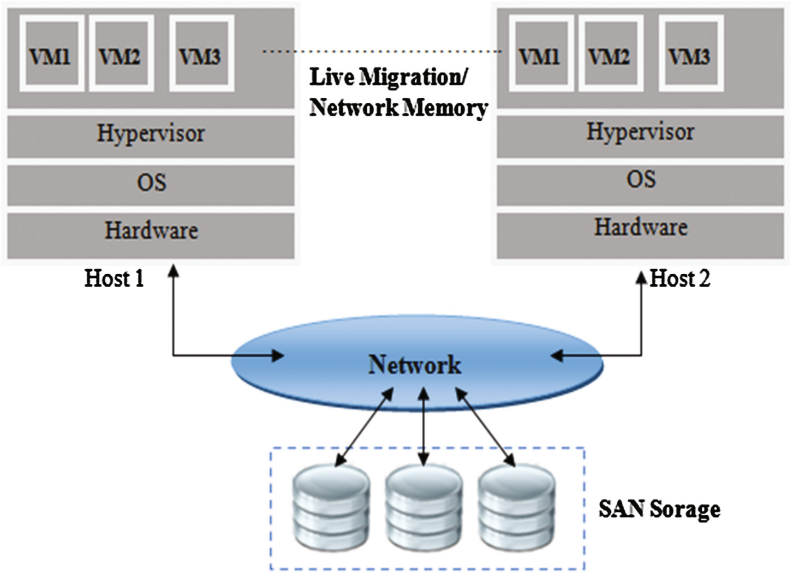


**Fig. 4.** General architecture of the cluster

If such an overload is detected on a $PM_i$ ($Free_i < OT$), the Elasticity Manager initiates the necessary processing. This processing is summarized in the pseudo code of Algorithm 1. This algorithm handles the overload due to oversubscription by live migration strategy where exploiting the global memory.

**Algorithm 1:** OverLoad_Mitigation (*Overloaded_hoste $PM_i$*)

```
Input: VMList = List of VMs hosted in $PM_i$ sorted by allocated memory
                (decreasing order).
       PMList = List of PMs excluding $PM_i$ sorted by CPU load (increasing
                order)
01: While ($Free_i$ < OT) do
02:     If $CpuL_i$ >= CpuT then      // Live Migration is applied
03:         CurrentVM = VMList.First
04:         k = 1 ; Flag = false ;
05:         While ( k <= PMList.Size & Flag == false) do
06:             if $Free_k$ >= AllocatedMemory(CurrentVM) Then
07:                 Live Migrate CurrentVM to $PM_k$
08:                 $Free_i$ = $Free_i$ + AllocatedMemory(CurrentVM)
09:                 $Free_k$ = $Free_k$ - AllocatedMemory(CurrentVM)
10:                 VMList.Remove(currentVM)
11:                 Flag = true
12:             endif
13:             if (Flag == false) & (CurrentVM $\neq$ PMList.Last) then
14:                 currentVM = CurrentVM.Next
15:                 Go To step 4
16:             endif
17:             k = k+1
18:         endWhile
19:         if Flag == false then  go to step 22
20:         endif
21:     else              // Network Memory is applied
22:         Repeat
23:             Find $PM_j$ among PMList whith the largest $Free_j$
24:             LackSpace = OT - $Free_i$
25:             Allocate $RP_j$ = Min (Lackspace, $Free_j$) to $PM_i$
                                as a remote paging
26:             $Free_i$ = $Free_i$ + RP
27:             $Free_j$ = $Free_j$ - RP
28:         Until  ($Free_i$ >= OT) or (All Hosts in PMList are browsed)
29:     endif
30: endWile
```

*Algorithm Description:*

- *Lines 2–20 relate to overload mitigation by live migration (CPU load is high):* We first take the VM with the largest size and then we brows all the remote hosts (line 5) until we find enough space (line 6), so the migration is done. It is noted here, that if a live migration is done, this doesn't mean that the overload is mitigated, and the main loop (line 01) allows continuing mitigation. The update of the free spaces - local and remote - (line 8, 9) and the deletion of the migrated VM (line 10) are then carried out. If we have not been able to find a free space, we pass to the VM with the size just lower (line 14, 15). If no migration is possible (line 19) the second strategy (Network Memory) is switched.
- *Lines 21–29 relate to overload mitigation by network memory (CPU load is low):* The host with the largest free space is chosen first to avoid fragmentation of the overloaded host (line 23). Intuitively, the lack space is calculated as "- $\text{Free}_i$", and as long as a threshold is used, it is calculated as "OT - $\text{Free}_i$" (line 24). In line 25 - inspired from [14] - we allocate to the overloaded host a space exactly equal to the lack space if there is sufficient space in the remote host. If the space is not sufficient, all the space is allocated and other hosts are browsed until the overload is no longer or all hosts are browsed (line 28). After the allocation is done, the free spaces -local and remote- (line 26, 27) are updated.

When network memory is applied, the target PM is selected with the largest space instead of the closest space, used in the naive migration algorithm [10]. The choice of the PM with the closest space minimizes the fragmentation of the remote host but, it can generate another overload on it, and thus, a cascade of overload will occur.

On the other hand, if the algorithm fails to mitigate overload, i.e. the overload remains for a certain duration - defined within the SLA - other strategies are explored by the cloud provider (ignition of new PMs, preemption of some VMs,).

## 5 Conclusion and Future Work

In this paper, we proposed an autonomous architecture for the management of vertical elasticity in an IaaS Cloud. This architecture attempts to maximize the memory utilization through the oversubscription technique. If the oversubscription ratio is not managed carefully, this technique quickly causes overloads. For this, our architecture adopts a dynamic ratio depending on the available memory on the local host and the one available across the network. In addition to the live migration, the network memory was exploited to mitigate overload caused by oversubscription.

Our architecture is based on a centralized structure where the elasticity manager is hosted on a dedicated node (master node). Therefore, our future work is to distribute this component on each node of the cluster, where each one makes its own decisions to mitigate the overload.

As another perspective, we tend to integrate our architecture with one of the open source cloud management platforms, such as OpenStack, OpenNebula or Eucalyptus, which do not support currently vertical elasticity.

## References

1. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. IEEE Comput. **36**(1), 41–50 (2003)
2. Householder, R., Arnold, S., Green, R.: On cloud-based oversubscription. Int. J. Eng. Trends Technol. **8**(8), 425–431 (2014)
3. Baset, S.A., Wang, L., Tang, C.: Towards an understanding of oversubscription in cloud. In: Presented as Part of the 2nd USENIX Workshop on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services (2012)
4. Caglar, F., Gokhale, A.: iOverbook: intelligent resource-overbooking to support soft real-time applications in the cloud. In: 2014 IEEE 7th International Conference on Cloud Computing (CLOUD), pp. 538–545. IEEE (2014)
5. Kim, S., Kim, H., Lee, J., Jeong, J.: Group-based memory oversubscription for virtualized clouds. J. Parallel Distrib. Comput. **74**(4), 2241–2256 (2014)
6. Reiss, C., Tumanov, A., Ganger, G.R., Katz, R.H., Kozuch, M.A.: Towards Understanding Heterogeneous Clouds at Scale: Google Trace Analysis. Intel Science and Technology Center for Cloud Computing, Technical report, 84 (2012)
7. Ghosh, R., Naik, V.K.: Biting off safely more than you can chew: predictive analytics for resource over-commit in iaas cloud. In: 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 25–32. IEEE (2012)
8. Tomás, L., Tordsson, J.: Improving cloud infrastructure utilization through overbooking. In: Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference, p. 5. ACM (2013)
9. Williams, D., Jamjoom, H., Liu, Y.-H., Weatherspoon, H.: Overdriver: handling memory overload in an oversubscribed cloud. In: Proceedings of the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE) (2011)
10. Zhang, X., Shae, Z.-Y., Zheng, S., Jamjoom, H.: Virtual machine migration in an over-committed cloud. In: Network Operations and Management Symposium (NOMS), pp. 196–203. IEEE (2012)
11. Moltó, G., Caballer, M., de Alfonso, C.: Automatic memory-based vertical elasticity and oversubscription on cloud platforms. Future Gener. Comput. Syst. **56**, 1–10 (2016)
12. Vijaya Krishna, D., Santhosh, G., Sammulal, P.: CPU and memory based cluster load balancing for jobs with bursts of loads. Int. J. Adv. Res. Comput. Commun. Eng. **2**(10), 3959–3963 (2013)
13. Sharifian, H., Sharifi, M.: Network RAM based process migration for HPC clusters. J. Inf. Syst. Telecommun. **1**, 47–53 (2013)
14. Xiao, L., Chen, S., Zhang, X.: Adaptive memory allocations in clusters to handle unexpectedly large data-intensive jobs. IEEE Trans. Parallel Distrib. Syst. **15**(7), 577–592 (2004)
15. Ying, L.: Sponge: an oversubscription strategy supporting performance interference management in cloud. China Commun. **12**(11), 1–14 (2015)

# A Security Framework for Cloud Data Storage(CDS) Based on Agent

Oussama Arki(✉) and Abdelhafid Zitouni

Lire Labs, Abdelhamid Mehri Constantine 2 University,
Ali Mendjli, 25000 Constantine, Algeria
{oussama.arki,abdelhafid.zitouni}@univ-constantine2.dz

**Abstract.** The Cloud has become a new Information Technology(IT) model for delivering resources such as computing and storage to customers on demand, it provides both high flexibility and resources use. However we are gaining these advantages at the cost of high security threats, which presents the major brake for the migration towards Cloud Computing.

Cloud Data Storage(CDS) is one of the Cloud services, it allows users to store their data in the Cloud, this service is very useful for companies and individuals, but data security remains the problem which makes customers worried about their data that reside in the Cloud. In this paper, we propose a framework of security to ensure the CDS, which is based on agents, it contains three layers: Cloud Provider layer, Customer layer and Trusted Third Party(TTP) layer.

**Keywords:** Cloud Computing · Cloud Data Storage · Trust · Data security · Multi-agent system · Integrity · Confidentiality

## 1 Introduction

In the last years, computing and storage technologies are rapidly developed, one of the major reasons is decreasing costs and increasing power of the computer resources beside to the success of Internet, which led to the situation where a big volume of data can be collected, stored and treated.

Cloud Computing is a new IT model, it provides storage and computation resources as a service to the Cloud Customers (CC), often through a network (typically the Internet).

Cloud Data Storage is one of the Cloud services. It allows users to store their data in the Cloud, by reserving a virtual space in the Cloud, it also provides the powerful way of managing data. Cloud Data Storage allows to store and manage the data remotely. Users do not have to buy the expensive hardware and to have policies to regulate and manage the data [1]. The major brake for the adoption of this service is the data security. The virtualisation and the colocation of the physical resources are the principal characteristics that distinguish the Cloud Computing, which make the data security in the Cloud more difficult

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_6

and complex than the traditional systems. The security of data in the Cloud is the biggest challenge of Cloud Providers (CP).

To ensure the CDS security and safety, we propose in this paper a framework of security, which contains three layers: the Customer layer, the Cloud Provider layer and the TTP layer. It is based on trust model and mobile agents.

This paper is organized as follows: Section two introduces Cloud Computing. Section three discusses the Information Security and Cloud Storage. Section four is about the related works. Section five presents our framework and the last section is a conclusion.

## 2 Cloud Computing Overview

According to the famous definition of NIST (National Institute of Standards and Technology): Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [2].

This Cloud model is composed of five essential characteristics, three service Models, and four deployment models [3], as in Fig. 1.

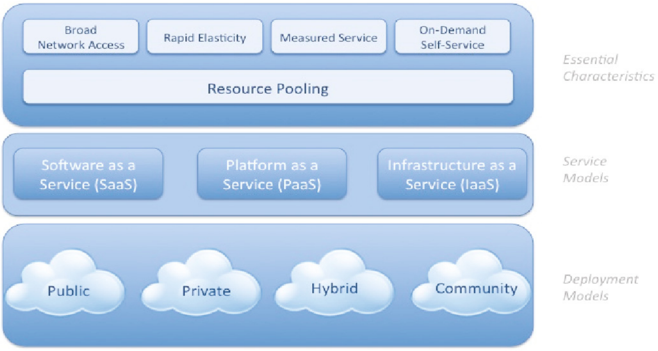


**Fig. 1.** NIST visual model of Cloud Computing definition [4]

## 3 Information Security and Cloud Storage

In this section, we speak about Information Security in the Cloud and the Cloud Storage concerns.

### 3.1 Information Security

The term Information Security means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide integrity, confidentiality and availability. Here, integrity means guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity. Confidentiality means preserving authorized restrictions on disclosure and access, including means for defending proprietary and personal privacy information. And availability means ensuring timely and reliable access to and use of information [5].

To understand this term in Cloud Computing environment, we must take into account the three main properties of Information Security [6]:

**Confidentiality:** It means keeping user data secret in the Cloud systems. It ensures that user data which reside in Cloud cannot be accessed by unauthorized person. There are two basic approaches to achieve such confidentiality, physical isolation and cryptography. Confidentiality can be achieved through proper encryption technique: symmetric and asymmetric algorithms.

**Data Integrity:** It means Keeping data integrity is a fundamental task. It means in Cloud system is to preserve information integrity. Data could be encrypted to provide confidentiality but it will not guarantee that data reside on Cloud has not been altered. There are two approaches which provide integrity Digital Signature and Message Authentication Code.

**Availability:** Data should be available when it is requested via authorized user. It ensure that user can be able to use the service any time from any place. Two strategies called Hardening and Redundancy are mainly used to enhance the availability.

### 3.2 Cloud Storage

Cloud storage is an important service of Cloud Computing, which offers services for data owners to host their data in the Cloud. This new paradigm of data hosting and data access services introduces two major security concerns [3]:

- **Protection of data integrity.** Data owners may not fully trust the Cloud server and worry that data stored in the Cloud could be corrupted or even removed.
- **Data access control.** Data owners may worry that some dishonest servers give data access to unauthorized users, such that they can no longer rely on the servers to conduct data access control.

## 4 Related Works

In the last years, many researchers used agents in their works, to ensure the security in Cloud Computing.

In [7] Shantanu et al., proposed a Cloud security two-tier framework, which is based on a trust model and the use of agents, they used two agents for the trust, one in Customer side and the other in Cloud Provider side, these agents analyse the user behavior, to ensure that the user remains always as a trusted entity before the interaction with the Cloud Provider. This model suffers from some weakness, because there is no mechanism to prevent malicious activity without Cloud service provider information about user activities, and the use of proxy server, which presents also a weak point, if it crashes, the customer can not communicate with the Cloud Provider.

In [8] Priyank et al., proposed an MAS framework to ensure the security of the customer resources in the Cloud, they used mobile agents to supervise the hyper-visor and to collect information from the virtual machines of the customer, these agents monitor the virtual machines, to keep their privacy and integrity. If there is any problem, they notify it. The weak point in this model is that the user side is ignored, like the user identification and the identity management.

In [9] Amir et al., proposed a framework of MAS to facilate security of Cloud Data Storage, the proposed framework has two layers: Customer layer and Cloud Data Storage layer. In the Customer layer an interface agent is used for the interaction with the customer, the other agents are distributed in the Cloud Data Storage layer, each one of them has a specific task and they communicate to achieve the global goal. The major problem in this work is that the security of the Customer layer is not dealt with, specialy the mechanism for the identification of the customer.

In [10] Alwesabi et al., proposed a MAS framework of Cloud security in the form of two-tier framework: the Virtual Server layer and the Cloud Provider layer. The communication between the two layers is ensured by the use of mobile agents. In the Virtual Server side they used an analyzer agent for the authentication of the customers, and in the Cloud Provider side a security agent is used to ensure the security, the problem in this framework is that the security mechanism used by the security agent is not clear.

In [11] Benabied et al., proposed a MAS framework to make safe Cloud environment, which contains two layers: Cloud Provider layer and the Customer layer. The communication between the two layers is ensured by the use of mobile agents, to ensure the trust at Cloud environment, this framework based on the use of Trust Model, they used two types of agents, one in Customer side and the other in Cloud Provider side, this model ensures the customer trust and guarantees that only the trusted customers can interact with the Cloud Provider. The weakness of this framework is the monitoring of the Cloud Provider side, which is ignored, like the supervising of the virtual machines.

In [12] Md. Rafiqul et al., proposed an agent based framework for providing security to Data Storage in Cloud, which contains three levels: Customer level, Cloud Provider level and Data Storage level. In their work, they considered that the degree of data security is depended on data sensitivity, because of that, they classified the sensitivity of data into five categories, for each category they used a different method of authentication, a different algorithm of encoding and a different function of hashing. By using this technique they increased the Cloud performance, but they risk the data of the customer.

In [13] Venkateshwaran et al., proposed a security framework for Agent-Based Cloud Computing, they based on the use of Trust Model. In their work, they used a security agent, which ensures the security, it is responsible for the authentication of the customers and the analysis of the customer's trust. By using this Trust Model, they guarantee that only trusted users can interact whith the Cloud Provider.

## 5 The Proposed Framework

To ensure the security of data in the Cloud, we proposed a framework of security based on agents, which contains three layers: the Customer layer, the Cloud Provider layer and the TTP layer. In this work, we used a Trust Model to manage the confidence of the customers, an Encryption Method to ensure the privacy and the confidentiality of data and an Integrity Technique to allow the user to check the integrity of his data in the Cloud.

The TTP Layer is an intermediary between the Customer layer and the Cloud Provider layer. It is responsible for querying the Cloud Provider on behalf of user. It has expertise and capabilities that users may not have [1]. It is responsible for the Encoding/Decoding operation of data, before storing them in the Cloud. It is also used to check the integrity of data, without recovering them by the customer. Figure 2 presents our framework.

### 5.1 Framework Description

To store his data in the Cloud, the customer must initially authenticate himself, for that it provides his information of identity to the Interface Agent. This last analyzes the information seized by the customer, then asks the Proxy Agent to check the identity of the customer. The Proxy Agent checks the identity of the customer through his database. If the customer is well authenticated, the Proxy Agent requests the Interface Agent to post the user interface to the customer, and asks the Customer Trust Agent(CTA) to calculate the trust degree of this customer. Only if the trust degree of the customer is enough to establish a connection with the Cloud, the Proxy Agent creates a Mobile Agent(MA1), to carry the request of the customer towards the TTP layer.

When the first phase is well established, if the customer wants to transfer his data towards the Cloud, the Administrator Agent obtains the data from MA1, which came from the Customer layer, and gives them to the Encoding Agent, which encodes them, to ensure that no one can read the data that will be transferred towards the Cloud (even for the Cloud Provider). After the encoding operation, the Administrator Agent creates another Mobile Agent(MA2), to carry the data towards the Cloud Provider.

When the MA2 arrives at the Cloud layer, it contacts the Provider Agent. If all is regulated on the Storage level, the Provider Agent asks the Executor Agent to execute the request and to store the data of the customer.

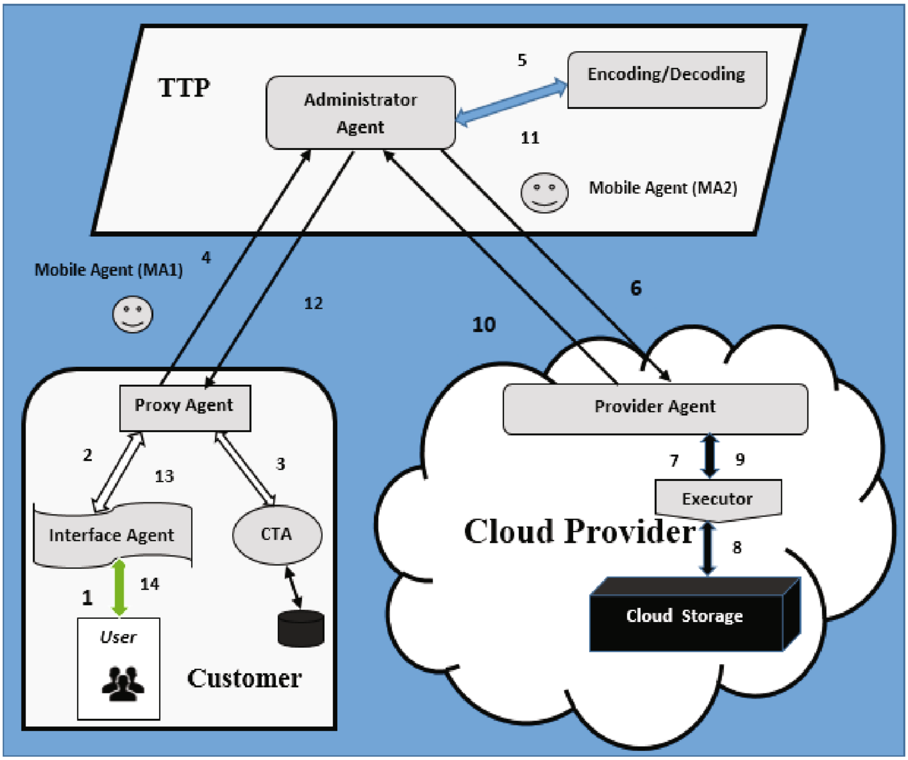


**Fig. 2.** Framework description

In the opposite direction, to turn over the data to the customer, the same procedure is carried out, the Provider Agent gives the data that is retrieved by the Executor Agent to the MA2, that came from the TTP layer, which moves towards the TTP layer, then the Encoding Agent decodes the data of the customer, after the decoding operation, the Administrator Agent gives the data to the MA1 that came from the customer layer, which moves to the Customer layer and gives the data to the Proxy Agent, this last gives them to the Interface Agent to display them to the customer.

The communication between the Customer layer and Cloud Provider layer through the TTP layer, can be summarized with the two sequence diagrams below, where the steps are as follows:

- **Step 01:** The customer sends his information of authentication to the Interface Agent, which analyzes them, if the information is normal, it sends them to the Proxy Agent.
- **Step 02:** The Proxy Agent checks the identity of the Customer, if the information is correct, it connects himself to the Interface Agent with SSL connection and asks him to post the user interface to the customer. If it is not, the access is rejected.

- **Step 03:** The Proxy Agent requests the consumer trust degree from the CTA. Which calculates the trust degree of the customer and indicates if the customer is a trusted entity or not.
- **Step 04:** The Proxy Agent receives the response from the CTA, if the customer's trust degree is more than the threshold for the connection, it creates the MA1, which carries the request of the customer towards the TTP layer. If it is not, the request is removed.
- **Step 05:** The Administrator Agent in the TTP layer receives the MA1, if the customer wants to store his data, it passes the data of the customer to the Encoding Agent, which encodes them and turns over them towards the Administrator Agent (Fig. 3).
- **Step 06:** The Administrator Agent creates MA2, which carries the request of the customer towards the Cloud Provider layer.
- **Step 07:** The Provider Agent receives the MA2, and asks the Executor Agent to carry out the request of the customer.
- **Step 08:** The Executor Agent carries out the request of the customer (for storing the data or to turn over them).
- **Step 09:** If the customer asked to recover his data, the Executor Agent turns over them and gives them to the Provider Agent.
- **Step 10:** The Provider Agent passes the data to the MA2, which moves towards the TTP layer and gives them to the Administrator Agent.
- **Step 11:** The Administrator Agent asks the Encoding Agent to decode the data of the customer. This last decodes them, and turns over them to the Administrator Agent.
- **Step 12:** The Administrator Agent gives the data to the MA1, which moves towards the Customer layer and passes them to the Proxy Agent.

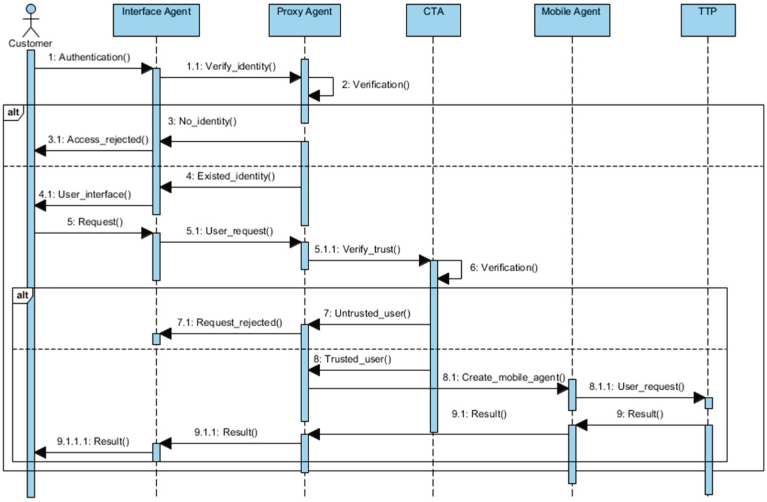


**Fig. 3.** The interaction between Customer and TTP

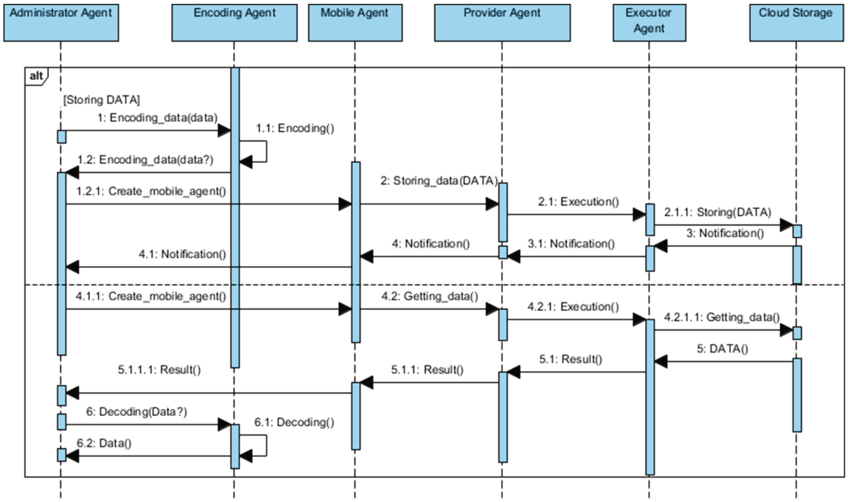


**Fig. 4.** The interaction between TTP and Cloud Provider

- **Step 13:** The Proxy Agent receives the data and gives them to the Interface Agent.
- **Step 14:** The Interface Agent displays the data to the customer, and the customer can recover his data in full security (Fig. 4).

### 5.2 Trust Evaluation

Trust is one of the most important means to improve security and privacy of Cloud platforms. While in fact trust is the most complex relationship between entities because it is extremely abstract, unstable and difficult to be measured and managed [14]. In our framework, we based on the work of [15], which is a pervasive trust management model for dynamic open environments. It based on fuzzy logic to calculate the trust degree of the customer, according to the actions carried out by them.

To calculate the trust degree of the customer, we based on the continuous examination of the customer behavior, the CTA uses a function of trust to indicate the trust degree of each customer. After each action doing by the customer, this agent recalculates the trust degree of the customer. The customer trust degree can be reduced or increased by the CTA according to actions carried out by the customer. At the end of each evaluation the customers are classified in three categories: trusted entity, innocent entity and untrusted entity [15].

We can also classify the actions carried out by the customers in two classes: positive actions and negative actions, the positive actions are correct actions carried out by the trusted entity, However, we assume that all negative actions are not the same, that is the reason because we distinguish between wrong actions and malicious actions, the wrong actions are bad actions which do not cause any damage to the system as the attempt to access to unauthorized resources, they are carried out by the innocent entity, but the malicious actions are harmful actions like attacks, they are carried out by the untrusted entity [15].

To calculate the action value Pa, we take into account the performed action weight, but this value is penalised or rewarded by the past behaviour. This function increases or decreases according to the performed positive and negative actions respectively, so [15]:

$$Pa = \left(1 - \frac{Na}{Totala}\right).Wa^{m} \qquad where \quad 0 <= Pa <= 1$$

- **(1-(Na/Totala)):** means the past behaviour of the customer, it goes towards **0** if the behavior is negative, and towards **1** for a positive behavior.
- **Na:** is the number of negative actions realized by the customer.
- **Totala:** is the total number of performed actions by the customer during the interaction with the Cloud Provider.
- **Wa:** is the action's weight according to it's nature (positive, wrong, and malicious). Wa for positive actions is **1**, for wrong actions is **0.5** and for malicious actions is **0**.
- **m:** the parameter m is the security level, where **m>=1**.

When the customer realizes a new action, Pa is recomputed, which reflects the present behavior of the customer, the new trust degree will take it into account and it will modify the current trust degree.

By the use of this model in our framework, we have an efficient method for the access management, we guarantee that only the authorised customers can interact with the Cloud Provider, we have also a kind of monitoring of the actions carried out by the customers. So we can stop any attempt of malicious action like the access for unauthorized resources, abuse of Cloud services or any attacks.

### 5.3 Data Encryption

For the data encryption phase, there is two steps, to guarantee the data confidentiality and privacy. The step of fragmentation, where we divide the data into many fragments, then a step of encoding, where RSA algorithm is used to find out the keys, these keys are used to encode and decode the file [16].

So the Encoding Agent divides initially the data of the customer into many fragments, then it encodes them by using the RSA algorithm. After the Encoding Operation, the TTP sends all of the fragments to the Cloud Provider to store them. So the Cloud Provider contains only a parts of data, which have no means. With this technique we guarantee that only the TTP can recover the data of the customer. Figure 5 shows the Encoding Operation.

For the retrieving of the data, the TTP carries out the opposite operation, to recover the real data of the customer from the fragments retrieved from the Cloud Provider. As in Fig. 6.

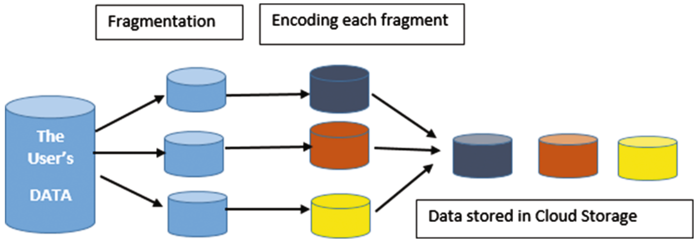


**Fig. 5.** Encoding operation

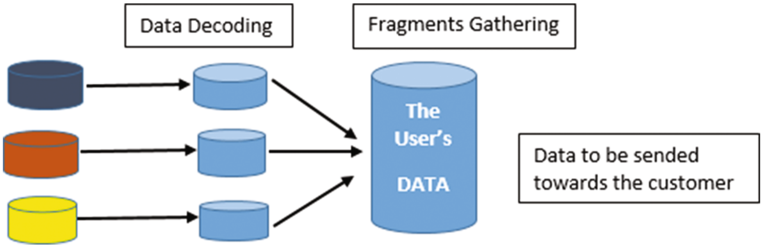


**Fig. 6.** Decoding operation

With the encryption of the file at the TTP layer, we guarantee that the data of the customers are safe, so the client will not worry about the privacy and the confidentiality of his data, which reside in the Cloud.

### 5.4 Data Intergrity Check

Integrity, in terms of data security, is nothing but the guarantee that data can only be accessed or modified by those authorized to do so, in simple word it is a process of verifying data. Data Integrity is very important among the other Cloud challenges. As data integrity gives the guarantee that data is of high quality, correct, unmodified [17].

To allow the customer to check the integrity of his data stored in the Cloud, we based on the Provable Data Possession(PDP) Scheme based on MAC(Message Authentication Code), to ensure data integrity of file F stored on Cloud storage in very simple way. The TTP computes a MAC for each fragment of the whole file with a set of secret keys and stores them locally before outsourcing it to the Cloud Service Provider(CSP). It Keeps only the computed MACs on his local storage, then sends the fragments of the file to the CSP, and deletes the local copy of the fragments. Whenever a customer needs to check the Data integrity of file F, the TTP sends a request to retrieve the file from CSP, reveals a secret keys to the CSP and asks to recompute the MACs of the fragments of the file, and compares the re-computed MACs with the previously stored values [17] (Fig. 7).

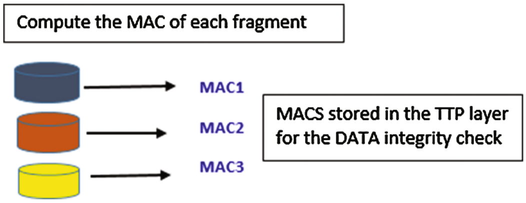

Compute the MAC of each fragment
MAC1
MAC2
MAC3
MACS stored in the TTP layer for the DATA integrity check

**Fig. 7.** Generating a MAC for each fragment

If the re-computed MACs at Cloud Provider and the MACs stored in the TTP are the same, then the data integrity is checked.

## 6 Conclusion

Cloud Computing security is very important for the continuity of this model, particularly in Cloud Storage service. In this paper we presented our proposed framework to ensure the security of Cloud Data Storage(CDS), which contains three layers. It is based on the use of agents, a Trust Model and the TTP which is used for the encryption and the integrity check of data.

Work is currently going on the framework implantation, where it will be applied to a specific case study. Further research could be realized to improve and to extend the present work, by including cognitive agents to make the interaction between these three layers more automatic, to increase Cloud performance.

## References

1. Desai, C.V., Jethava, G.B.: Survey on data integrity checking techniques in cloud data storage. Int. J. Adv. Res. Comput. Sci. Softw. Eng. (IJARCSSE) **4**(12), 292 (2014)
2. NIST. Nist cloud computing standards roadmap. Technical report, National Institute of Standards and Technology(NIST) (2013)
3. Yang, K., Jia, X.: Security for cloud storage systems. Technical report. Springer (2014)
4. CSA. Security guidance for critical areas of focus in cloud computing v2.1. Technical report, Cloud Security Alliance (CSA) (2009)
5. Akter, L., Monzurur Rahman, S.M., Hasan, Md.: Information security in cloud computing. Int. J. Inf. Technol. Converg. Serv. (IJITCS) **3**(4), 18 (2013)
6. Yadav, P., Sujata, : Security issues in cloud computing solution of DDOS and introducing two-tier captcha. Int. J. Cloud Comput. Serv. Archit. (IJCCSA) **3**(3), 29 (2013)
7. Pal, S., Khatua, S., Chaki, N., Sanyal, S.: A new trusted and collaborative agent based approach for ensuring cloud security. Ann. Faculty Eng. Hunedoara Int. J. Eng. **10**(1) (2012)
8. Hada, P.S., Singh, R., Manmohan, M.: Security agents: a mobile agent based trust model for cloud computing. Int. J. Comput. Appl. **36**(12), 12–15 (2011)

9. Talib, A.M., Atan, R., Abdullah, R., Murad, M.A.A.: A framework of multi agent system to facilitate security of cloud data storage. In: Annual International Conference on Cloud Computing and Virtualization (2010)
10. Alwesabi, A., Okba, K.: Security method: cloud computing approach based on mobile agents. Int. J. New Comput. Archit. Appl. (IJNCAA) **4**(1), 17–29 (2014)
11. Benabied, S., Zitouni, A., Djoudi, M.: A cloud security framework based on trust model and mobile agent. In: IEEE (2015)
12. Islam, Md.R., Habiba, M.: Agent based framework for providing security to data storage in cloud. In: IEEE (2012)
13. Venkateshwaran, K., Malviya, A., Dikshit, U., Venkatesan, S.: Security framework for agent-based cloud computing. Int. J. Artif. Intell. Interactive Multimed. **3**(3), 37–40 (2015)
14. Li, W., Ping, L.: Trust model to enhance security and interoperability of cloud environment. In: Proceedings of CloudCom 2009. Springer, Heidelberg (2009)
15. Almenarez, F., Marin, A., Campo, C., Garcia, R.C.: PTM: a pervasive trust management model for dynamic open environments. In: Proceedings of First Workshop on Pervasive Security, Privacy and Trust PSPT (2004)
16. Tikore, S.V., Pradeep, K.D., Prakash, B.D.: Ensuring the data integrity and confidentiality in cloud storage using hash function and TPA. Int. J. Recent Innov. Trends Comput. Commun. (IJRITCC) **3**(5), 2738 (2015)
17. Giri, M.S., Gaur, B., Tomar, D.: A survey on data integrity techniques in cloud computing. Int. J. Comput. Appl. **122**(2) (2015)

# Proposal for the Design of a New Technological Infrastructure for the Efficient Management of Network Services and Applications in a High Complexity Clinic in Colombia

Leonel Hernandez(✉), Humberto Villanueva, and Sandra Estrada

Department of Telematic Engineering, Engineering Faculty, Institución Universitaria ITSA, Barranquilla, Colombia
lhernandezc@itsa.edu.co, msn_huvi_15@hotmail.com, sandraestradahurtado@gmail.com

**Abstract.** Characterization of the information collected from the CLINIC CRC SITE – BARRANQUILLA (Headquarter) network infrastructure will serve as a foundation for building the new proposal of technological infrastructure for the new site of the clinic located in Baranoa. It will be designed according to the guidelines of the methodologies of design Top-Down and PPDIOO, which are: analyze requirements, develop a logical design, develop a physical design, test, optimize and document design, testing the network, and monitor performance. Following each of the previous stages 4 scenarios were designed and tested, which simulate the administration and performance of the network services.

**Keywords:** Infrastructure · Servers · Network services · Network design · VLANs

## 1 Introduction

At present the technological field is advancing faster than expected. During the last two decades, there has been a huge growth in the size of networks and this is due to the inevitable need to stay in communication. Everyday tasks are performed through technological devices that allow automating the simplest processes performed by the human being. However, the misuse of different technologies used in a functional network infrastructure, reduces the ability of the various devices built into the network design.

From this situation, every day in companies or other organizations it becomes necessary a good network design that guarantees the security, availability, scalability and performance of the network environment, this balance offers security to one of the most valuable assets in any organization, information.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_7

For this reason, the following project shows the importance of a flexible design that can adapt to changes in traffic pattern and other requirements, changes that may come from new protocols, and avoid incorporating design elements that would make it difficult to implement new Technologies in the future. To this end, the new network infrastructure of Clinic CRC – Site BARANOA branch will be taken, with the purpose of providing efficient connectivity to the numerous services and network applications to the various users.

## 2 Networking Concepts

Among all the essentials for human existence, the need to interact is right after the need to sustain life. Communication is almost as important to us as air, water, food and a place to live.

The methods we use to share ideas and information are constantly changing and evolving. While the human network was once limited to face-to-face, the media advance continues to expand the scope of our communications. From the press to television, each new development has improved and strengthened our communication.

With every advance in communication technology, the creation and networking of solid data are having a deep effect.

### 2.1 Computers Networks

Gerardo defines the Computer Network as a system consisting of multiple computers that are linked by some means of data communication, for example: coaxial cable (like cable TV), twisted pair, optical fiber, radio signal, Satellite, etc. Figure 1 shows a scenario of computer networks where a client and a server communicate [1, 2]:

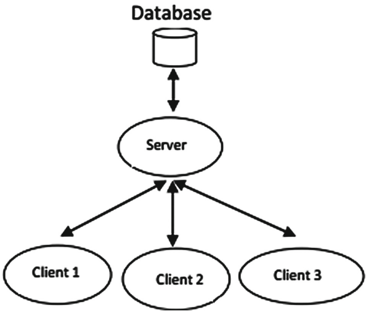


**Fig. 1.** Client – Server Scenario

Computer networks offer several advantages for the administration of an organization, which are important to highlight in Table 1:

**Table 1.** Advantages of computer networks

| Advantages | Description |
|---|---|
| Accessibility | It provide access to the members of the organization to an Institutional Database, where important events that occur to the company are recorded, no matter where they take place |
| Security | In a network, only those who have authorization for it can participate and the type of activities they can do (consult, register, modify, delete, etc.) can be defined by the organization |
| Efficiency and effectiveness | Networks can eliminate repetitive operations within an organization's processes and even innovate the minimum processes, generating added value and increasing the strength of an organization |
| Savings | A network can allow an organization to decrease its investments in computers and programs when sharing a shared resource, allowing us to share a network printer, CD players, etc. |

### 2.1.1 LAN

To define the concept of LAN, Andrew expresses that Local Area Networks (generally known as LANs) are privately owned networks that are in a single building or on a campus of a few kilometers in length. Figure 2 shows the interconnection of several computers or peripherals, an example of a LAN [3].

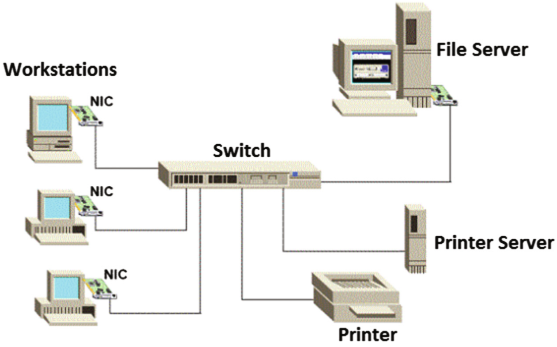


**Fig. 2.** Local area network – LAN

### 2.1.2 VLAN

A VLAN allows a network administrator to logically create groups of devices connected to the network that act as if they are in their own separate network, even if they share a common infrastructure with other VLANs. Figure 3 shows how it is possible to use a VLAN to geographically structure the network and manage access and security policies for user groups [4, 5]

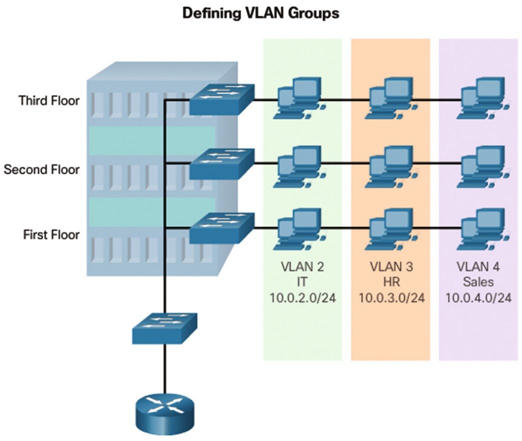


**Fig. 3.** Virtual local area network – VLAN

## 2.2 File Service

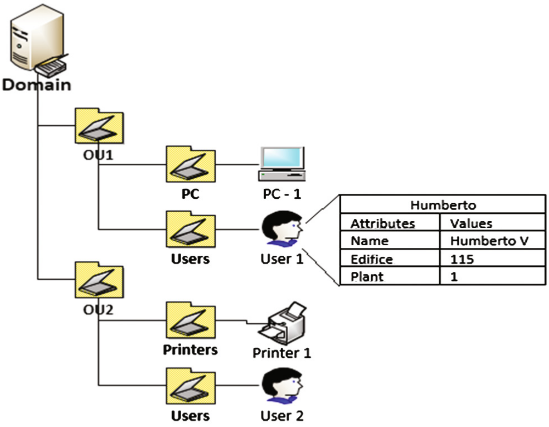
According to Marty, a File Server provides a network resource for storing information (files or programs) that other network members can use, as long as they have the appropriate permissions. The file server also facilitates the backup of network storage resources. Allows sharing directories, drives and other objects [6].

## 2.3 Active Directory

Simmons explains Active Directory as a domain controller service that allow to store and manage directory information and login processes for computers that are involved in a domain. It is a service established on one or more servers where objects such as users, computers or groups are created, to manage the logins on the computers connected to the network, as well as the administration of policies throughout the network. Its hierarchical structure allows to maintain a series of objects related to components of a network, such as users, user groups, permissions and resource allocation and access policies [7].

### 2.3.1 Active Directory Objects

Honeycutt states that the objectives of the Active Directory are to share and control resources, equipment, information and programs that are locally or geographically dispersed. Provide reliability to the information, having storage alternatives. Get a good cost/benefit ratio. Provide privileges to users of information Transmit information between distant users in the fastest and most efficient way possible. Figure 4 shows an example of the Active Directory architecture [8].



| Humberto | |
|---|---|
| Attributes | Values |
| Name | Humberto V |
| Edifice | 115 |
| Plant | 1 |

**Fig. 4.** Active directory architecture example

### 2.3.2 Organizational Unit

Stanek defines an Organizational Unit as a container where it can put different Active Directory objects as Users, Computers, Groups and even other OUs. Within the same, it can delegate Administration permissions on the objects that it has inside and it can attach domain policies, to create and to apply different configurations on the types of objects that we have inside. Figure 5 shows the creation of an Organizational Unit from the Active Directory Users and Computers console [9].

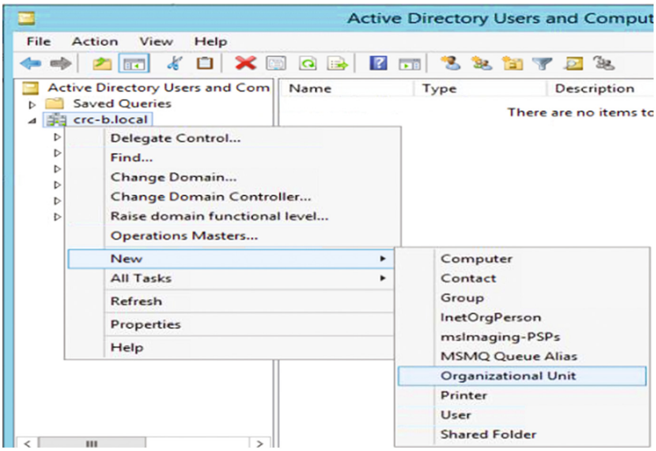


**Fig. 5.** Organizational unit creation

### 2.4 Network Topologies

The topology of a network is the relationship of the network devices and the interconnections between them. The word topology means basically form; The term network topology refers to the shape of a network, that is, how all nodes (points) in a network are wired [10].

## 3 Network Design: A Literature Review

Data networks and communications, in general, must meet the following basic conditions: security, scalability, high performance, redundancy, manageable and easy to maintain. The key to the fulfillment of these conditions is to design a hierarchical network of interconnectivity enabling users to various network services and applications that support it [11]. There are important studies and research about green data center, it is found in the scientific literature. In this literature review, some of these investigations are mentioned.

Marugan in his research "Design of Network Infrastructure and Computer Support for a Public Education Center" established the guidelines for modeling a technological endowment for a kindergarten, taking into consideration several requirements for the design of networks [12]. Ruile in his research "Design and implementation of IT service management system of college or university campus network" analyzes the design and implementation of a network management service on a university campus [13]. Haitao in his study "Architecture design and implementation methods of heterogeneous emergency communication network" shows how to design and build an emergency communication network, using methods of implementation with WiMax [14]. Hernandez, in his applied research "Distributed Infrastructure For Efficient Management of Network Services. Case: Large Company In Mining Sector in Colombia", explain the process to design and implement a distributed network infrastructure for a mining company [15]. Cadena in his study "Analysis and Design for Health Area No 1 located in the Historic Center of the city of Quito", shows the stages of the design and implementation of a new network infrastructure of a Clinic in Quito [16].

## 4 Proposal of Design of Network Topologies for the New Infrastructure

According to the information provided by the physical infrastructure plans of the Reina Catalina Clinic - Baranoa Headquarters, the topology of the LAN and Data Center has been designed. The PPDIOO and Top-Down network design methodologies were used to design and implement the new network infrastructure of the clinic [17].

### 4.1 LAN Topology

As for the local network of the Clinic, the LAN topology has been designed, indicating the equipment that will be part of the Local network, where it will access the users of different departments. Also, in Fig. 6, the LAN topology for the new network infrastructure is shown:

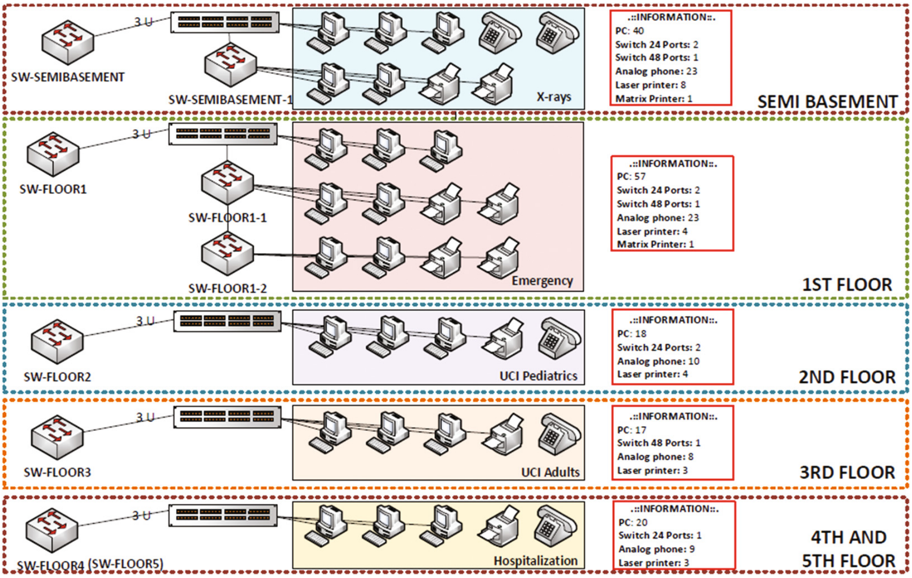


**Fig. 6.** LAN topology – CRC Baranoa

Similarly, Table 2 shows the total number of Assets in the local network

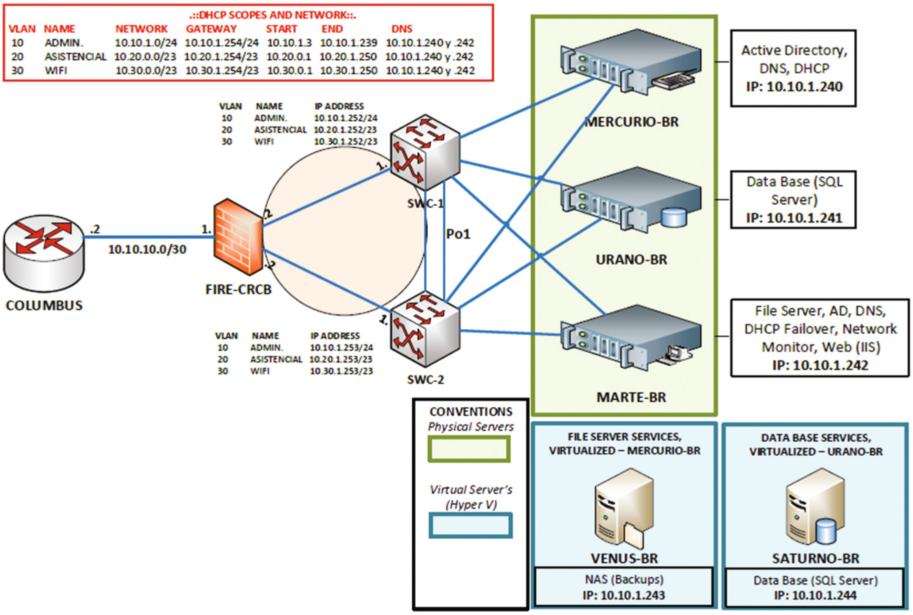
**Table 2.** Total Number of Assests

| Number of network active devices | |
|---|---|
| 24 Ports switch | 8 |
| 48 Ports switch 48 | 5 |
| Analog phones | 67 |
| Desktops | 188 |
| Laser printers | 22 |
| Printers spot matrix | 2 |

## 4.2 Data Center Topology

According to the information obtained in the interviews conducted to the head of systems of the Clinic Reina Catalina - Sede Barranquilla, it was possible to determine the services that are required for the new network infrastructure. Concepts were also used about the design of data centers [18].

Similarly, the addressing used for the simulation of Data Center equipment will be segmented into VLAN's and other important conventions indicated in the Topology below in Fig. 7:



.::DHCP SCOPES AND NETWORK::.

| VLAN | NAME | NETWORK | GATEWAY | START | END | DNS |
|---|---|---|---|---|---|---|
| 10 | ADMIN. | 10.10.1.0/24 | 10.10.1.254/24 | 10.10.1.3 | 10.10.1.239 | 10.10.1.240 y .242 |
| 20 | ASISTENCIAL | 10.20.0.0/23 | 10.20.1.254/23 | 10.20.0.1 | 10.20.1.250 | 10.10.1.240 y .242 |
| 30 | WIFI | 10.30.0.0/23 | 10.30.1.254/23 | 10.30.0.1 | 10.30.1.250 | 10.10.1.240 y .242 |

| VLAN | NAME | IP ADDRESS |
|---|---|---|
| 10 | ADMIN. | 10.10.1.252/24 |
| 20 | ASISTENCIAL | 10.20.1.252/23 |
| 30 | WIFI | 10.30.1.252/23 |

| VLAN | NAME | IP ADDRESS |
|---|---|---|
| 10 | ADMIN. | 10.10.1.253/24 |
| 20 | ASISTENCIAL | 10.20.1.253/23 |
| 30 | WIFI | 10.30.1.253/23 |

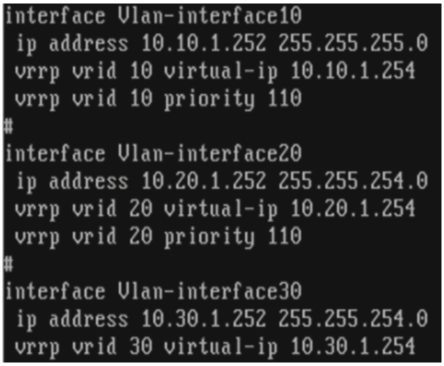
**Fig. 7.** Data center topology

# 5 Simulated Scenarios of the Various Network Services of the New Technological Infrastructure

For the development of the Project, four scenarios have been determined where it is intended to test the functionality of network services and devices, server administration, traffic monitoring and other key services for this solution. All this will be done through simulations and virtualization technologies.

## 5.1 Scenario 1: Switch and Firewall Operation and Management

During the simulation of this scenario, it was verified the operation and initial configuration of the networks, segmenting in three VLAN's. In addition, there is evidence of link aggregation, fault tolerance, and load balancing of the subnetworks with the VRRP protocol. Remote management of Switches with SSH was also performed. Figure 8 shows the basic configuration of the routing (VLAN's) and the VRRP [19, 20] in some of the switches of this scenario:

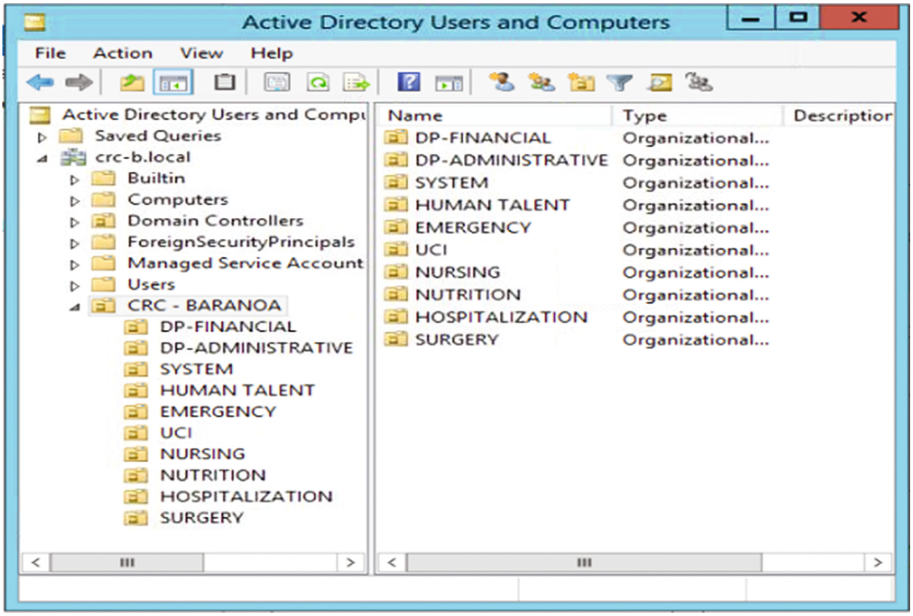


```
interface Vlan-interface10
 ip address 10.10.1.252 255.255.255.0
 vrrp vrid 10 virtual-ip 10.10.1.254
 vrrp vrid 10 priority 110
#
interface Vlan-interface20
 ip address 10.20.1.252 255.255.254.0
 vrrp vrid 20 virtual-ip 10.20.1.254
 vrrp vrid 20 priority 110
#
interface Vlan-interface30
 ip address 10.30.1.252 255.255.254.0
 vrrp vrid 30 virtual-ip 10.30.1.254
```

**Fig. 8.** VLAN and VRRP basic configuration

## 5.2 Scenario 2: Primary Services

During the simulation of this scenario, it is checked the administration of the users by Active Directory and services of networks like DHCP and DNS in servers with Operating System Windows Server [21–23]. In addition, the verification of these services is carried out checking the connectivity of the users with routing delivered from the server and the access of the users through the domain controller. Similarly, the configuration of the switches is preserved in the same configuration of the previous scenario, with the



- Active Directory Users and Compu
  - Saved Queries
  - crc-b.local
    - Builtin
    - Computers
    - Domain Controllers
    - ForeignSecurityPrincipals
    - Managed Service Account
    - Users
    - CRC - BARANOA
      - DP-FINANCIAL
      - DP-ADMINISTRATIVE
      - SYSTEM
      - HUMAN TALENT
      - EMERGENCY
      - UCI
      - NURSING
      - NUTRITION
      - HOSPITALIZATION
      - SURGERY

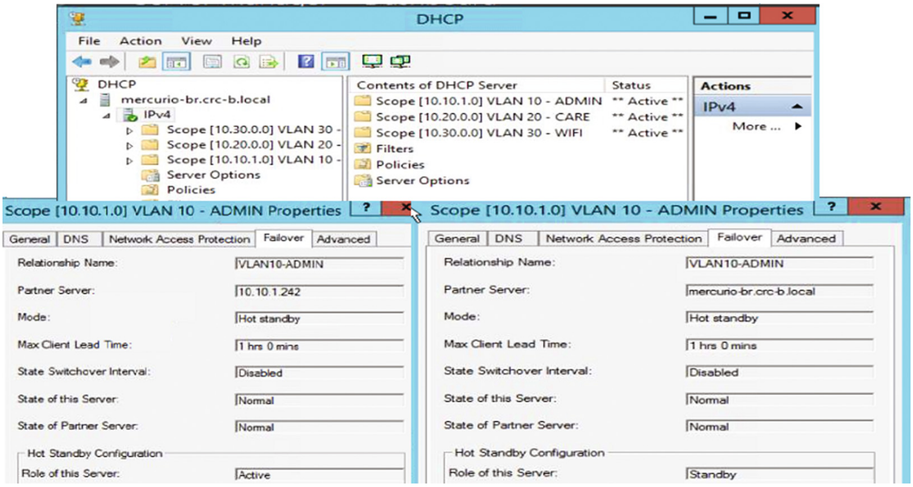
| Name | Type | Descriptior |
|---|---|---|
| DP-FINANCIAL | Organizational... | |
| DP-ADMINISTRATIVE | Organizational... | |
| SYSTEM | Organizational... | |
| HUMAN TALENT | Organizational... | |
| EMERGENCY | Organizational... | |
| UCI | Organizational... | |
| NURSING | Organizational... | |
| NUTRITION | Organizational... | |
| HOSPITALIZATION | Organizational... | |
| SURGERY | Organizational... | |

**Fig. 9.** Active directory users and computers - CRCB

difference that DHCP Relay is configured in the other VLANs where the DHCP Server is not located. Figure 9 shows the Active Directory Users and Computers console:

### 5.3 Scenario 3: Backup Services

In this scenario, the fault tolerance is checked against the services that were tested in the previous scenario. Retaining Active Directory and DHCP configuration [24, 25], a replica of these services was performed, making the configuration of a domain controller to an existing domain by the Active Directory and DHCP Failover for the DHCP service by setting the standby mode on the associated server. In addition, we have monitored the network services for solving problems that we have in the network. Figure 10 shows the DHCP service running on the Server and the replication of a network scope to the associated server:



| Contents of DHCP Server | Status |
|---|---|
| Scope [10.10.1.0] VLAN 10 - ADMIN | ** Active ** |
| Scope [10.20.0.0] VLAN 20 - CARE | ** Active ** |
| Scope [10.30.0.0] VLAN 30 - WIFI | ** Active ** |
| Filters | |
| Policies | |
| Server Options | |

Scope [10.10.1.0] VLAN 10 - ADMIN Properties

Relationship Name: VLAN10-ADMIN
Partner Server: 10.10.1.242
Mode: Hot standby
Max Client Lead Time: 1 hrs 0 mins
State Switchover Interval: Disabled
State of this Server: Normal
State of Partner Server: Normal
Hot Standby Configuration
Role of this Server: Active

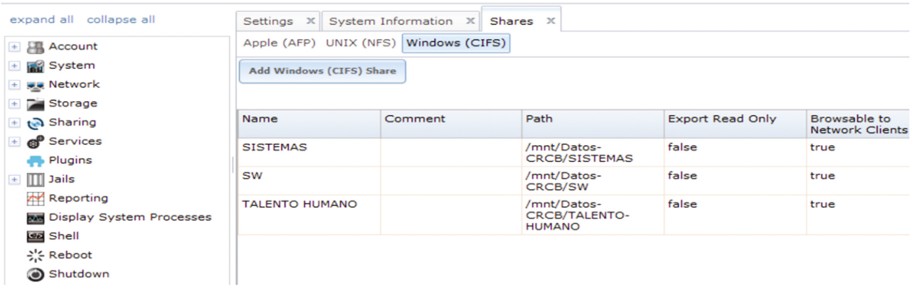
Scope [10.10.1.0] VLAN 10 - ADMIN Properties

Relationship Name: VLAN10-ADMIN
Partner Server: mercurio-br.crc-b.local
Mode: Hot standby
Max Client Lead Time: 1 hrs 0 mins
State Switchover Interval: Disabled
State of this Server: Normal
State of Partner Server: Normal
Hot Standby Configuration
Role of this Server: Standby

**Fig. 10.** DHCP server and DHCP failover

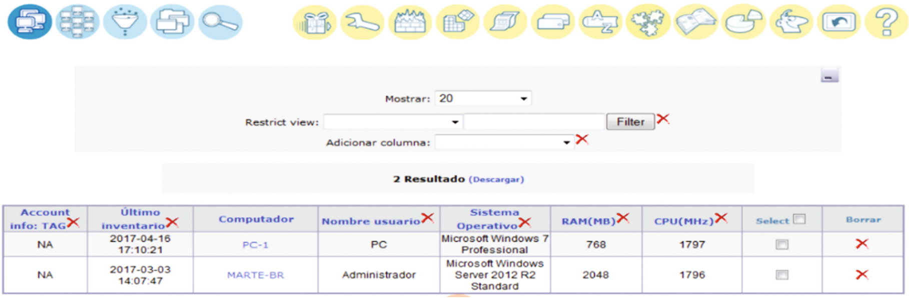
### 5.4 Scenario 4: Data Base and Files Services

In this scenario, the operation of the Storage services of this Network Infrastructure will be checked, obtaining a file server and another dedicated server to back up the data that will be handled in the clinic. On the backup file server with the FreeNAS Operating System, file sharing was established with the SMB, FTP and TFTP protocol. In the same way, it was verified the operation of a database service in charge of carrying out the inventory of the computer equipment with all the information of Hardware and Software. Figures 11 and 12 show the backup server with some shared folders and the OCS Inventory Inventory database service:



| Name | Comment | Path | Export Read Only | Browsable to Network Clients |
|---|---|---|---|---|
| SISTEMAS | | /mnt/Datos-CRCB/SISTEMAS | false | true |
| SW | | /mnt/Datos-CRCB/SW | false | true |
| TALENTO HUMANO | | /mnt/Datos-CRCB/TALENTO-HUMANO | false | true |

**Fig. 11.** FreeNAS backup file services



Mostrar: 20

Restrict view:

Adicionar columna:

2 Resultado (Descargar)

| Account info: TAG | Último inventario | Computador | Nombre usuario | Sistema Operativo | RAM(MB) | CPU(MHz) | Select □ | Borrar |
|---|---|---|---|---|---|---|---|---|
| NA | 2017-04-16 17:10:21 | PC-1 | PC | Microsoft Windows 7 Professional | 768 | 1797 | □ | ✕ |
| NA | 2017-03-03 14:07:47 | MARTE-BR | Administrador | Microsoft Windows Server 2012 R2 Standard | 2048 | 1796 | □ | ✕ |

**Fig. 12.** Inventory database service – OCS inventory

## 6 Conclusions

During the development of the proposal, the initial information was collected through an interview, to analyze the requirements to be able to formulate a feasible Network Design proposal that will adapt to the changes made in the future. In the development of the research was used as a descriptive methodology for the collection of documentation of the current network of the Main Clinic, and applied where the proposed design of the new infrastructure was proposed. In addition, network topologies were designed for the Data Center and LAN, network services were also chosen that were simulated in four scenarios to verify the development and administration of these services.

This paper hopes to be a contribution to the scientific literature related to the design of network infrastructure and an important contribution for a healthcare provider to function efficiently at all levels, thanks to a modern network design that supports the latest protocols and technological trends.

## References

1. Vilet, G.: La Tecnología y los Sistemas de Información aplicados en los negocios y la Educación. San Luis Potosi, p. 116 (2008)
2. Oluwatosin, H.S.: Client-server model. IOSR J. Comput. Eng. Ver. **IX**(1), 2278–8727 (2014)

3. Tanenbaum, A.S., Wetherall, D.J.: Redes De Computadoras (2012)
4. Cisco Networking Academy, "Cisco Networking Academy" (2015). http://www.cisco.com/web/learning/netacad/index.html
5. Garimella, P., Sung, Y.W.E., Zhang, N., Rao, S.: Characterizing VLAN usage in an operational network. In: INM 2007, pp. 305–306 (2007)
6. Matthews, M.: Microsoft Windows Server 2008: A Beginner's Guide, p. 598 (2008)
7. Simmons, C.: Active Directory Bible (2001)
8. Honeycutt, J.: Introducing Microsoft Windows Server 2003, Ilustrated. Microsoft Press, 2003 (2011)
9. Stanek, W.: Microsoft Windows 2000. Manual del Administrador (2000)
10. Pandya, K.: Network Structure or Topology. Netw. Struct. or Topol. **1**(2), 22–27 (2013)
11. Dye, M.: Network fundamentals: CCNA exploration companion guide (2008)
12. Marugan Merinero, J.: Diseño de Infraestructura de red y soporte informatico para un centro publico de educación infantil y primaria, p. 180 (2010)
13. Ruile, L.: Design and implementation of IT service management system of college or university campus network. In: 2012 7th International Conference on Computer Science and Education (ICCSE), pp. 299–304 (2012)
14. Wang, H., Song, L.: Architecture design and implementation methods of heterogeneous emergency communication network. In: Communications in Computer and Information Science. CCIS, vol. 143, Part 1, pp. 122–127 (2011)
15. Hernandez, L.: Distributed infrastructure for efficient management of network services. Case: large company in mining sector in Colombia. In: 2016 2nd International Conference Science Information Technology Proceedings, pp. 63–68. IEEE (2016)
16. Cadena, L.: Análisis y Diseño de la red de datos para el área de salud No 1 ubicada en el centro histórico de la ciudad de Quito. p. 95. Quito (2016)
17. Oppenheimer, P.: Top-down Network Design (1999)
18. Bruno, A., Jordan, S.: CCDA 640-864 Official Certification Guide. Indianapolis (2011)
19. Hinden, R.: Virtual Router Redundancy Protocol (VRRP), RFC3768, no. 3768 (2004)
20. Pavlik, J., Komarek, A., Sobeslav, V., Horalek, J.: Gateway redundancy protocols. In: Proceedings of CINTI 2014 – 15th IEEE International Symposium on Computational Intelligence and Informatics, pp. 459–464 (2014)
21. Smith, R.: DNS in windows server 2008 R2. (cover story). Wind. IT Pro **17**(3), 27–30 (2011)
22. Lane, R., Muggli, N., Bhai, S.: Active Directory Domain Services in the Perimeter Network (Windows Server 2008). Computer, 52, April 2009
23. Svidergol, B., Allen, R.: Active directory cookbook. Saudi Med. J. **33**, 832 (2013)
24. Droms, R.: Automated configuration of TCP/IP with DHCP. IEEE Internet Comput. **3**(4), 45–53 (1999)
25. Lin, C., Su, T., Wang, Z.: Summary of high-availability DHCP service solutions. In: Proceedings – 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology, IC-BNMT 2011, pp. 12–17 (2011)

# Initial Centroid Selection Optimization for K-Means with Genetic Algorithm to Enhance Clustering of Transcribed Arabic Broadcast News Documents

Ahmed Mohamed Maghawry[1(✉)], Yasser Omar[1], and Amr Badr[2]

[1] Department of Computer Science,
College of Computers and Information Systems,
Arab Academy for Science and Technology (AAST), Cairo, Egypt
ahmed.mg.mohamed@gmail.com, dr_yaser_omar@yahoo.com
[2] Department of Computer Science, Faculty of Computers and Information,
Cairo University, Cairo 12613, Egypt
A.badr.fci@gmail.com

**Abstract.** In this research a collection of artificial intelligence techniques are combined together to optimize the process of clustering textual transcripts obtained from audio sources. Since clustering techniques have drawbacks that if not taken care of will produce sub optimal clustering solutions, it's essential to attempt to optimize the clustering algorithms to avoid sub optimal solutions. As an attempt to overcome this problem, different artificial intelligence techniques are applied to avoid clustering problems. The main objectives of this research is to optimize automatic topic clustering of transcribed speech documents, and investigate the impact of applying genetic algorithm optimization and initial centroid selection optimization (ICSO) in combination with K-means clustering algorithm using Chi-Square similarity measure on the accuracy and the sum of square distances (SSD) of the selected clustering algorithm. The evaluation showed that using ICSO with genetic algorithm and K-means clustering algorithm with Chi-square similarity measure achieved the highest accuracy with the least SSD.

**Keywords:** Clustering · K-means · Genetic algorithm · Speech transcripts · Text clustering · Topic identification · Optimization · Centroid selection

## 1 Introduction

### 1.1 Dealing with Rapidly Growing Audible News

Audible news broadcasted on radio stations, television and on the internet is growing exponentially, to facilitate future search and retrieval, massive amount of data must be organized and stored, hence it demands rapid and robust techniques to organize and store these massive amounts of data. There are many challenges that still confront the field of multimedia information retrieval field despite its rapid advance in the past decade.

The asymmetric nature of audio and video is the main problem challenging researchers on this field. Regarding audio, the analysis of audio documents has focused on

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_8

two main directions. The first approach was to develop audio data classification schemes to segment an audio document into coherent chunks of different types of audio classes — music, speech, speech and music etc. [1, 2]. The second approach focused on transcribing audio streams into text documents using Automatic Speech Recognition (ASR) — a computer technology that identifies words spoken by a person into a microphone or telephone and convert them to written text — then perform analysis for automatic indexing, retrieval, and other tasks [3, 4]. These attempts have shown the availability of applying ASR to audio streams to achieve indexing, retrieval and other tasks with significant degree of success. However, there are other factors that affect the degree of success of those approaches including the size and quality of the transcription process.

This research is organized as follows:

Section 2 Background, Sect. 3, K-Means challenges is discussed. Section 4 the Proposed Model. Section 5, experimental results are evaluated. Section 6 concludes the research, Sect. 7 future work.

## 2 Background

### 2.1 K-Means Clustering Algorithm

The K-means clustering algorithm will be used in this research, not only because it's one of the most commonly used clustering techniques but also because it has been applied in many scientific and technological fields [15–24]. The K-means method has not only suffered from a major problem of which the algorithm produces empty clusters [18] added to that the problem produced by the random nature of cluster's initial centers selection that causes the algorithm to tend to sub optimal solutions. K-means clustering algorithm will be used to group transcribed textual documents obtained from audio sources into topics by applying a similarity measure based on the Chi-square method, which is designed to eliminate non informative words that will more likely be erroneous words when applied on transcribed documents [4].

The K-means clustering algorithm belongs to the partitioning based and nonhierarchical clustering techniques [16]. The algorithm starts with a set of numeric objects X and an integer number k, then attempts to find the partition of X into k clusters while minimizing the sum of squared errors [19]. First the K-means algorithm initializes the k cluster centers. Second, the algorithm attempts to allocate each of the input data points to the closest centers according to the square of the Euclidean distance from the cluster [26]. Third, the mean value of each cluster is computed in order to update the cluster center. This updating process happens because of the change in the membership of each cluster [20]. Re-assigning the membership of the input vectors and the continuous update of the cluster centers is repeated until no more changes in the value of any of the cluster centers occurs.

K-means is commonly used because of its simplicity and the ability of applying it on a wide variety of data types. However, it's quite sensitive to the initial positions of cluster centers. Listed below are the steps of the K-means algorithm:

1. Initialization: K data points are chosen randomly to initialize the K cluster centers.
2. Nearest-neighbor search: for each data point, the data point will be assigned to a cluster center if this cluster center is the closest to that data point.

How near the data vector is close to a centroid is calculated using formula (1).

$$\mathrm{d}(\boldsymbol{z_p}, \boldsymbol{a_j}) = \sqrt{\sum\nolimits_{k=1}^{d} \left(z_{pk} - a_{jk}\right)^2} \tag{1}$$

where d represents the dimension of the data point vector, $Z_p$ represents the centroid of the cluster P and $a_j$ is the data point's vector. 3. Updating the mean: for each cluster, calculate the mean of the input vectors assigned to that cluster to find the new cluster's center. 4. Stopping criteria: step 2 and step 3 are repeated until there's no change in the value of the calculated means.

### 2.2 Genetic algorithm

On the other hand, genetic algorithms where introduced by Holland [6] and further described by Goldberg [7] as optimization techniques to search for global or near global optimal solutions, it's a smart exploitation of the random search used to solve optimization problems. To overcome the transcription errors produced by the common drawbacks of ASR, root-based stemming technique is applied. To achieve topic identification, K-means [5] clustering technique is utilized.

### 2.3 K-Means with GA and Optimized Initial Centroid Selection

This work embraces the approach of applying ASR technology to Arabic news audio documents, and then applying preprocessing techniques and clustering algorithm on the transcribed textual documents produced by the ASR as in Fig. 1, then attempt to optimize the operation of the K-means initial centroid selection using Initial Centroid Selection Optimization (ICSO) an approach presented in this research, which should enhance the quality of the randomly selected centroids as in Fig. 2.

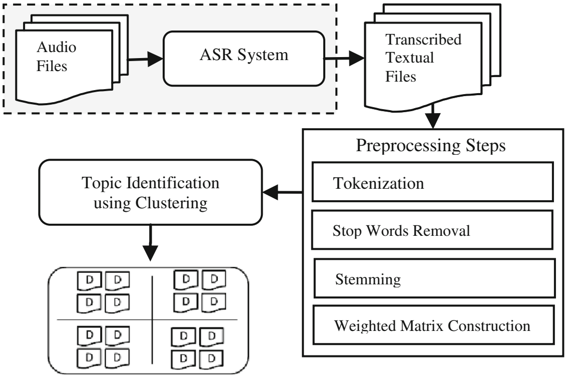

Audio Files
ASR System
Transcribed Textual Files
Preprocessing Steps
Tokenization
Stop Words Removal
Stemming
Weighted Matrix Construction
Topic Identification using Clustering
D

**Fig. 1.** The procedure of clustering-based topic identification of transcribed textual files obtained out of audio files.

Finally introduce these centroids for the K-means algorithm and produce a number of clustering solutions, and deliver these solutions as the initial population for the genetic algorithm to attempt to find the global or near global optimal solution. The topic clustering accuracy is evaluated for the selected clustering algorithm in four situations: When the transcribed documents are clustered using pure K-means without the use of neither ICSO nor GA, when clustered with ICSO support, and when clustered using ICSO and GA optimization as shown in Table 1.

**Table 1.** Testing scenarios

| Case ID | K-means | Centroid optimization | GA |
|---|---|---|---|
| A | ✓ | ✗ | ✗ |
| B | ✓ | ✓ | ✗ |
| C | ✓ | ✗ | ✓ |
| D | ✓ | ✓ | ✓ |

## 3 K-Means Challenges

Despite the simplicity of k-means and its wide scale of usage in different fields, there are some challenges related to it, one of the most important drawbacks of k-means is that the algorithm's final clustering result is extremely sensitive to one of the basic and mandatory steps of k-means which is the initial random centroids selection [22]. As a result, for a given clustering problem, different algorithm runs can output different clustering solutions for the same problem depending on the initial centroids selection [24], that's why in many previous researches and applications, k-means results in terms of accuracy where not on the top because the algorithm may tend to sub optimal solutions [12].

Therefore we propose in this paper that if we provided the k-means algorithm with high quality initial centroids, the algorithm will show significant results. Furthermore, if genetic algorithm optimization was applied alongside k-means with high quality initial centroids, the algorithm will show results that might exceed other clustering techniques.

## 4 Proposed Model

The K-means is quite sensitive to the initial selection of random centroids, as it depends on the initial centroids to compute distances between them and the data set elements targeted for clustering and assign each element to the closest centroid, then compute the mean of each formed cluster and recalculate the centroid value.

Assume we have a test data set of 100 objects that we have a prior knowledge that they can be divided into 4 clusters each containing 25 elements. If we pass this data set to the k-means to cluster it and provided the algorithm with k = 4, the algorithm will start by randomly picking 4 centroids, the problem will happen if the algorithm picked more than one centroid that belong to the "same" category, moreover the algorithm might pick all 4 initial random centroids from the same category, because of that, obviously the algorithm will out put a solution with very bad accuracy. That's why in the Initial Centroid Selection Optimization phase we will guide the k-means algorithm to pick high quality initial centroids, in other words, maximize the probability that the algorithm will pick 4 initial random centroids that doesn't belong to the same cluster. Our data set will be text transcripts gained from transcribing Arabic audio news files.

### 4.1 Vector Representation Model

Initially, all files will be represented using Vector Representation Model (VRM) [12]. Then, these vectors will be sorted by each word's weight either ascending or descending both will achieve our objective which is, by sorting these vectors, those who are similar will be grouped together. Each vector will be a row matrix 1 × n where n is the number of all unique words that are present in all the transcribed files, and all words will be grouped into zones within the vector so the summation of all the weights the belong to a specific zone will describe the weight of each document regarding that zone.

### 4.2 Initial Centroid Selection Optimization

Then divide these vectors into k zones according to the user specified k, and direct the algorithm to choose the k initial random centroids one from each zone. Thus each initial centroid will be random and at the same time each initial centroid will be more likely different than the others hence doesn't belong to the same cluster, hence provide high quality initial centroids to the k-means algorithm to start with as visualized in Fig. 2.

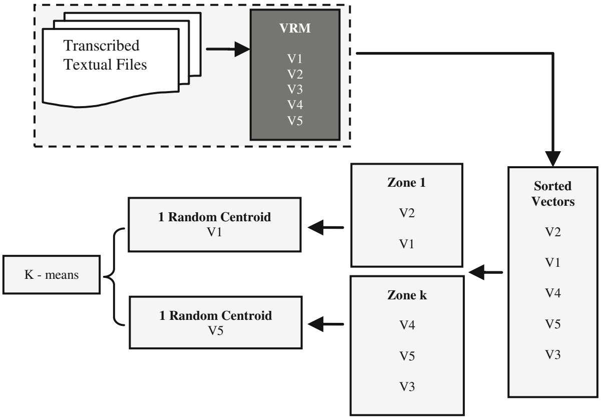


**Fig. 2.** The procedure used to provide high quality centroids to the k-means, using k = 2

After the n vectors are sorted, they will be divided using formula (2):

$$Z = n/k \tag{2}$$

where n is the total number of vectors and k is the user specified number of clusters and Z is the number of zones.

### 4.3 Genetic Algorithm to Optimize K-Means

A genetic algorithm is a randomized search and optimization technique which is guided by principles of evolution and natural genetics, having a large amount of parallelism [9, 10].

In order to apply genetic algorithm based data clustering we first need to specify how an individual will be represented then initialize starting population then pass it to the fitness function then select fit chromosomes then apply crossover to mix good solutions together in hope of a better solution might arise from this mixture, finally apply mutation to prevent the chromosomes to be trapped in a local minimum value in one of its genes. Each individual represents one feature subspace. Its fitness represents the clustering result and how good it is regarding the feature space that the individual represents. The larger the fitness, denser the data in such feature subspace, the better the clustering results will be [11].

The proposed algorithm will be as following:

- **Input:**
  - P: Population size.
  - PM: Population means.
  - K: Number of clusters.
  - D: Data set in VRM.
  - MaxGen: Maximum number of generations.
  - TSSD: Targeted Sum of squared distances.
- **Output:**
  - Result: The fittest chromosome.
  - Mean: mean of the fittest chromosome
- **Steps:**

```
[0] START

[1] Sort the vectors either ascending or descending

[2] For i = 0 : P do the following:
    • Generate K random optimized centroids using (ICSO) and pass them to K-Means to get P clustering solutions.
        ○ For each P, Loop until convergence
            ▪ Save each result and its updated means in P and PM at the same index.

[3] Pass the P solutions gained from step 1 to the genetic algorithm as the initial population

[4] For each individual do the following:

    (a) Calculate the fitness of each individual with each mean in PM
            - Result = most fit individual
            - Means = means of Result

      (b) If MaxGen or TSSD is reached go to [5] END and deliver Result and Means as the optimal solution.

      (c) Apply selection

      (d) Apply Crossover

      (e) Apply mutation.

    (f) - Pass off spring to (a) –Loop
[5] END
```

## 4.4 Algorithm Explanation

Provide the algorithm with the following inputs:

1. Population size
2. Number of clusters
3. Data set
4. Maximum number of generations
5. Targeted sum of square distances.

Population size will indicate the maximum number of individuals (possible solutions) that will be generated initially for the genetic algorithm to use. Number of clusters as one of the basic need of the K-means algorithm to start, will be provided by the user to direct the K-means algorithm to split the input data set into a number of clusters. Data set which is the data set to be used as subject of the test.

Maximum number of generations is a criterion value which will indicate the maximum number of generations available for the genetic algorithm to go through, it will be used as a stopping criterion which if reached the whole composed algorithm will halt and deliver final results. Targeted sum of square distances is also a criterion value which indicates a desired sum of squared distances value that if reached by the composed algorithm will cause it to halt and deliver final results.

The first step is to sort all the vectors either ascending or descending, by doing that the scrambled vectors will be rearranged and we assume that vectors with close characteristics in terms of words weights will be grouped together. The second step is to generate *K* initial random centroids using the initial centroid selection optimization method shown in Fig. 3.

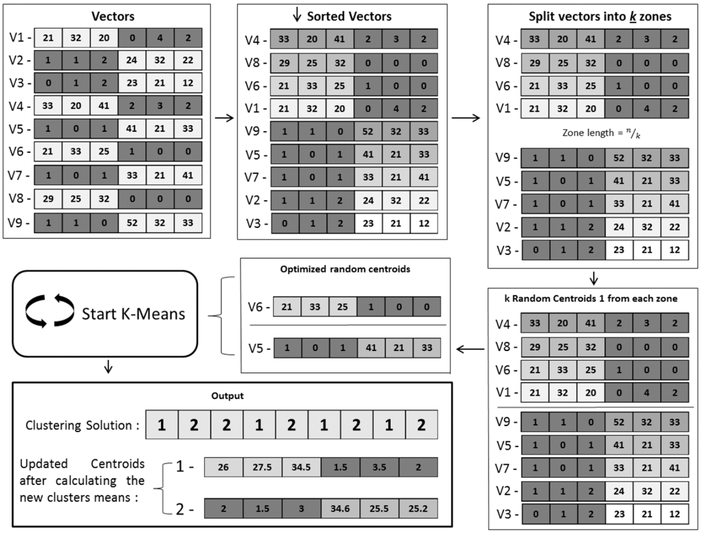


**Fig. 3.** Optimizing the selected initial centroids using dump vectors.

We assume that by using the initial centroid optimization method to generate k initial centroids for each chromosome of P, and letting k-means produce P clustering solutions and update the centroids until convergence as in Fig. 3, all clustering solutions acquired in this step as in Fig. 4, will already be a good clustering solutions that might face the problem of being not the optimal solution, thus, passing them to genetic algorithm for optimization.

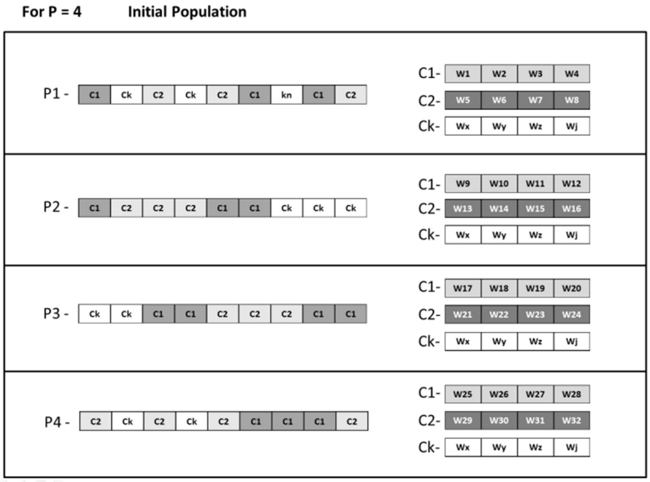


**Fig. 4.** Initial population (P1, P2, P3, and P4) attached with their centroids.

The third step, for each *P and PM* concatenate the updated centroids with their solution at *P* to form the final structure of the chromosome as shown in Fig. 5.

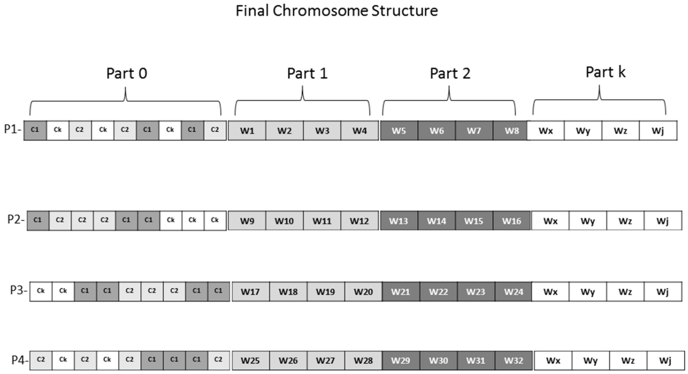

Final Chromosome Structure
Part 0
Part 1
Part 2
Part k
P1- C1 Ck C2 Ck C2 C1 Ck C1 C2 W1 W2 W3 W4 W5 W6 W7 W8 Wx Wy Wz Wj
P2- C1 C2 C2 C2 C1 C1 Ck Ck Ck W9 W10 W11 W12 W13 W14 W15 W16 Wx Wy Wz Wj
P3- Ck Ck C1 C1 C2 C2 C2 C1 C1 W17 W18 W19 W20 W21 W22 W23 W24 Wx Wy Wz Wj
P4- C2 Ck C2 Ck C2 C1 C1 C1 C2 W25 W26 W27 W28 W29 W30 W31 W32 Wx Wy Wz Wj

**Fig. 5.** Final chromosome structure.

Now that we have *P* number of clustering solutions concatenated with their updated centroids acquired from the previous step, the third step is to pass these chromosomes to the genetic algorithm to start operating on them to attempt to search for the most optimum clustering solution. Fourth, the genetic algorithm will calculate the fitness of all chromosomes regarding their centroids by calculating the sum of square distances between cluster elements and their centroid as in Eq. (3):

$$f(C1, C2, \ldots, Cn) = \sum\nolimits_{i=1}^{K} \sum\nolimits_{x_j \in C_i} ||x_j - z_i|| \tag{3}$$

Then keep the fittest chromosome and its means, and then apply genetic operators only on *parts 1 to k as in* Fig. 5 of each chromosome. To maintain chromosome integrity during the crossover operation we must crossover corresponding parts of each chromosome for example "part 1 from chromosome 1 with part 1 from chromosome 2", because each corresponding parts are generated from the same zone in the initial centroid selection optimization phase as shown in Fig. 3.

That's why we assume that part 1 from chromosome X is at the same context with part 1 from chromosome Y. Then pass the offspring to the fitness function and loop until maximum number of generations or a targeted SSD is reached. Finally we will acquire a clustering solution to a problem to calculate the average accuracy and to compare it to previous results.

## 5 Experimental Results Evaluation

The proposed algorithm and techniques will be tested on a data set combined of 1000 transcribed Arabic news broadcast videos, 18% of the transcripts where categorized into 4 sets of news categories (Politics, Weather, Business, Sport), then a collection of text files preprocessing procedures were made on them as following:

Tokenization, word grouping, words suspension, all these steps are done on the data set to prepare it to be presented in Vector Representation Model to get the weighted matrix of all documents regarding the constructed vector, then start our implementation (Fig. 6).

Now that we got the vector designed and words from the same categories grouped together in regions within that vector as mentioned before, the remaining 82% of the data set will be represented in VRM using the same vector.

Four different algorithms of K-means were used [14], (Lloyd, Forgy's, McQueen, Hartigan-Wong), advantages and disadvantages of each is listed in Table 2.

All files of the data set where randomly shuffled and given a standard name from D1 to Dn where n is the total number of documents and the following test cases were applied. First test case, a pure K-Means was applied on the data set which we know in advance that it has Business, Politics, Sport, Weather, 250 file in each category. The Second test case was applying K-Means with centroid optimization on the same data set, third was applying K-Means with genetic algorithm optimization, and finally applying K-Means with initial centroid selection optimization and genetic algorithm optimization, all previous scenarios are repeated 4 times, one for each version of k-means and the following results were acquired:

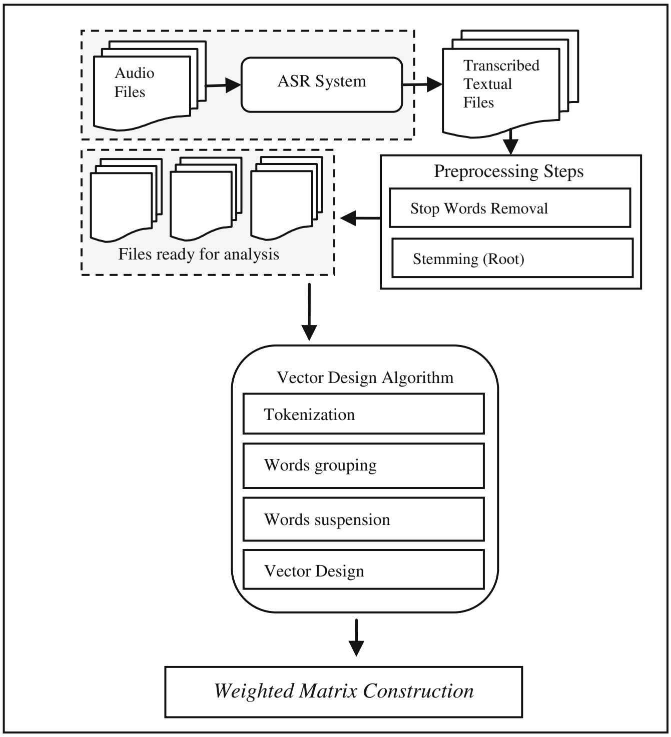


**Fig. 6.** The process of representing all of the data set into VRM.

**Table 2.** K-means algorithms

| Algorithm | Advantages | Disadvantages |
|---|---|---|
| Lloyd | - For large data sets<br>- Discrete data distribution<br>- Optimize total SSD | - Slower convergence<br>- Possible to create empty clusters |
| Forgy's | - For large data sets<br>- Continuous data distribution<br>- Optimize total SSD | - Slower convergence<br>- Possible to create empty clusters |
| McQueen | - Fast initial convergence<br>- Optimize total SSD | - Need to store the two nearest-cluster computations for each case<br>- Sensitive to the order the algorithm is applied to the cases |
| Hartigan-Wong | - Fast initial convergence<br>- Optimize within-cluster SSD | - Need to store the two nearest-cluster computations for each case<br>- Sensitive to the order the algorithm is applied to the cases |

## 5.1 Applying K-Means Only

**Table 3.** Results for case 5.1

| # | K-means algorithm | Avg accuracy | Avg iterations | Avg SSD |
|---|---|---|---|---|
| 1 | Hartigan-Wong | 83.3% | 2.67 | 10,714,603.925 |
| 2 | Lloyd | 76% | 3.6 | 9,155,058.925 |
| 3 | Forgy | 80.67% | 4 | 9,210,684.500 |
| 4 | McQueen | 83.3% | 3.3 | 4,857,707.525 |
| – | Average SSD | 3,393,8054.875 | | |
| – | Average Accuracy | 80.81% | | |
| – | Average Iterations | 3.39 | | |

## 5.2 Applying K-Means with GA

**Table 4.** Results for case 5.2

| # | K-means algorithm | Avg accuracy | Avg MaxGen | Avg iterations | Avg SSD |
|---|---|---|---|---|---|
| 1 | Hartigan-Wong | 84.7% | 17 | 2.67 | 10,524,203.67 |
| 2 | Lloyd | 78% | 24 | 3.6 | 9,155,058.49 |
| 3 | Forgy | 81.41% | 28 | 4 | 9,210,622.500 |
| 4 | McQueen | 84.3% | 18 | 3.3 | 4,857,693.525 |
| – | Average SSD | 3,374,7617.912 | | | |
| – | Average Accuracy | 82.10% | | | |
| – | Average MaxGen | 21.75 | | | |
| – | Average Iterations | 3.39 | | | |

## 5.3 Applying K-Means with Initial Centroid Selection Optimization

**Table 5.** Results for case 5.3

| # | K-means algorithm | Avg accuracy | Avg iterations | Avg SSD |
|---|---|---|---|---|
| 1 | Hartigan-Wong | 100% | 1.67 | 2,115,090.21 |
| 2 | Lloyd | 86.3% | 2.6 | 2,238,123.64 |
| 3 | Forgy | 84.67% | 2.6 | 2,238,123.64 |
| 4 | McQueen | 86.3% | 1.67 | 2,238,123.64 |
| – | Average SSD | 2,207,365.28 | | |
| – | Average Accuracy | 89.31% | | |
| – | Average Iterations | 2.135 | | |

### 5.4 Applying K-Means with Initial Centroid Selection Optimization and GA

**Table 6.** Results for case 5.4

| # | K-means algorithm | Avg accuracy | Avg MaxGen | Avg iterations | Avg SSD |
|---|---|---|---|---|---|
| 1 | Hartigan-Wong | 100% | 13 | 1.67 | 2,115,046.21 |
| 2 | Lloyd | 86.3% | 21 | 2.6 | 22,380,75.24 |
| 3 | Forgy | 84.67% | 25 | 2.6 | 2,238,075.24 |
| 4 | McQueen | 86.3% | 14 | 1.67 | 2,238,075.24 |
| – | Average SSD | 2,207,317.98 | | | |
| – | Average Accuracy | 89.31% | | | |
| – | Average MaxGen | 18.25 | | | |
| – | Average Iterations | 2.135 | | | |

### 5.5 Final Results

Results from Tables 3, 4, 5 and 6 are acquired using “speechnotes” [13] transcriber with WER = 10.2% for 40,435 reference words on our data set that consists of 1000 transcribed audio files using “speechnotes”.

**Table 7.** Final results

| Test case# | Avg SSD | Avg Acc (%) | Avg MaxGen | Avg Iter |
|---|---|---|---|---|
| 1 | 33,938,054.875 | 80.81 | – | 3.39 |
| 2 | 33,747,617.912 | 82.10 | 21.75 | 3.39 |
| 3 | 2,207,365.28 | 88.31 | – | 2.135 |
| 4 | 2,207,317.98 | 88.31 | 18.25 | 2.135 |

### 5.6 The Average Sums of Squared Distances as in Fig. 7

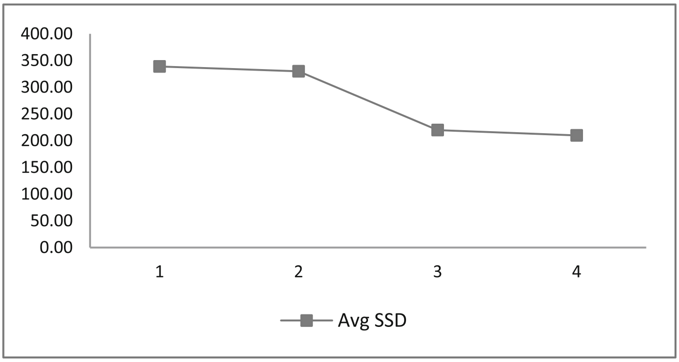


**Fig. 7.** Average sum of squared distances for all test cases

### 5.7 The Average Accuracy as in Fig. 8

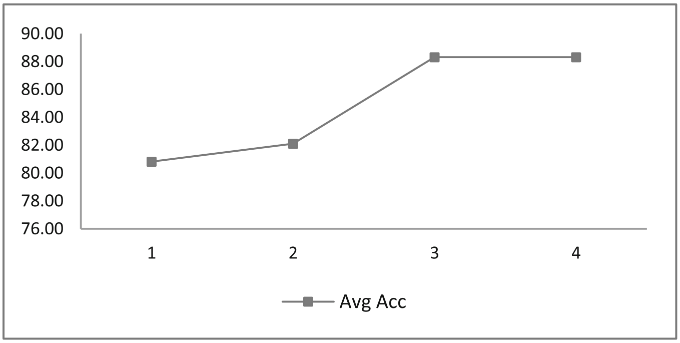


**Fig. 8.** Average accuracy

### 5.8 The Average Maximum Number of Generations as in Fig. 9

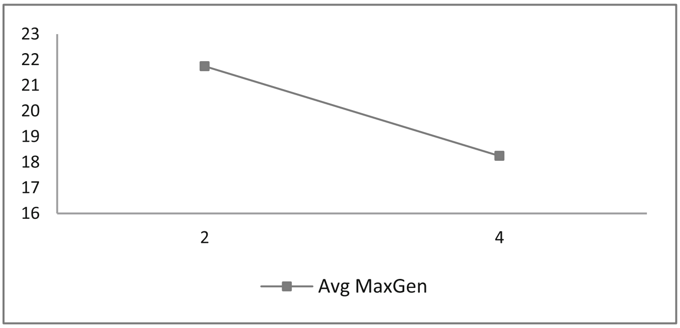


**Fig. 9.** Average maximum number of generations

### 5.9 The Average Number of Iterations as in Fig. 10

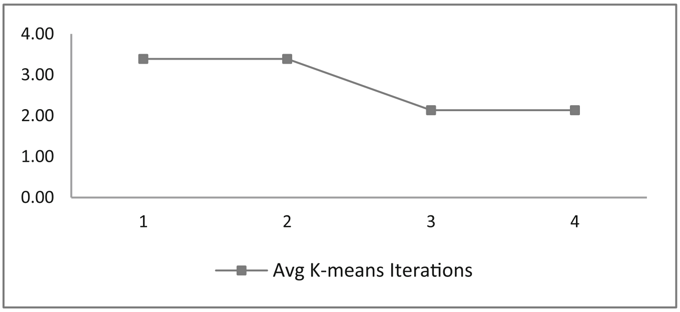


**Fig. 10.** Average number of K-means iterations

### 5.10 Previous Results [12]

Previous results acquired using Dragon Dictation Recognition System with WER = 20.6 for 30,040 reference words on previous data set consisting of 1000 transcribed audio files using Dragon Dictation Recognition System (Table 8).

**Table 8.**

| Clustering approach | Average accuracy |
|---|---|
| | Chi-square (%) |
| K-means | 79.05 |
| Spectral | 87.21 |

### 5.11 Results on Previous Data Set

Our technique was applied on previous data set used in [12] to compare results (Table 9).

**Table 9.**

| Clustering approach | Average accuracy |
|---|---|
| | Chi-square (%) |
| K-means + GA + ICSO | 87.91 |

## 6 Conclusion

Four test cases were implemented and the results were acquired in Tables 3, 4, 5 and 6, and then compared against each other in Table 7. Comparison shows that clustering using genetic algorithm has slightly improved the average accuracy by 1.29% and the average SSD by 190,436.963 with no change in the average iterations, while the dramatic change appeared when the initial centroid selection optimization technique was applied, which improved the average accuracy by 7.5% and the SSD by 31,730,689.595 and the average iterations by 1.25 iteration. Applying genetic algorithm after the ICSO technique has slightly improved the SSD by 47.3, but neither improved the average accuracy nor the average iterations.

We conclude that the improving impact of genetic algorithm on k-means is not as dramatic as initial centroid selection optimization, applying genetic algorithm optimization with k-means alone will result in a slight improvement in terms of average accuracy and the sum of square distances, while applying initial centroid selection optimization alone with k-means will result in a significant improvement in terms of average accuracy and average iterations for k-means to converge, finally applying genetic algorithm on the results of k-means and initial centroid selection optimization will result in a slight improvement in terms of sum of square distances. Finally running k-means with ICSO and GA on the dataset of [12] has resulted –as expected- in a significant improvement in terms of accuracy by 8.86%, while slightly exceeded the spectral algorithm implementation by 0.7%.

## 7 Future Work

In this research we found out that ICSO had a dramatic impact in terms of accuracy and SSD for the clustering process while Genetic algorithm didn't had as much impact as ICSO. Future research will be conducted in order to find better ways to utilize GA to get the most efficient use for the algorithm to optimize k-means results.

## 8. References

1. Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-based classification, search, and retrieval of audio. IEEE Multimed. **3**, 27–36 (1996). doi:10.1109/93.556537
2. Li, D., Sethi, I.K., Dimitrova, N., Mcgee, T.: Classification of general audio data for content-based retrieval. Pattern Recognit. Lett. **22**, 533–544 (2001). doi:10.1016/s0167-8655(00)00119-7
3. Coden, A., Brown, E.: Speech transcript analysis for automatic search. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences. doi:10.1109/hicss.2001.926473
4. Ibrahimov, O.V., Sethi, I.K., Dimitrova, N.: Data mining and knowledge discovery: theory. Tools Technol. IV (2002). doi:10.1117/12.460239
5. A comparison of document clustering algorithms. In: Proceedings of the 5th International Workshop on Pattern Recognition in Information Systems (2005). doi:10.5220/0002557501860191
6. Hayes-Roth, F.: Review of "Adaptation in Natural and Artificial Systems by John H. Holland", The U. of Michigan Press, 1975. ACM SIGART Bull. **53**, 15 (1975). doi:10.1145/1216504.1216510
7. Genetic algorithms in search, optimization, and machine learning. Choice Rev. Online (1989). doi:10.5860/choice.27-0936
8. Nazeer, K.A.A., Sebastian, M.P., Kumar, S.D.M.: A heuristic k-means algorithm with better accuracy and efficiency for clustering health informatics data. J. Med. Imaging Health Inform. **1**, 66–71 (2011). doi:10.1166/jmihi.2011.1010
9. Banerjee, A., Louis, S.J.: A recursive clustering methodology using a genetic algorithm. In: 2007 IEEE Congress on Evolutionary Computation (2007). doi:10.1109/cec.2007.4424740
10. Sun, H.-J., Xiong, L.-H.: Genetic algorithm-based high-dimensional data clustering technique. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery (2009). doi:10.1109/fskd.2009.215
11. Jian-Xiang, W., Huai, L., Yue-Hong, S., Xin-Ning, S.: Application of genetic algorithm in document clustering. In: 2009 International Conference on Information Technology and Computer Science (2009). doi:10.1109/itcs.2009.269
12. Jafar, A.A., Fakhr, M.W., Farouk, M.H.: Clustering-based topic identification of transcribed Arabic broadcast news. In: New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering. Lecture Notes in Electrical Engineering, pp. 253–260 (2014). doi:10.1007/978-3-319-06764-3_32
13. Speech to Text Online Notepad. Free. In: Speechnotes. https://speechnotes.co/#app. Accessed 12 May 2017
14. Morissette, L., Chartier, S.: The k-means clustering technique: general considerations and implementation in mathematica. Tutor. Quant. Methods Psychol. **9**, 15–24 (2013). doi:10.20982/tqmp.09.1.p015

15. Xu, R., Wunschii, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**, 645–678 (2005). doi:10.1109/tnn.2005.845141
16. Survey report on K-means clustering algorithm. Int. J. Mod. Trends Eng. Res. **4**, 218–221 (2017). doi:10.21884/ijmter.2017.4143.lgjzd
17. Na, S., Xumin, L., Yong, G.: Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (2010). doi:10.1109/iitsi.2010.74
18. Agarwal, S.: Data mining: data mining concepts and techniques. In: 2013 International Conference on Machine Intelligence and Research Advancement (2013). doi:10.1109/icmira.2013.45
19. Hamerly, G., Drake, J.: Accelerating Lloyd's algorithm for k-means clustering. Partitional Clust. Algorithms (2014). doi:10.1007/978-3-319-09259-1_2
20. Lei, X.-F.: An efficient clustering algorithm based on local optimality of K-means. J. Softw. **19**, 1683–1692 (2008). doi:10.3724/sp.j.1001.2008.01683
21. Tiwari, A.K., Sharma, L.K., Krishna, G.R.: Entropy weighting genetic k-means algorithm for subspace clustering. Int. J. Comput. Appl. **7**, 27–30 (2010). doi:10.5120/1263-1628
22. Zheng, D., Wang, Q.-P.: Selection algorithm for K-means initial clustering center. J. Comput. Appl. **32**, 2186–2188 (2013). doi:10.3724/sp.j.1087.2012.02186
23. Wu, J.: Cluster analysis and K-means clustering: an introduction. In: Advances in K-Means Clustering. Springer Theses, pp. 1–16 (2012). doi:10.1007/978-3-642-29807-3_1
24. An Introduction to Classification and Clustering. Cluster Analysis Wiley Series in Probability and Statistics, pp. 1–13 (2011). doi:10.1002/9780470977811.ch1
25. Wu, J.: The Uniform Effect of K-means Clustering. In: Advances in K-Means Clustering. Springer Theses, pp. 17–35 (2012). doi:10.1007/978-3-642-29807-3_2
26. Shrivastava, P., Kavita, P., Singh, S., Shukla, M.: Comparative analysis in between the k-means algorithm, k-means using with Gaussian mixture model and fuzzy c means algorithm. Commun. Comput. Syst. (2016). doi:10.1201/9781315364094-186

# An Imperialist Competitive Algorithm to Solve the Manufacturing Cell Design Problem

Ricardo Soto[1], Broderick Crawford[1], Rodrigo Olivares[1,2(✉)], Héctor Ortega[1], and Boris Almonacid[1]

[1] Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
{ricardo.soto,broderick.crawford}@pucv.cl,
{hector.ortega.m,boris.almonacid.g}@mail.pucv.cl
[2] Universidad de Valparaíso, Valparaíso, Chile
rodrigo.olivares@uv.cl

**Abstract.** The manufacturing cell design problem is part of the cellular manufacturing system and it has been widely studied as an optimization problem. It consists of grouping machines in parts into manufacturing cells in order to minimize the inter-cell movements. In recent years, different approximate methods have been used to solve this problem. In this paper, we propose a new approximate method inspired on the phenomenon of the colonial age, called imperialist competitive algorithm. In the colonial age, the most powerful countries competed to conquer colonies for increasing their power, where the country with highest power was considered the imperialist one. We performed several experiments on a set of 90 instances, where the proposed approach is able to produce optimal values for the whole set of tested instances.

**Keywords:** Manufacturing cell design problem · Imperialist competitive algorithm · Metaheuristics

## 1 Introduction

The Manufacturing Cell Design Problem (MCDP) is part of Group Technology and it consists of grouping machines in parts or products into families, which are processed in a miniature factory, called cell [2]. In general terms, it is design to reduce the production costs by grouping machine and part families. In this context, the main aim is to build manufacturing process in a way that minimizes movements of parts from one cell to another finding machine-part's associations with the least amount of part movements between cells.

Many researches can be found for solving MCDP, such as: a production flow analysis for planning group technology [3], the part families problem in flexible manufacturing systems [7], an evaluation of search algorithms and clustering efficiency measures for machine-part matrix clustering [8] and a linear formulation of the machine-part cell formation problem [2], which it is used as the mathematical model in this research. Due to MCDP is an NP-hard problem, several researchers consider applying approximation methods as metaheuristics: a migrating birds

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_9

optimization algorithm for machine-part cell formation problems [9,10], solving the MCDP via invasive weed optimization [12], solving MCDP by using a dolphin echolocation algorithm [11] and solving MCDP using a shuffled frog leaping algorithm [13].

In this work, we propose the Imperialist Competitive Algorithm (ICA) to solve the MCDP. This metaheuristic is a population-based method inspired on phenomenon of the colonial age, where the most powerful countries competed to conquer colonies for increasing their power [1,6]. This approach has three stages: initialization, competition and elimination. Then, in the initialization stage, countries with the best cost are determined as the imperialist countries. In the competition stage, every imperialist tries to conquer more colonies. This chance is determined by the power of each imperialist that is related to its cost. In the final stage, the weakest empires collapse and the strongest empires win colonies [5]. The algorithm reach a solution once an unique empire governs. We present interesting results on set of 90 instances taken from Boctor's experiments, where the incorporation of the above-mentioned elements clearly improves the results.

This paper is organized as follows: In Sect. 2, we describe the manufacturing cell design problem. Section 3 presents the algorithm used to solve the problem. Section 4 provides the experimental results. Finally, conclusions of the results and future works are detailed in Sect. 5.

## 2 Problem Description

The Manufacturing Cell Design Problem (MCDP) proposes to divide a plant of industrial production in a number of cells. Each cell contains machines with similar types of processes or families of parts. The goal is to identify an organization of cells so that the transport of different parts between cells is minimized. In this paper, we define the manufacturing cell design problem by using an array-based grouping approach. The main idea is to represent the processing requirements of parts on machines through an incidence matrix named machine-part $(M \times P)$. This matrix holds binary domains and is denoted as $A = a_{ij}$, where:

$$a_{ij} = \begin{cases} 1 \text{ if part } j \text{ visits machine } i \text{ for the processing;} \\ 0 \text{ otherwise.} \end{cases}$$

For the resolution of MCDP three matrices are used. The first matrix is the incidence matrix $M \times P$ (see Fig. 1), which indicates the parts that are processed by machines. The second and third matrix are solutions $M \times C$ and $P \times C$ (see Fig. 2), which together with the incidence matrix form a new distribution of machines and parts in cells (see Fig. 3). The parameters used are: 5 machines, 7 parts, $M_{max} = 3$ for 2 cells. Finally, we can determined that:

- The new distribution is described in Fig. 3.
- The cell 1 is composed of the machines 2, 3, 5 and of parts 1, 3, 7.
- The cell 2 is composed of the machines 1, 2 and the parts 2, 4, 5, 6.

| | Part 1 | Part 2 | Part 3 | Part 4 | Part 5 | Part 6 | Part 7 |
|---|---|---|---|---|---|---|---|
| **Machine 1** | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| **Machine 2** | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **Machine 3** | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **Machine 4** | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| **Machine 5** | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

**Fig. 1.** Initial incidence matrix $M \times P$

| | Cell 1 | Cell 2 |
|---|---|---|
| **Machine 1** | 0 | 1 |
| **Machine 2** | 1 | 0 |
| **Machine 3** | 1 | 0 |
| **Machine 4** | 0 | 1 |
| **Machine 5** | 1 | 0 |

| | Cell 1 | Cell 2 |
|---|---|---|
| **Part 1** | 1 | 0 |
| **Part 2** | 0 | 1 |
| **Part 3** | 1 | 0 |
| **Part 4** | 0 | 1 |
| **Part 5** | 0 | 1 |
| **Part 6** | 0 | 1 |
| **Part 7** | 1 | 0 |

**Fig. 2.** Matrix $M \times C$ and matrix $P \times C$

| | | Cell 1 | | | Cell 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Part 1** | **Part 3** | **Part 7** | **Part 2** | **Part 4** | **Part 5** | **Part 6** |
| **Cell 1** | **Machine 2** | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| | **Machine 3** | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| | **Machine 5** | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **Cell 2** | **Machine 1** | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| | **Machine 4** | 0 | 0 | 0 | 1 | 1 | 0 | 1 |

**Fig. 3.** Rearranged incidence matrix $M \times P$

- The optimum value obtained is 1 (see exception in Fig. 3: machine 1 with part 1 without assigned cell).

A mathematical formulation of machine-part cell formation problem is given by Boctor [2]. The optimization model is stated as follows, Let $M$ the number of machines, $P$ the number of parts, $C$ the number of cells, $i$ the index of machines ($i = \{1, \ldots, M\}$), $j$ the index of parts ($j = \{1, \ldots, P\}$), $k$ the index of cells ($k = \{1, \ldots, C\}$), $M_{max}$ the maximum number of machines per cell, $A = a_{ij}$ the $M \times P$ machine-part incidence matrix, $y_{ik}$ the $M \times C$ machine-cell matrix ($y_{ik} = 1$, if machine $i \in$ cell $k$; 0 otherwise), and $z_{jk}$ the $P \times C$ part-cell matrix ($z_{jk} = 1$, if part $j \in$ cell $k$; 0 otherwise).

The problem is represented by the following mathematical model:

$$\texttt{minimize} \sum_{k=1}^{C} \sum_{i=1}^{M} \sum_{j=1}^{P} a_{ij} z_{jk} (1 - y_{ik}) \tag{1}$$

subject to:

$$\sum_{k=1}^{C} y_{ik} = 1 \tag{2}$$

$$\sum_{k=1}^{C} z_{jk} = 1 \tag{3}$$

$$\sum_{i=1}^{M} y_{ik} \leq M_{max} \tag{4}$$

## 3 Imperialist Competitive Algorithm

Similar to other evolutionary algorithms, the Imperialist Competitive Algorithm (ICA) starts with an initial random population where each individual in the population represents a country. The first step is to initialize the empires, to which should be selected with lower cost countries as imperialist countries, the other countries will be considered colonies of the imperialists. The allocation of the colonies for the imperialists is based on the power of the latter. After this initialization, the next step is to move the colonies to the imperialist. An imperialist and its colonies form an empire, which competes with other empires in the world (search space). The survival of an empire relies on the power of itself to start taking colonies of the competitors, while the power of the great empires increases and the less powerful empires collapse. The extreme case of the imperialist competition is when only one imperialist controls all colonies around the world [6].

### 3.1 Initialization

The vector representing to one country is defined as follows: $c = [p_1, p_2, \ldots, p_n]$, where $n$ is the number of variables and $p_i$ represents the value of the $i$th variable of the country. For purposes of this work, a country represents a machine-part matrix making the country a variable represented in a box of said matrix. The cost of a country is calculated as a function of the variables: $cost = f(c) = f(p_1, p_2, \ldots, p_n)$.

At this stage, should be generated an initial population of $N_{pop}$ size. Next, we have to select some of the best countries, having the lowest cost function values, with the size of $N_{imp}$ from $N_{pop}$, and set them to be imperialists. The rest of countries are set to be colonies $N_{col} = N_{pop} - N_{imp}$.

To form empires, the colonies should be divided among the imperialist according to the power of the imperialists. The normalized cost of each imperialist is determined as follows: $C_n = max\{c_i\} - c_n$, where $c_n$ is the $n$th imperialist's cost, and $C_n$ is the normalized cost of $n$th imperialist. An imperialist with larger cost (i.e. a weaker imperialist country) has smaller normalized cost. Therefore, the power of each imperialist is calculated based on the normalized cost:

$$p_n = \left| \frac{c_n}{\sum_{i=1}^{N_{imp}} c_i} \right|$$

where $p_n$ is the power of $n$th imperialist. The normalized power of $n$th imperialist is the number of colonies that are possessed by that imperialist, calculated by: $NC_n = round\{p_n \cdot (N_{col})\}$, where $NC_n$ is the number of initial colonies possessed by the $n$th imperialist; $N_{col}$ is the total number of colonies in the initial population, and round is a function that gives the nearest integer of a fractional number.

### 3.2 Assimilation

Assimilation is the movement of the colony towards its imperialist. The $x$ movement carried out by a colony is generated by a random distribution domain [0, $\beta * d$]: $x \sim U(0, \beta * d)$, where the value of $\beta$ is between 1 and 2. Set $\beta > 1$ generates the colony moves toward the imperialist. The movement of the colony has a $\theta$ deviation parameter that takes values of uniform distribution. Thus: $\theta \sim U(-\varphi, \varphi)$, where $\varphi$ is an arbitrary parameter, a large value of this parameter facilitates global exploration and a small value will produce a local search.

### 3.3 Revolution

According to imperial history, colonies of an empire are absorbed by imperialist in terms of social, cultural, economic, and political characteristics; however, there might be some colonies that resist to be absorbed by imperialists. In fact, those colonies perform some sort of reformations in their characteristics. In ICA, this operation is called revolution. Revolution brings sudden random changes in the position of some the colonies in the search space and it increases exploration preventing the early convergence of countries to local optima.

### 3.4 Exchange Imperialist-Colonial Position

Once assimilation and revolution operations are performed on colonies of an empire, the cost functions of new position of colonies are then compared with cost function of position of imperialist. If we find any colony whose cost function is less than cost function of imperialist, then we swap imperialist with that colony.

### 3.5 Total Power of an Empire

The total power of an empire is based on the power of the imperialist and a fraction of the power of their colonies:

$$TC_n = cost(imperialist) + \zeta * mean\{cost(colonies\ of\ empire_n)\}$$

where $TC_n$ is the total cost of the $n$th empire, and $\zeta$ is a positive number between 0 and 1, usually close to 0. A small value of $\zeta$ emphasizes the influence of imperialist power in the total power the empire, while a small value of this parameter indicates the influence of the colonies to determine the total power of the empire.

During competition among the imperialist countries, weaker empires will be collapsed gradually. This means that the weaker empires will lose their colonies over time, while stronger empires will possess the colonies of weaker empires, thereby increasing their power. Therefore, one or some of the weakest colonies belonging to the weakest empire will be given to a different empire based on competition that occurs among all empires. Stronger empires have a greater chance to possess the weakest colony

The weakest colony in weakest empire is subject to competition among empires 2 to $n$. In order to model the competition process among the empires, we need to compute the normalized total cost of empire by $NTC_n = max\{TC_i\} - TC_n$, where $TC_n$ is the total cost of the $n$th empire and $NTC_n$ is the normalized total cost of corresponding $n$th empire. Then, the probability of possessing a colony is determined by:

$$p_n = \left| \frac{NTC_n}{\sum_{i=1}^{N_{imp}} NTC_i} \right|$$

where $\sum_{i=1}^{N_{imp}} p_i = 1$. There is a need for a mechanism to distribute the weakest colonies among the empires based on their possession probabilities. The ICA introduces a new distribution mechanism, which requires a Probability Density Function (PDF), addressed below:

Vector $P$ with the size of $1 * N_{imp}$ contains the possession probability of a colony by empires as follows: $P = [p_1, p_2, \ldots, p_{N_{imp}}]$.

Then, vector $R$ with the same size of $P$ is formed in which its elements are generated using uniform distribution within interval of [0,1] as follows: $R = [r_1, r_2, \ldots, r_{N_{imp}}]$, where $r_i \sim U(0,1)$.

Finally, the vector $D$ is defined as: $D = P - R = [d_1, d_2, \ldots, d_{N_{imp}}] = [p_1 - r_1, p_2 - r_2, \ldots, p_{N_{imp}} - r_{N_{imp}}]$. Once vector $D$ is calculated, the weakest colony is assigned to the empire with the larger index.

### 3.6 Collapsing the Weaker Empires

Weaker empires lose their colonies gradually to stronger empires, which in turn grow more powerful and cause the weaker empires to collapse over time.

### 3.7 Convergence

Similar to other evolutionary algorithms, ICA continues until stopping criteria are met, such as predefined running time or a certain number of iterations. The ideal stopping criterion is when all empires have collapsed and only one (grand empire) remains.

Finally, the procedure describing ICA can be summarized as the pseudo-code shown in Algorithm 1.

**Algorithm 1.** Imperialist Competitive Algorithm

```
initialize parameters
generate population
{initialize the empire}
for ($i = 1 : N_{pop}$) do
    compute the cost $c_i$
    sort the cost $c_i$ in descending order
    select $N_{imp}$ out of $N_{pop}$
    normalize the cost of each imperialist $C_n$
    compute the normalized power of each imperialist $P_n$
    assign $N_{col}$ remained countries to the imperialist
end for
{assimilation, revolution, imperialist competition processes}
while stopping condition is not reached do
    for ($j = 1 : N_{imp}$) do
        move the colony toward the relevant imperialist (assimilation)
        compute the costs of assimilated countries
        perform revolution on new colony
        if cost(new colony) $<$ cost(imperialist) then
            exchange the position of colony and imperialist
        end if
        pick the weakest colony from the weakest empire and assign it to the empire
        that has most likelihood to possess it
    end for
    elimination process
    if there is imperialist with no colonies then
        eliminate the imperialist
    end if
end while
```

## 4 Experimental Results

We have performed an experimental evaluation of the proposed approach on different instances taken from Boctor's experiments. Imperialist Competitive Algorithm was development in Java SE 1.7 and the experiments have been launched on a 3.30 GHz Intel Core i5 with 4 Gb RAM running Windows 7 Professional 32 bits.

The configuration for approaches is detailed as follows: *number of iterations* $T = 100000$; *number of population* $N_{pop} = 1000$; *number of imperialists* $N_{imp} = 50$; *assimilation deviation* $\theta = 0.7$; *assimilation address* $\beta = 0.5$; *influence coefficient of colonies* $\zeta = 0.1$; and *rate of revolution* $r = 0.3$.

Results are evaluated using the relative percentage deviation ($RPD$). $RPD$ value quantifies the deviation of the objective value $Z$ from $Z_{opt}$ that in our case

**Table 1.** Experiments using $C = 2$:

| Instance | Boctor problem | $M_{max}$ | Opt. value | ICA | | | SA | PSO |
|---|---|---|---|---|---|---|---|---|
| | | | | Opt | Avg. | RPD% | Opt. | Opt. |
| 1 | 1 | 8 | **11** | **11** | 12.81 | 0.00 | **11** | **11** |
| 2 | 1 | 9 | **11** | **11** | 11.42 | 0.00 | **11** | **11** |
| 3 | 1 | 10 | **11** | **11** | 11.27 | 0.00 | **11** | **11** |
| 4 | 1 | 11 | **11** | **11** | 11.65 | 0.00 | **11** | **11** |
| 5 | 1 | 12 | **11** | **11** | 12.95 | 0.00 | **11** | **11** |
| 6 | 2 | 8 | **7** | **7** | 7.82 | 0.00 | **7** | **7** |
| 7 | 2 | 9 | **6** | **6** | 7.3 | 0.00 | **6** | **6** |
| 8 | 2 | 10 | **4** | **4** | 5.43 | 0.00 | 10 | 5 |
| 9 | 2 | 11 | **3** | **3** | 3.86 | 0.00 | 4 | 4 |
| 10 | 2 | 12 | **3** | **3** | 3.73 | 0.00 | **3** | 4 |
| 11 | 3 | 8 | **4** | **4** | 5.22 | 0.00 | 5 | 5 |
| 12 | 3 | 9 | **4** | **4** | 5.29 | 0.00 | **4** | **4** |
| 13 | 3 | 10 | **4** | **4** | 5.19 | 0.00 | **4** | 5 |
| 14 | 3 | 11 | **3** | **3** | 3.95 | 0.00 | 4 | 4 |
| 15 | 3 | 12 | **1** | **1** | 2.62 | 0.00 | 4 | 3 |
| 16 | 4 | 8 | **14** | **14** | 15.1 | 0.00 | **14** | 15 |
| 17 | 4 | 9 | **13** | **13** | 13.37 | 0.00 | **13** | **13** |
| 18 | 4 | 10 | **13** | **13** | 13.47 | 0.00 | **13** | **13** |
| 19 | 4 | 11 | **13** | **13** | 13.68 | 0.00 | **13** | **13** |
| 20 | 4 | 12 | **13** | **13** | 13.65 | 0.00 | **13** | **13** |
| 21 | 5 | 8 | **9** | **9** | 9.92 | 0.00 | **9** | 10 |
| 22 | 5 | 9 | **6** | **6** | 7.1 | 0.00 | **6** | 8 |
| 23 | 5 | 10 | **6** | **6** | 6.98 | 0.00 | **6** | **6** |
| 24 | 5 | 11 | **5** | **5** | 5.9 | 0.00 | 7 | **5** |
| 25 | 5 | 12 | **4** | **4** | 5.06 | 0.00 | **4** | 5 |
| 26 | 6 | 8 | **5** | **5** | 6.69 | 0.00 | **5** | **5** |
| 27 | 6 | 9 | **3** | **3** | 3.77 | 0.00 | **3** | **3** |
| 28 | 6 | 10 | **3** | **3** | 4.08 | 0.00 | 5 | **3** |
| 29 | 6 | 11 | **3** | **3** | 4.03 | 0.00 | **3** | 4 |
| 30 | 6 | 12 | **2** | **2** | 2.74 | 0.00 | 3 | 4 |
| 31 | 7 | 8 | **7** | **7** | 7.81 | 0.00 | **7** | **7** |
| 32 | 7 | 9 | **4** | **4** | 6.02 | 0.00 | **4** | 5 |
| 33 | 7 | 10 | **4** | **4** | 5.26 | 0.00 | **4** | 5 |
| 34 | 7 | 11 | **4** | **4** | 5.32 | 0.00 | **4** | 5 |
| 35 | 7 | 12 | **4** | **4** | 5.24 | 0.00 | **4** | 5 |

(*Continued*)

**Table 1.** (*Continued*)

| Instance | Boctor problem | $M_{max}$ | Opt. value | ICA | | | SA | PSO |
|---|---|---|---|---|---|---|---|---|
| | | | | Opt | Avg. | RPD% | Opt. | Opt. |
| 36 | 8 | 8 | **13** | **13** | 13.7 | 0.00 | **13** | 14 |
| 37 | 8 | 9 | **10** | **10** | 11.66 | 0.00 | 20 | 11 |
| 38 | 8 | 10 | **8** | **8** | 9.19 | 0.00 | 15 | 10 |
| 39 | 8 | 11 | **5** | **5** | 6.22 | 0.00 | 11 | 6 |
| 40 | 8 | 12 | **5** | **5** | 7 | 0.00 | 7 | 6 |
| 41 | 9 | 8 | **8** | **8** | 9.9 | 0.00 | 13 | 9 |
| 42 | 9 | 9 | **8** | **8** | 9.85 | 0.00 | **8** | **8** |
| 43 | 9 | 10 | **8** | **8** | 9.64 | 0.00 | **8** | **8** |
| 44 | 9 | 11 | **5** | **5** | 6.77 | 0.00 | 8 | **5** |
| 45 | 9 | 12 | **5** | **5** | 6.91 | 0.00 | 8 | 8 |
| 46 | 10 | 8 | **8** | **8** | 8.95 | 0.00 | **8** | 9 |
| 47 | 10 | 9 | **5** | **5** | 6.31 | 0.00 | **5** | 8 |
| 48 | 10 | 10 | **5** | **5** | 6.29 | 0.00 | **5** | 7 |
| 49 | 10 | 11 | **5** | **5** | 5.84 | 0.00 | **5** | 7 |
| 50 | 10 | 12 | **5** | **5** | 6.41 | 0.00 | **5** | 6 |

is the best known value for each instance ($Z_{opt}$ in Table 1, and it is calculated as follows:

$$RPD = \left( \frac{Z - Z_{opt}}{Z_{opt}} \right) \times 100 \tag{5}$$

The minimum (Min), maximum (Max) and average (Avg) of the solutions obtained were achieved running 30 executions over each one of the test instances. To calculate $RPD$ value, we used $Z = Z_{min}$.

The other instances of Boctor are show in the Table 2. As the above table, the optimum achieved in the investigation of Boctor [2] are presented in the column as Boctor, while the results obtained by our research are under the title ICA. As shown in Table 2, ICA reached all the problems global optimums configured with $C = 3$ and $M_{max} = \{6, 7, 8, \text{ and }, 9\}$ respectively.

To compare the ICA performances, we use two-known algorithms: Simulated Annealing (SA) and Particle Swarm Optimization (PSO). Results of both approaches can be find in [2] and [4], respectively.

Table 1, shown the results obtained for 50 instances of the problems studied. The optimum achieved in the research of Boctor [2] are presented in the column as Boctor, while the results obtained by our research are under the title ICA. ICA reached all the problems global optimums configured with $C = 2$ and $M_{max} = \{8, 9, 10, 11 \text{ and } 12\}$ respectively.

**Table 2.** Experiments using $C = 3$

| Instance | Boctor problem | $M_{max}$ | Opt. value | ICA | | | SA | PSO |
|---|---|---|---|---|---|---|---|---|
| | | | | Opt | Avg. | RPD% | Opt. | Opt. |
| 51 | 1 | 6 | **27** | **27** | 29.44 | 0.00 | 28 | - |
| 52 | 1 | 7 | **18** | **18** | 20.77 | 0.00 | **18** | - |
| 53 | 1 | 8 | **11** | **11** | 13.22 | 0.00 | **11** | - |
| 54 | 1 | 9 | **11** | **11** | 12.23 | 0.00 | **11** | - |
| 55 | 2 | 6 | **7** | **7** | 9.08 | 0.00 | **7** | - |
| 56 | 2 | 7 | **6** | **6** | 7.42 | 0.00 | **6** | - |
| 57 | 2 | 8 | **6** | **6** | 7.01 | 0.00 | 7 | - |
| 58 | 2 | 9 | **6** | **6** | 7.33 | 0.00 | **6** | - |
| 59 | 3 | 6 | **9** | **9** | 10.08 | 0.00 | 12 | - |
| 60 | 3 | 7 | **4** | **4** | 6.67 | 0.00 | 8 | - |
| 61 | 3 | 8 | **4** | **4** | 5.51 | 0.00 | 8 | - |
| 62 | 3 | 9 | **4** | **4** | 4.84 | 0.00 | **4** | - |
| 63 | 4 | 6 | **27** | **27** | 28.03 | 0.00 | **27** | - |
| 64 | 4 | 7 | **18** | **18** | 20 | 0.00 | **18** | - |
| 65 | 4 | 8 | **14** | **14** | 15.71 | 0.00 | **14** | - |
| 66 | 4 | 9 | **13** | **13** | 14.42 | 0.00 | **13** | - |
| 67 | 5 | 6 | **11** | **11** | 12.29 | 0.00 | **11** | - |
| 68 | 5 | 7 | **8** | **8** | 9.55 | 0.00 | 9 | - |
| 69 | 5 | 8 | **8** | **8** | 9.53 | 0.00 | 9 | - |
| 70 | 5 | 9 | **6** | **6** | 7.82 | 0.00 | 8 | - |
| 71 | 6 | 6 | **6** | **6** | 6.97 | 0.00 | 8 | - |
| 72 | 6 | 7 | **4** | **4** | 5.72 | 0.00 | 5 | - |
| 73 | 6 | 8 | **4** | **4** | 5.39 | 0.00 | 5 | - |
| 74 | 6 | 9 | **3** | **3** | 4.64 | 0.00 | 4 | - |
| 75 | 7 | 6 | **11** | **11** | 13.21 | 0.00 | **11** | - |
| 76 | 7 | 7 | **5** | **5** | 6.37 | 0.00 | **5** | - |
| 77 | 7 | 8 | **5** | **5** | 7.1 | 0.00 | **5** | - |
| 78 | 7 | 9 | **4** | **4** | 6.35 | 0.00 | 5 | - |
| 79 | 8 | 6 | **14** | **14** | 15.11 | 0.00 | **14** | - |
| 80 | 8 | 7 | **11** | **11** | 12.71 | 0.00 | **11** | - |
| 81 | 8 | 8 | **11** | **11** | 13.23 | 0.00 | **11** | - |
| 82 | 8 | 9 | **10** | **10** | 11.69 | 0.00 | **10** | - |
| 83 | 9 | 6 | **12** | **12** | 14.39 | 0.00 | **12** | - |
| 84 | 9 | 7 | **12** | **12** | 13.42 | 0.00 | **12** | - |
| 85 | 9 | 8 | **8** | **8** | 10.73 | 0.00 | 13 | - |
| 86 | 9 | 9 | **8** | **8** | 9.68 | 0.00 | **8** | - |
| 87 | 10 | 6 | **10** | **10** | 13 | 0.00 | 12 | - |
| 88 | 10 | 7 | **8** | **8** | 9.32 | 0.00 | 14 | - |
| 89 | 10 | 8 | **8** | **8** | 9.14 | 0.00 | **8** | - |
| 90 | 10 | 9 | **5** | **5** | 7.45 | 0.00 | 8 | - |

## 5 Conclusions

In this paper, we present a version of the algorithm imperialist competitive algorithm to solve the problem design cell manufacturing. This was tested with different instances of well-known problems taken from Boctor's experiments.

Experiment results illustrated an excellent performance by the algorithm, obtaining global optimal in all instances, exposing the robustness of the approach.

As future work, we plan to experiment with additional modern metaheuristics and to provide a larger comparison of recent bio-inspired techniques to solve the MCDP. The integration of autonomous search to the presented approach would be another direction of research to follow as well, for instance to dynamically choice the best parameter setting during solving according to performance indicators, such as fitness, variability of solutions, search space reduce, among others.

**Acknowledgments.**
Broderick Crawford is supported by Grant CONICYT/FONDECYT/REGULAR/1171243. Ricardo Soto is supported by Grant CONICYT/FONDECYT/REGULAR/1160455. Rodrigo Olivares is supported by CONICYT/FONDEF/IDeA/ID16I10449 and Postgraduate Grant Pontificia Universidad Católica de Valparaíso (INF-PUCV 2015-2017). Finally, Boris Almonacid are supported by Postgraduate Grant Pontificia Universidad Católica de Valparaíso, Chile (INF-PUCV 2015).

## References

1. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. In: 2007 IEEE Congress on Evolutionary Computation. Institute of Electrical & Electronics Engineers (IEEE), September 2007
2. Boctor, F.F.: A jinear formulation of the machine-part cell formation problem. Int. J. Prod. Res. **29**(2), 343–356 (1991)
3. Burbidge, J.L.: Production flow analysis for planning group technology. J. Oper. Manag. **10**(1), 5–27 (1991)
4. Durán, O., Rodriguez, N., Consalter, L.A.: Collaborative particle swarm optimization with a data mining technique for manufacturing cell design. Expert Syst. Appl. **37**(2), 1563–1567 (2010)
5. Forouharfard, S., Zandieh, M.: An imperialist competitive algorithm to schedule of receiving and shipping trucks in cross-docking systems. Int. J. Adv. Manuf. Technol. **51**(9–12), 1179–1193 (2010)
6. Hosseini, S., Al Khaled, A.: A survey on the imperialist competitive algorithm metaheuristic: implementation in engineering domain and directions for future research. Appl. Soft Comput. J. **24**, 1078–1094 (2014)
7. Kusiak, A.: The part families problem in flexible manufacturing systems. Ann. Oper. Res. **3**(6), 277–300 (1985)
8. Shargal, M., Shekhar, S., Irani, S.: Evaluation of search algorithms and clustering efficiency measures for machine-part matrix clustering. IIE Trans. **27**(1), 43–59 (1995)

9. Soto, R., Crawford, B., Almonacid, B., Paredes, F.: A migrating birds optimization algorithm for machine-part cell formation problems. In: Mexican International Conference on Artificial Intelligence, pp. 270–281. Springer (2015)
10. Soto, R., Crawford, B., Almonacid, B., Paredes, F.: Efficient parallel sorting for migrating birds optimization when solving machine-part cell formation problems. Sci. Program. (2016)
11. Soto, R., Crawford, B., Carrasco, C., Almonacid, B., Reyes, V., Araya, I., Misra, S., Olguín, E.: Solving manufacturing cell design problems by using a dolphin echolocation algorithm. In: International Conference on Computational Science and Its Applications, pp. 77–86. Springer (2016)
12. Soto, R., Crawford, B., Castillo, C., Paredes, F.: Solving the manufacturing cell design problem via invasive weed optimization. In: Artificial Intelligence Perspectives in Intelligent Systems, pp. 115–126. Springer (2016)
13. Soto, R., Crawford, B., Vega, E., Johnson, F., Paredes, F.: Solving manufacturing cell design problems using a shuffled frog leaping algorithm. In: The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), 28–30 November, 2015, Beni Suef, Egypt, pp. 253–261. Springer (2016)

# Optical Character Recognition System for Czech Language Using Hierarchical Deep Learning Networks

Arindam Chaudhuri[1](✉) and Soumya K. Ghosh[2]

[1] Samsung R & D Institute Delhi, Noida 201304, India
arindam_chau@yahoo.co.in
[2] Department of Computer Science Engineering,
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
skg@iitkgp.ac.in

**Abstract.** Optical character recognition (OCR) systems play vital role in pattern recognition research. With rapid growth of OCRs for different languages developing OCR for Czech language is looked upon as positive aspect for people speaking Czech language. In this paper, we develop OCR system for Czech language using hierarchical fuzzy convolutional neural networks (HFCNN). We present end-to-end framework that includes pre-processing activities, segments text image, classifies characters and performs recognition. The feature extraction is performed through fuzzy Hough transform. The feature based classification is performed through HFCNN. A comprehensive assessment of proposed method is performed through publicly available Czech language dataset. OCR recognition accuracy is a major concern. There is always an inherent degree of vagueness and impreciseness present in reallife data. Due to this recognition system is treated here through fuzzy sets encompassing indeterminate uncertainty. The simulation studies reveal that deep learning based OCR for Czech language performs consistently better than traditional models. The experimental results demonstrate efficiency of proposed approach.

**Keywords:** OCR · Czech language · Deep learning · CNN · FCNN · HFCNN

## 1 Introduction

Optical character recognition (OCR) is challenging research area in pattern recognition [1, 2] because of its immense application potential. OCR was initially studied in early 1930s [2] by Gustav Tauschek. It is mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. It is widely used form of data entry from printed paper data records. The image captured by digital camera is converted into suitable form required by machine. It is common method of digitizing printed texts which can be electronically edited and stored more compactly such that it can be used in machine learning and intelligence processes. Availability of huge data in several languages has created an opportunity to analyse OCR systems analytically. There has not been any significant development towards an end-to-end OCR system for Czech language. The availability of huge corpus of scanned documents online requires

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_10

the necessity for Czech OCR system. Building an OCR system on real-world documents containing noise and erasure is more complex. The task of an OCR system is divided into segmentation and recognition where design of each influences the other. As segmentation robustness increases, recognizer task becomes simple and vice-versa. The segmentation techniques are similar across languages because connected components can be extracted to give written text units. The Czech language is very much similar to Latin languages [3]. It is of intermediate complexity with consonant-vowel pairs written as unit. The major concern in OCR systems is recognition accuracy. The deep learning techniques have shown considerable success in feature based classification [4]. One promising network which has emerged in this direction is convolutional neural network (CNN). It has shown considerable success for several recognition tasks viz digit recognition, image classification, hand-written character recognition etc. [4].

With this motivation, hierarchical fuzzy convolutional neural network (HFCNN) [5] is used here for Czech language character recognition task. It takes full advantage of deep CNN towards modeling long-term information of data sequences. The performance of HFCNN is improved by fine tuning parameters of network in hierarchical fashion. The pre-processing is performed through text region extraction, skew detection and correction, binarization, noise removal, character segmentation and thinning activities. Fuzzy Hough transform performs feature extraction. The feature based classification is performed through HFCNN. The character recognition is performed by HFCNN at review level. The evaluation is done on different Czech character data publicly available at [6]. The experimental results show superiority over traditional models. This paper is organized as follows. In Sect. 2 computational method of HFCNN based OCR for Czech language is highlighted. This is followed by experiments and results in Sect. 3. Finally in Sect. 4 conclusions are given.

## 2 Computational Method

In this section mathematical framework of proposed HFCNN model [5] is presented. The different components of OCR system for Czech language is given in Fig. 1.

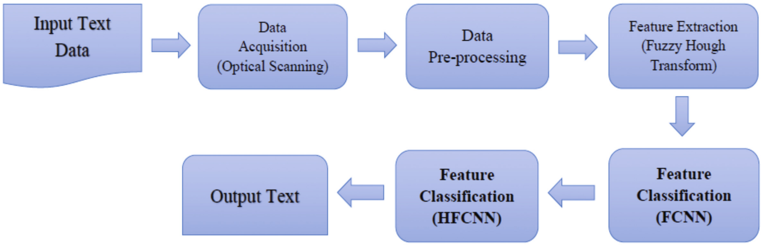

Input Text Data
Data Acquisition (Optical Scanning)
Data Pre-processing
Feature Extraction (Fuzzy Hough Transform)
Feature Classification (FCNN)
Feature Classification (HFCNN)
Output Text

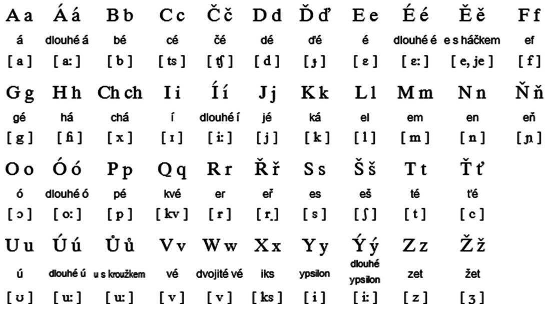
**Fig. 1.** The components of OCR system for Czech language

### 2.1 Problem Description

The research problem entails in developing OCR for Czech language. To achieve this character recognition task, we propose HFCNN to recognize Czech character data patterns at [6]. The character data patterns are subjected to pre-processing and features extraction. The extracted features are passed to FCNN classifier which performs recognition. The deep CNN demystifies training big and deep neural networks. It solves both image classification problem as well as integrates text recognition framework.

### 2.2 Datasets

The experimental data for performing OCR experiments is taken from Czech language website [6] shown in Fig. 2. The dataset contains Czech text used for training and testing. The database contains unconstrained handwritten text at resolution of 300 dpi as PNG images with 256 gray levels. Czech language database is structured through handwriting samples containing digits, lower and upper case letters and words.



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A a | Á á | B b | C c | Č č | D d | Ď ď | E e | É é | Ě ě | F f |
| á | dlouhé á | bé | cé | čé | dé | ďé | é | dlouhé é | e s háčkem | ef |
| [ a ] | [ aː ] | [ b ] | [ ts ] | [ tʃ ] | [ d ] | [ ɟ ] | [ ɛ ] | [ ɛː ] | [ e, je ] | [ f ] |
| G g | H h | Ch ch | I i | Í í | J j | K k | L l | M m | N n | Ň ň |
| gé | há | chá | í | dlouhé í | jé | ká | el | em | en | eň |
| [ g ] | [ ɦ ] | [ x ] | [ ɪ ] | [ iː ] | [ j ] | [ k ] | [ l ] | [ m ] | [ n ] | [ ɲ ] |
| O o | Ó ó | P p | Q q | R r | Ř ř | S s | Š š | T t | Ť ť | |
| ó | dlouhé ó | pé | kvé | er | eř | es | eš | té | ťé | |
| [ ɔ ] | [ oː ] | [ p ] | [ kv ] | [ r ] | [ r̝ ] | [ s ] | [ ʃ ] | [ t ] | [ c ] | |
| U u | Ú ú | Ů ů | V v | W w | X x | Y y | Ý ý | Z z | Ž ž | |
| ú | dlouhé ú | u s kroužkem | vé | dvojité vé | iks | ypsilon | dlouhé ypsilon | zet | žet | |
| [ ʊ ] | [ uː ] | [ uː ] | [ v ] | [ v ] | [ ks ] | [ i ] | [ iː ] | [ z ] | [ ʒ ] | |

**Fig. 2.** The components of OCR system for Czech language

### 2.3 Data Acquisition

OCR systems progress in Czech language is motivated from online or offline data acquisition [2]. Following lines of other European languages [2], data acquisition here can be online or offline. The offline character recognition captures data from paper through optical scanners or cameras whereas online recognition systems utilize digitizers which directly capture writing with strokes order, speed, pen up and down etc. The scope of this work is restricted to OCR systems and so it is confined to offline character recognition [2] for Czech language.

### 2.4 Data Pre-Processing

After data is acquired, it is pre-processed. In pre-processing [7] operations such as text region extraction, skew detection and correction, binarization, noise removal, character segmentation and thinning are performed. The main objective here is to organize information so that subsequent character recognition task becomes simpler. It essentially enhances image rendering it suitable for segmentation.

*Text region extraction*: In text region input image $I_{P \times Q}$ is partitioned into $m$ blocks $B_i; i = 1, 2, \ldots\ldots, m$ such that $B_i \cap B_j = \emptyset$ and $I_{P \times Q} = \bigcup_{i=1}^{m} B_i$. A block is pixels set $B_i = [f(x, y)]_{H \times W}$ where $H$ and $W$ are height and width of block respectively. Each individual block $B_i$ is classified as either information or background block based on intensity variation. After removal of background blocks adjacent information blocks constitute isolated component regions $R_i; i = 1, 2, \ldots\ldots, n$ such that $R_i \cap R_j = \emptyset\ \forall$ but $\bigcup_{i=1}^{n} R_i \neq I_{P \times Q}$ because some background blocks are removed. The region area is multiple of blocks area. These regions are classified as text region or non-text region using various feature characteristics such as dimensions, aspect ratio, information pixel density, region area, coverage ratio, histogram, etc. This technique is presented in [7].

*Skew detection and correction*: When text document is fed into scanner few skew degrees are unavoidable. The skew angle text lines in image make horizontal angle. These images suffer from skew and perspective distortion [7]. They occur due to non-parallel axes at image capturing. The skewness of different portions of image vary between $+\alpha$ to $-\beta$ degrees. The image cannot be deskewed at single pass. The perspective distortion effect is distributed throughout image and is not visible within small region. The segmentation module generates only few text regions. These text regions are deskewed using fast skew correction [7]. Every text region has dark and gray pixels. The dark pixels are texts and gray pixels constitute background. For four sides of virtual bounding rectangle of text region, there are four profile sets. If length and breadth of bounding rectangle are $M$ and $N$ then two profiles have $M$ and other two have $N$ values each. These values are pixel distances from side to first gray or black pixel of text region. The profiles from bottom side of text region is considered for estimating skew angle. This bottom profile is $\{h_i; i = 1, 2, \ldots\ldots, M\}$. The mean is $\mu = \frac{1}{M}\sum_{i=1}^{M} h_i$ and first order moment is $\tau = \frac{1}{M}\sum_{i=1}^{M} |\mu - h_i|$. The profile size is reduced by excluding $h_i$ that are not within $\mu \pm \tau$. From remaining profile leftmost $h_1$, rightmost $h_2$ and middle $h_3$ elements are chosen. The final skew angle is computed by averaging three skew angles obtained from three pairs $h_1$–$h_3$, $h_3$–$h_2$ and $h_1$–$h_2$. The skew angle is estimated by same angle rotation.

*Binarization*: It is next step in OCR process. Here goal is to keep relevant information in image. The binarization techniques of gray scale images are classified as overall and local threshold. The skew corrected text region is then binarized [7]. The arithmetic mean of maximum $G_{max}$ and minimum $G_{min}$ gray levels around pixel is taken as threshold for binarizing pixel. In present algorithm eight immediate neighbors around pixel subject to binarization are also taken as deciding factors for binarization.

*Noise removal*: The scanned text documents contain noise that arises due to printer, scanner, document age etc. Therefore, it is necessary to filter noise before image is processed. Here low-pass filter is used to process image [7] used for later processing.

*Character segmentation*: When text image is skew corrected, binarized and noise removed, actual text content is extracted. This leads to character segmentation [7]. After binarizing noise free text region, horizontal histogram profile $\{f_i; i = 1, 2, \ldots\ldots, H_R\}$ of region is analyzed for segmenting region into text lines. Here $f_i$ is number of black pixel along $i$th of text region and hidden region denotes height of deskewed text region.

All possible line segments are determined by thresholding profile values. The threshold is chosen to allow over segmentation. Text line boundaries are referred by $i$ values for which $f_i$ is less than threshold. After this inter segment distances are analyzed and some segments are rejected based on distance between two lines. Using vertical histogram profile of each individual text lines, words and characters are segmented.

*Thinning*: The character segmentation process follows thinning. In thinning one-pixel-width representation or object skeleton is obtained by preserving object connectedness [7] and its end points. This process reduces image components to essential information so that further analysis and recognition are facilitated. This enables easier subsequent detection of pertinent features. Here hilditch algorithm is used for thinning.

## 2.5 Feature Extraction Through Fuzzy Hough Transform

The heart of OCR system is feature vector formation in recognition stage. Here features are extracted from segmented image areas containing characters to be recognized. The feature extraction phase is visualized as set of features that define character shape as precise and unique. The feature extraction selects best input feature subset. These methods create new features based on transformations [7]. The selected features help discriminating characters. Achieving high recognition performance is attributed towards selection of appropriate feature extraction. The features capturing topological and geometrical information are most desired. Hough transform based feature extraction through fuzzy probability is applied for Czech language OCR [5]. This method is used for detection of lines and curves from images. It treats image as fuzzy points. For line detection it uses mapping $r = xcos\theta + ysin\theta$ which provides three important line characteristics. The parameters $r$ and $\theta$ specify position and line orientation. The count of $(r,\ \theta)$ accumulator cell used in Hough transform specifies number of black pixels lying on it. The fuzzy set definitions used here are available in [7] for first quadrant values which can be extended to other $\theta$ values.

## 2.6 Feature Classification Through Fuzzy Convolutional Neural Networks

After extracting essential features from pre-processed character image, concentration pointer turns on feature based classification methods. Here Czech language characters are recognized through FCNN which is deep learning based technique [5]. FCNN architecture is discussed here followed by HFCNN in next subsection.

FCNN exploits 2D correlation structure of image data. The convolutional and pooling layers that operate on 2D or 3D image data forms heart of FCNN architecture. The input to FCNN is 3D image with two spatial and one frequency dimension. The gray-scale image sample have one map stack. The input here is 48 × 48 single-map binary image. The convolution operation is performed on 3D image which is stack of 2D images. The intermediate representations are 3D images with thousands of maps which are of smaller size. Each constituent map of 3D input image is convolved with 2D kernel. The dot-products are summed to generate single output map. An input with $d_{in}$ maps gets convolved with 3D kernel size $k \times k$ to produce single output map. When $d_{out}$ output maps are required as many such kernels are used. A 3D to 3D convolution operation has kernel size $d_{in} \times d_{out} \times k \times k$. 3D convolution kernel has $k^2 \times d_{in} \times d_{out}$ parameters. This operation is series of $d_{out}$ feature extractors. The fuzzy convolution operation $\widetilde{Cov}'_{\tilde{W}}$ [5] on 3D input tensor $A(d_{in} \times s \times s)$ with fuzzy kernel $\tilde{W}$ is $\widetilde{Cov}'_{\tilde{W}} : \tilde{\mathbb{R}}^{d_{in} \times s \times s} \rightarrow \tilde{\mathbb{R}}^{d_{out} \times s \times s}$; $\widetilde{Cov}'_{\tilde{W}}(A)_{r,p,q} = \sum_{m=1}^{d_{in}} \sum_{i=-l}^{l} \sum_{j=-l}^{l} \tilde{W}^{r}_{m,i,j} A_{m,p+i,q+j}$. It assumes that for any indices out of range $A$ is zero. A nonlinearity $R$ is applied after convolution $\widetilde{Cov}'_{\tilde{W}}$ to give $\widetilde{Cov}_{\tilde{W}}$ convolutional layer operation $\widetilde{Cov}_{\tilde{W}}(A) = R\left(\widetilde{Cov}'_{\tilde{W}}(A)\right)$. Here $\tilde{W}$ is fuzzy weight tensor that is to be learned by training network. The fuzzy real numbers are modeled through trapezoidal membership function [8] to handle inherent impreciseness and vagueness. A typical convolution layer outputs many more maps than it takes in. There is also high correlation between adjacent output values in map. Hence it pools or scales maps down by factor of two in each co-ordinate. Pooling is done over 2 × 2 grid. The maximum of four pixels in grid is extracted as output. The number of maps four-fold at each convolutional layer is increased and area of mage-maps by factor of four at succeeding pool layer is decreased. This preserves image size while transforming to different and useful space. The 2 × 2 max-pooling operation on tensor $A$ is $P_2(A)_{r,p,q} = \max_{\substack{i = 2p-1, 2p \\ j = 2q-1, 2q}} A_{r,i,j}$. FCNN has two to twenty convolutional and pooling layers. The final output is 3D image is flattened into vector followed by one or more fully-connected layers. A fully-connected layer is simple matrix multiplication followed by fuzzy non-linearity $\tilde{F}_{\tilde{W}} : \tilde{\mathbb{R}}^{n_1} \rightarrow \tilde{\mathbb{R}}^{n_2}$ such that $\tilde{F}_{\tilde{W}}(A) = \tilde{R}(\tilde{W}A)$. Here $\tilde{R}$ is fuzzy non-linearity applied element-wise. This is hidden layer of network. The last of such fully connected layers is output layer $F^S_W$. It has as many nodes as number of classes $K$. It employs specific form of fuzzy non-linearity, *softmax* function $\tilde{S}, \tilde{F}^{\tilde{S}}_{\tilde{W}} : \tilde{\mathbb{R}}^{n_2} \rightarrow \tilde{\mathbb{R}}^{K}$ such that $\tilde{F}^{\tilde{S}}_{\tilde{W}}(A) = \tilde{S}(\tilde{W}A)$ with $\tilde{S}_k(A) = \frac{e^{A_k}}{\sum_{j=1}^{k} e^{A_j}}$. This transformation used in multi-logit model and produces positive values that sum to one. The $K$ vector obtained from matrix multiplication is exponentiated to make it positive and then normalized. These $K$ values are interpreted as class probabilities. Here fuzzy real numbers are modeled through trapezoidal membership function [8] to handle inherent impreciseness and vagueness. The non-linearities are applied to each network output layer. A network with non-linear hidden layer acts as universal function approximator. The sigmoid function $\text{sigmod}(x) = \frac{1}{1+e^{-x}}$ [5] is used here to tackle nonlinearity. Each sigmoid node in hidden layer simulates step function of different size at different location. The output node combines these steps to approximate desired mapping. Without non-linear activations network acts as linear transformer regardless of its depth. Let $X$, $Y$ denote random image and its class label respectively. The fuzzy likelihood is: $\widetilde{Prob}(Y = y|X = x; \mathcal{W}) = \widetilde{prob}_y(x; \mathcal{W}) = \tilde{S}_y(\tilde{W}_1 A(x))$. Here $A(x)$ is input to *softmax* layer for given set of network parameters $\mathcal{W}$. The networks are trained to maximize log of likelihood over training data to find optimum network parameters $\mathcal{W}^*$. The results obtained from earlier phase are scaled up or down to 48 × 48 square and fed to FCNN. The aspect ratio is preserved while scaling. In addition to 2304 binary pixel values there are two numbers representing top and baselines

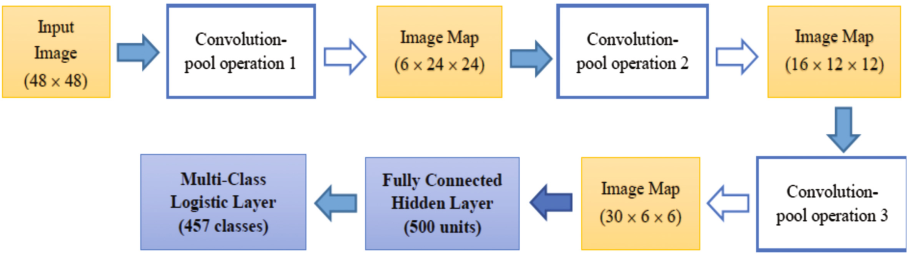

Input Image (48 × 48)
Convolution-pool operation 1
Image Map (6 × 24 × 24)
Convolution-pool operation 2
Image Map (16 × 12 × 12)
Convolution-pool operation 3
Image Map (30 × 6 × 6)
Fully Connected Hidden Layer (500 units)
Multi-Class Logistic Layer (457 classes)

**Fig. 3.** The FCNN model for feature classification

location. These are later incorporated into network. The traditional architecture by LeCun et al. [5] is used as reference to compare various design choices and regularizations. It has three pairs of convolutional pool layers which employ 3 × 3 convolution kernel and 2 × 2 max-pooling window. These are followed by two fully connected layers. The last layer's *softmax* activation yields final class probabilities. FCNN model now becomes:$\widetilde{prob}(x, \mathcal{W}) = \tilde{F}^{\tilde{S}}_{\tilde{W}_5} * \tilde{F}_{\tilde{W}_4} * P_2 * \widetilde{Cov}_{\tilde{W}_3} * P_2 * \widetilde{Cov}_{\tilde{W}_2} * P_2 * \widetilde{Cov}_{\tilde{W}_1}(x)$. Here $\mathcal{W}$ is network parameters to be learned. The Fig. 3 shows FCNN model. The implementation of FCNN is performed in MATLAB [5].

### 2.7 Feature Classification Hierarchical Fuzzy Convolutional Neural Networks

The hierarchical version of FCNN viz HFCNN [5] is proposed here. The computational benefits [5] serve major motivation. HFCNN is different from FCNN in terms of classification accuracy based on similarities and running time when data volume grows [5]. The model architecture is shown in Fig. 4. HFCNN architecture correlates data behavior across multiple relevant classification features. This allows computational overhead distribution towards OCR construction. At 1st, 2nd and 3rd layers small size FCNNs are utilized. The number of FCNNs are increased at layers 4 and 5. FCNNs in last layer are constructed over subset examples for which neuron in 5th layer is convolution matching unit (CMU) [5]. FCNNs at last layer is larger in size than 1st to 5th layers. It improves resolution and discriminatory power of FCNN with less training overhead. Building HFCNN requires several normalization operations. This provides for initial pre-processing and inter-layer quantization between 1st to 5th layers. The temporal pre-processing provides suitable data representation and supports time based representation. The 1st CNN layer treats each feature independently with each data instance mapped to sequence values. For temporal representation standard FCNN has no capacity to recall histories of patterns directly. A shift register of length $h$ is used where tap is taken at predetermined repeating interval $v$ so that $h\%v = 0$ where % is modulus operator. The 1st level FCNNs receive values from shift register. Thus, as each new connection is encountered at left, content of each shift register location is transferred one location to right with previous item at $h$th location being lost. In case of $n$-feature architecture it is necessary to quantize number of neurons between 1st to 5th level FCNNs. The purpose of 2nd to 5th level FCNN is to provide an integrated view of input feature specific FCNNs developed in 1st layer. There is potential for each neuron in 2nd to 5th layer FCNN to have an input dimension defined by total neuron count across all 1st layer FCNNs. This is brute force solution that does not scale computationally. Given topological ordering provided by FCNN, neighboring neurons respond to similar stimuli. The topology of each 1st layer CNN is quantized in terms of fixed number of neurons using potential function classification algorithm [5]. This reduces number of inputs in 2nd to 5th layers of FCNN. The neurons in 4th layer acts as CMU for examples with same class label that maximizes detection rate and minimizes false positives. However, there is no guarantee for this. In order to resolve this 4th layer CNN neurons that act as CMU for examples from more than one class are used to partition data. The 5th layer FCNNs are trained on subsets of original training data. This enables 5th layer FCNNs size to increase which improves class specificity with reasonable computational cost. Once training is complete 4th layer CMUs act to identify which examples are forwarded to corresponding 5th layer FCNNs on test dataset. A decision rule determines under what conditions classification performance of CMU at 4th layer FCNN is judged sufficiently poor for association with 5th layer FCNN. There are several aspects that require attention such as minimum acceptable misclassification rate of 4th layer CMU relative to number of examples labeled at 4th layer CMU and number of examples 4th layer CMU represent. The basic implication is that there must be optimal number of connections associated with 4th layer CMU for training of corresponding 5th layer FCNN and misclassification rate over examples associated with 4th layer CMU exceeds threshold. HFCNN is characterized in terms of success probability in recovering true hierarchy $H^*$ and runtime complexity. Some restrictions are placed to similarity function $S$ [5] such that similarities scale with hierarchy upto some random noise: (a) for each $y_i \in Cs_j \in Cs^*$ and $j' \neq j$ : $\min_{y_p \in Cs_j} \mathbb{Exp}[S(y_i, y_p)] - \max_{y_p \in Cs_j'} \mathbb{Exp}[S(y_i, y_p)] \geq \gamma > 0$, here expectations are taken with respect to noise on $S$; (b) for each $y_i \in Ct_j$, a set of $V_j$ words of size $v_j$ drawn uniformly from $Cs_j$ satisfies $\mathbb{Prob}\left(\min_{y_p \in Cs_j} \mathbb{Exp}[S(y_i, y_p)] - \sum_{y_p \in V_j} \frac{S(y_i, y_p)}{v_j} > \epsilon\right) \leq 2e^{\left\{\frac{-2v_j\epsilon^2}{\sigma^2}\right\}}$, here $\sigma^2 \geq 0$ parameterizes noise on similarity function $S$. From viewpoint of feature learning stacked FCNNs extracts features of sequences in character datasets. Various trade-offs are done towards improving representation ability and avoiding over fitting. It is easy to overfit network with limited data training sequences. This algorithm can be fine-tuned with certain dynamic heuristics.

## 3 Experiments and Results

In this section experimental results are presented for Czech language OCR. In order validate HFCNN classification supremacy experiments are performed with traditional techniques viz k-nearest neighbor (kNN), multiclass logistic regression (MLR), linear support vector machines (SVM), conventional artificial neural network (ANN) with hidden layers, CNN and FCNN on dataset [5]. The results are highlighted in Table 1. The kNN classifier gives poor performance with 34.65% test error. With MLR test error reduces to 31.06%. Linear SVM gives test error of 27.86%. A conventional ANN with two hidden layers each with two thousand nodes reduces error considerably using nearly 10 times more parameters than MLR and linear SVM. These architectures ignore

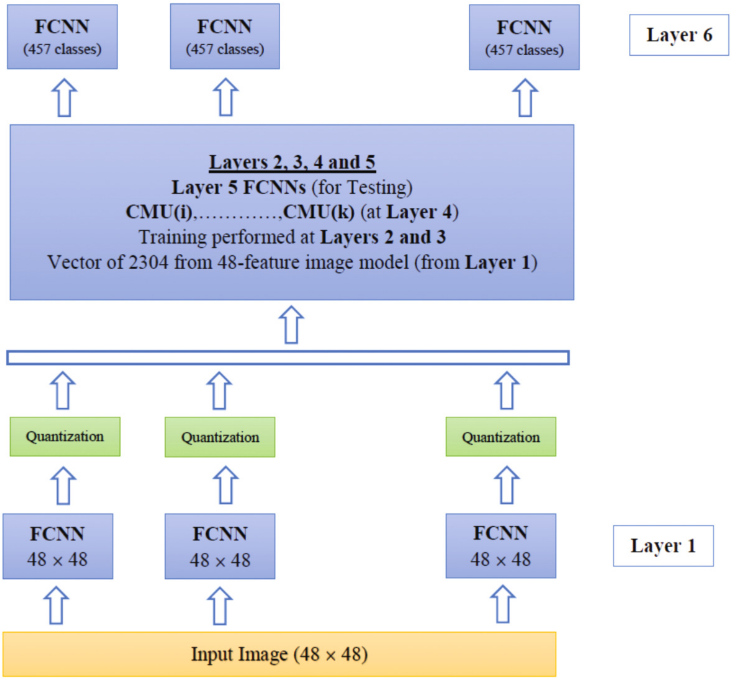


**Fig. 4.** The architecture of proposed HFCNN model

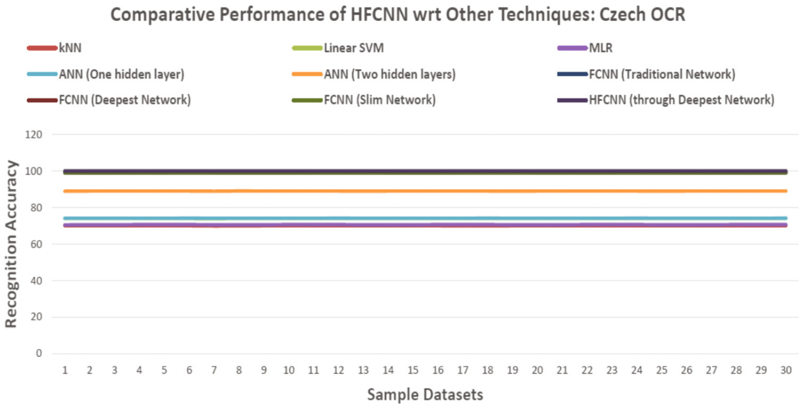


**Fig. 5.** The comparative performance of HFCNN vs other techniques for Czech language

**Table 1.** Comparative performance of various network architectures on Czech language dataset

| Network architecture | Training error % | Test error % | Network speed (ms) | Network size (parameters millions) |
|---|---|---|---|---|
| kNN (k = 10) | 25.48 | 34.65 | ≫10 | – |
| Linear SVM | 24.28 | 27.86 | 0.025 | 1.05 |
| MLR (48 × 48 – 457 softmax output) | 28.08 | 31.06 | 0.039 | 1.05 |
| ANN (one hidden layer: 48 × 48 – 1000 nodes – 457 softmax output) | 24.55 | 27.39 | 0.057 | 2.76 |
| ANN (two hidden layers: 48 × 48 – 2000 nodes – 2000 nodes – 457 softmax output) | 8.28 | 14.00 | 0.130 | 9.55 |
| FCNN (traditional network) [48 × 48 – 3 × 3 convolutional layers (8 output maps) – 2 × 2 max pooling layer – 3 × 3 convolutions layers (24 output maps) – 2 × 2 max pooling layer – 3 × 3 convolutional layers (72 output maps) – 2 × 2 max-pooling layer – 500 nodes – 457 softmax output] | 0.55 | 1.50 | 0.269 | 1.14 |
| FCNN (deepest network) [48 × 48 – 3 × 3 convolutional layers (6 output maps) – 3 × 3 convolutional layers (6 output maps) – 2 × 2 max-pooling layer – 3 × 3 convolutions layers (18 output maps) –3 × 3 convolutional layers (18 output maps) – 2 × 2 max-pooling layer – 3 × 3 convolutional layers (54 output maps) – 3 × 3 convolutions layers (54 output maps) – 2 × 2 max-pooling layer – 3 × 3 convolutional layers (162 output maps) – 3 × 3 convolutional layers (162 output maps) –2 × 2 max-pooling layer – 457 softmax output] | 0.08 | 0.75 | 0.648 | 0.75 |
| FCNN (slim network) [48 × 48 – 3 × 3 convolutional layers (4 output maps) – 2 × 2 max-pooling layer – 3 × 3 convolutional layers (6 output maps) – 3 × 3 convolutional layers (8 output maps) – 2 × 2 max-pooling layer – 3 × 3 convolutional layers (32 output maps) – 3 × 3 convolutional layers (50 output maps) -3 × 3 convolutional layers (50 output maps) – 2 × 2 max-pooling layer – 457 softmax output] | 0.41 | 1.36 | 0.169 | 0.25 |
| HFCNN (through deepest network) | 0.04 | 0.60 | 0.848 | 3.00 |

2D data structure. A structure exploiting CNN with three convolutional pool operations followed by two fully connected layers gives an error rate of about 1.50% with same number of parameters as MLR. The deeper networks with more convolutional layers and only one fully connected layer give better recognition accuracies. The network size can be made small with $10^5$ parameters and test error rates below 1.55%. For deeper networks with only one fully connected layer half the parameters are in last classification layer. The remaining network with its six to eight convolutional layers uses other half parameters. The network parameters are equally split between classification and feature extraction tasks. However, convolution operation is expensive and can significantly increase time required to classify character but this provides other benefits. The fourteen layered ANN from Table 1 requires specification of nearly hundred hyper parameters. The slim architecture from Table 1 has been designed to improve both on size and speed. Some important design aspects are briefly highlighted to demystify the process. Traditional tanh activation increases computational complexity and reduces performance considerably. The results are summarized in Table 2. Traditional CNN architecture based on LeCun et al. [5] has pool layer after each convolutional layer. The kernel size decreases as image size decreases with pooling. The present architectures tend to use 3 × 3 kernels all through. Where bigger kernels are needed multiple convolutional layers are used instead with increased computational complexity. Considering fitting model $\{0,1\}^{48\times48} \rightarrow (0,1)^{457}$ using over million parameters, overfitting process may crop up. The Table 3 summarizes effects of various regularization forms on testing and training errors. The Fig. 5 gives comparative performance of HFCNN over other techniques for Czech language dataset samples.

**Table 2.** The effect of various design aspects on testing and training errors

| Network design choices | Training error % | Test error % |
|---|---|---|
| HFCNN (through deepest network) | 0.04 | 0.60 |
| FCNN (traditional network) | 0.55 | 1.50 |
| No input inversion | 0.54 | 1.48 |
| tanh activations | 1.10 | 2.05 |

**Table 3.** The effect of various regularization forms on testing and training errors

| Regularization variations | Training error % | Test error % |
|---|---|---|
| HFCNN (through deepest network) | 0.04 | 0.60 |
| FCNN (traditional network) | 0.55 | 1.50 |
| Without input distortion | 0.04 | 1.77 |
| Without dropout | 0.25 | 1.14 |
| Without input distortion and dropout | 0.04 | 2.78 |
| Without depth (third convolutional pool layer) | 1.19 | 2.69 |
| Without input distortion, dropout and depth | 0.03 | 4.10 |
| Without the elastic aspect of distortion | 0.48 | 1.16 |

## 4 Conclusion

OCR system of Czech language using hierarchical version of deep learning networks is proposed here. The computational framework segments text image, classifies characters using HFCNN and performs recognition. It uses deep CNN towards modeling long-term data sequences. HFCNN performance is enhanced by fine tuning network parameters. The feature extraction is achieved through fuzzy Hough transform. HFCNN recognizes characters at review level. OCR system is evaluated using different types of Czech character datasets. The experimental results and simulation studies demonstrate efficiency of proposed approach.

## References

1. Yu, F.T.S., Jutamulia, S. (eds.): Optical Pattern Recognition. Cambridge University Press, Cambridge (1998)
2. Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K.: Introduction, book chapter: optical character recognition systems for different languages using soft computing. Stud. Fuzziness Soft Comput. **352**, 1–6 (2017)
3. Czech Language: https://en.m.wikipedia.org/wiki/Czech_language
4. Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K.: Soft computing techniques for optical character recognition systems, book chapter: optical character recognition systems for different languages using soft computing. Stud. Fuzziness Soft Comput. **352**, 43–83 (2017)
5. Chaudhuri, A.: Some Experiments on Optical Character Recognition Systems for Different Languages Using Soft Computing Techniques, Technical Report, Birla Institute of Technology Mesra, Patna Campus, India (2010)
6. Czech Language Dataset: http://www.czech-language.cz
7. Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K.: Optical character recognition systems, book chapter: optical character recognition systems for different languages using soft computing. Stud. Fuzziness Soft Comput. **352**, 9–41 (2017)
8. Zimmermann, H.J.: Fuzzy Set Theory and its Applications, 4th edn. Kluwer Academic Publishers, Boston (2001)

# A Percentile Transition Ranking Algorithm Applied to Knapsack Problem

José García[1,2](✉), Broderick Crawford[2], Ricardo Soto[2], and Gino Astorga[2,3]

[1] Centro de Investigación y Desarrollo Telefónica, 7500961 Santiago, Chile
joseantonio.garcia@telefonica.com
[2] Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
{broderick.crawford,ricardo.soto}@pucv.cl
[3] Universidad de Valparaíso, 2361864 Valparaíso, Chile
gino.astorga@uv.cl

**Abstract.** The binarization of Swarm Intelligence continuous metaheuristics is an area of great interest in operational research. This interest is mainly due to the application of binarized metaheuristics to combinatorial problems. In this article we propose a general binarization algorithm called Percentile Transition Ranking Algorithm (PTRA). PTRA uses the percentile concept as a binarization mechanism. In particular we will apply this mechanism to the Cuckoo Search metaheuristic to solve the set multidimensional Knapsack problem (MKP). We provide necessary experiments to investigate the role of key ingredients of the algorithm. Finally to demonstrate the efficiency of our proposal, we solve Knapsack benchmark instances of the literature. These instances show PTRA competes with the state-of-the-art algorithms.

**Keywords:** Combinatorial optimization · Multidimensional knapsack problem · Metaheuristics

## 1 Introduction

In recent years, the areas of physics and swarm intelligence have generated a large number of algorithms, many of which have been effective and efficient in solving complex optimization problems. Examples of these algorithms are Ant Colony Optimization [8], Firefly Algorithm [17], Gravitational Search Algorithm [14], Cuckoo Search Algorithm [18], Particle Swarm Optimization [9]. Many of these algorithms have the characteristic that the movement of the particles are performed in a continuous space. On the other hand, combinatorial problems arise in many areas of computer science and application domains. For example in protein structure prediction, grouping routing, planning, scheduling and timetabling problems. It is natural to try to apply algorithms inspired by physics and swarm intelligence in these combinatorial problems [4]. In the process of adaptation a series of difficulties arise when moving from continuous spaces to discrete spaces. Examples of these difficulties are spacial disconnect, hamming

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_11

cliffs, loss of precision and the curse of dimension [12]. This has the consequence that binarizations are not always effective and efficient [11].

In this paper, a general binarization technique called Percentile Transition Ranking Algorithm (PTRA) is proposed to binarize continuous swarm intelligence metaheuristics. The main operator corresponds to the percentile ranking transition operator. This operator performs the binarization using percentiles grouping process and it is complemented with local search and perturbation operators. The main goal of this work corresponds to evaluate our algorithm when dealing with an NP-hard combinatorial optimization problem such as the MKP. To develop the evaluation, we used the metaheuristic Cuckoo Search. The metaheuristic Cuckoo Search was chosen because it is a swarm intelligence continuous metaheuristic that has been widely used in combinatorial problems [5,15].

Experiments were developed that shed light on the contribution of the different operators to the effectiveness of the algorithm. Moreover, our algorithm was compared with recent algorithms that use transfer functions as binarization method. For this purpose we use tests problems from the OR-Library.[1] We compared our framework with the Binary Artificial Algae Algorithm (BAAA) published by [19]. The numerical results show that PTRA achieves highly competitive results.

The remainder of this paper is organized as follows. Section 2 briefly introduces the Knapsack problem. In Sect. 3 we explain the transition ranking binarization algorithm. The results of numerical experiments are presented in Sect. 4. Finally we provide the conclusions of our work.

## 2 KnapSack Problem

The MKP is a combinatorial problem that has multiple applications in science and engineering. For example capital budgeting and project selection applications. The MKP has also been introduced to model problems like cutting stock, loading problems, allocation of processors in a distributed data processing [7], and delivery in vehicles with multiple compartments [3].

Numerous methods have been developed to solve the MKP. The exact methods were applied in the 80's to solve MKP. They generate a variety of methods including dynamic programming, branch-and-bound, network approach and reduction schemes. The exact methods have made possible the solution of middle size MKP instances. The major drawback of these methods remains the temporal complexity when dealing with large instances. Therefore, many researchers focus on heuristic and meta-heuristic search methods which can produce solutions of good qualities in a reasonable amount of time. In recent years, many bio-inspired and physics based algorithms, such Swarm Optimization [2], Firefly algorithm [1], Binary Black Hole [6] and Binary Fruitfly [16] have been proposed to solve large instances of the MKP.

[1] OR-Library: http://www.brunel.ac.uk/mastjjb/jeb/orlib/mknapinfo.html.

The MKP problem belongs to the class of NP-hard problems. MKP corresponds to a model of resource allocation, whose objective is to select a subset of objects that produce the greatest benefit considering certain capacity constraints. Each object $j$ consumes a different amount of resources in each dimension. Also each object has a profit associated. Formally the MKP can be set as:

$$\text{maximize} \sum_{j=1}^{n} p_j x_j \tag{1}$$

$$\text{subjected to} \sum_{j=1}^{n} c_{ij} x_j \leq b_i \ , \ i \in \{1, ..., m\} \tag{2}$$

$$\text{with } x_j \in \{0, 1\} \ , \ j \in \{1, ..., n\} \tag{3}$$

where $p_j$ is the profit for the item $j$, $c_{ij}$ corresponds to the consumption of resources of item $j$ in the dimension $i$, and $b_i$ is the capacity constraint of each dimension $i$. The representation of a solution of the problem is modelled naturally in binary form where 0 in the $j$-th position means that the $j$ item is not included in the Knapsack and 1 indicates that $j$ is included.

## 3 Percentile Transition Ranking Algorithm

The first step of PTRA corresponds to the initialization of the feasible solutions Sect. 3.2. Once the initialization of the particles is performed, it is consulted if the detention criterion is satisfied. This criterion includes a maximum of iterations. Subsequently, if the criterion is not satisfied, the percentile transition ranking operator is executed (Sect. 3.2). This operator is responsible for performing the iteration of solutions. Once the transitions of the different solutions are made, we compare the resulting solutions with the best solution previously obtained. In the event that a superior solution is found, this replaces the previous one. When a replacement occurs, the new solution is subjected to a local search operator. Finally, having met a number of iterations where there has not been a replacement for the best solution, a perturbation operator is used. The general algorithm scheme is detailed in Fig. 1. In the following subsection we will explain in detail the initialization method, the percentile transition ranking operator and the repair operator. The explanation of the other operators will be left for an extended version.

### 3.1 Initialization and Element Weighting

PTRA uses a binarization of swarm-intelligence metaheuristics to try to find the optimum. Each of these possible solutions, is generated as follows: First we select an item randomly. Subsequently we consulted the constraints of our problem if there are other elements that can be incorporated. The list of possible elements to be incorporated is obtained, the weight for each of these elements is calculated

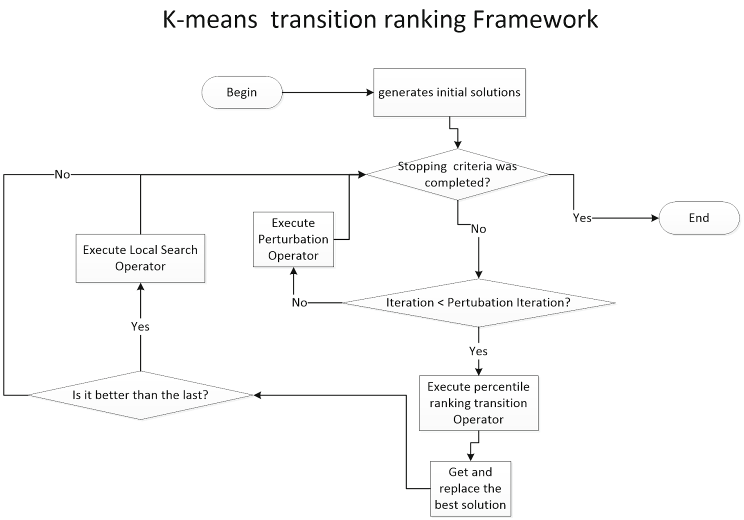

K-means transition ranking Framework
Begin
generates initial solutions
Stopping criteria was completed?
Yes
End
No
Iteration < Pertubation Iteration?
No
Execute Perturbation Operator
Yes
Execute percentile ranking transition Operator
Get and replace the best solution
Is it better than the last?
Yes
Execute Local Search Operator
No

**Fig. 1.** Flowchart of the percentile transition ranking algorithm.

and the best element is selected. The procedure continues until no more elements can be incorporated. The initialization algorithm is detailed in Fig. 2.

Several techniques were proposed in the literature, to calculate the weight of each element. For example [13] introduced the pseudo-utility in the surrogate duality approach. The pseudo-utility of each variable was given in Eq. 4. The variable $w_j$ is the surrogate multiplier between 0 and 1 which can be viewed as shadow prices of the $j$-th constraint in the linear programming (LP) relaxation of the original MKP

$$\delta_i = \frac{p_i}{\sum_{j=1}^{m} w_j c_{ij}} \tag{4}$$

Another more intuitive measure is proposed by [10]. This measure is focused on the average occupancy of resources. Its equation is shown in 5.

$$\delta_i = \frac{\sum_{j=1}^{m} \frac{c_{ij}}{m b_j}}{p_i} \tag{5}$$

In this paper, we propose a variation of this last measure focused on the average occupation. However this variation considers the elements that exist in backpacks to calculate the average occupancy. In each iteration depending on the selected items in the solution the measure is calculated again. The equation of this new measure is shown in Eq. 6.

$$\delta_i = \frac{\sum_{j=1}^{m} \frac{c_{ij}}{m(b_j - \sum_{i \in S} c_{ij})}}{p_i} \tag{6}$$

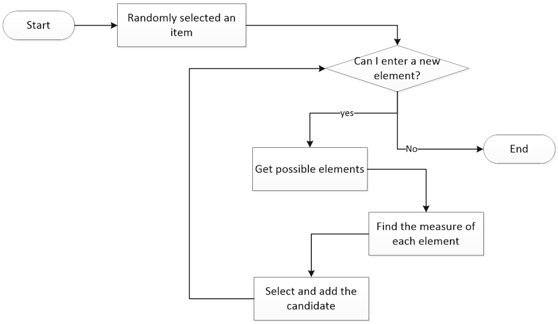

Start
Randomly selected an item
Can I enter a new element?
yes
No
End
Get possible elements
Find the measure of each element
Select and add the candidate

**Fig. 2.** Flowchart of generation of a new solution.

### 3.2 Percentile Transition Ranking Operator

Considering that our metaheuristic is a continuous and swarm intelligence. Due to its iterative nature, it needs to update the position of particles at each iteration. When the metaheuristic is continuous, this update is performed in $\mathbb{R}^n$ space. In Eq. 7, the position update is presented in a general form. The $x_{t+1}$ variable represents the $x$ position of the particle at time $t$+1. This position is obtained from the position $x$ at time $t$ plus a $\varDelta$ function calculated at time $t$+1. The function $\varDelta$ is proper to each metaheuristic and produces values in $\mathbb{R}^n$. For example in Cuckoo Search $\varDelta(x) = \alpha \oplus Levy(\lambda)(x)$, in Black Hole $\varDelta(x) = \text{rand} \ \times (x_{bh}(t) - x(t))$ and in the Firefly, Bat and PSO algorithms $\varDelta$ can be written in simplified form as $\varDelta(x) = v(x)$.

$$x_{t+1} = x_t + \varDelta_{t+1}(x(t)) \tag{7}$$

In the percentile transition ranking operator, we considering the movements generated by the metaheuristic in each dimension for all particles. $\varDelta^i(x)$ corresponds to the magnitude of the displacement $\varDelta(x)$ in the i-th dimension for the particle x. Subsequently these displacement are grouped using $\varDelta^i(x)$, the magnitude of the displacement. This grouping is done using the percentile list. In our case the percentile list used the values {20, 40, 60, 80, 100}.

The percentile operator has as entry the parameters percentile list (percentileList) and the list of values (valuesList). Given an iteration, the list of values corresponds to the magnitude $\varDelta^i$ of all particles in all dimensions. As a first step the operator uses the valueList and obtains the values of the percentiles given in the percentileList. Later, each value in the valueList is assigned the group of the smallest percentile to which the value belongs. Finally, the list of the percentile to which each value belongs is returned (percentileGroupValue).

A transition probability through the function $P_{tr}$ is assigned to each element of the valueList. This assignment is done using the percentile group assigned to each value (percentileGroupValue). For the case of this study, we particularly use the Step function given in rule 8.

$$P_{tr}(x^i) = \begin{cases} 0.1, & \text{if } x^i \in \text{group } \{0, 1\} \\ 0.5, & \text{if } x^i \in \text{group } \{2, 3, 4\} \end{cases} \tag{8}$$

Afterwards the transition of each particle is performed. In the case of Cuckoo search the rule 9 is used to perform the transition, where $\hat{x}^i$ is the complement of $x^i$. Finally, each solution is repaired using the repair operator. The algorithm is shown in 1.

$$x^i(t+1) := \begin{cases} \hat{x}^i(t), & \text{if } rand < P_{tg}(x^i) \\ x^i(t), & \text{otherwise} \end{cases} \tag{9}$$

**Algorithm 1.** Percentile ranking operator

```
Function percentileRankingTransition(valueList, percentileList)
Input valueList, percentileList
Output percentileGroupValue
percentileValue = getPercentileValue(valueList, percentileList)
for each value in valueList do
    percetileGroupValue = getPercentileGroupValue(percentileValue,valueList)
end for
return percetileGroupValue
```

### 3.3 Repair Operator

In each movement performed by operators: transition ranking, local search and perturbation, it is possible to generate solutions that are infeasible. Therefore, each candidate solution must be checked and modified to meet every constraint. This verification and subsequent repairing is performed using the measure defined in Sect. 3.1 Eq. 6. The procedure is shown in Algorithm 2. As input the repair operator receives the solution $S_{in}$ to repair, and the output of the repair operator gives the repaired solution $S_{out}$. As a first step, the repair algorithm asks whether the solution needs to be repaired. In the case that the solution needs repair, a weight is calculated for each element of the solution using the measure defined in Eq. 6. The element of the solution with the largest measure is returned and removed from the solution. This element is named $s_{max}$. This process is iterated until our solution does not require repair. The next step is to improve the solution. The Eq. 6 is again used for obtaining the element with the smallest measure that meets the constraints $s_{min}$ and add $s_{min}$ to the solution. In the case of absence of elements, empty is returned. The algorithm iterates until there are no elements that satisfy the constraints.

**Algorithm 2.** Repair Algorithm

```
**Function** Repair($S_{in}$)
**Input** Input solution $S_{in}$
**Output** The Repair solution $S_{out}$
$S \leftarrow S_{in}$
**while** needRepair($S$) == True **do**
   $s_{max} \leftarrow$ getMaxWeight($S$)
   $S \leftarrow$ removeElement($S$, $s_{max}$)
**end while**
state ← False
**while** state == False **do**
   $s_{min} \leftarrow$ getMinWeight($S$)
   **if** $s_{min} == \emptyset$ **then**
      state ← True
   **else**
      $S \leftarrow$ addElement($S$, $s_{min}$)
   **end if**
**end while**
$S_{out} \leftarrow S$
**return** $S_{out}$
```

# 4 Results

## 4.1 Insight of PTRA Algorithm

In this section we investigated some important ingredients of PTRA to get insight into the behavior of the proposed algorithm. To carry out this comparison the first 10 problems of the set cb.5.250 of the OR library were chosen. The contribution of the percentile transition ranking operator on the final performance of the algorithm was studied. The contribution of the perturbation and local search operators will be developed in an extended version. To compare the distributions of the results of the different experiments we use violin Chart. The horizontal axis X corresponds to the problems, while Y axis uses the measure % - Gap defined in Eq. 10

$$\% - Gap = 100\frac{BestKnown - SolutionValue}{BestKnown} \tag{10}$$

Furthermore, a non-parametric test, Wilcoxon signed-rank test is carried out to determine if the results of PTRA with respect to other algorithms have significant difference or not. The parameter settings and browser ranges are shown in Table 1.

**Evaluation of Percentile Transition Ranking Operator.** To evaluate the contribution of the percentile transition ranking operator to the final result. We designed a random operator. This random operator executes the transition with

**Table 1.** Setting of parameters for Cuckoo Search Algorithm.

| Parameters | Description | Value | Range |
|---|---|---|---|
| $\nu$ | Coefficient for the perturbation operator | 3% | [2, 3, 4] |
| N | Number of Nest | 20 | [15, 20, 25] |
| G | Number of percentiles | 5 | [4, 5, 6] |
| $\gamma$ | Step Length | 0.01 | [0.009,0.01,0.011] |
| $\kappa$ | Levy distribution parameter | 1.5 | [1.4,1.5,1.6] |
| Iteration Number | Maximum iterations | 1000 | [1000] |

a fixed probability (0.5) without considering the ranking of the particle in each dimension. Two scenarios were established. In the first one the perturbation and local search operators are included. In the second one these operators are excluded. PTRA corresponds to our standard algorithm. *05.pe* is the random variant that includes the perturbation and local search operators. *wpe* corresponds to the version with percentile transition operator without perturbation and local search operators. Finally *05.wpe* describes the random algorithm without perturbation and local search operators.

When we compared the Best Values between PTRA and *05.pe* which are shown in Table 2. PTRA outperforms to *05.pe*. However the Best Values between both algorithms are very close. In the Average comparison, PTRA outperforms *05.pe* in all problems. The comparison of distributions is shown in Fig. 3. We see the dispersion of the *05.pe* distributions are bigger than the dispersions of PTRA. In particular this can be appreciated in the problems 1, 4, 5, 6, and 9. Therefore, the percentile transition ranking operator together with perturbation

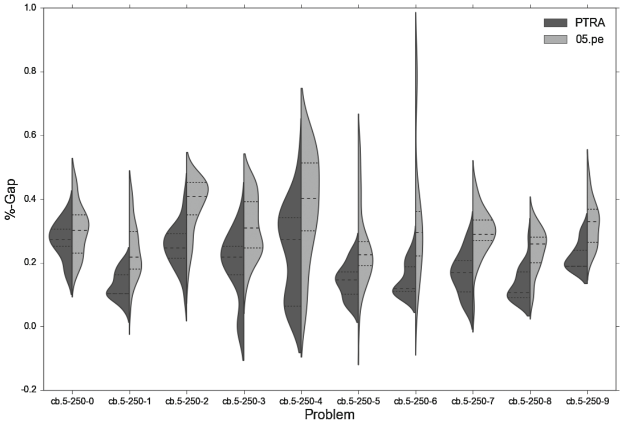


**Fig. 3.** Evaluation of percentile transition operator with perturbation and Local Search operators

**Table 2.** Evaluation of percentile transition ranking operator

| Set | Best<br>known | Best<br>*05.pe* | Best<br>PTRA | Best<br>*05.wpe* | Best<br>*wpe* | Avg<br>*05.pe* | Avg<br>PTRA | Avg<br>*05.wpe* | Avg<br>*wpe* |
|---|---|---|---|---|---|---|---|---|---|
| cb.5.250-0 | 59312 | 59211 | 59211 | 59158 | 59175 | 59132.1 | 59151.8 | 59071.8 | 59134.5 |
| cb.5.250-1 | 61472 | 61435 | 61435 | 61409 | 61409 | 61324.6 | 61393.1 | 61288.3 | 61380.3 |
| cb.5.250-2 | 62130 | 62036 | 62074 | 61969 | 61990 | 61894.4 | 61974.4 | 61801.6 | 61921.3 |
| cb.5.250-3 | 59463 | 59367 | 59446 | 59365 | 59349 | 59257.8 | 59331.2 | 59136.1 | 59275.6 |
| cb.5.250-4 | 58951 | 58914 | 58951 | 58883 | 58914 | 58725.6 | 58812.4 | 58693.6 | 58761.5 |
| cb.5.250-5 | 60077 | 60015 | 60015 | 59990 | 60015 | 59904.6 | 59970.4 | 59837.8 | 59951.2 |
| cb.5.250-6 | 60414 | 60355 | 60355 | 60348 | 60355 | 60208.2 | 60324.9 | 60230.6 | 60315.7 |
| cb.5.250-7 | 61472 | 61436 | 61436 | 61407 | 61401 | 61290.8 | 61341.8 | 61233.9 | 61343.9 |
| cb.5.250-8 | 61885 | 61829 | 61829 | 61790 | 61829 | 61737.1 | 61803.4 | 61644.9 | 61743.9 |
| cb.5.250-9 | 58959 | 58832 | 58866 | 58822 | 58851 | 58769.1 | 58786.9 | 58653.7 | 58782.8 |
| Average<br>p-value | 60413.5 | 60343 | 60361.8 | 60314.1 | 60328.8 | 60224.4 | 60289.0<br>5.27 e-06 | 60159.2 | 60261.1<br>1.85 e-05 |

and local search operators, contribute to the precision of the results. Finally, the PTRA distributions are closer to zero than *05.pe* distributions, indicating that PTRA has consistently better results than *05.pe*. When we evaluate the behaviour of the algorithms through the Wilcoxon test, this indicates that there is a significant difference between the two algorithms.

Our next step is trying to separate the contribution of local search and perturbation operator from the percentile transition operator. For this, we compared the algorithms *wpe* and *05.wpe*.

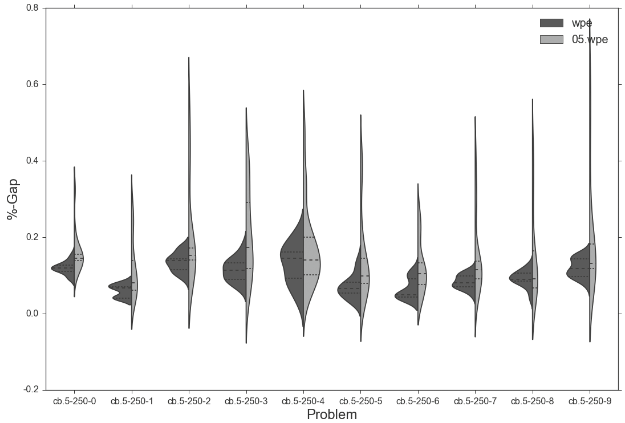


**Fig. 4.** Evaluation of percentile transition operator without perturbation and Local Search operators

When we check the Best Values shown in the Table 2, we note that *wpe* performs better than *05.wpe* in all problems except 3 and 7. However the results are quite close. In the case of the average indicator, *wpe* outperforms in all problems to *05.wpe*. The Wilcoxon test indicates that the difference is significant. This suggests that *wpe* is consistently better than *05.wpe*. In the violin chart

**Table 3.** OR-Library benchmarks MKP cb.5.500

| Instance | Best known | BAAA best | Avg | PTRA best | Avg | Time(s) | Std |
|---|---|---|---|---|---|---|---|
| 0 | 120148 | 120066 | 120013.7 | **120070** | 120022.6 | 343 | 26.8 |
| 1 | 117879 | **117702** | 117560.5 | 117690 | **117609.5** | 356 | 52.8 |
| 2 | 121131 | 120951 | 120782.9 | **121011** | **120918.0** | 354 | 39.5 |
| 3 | 120804 | 120572 | 120340.6 | **120609** | **120525.7** | 436 | 43.8 |
| 4 | 122319 | 122231 | 122101.8 | **122280** | **122151.5** | 418 | 48.5 |
| 5 | 122024 | 121957 | 121741.8 | **121982** | **121874.4** | 444 | 43.7 |
| 6 | 119127 | **119070** | 118913.4 | 119000 | **118931.0** | 449 | 30.1 |
| 7 | 120568 | 120472 | 120331.2 | **120487** | **120342.6** | 348 | 63.1 |
| 8 | 121586 | 121052 | 120683.6 | **121295** | **121196.5** | 326 | 63.5 |
| 9 | 120717 | **120499** | 120296.3 | 120485 | **120387.0** | 317 | 50.5 |
| 10 | 218428 | 218185 | 217984.7 | **218251** | **218200.2** | 339 | 32.6 |
| 11 | 221202 | 220852 | 220527.5 | **220946** | **220863.2** | 338 | 55.6 |
| 12 | 217542 | 217258 | 217056.7 | **217388** | **217295.9** | 315 | 43.5 |
| 13 | 223560 | 223510 | 223450.9 | **223526** | **223459.2** | 317 | 41.7 |
| 14 | 218966 | 218811 | 218634.3 | **218890** | **218814.4** | 318 | 36.9 |
| 15 | 220530 | **220429** | **220375.9** | 220410 | 220361.9 | 384 | 35.5 |
| 16 | 219989 | 219785 | 219619.3 | **219885** | **219767.2** | 369 | 60.0 |
| 17 | 218215 | **218032** | 217813.2 | 218027 | **217956.6** | 315 | 50.6 |
| 18 | 216976 | **216940** | **216862.0** | 216878 | 216840.0 | 354 | 23.1 |
| 19 | 219719 | 219602 | 219435.1 | **219622** | **219572.6** | 287 | 30.0 |
| 20 | 295828 | 295652 | 295505.0 | **295722** | **295662.9** | 270 | 32.2 |
| 21 | 308086 | 307783 | 307577.5 | **307972** | **307918.0** | 319 | 28.0 |
| 22 | 299796 | **299727** | 299664.1 | 299715 | **299673.8** | 246 | 22.4 |
| 23 | 306480 | **306469** | 306385.0 | 306439 | **306393.8** | 298 | 21.9 |
| 24 | 300342 | 300240 | 300136.7 | **300291** | **300221.6** | 273 | 29.5 |
| 25 | 302571 | 302492 | 302376.0 | **302503** | **302459.8** | 278 | 22.8 |
| 26 | 301339 | 301272 | 301158.0 | **301284** | **301257.2** | 267 | 21.5 |
| 27 | 306454 | 306290 | 306138.4 | **306385** | **306311.2** | 234 | 35.2 |
| 28 | 302828 | 302769 | 302690.1 | **302771** | **302723.4** | 245 | 29.6 |
| 29 | 299910 | 299757 | 299702.3 | **299844** | **299773.7** | 275 | 47.1 |
| Average | 214168.8 | 214014.2 | 213861.9 | 214055.3 | 213982.8 | 327.7 | 38.7 |

shown n the Fig. 4 it is further observed that the dispersion of the solutions for the case of *05.wpe* is much larger than in the case of *wpe*. This indicates that the operator percentile transition ranking plays an important role in the precision of the results.

### 4.2 PTRA Compared with BAAA

In this section we evaluate the performance of our PTRA with the algorithm BAAA developed in [19]. BAAA uses transfer functions as a general mechanism of binarization. In particular BAAA used the tanh $= \frac{e^{\tau|x|}-1}{e^{\tau|x|}+1}$ function to perform the transference. The parameter $\tau$ of the tanh function was set to a value 1.5. Additionally, a elite local search procedure was used by BAAA to improve solutions. As maximum number of iterations BAAA used 35000. The computer configuration used to run the BAAA algorithm was: PC Intel Core(TM) 2 dual CPU Q9300@2.5GHz, 4GB RAM and 64-bit Windows 7 operating system. In our PTRA algorithm, the configurations are the same used in the previous experiments. These are described in the Table 1.

The results are shown in Table 3. The comparison was performed for the set cb.5.500 of the OR-library. The results for PTRA were obtained from 30 executions for each problem. In black, the best results are marked for both indicators the Best Value and the Average. In the Best Value indicator, BAAA was higher in eight instances and PTRA in twenty two. In the averages indicator BAAA was higher in two instances, and PTRA in eighteen. We should also note that the standard deviation in most problems was quite low, indicating that PTRA has good accuracy.

## 5 Conclusions and Future Work

In this article, we proposed a algorithm whose main function is to binarize continuous swarm intelligence metaheuristics. To evaluate the performance of our algorithm, the multidimensional Knapsack problem was used together with the Cuckoo Search metaheuristic. The contribution of the main operator of the algorithm was evaluated, finding that the percentile transition ranking operator contributes significantly to improve the precision of the solutions. Finally, in comparison with state of the art algorithms our algorithm showed a good performance.

As a future works we want to investigate the behaviour of other metaheuristics. Furthermore, the algorithm must be verified with other NP-hard problems. Moreover to simplify the choice of the appropriate configuration, it is important to explore adaptive techniques. From an understanding point of view of how the framework performs binarization, it is interesting to understand how the algorithm alters the properties of exploration and exploitation. Also is interesting to study how the velocities and positions generated by continuous metaheuristics are mapped to positions in the discrete space.

**Acknowledgments.** Broderick Crawford is supported by grant CONICYT/ FONDECYT/REGULAR 1171243, Ricardo Soto is supported by Grant CONICYT /FONDECYT /REGULAR /1160455, and José García is supported by INF-PUCV 2016.

## References

1. Baykasoğlu, A., Ozsoydan, F.B.: An improved firefly algorithm for solving dynamic multidimensional knapsack problems. Expert Syst. Appl. **41**(8), 3712–3725 (2014)
2. Bhattacharjee, K.K., Sarmah, S.P.: Modified swarm intelligence based techniques for the knapsack problem. Appl. Intell. **46**, 158–179 (2016)
3. Chajakis, E., Guignard, M.: A model for delivery of groceries in vehicle with multiple compartments and lagrangean approximation schemes. In: Proceedings of Congreso Latino Ibero-Americano de Investigación de Operaciones e Ingeniería de Sistemas (1992)
4. Crawford, B., Soto, R., Astorga, G., García, J., Castro, C., Paredes, F.: Putting continuous metaheuristics to work in binary search spaces. Complexity **2017**, 19 (2017)
5. García, J., Crawford, B., Soto, R., Carlos, C., Paredes, F.: A k-means binarization framework applied to multidimensional knapsack problem. Appl. Intell., 1–24 (2017)
6. García, J., Crawford, B., Soto, R., García, P.: A multi dynamic binary black hole algorithm applied to set covering problem. In: International Conference on Harmony Search Algorithm, pp. 42–51. Springer (2017)
7. Gavish, B., Pirkul, H.: Allocation of databases and processors in a distributed computing system. Manage. Distrib. Data Process. **31**, 215–231 (1982)
8. Glover, F., Kochenberger, G.A.: The ant colony optimization metaheuristic: Algorithms, applications, and advances. In: Handbook of Metaheuristics, pp. 250–285 (2003)
9. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer (2011)
10. Kong, X., Gao, L., Ouyang, H., Li, S.: Solving large-scale multidimensional knapsack problems with a new binary harmony search algorithm. Comput. Oper. Res. **63**, 7–22 (2015)
11. Lanza-Gutierrez, J.M., Crawford, B., Soto, R., Berrios, N., Gomez-Pulido, J.A., Paredes, F.: Analyzing the effects of binarization techniques when solving the set covering problem through swarm optimization. Expert Syst. Appl. **70**, 67–82 (2017)
12. Leonard, B.J., Engelbrecht, A.P., Cleghorn, C.W.: Critical considerations on angle modulated particle swarm optimisers. Swarm Intell. **9**(4), 291–314 (2015)
13. Pirkul, H.: A heuristic solution procedure for the multiconstraint zero? one knapsack problem. Naval Res. Logist. **34**(2), 161–172 (1987)
14. Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.: Gsa: a gravitational search algorithm. Inf. Sci. **179**(13), 2232–2248 (2009)
15. Soto, R., Crawford, B., Olivares, R., Barraza, J., Figueroa, I., Johnson, F., Paredes, F., Olguín, E.: Solving the non-unicost set covering problem by using cuckoo search and black hole optimization. Natural Comput., January 2017
16. Wang, L., Zheng, X., Wang, S.: A novel binary fruit fly optimization algorithm for solving the multidimensional knapsack problem. Knowl.-Based Syst. **48**, 17–23 (2013)

17. Yang, X.-S.: Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Comput. **2**(2), 78–84 (2010)
18. Yang, X.-S., Deb, S.: Cuckoo search via lévy flights. In: 2009 World Congress on Nature & Biologically Inspired Computing, NaBIC 2009, pp. 210–214. IEEE (2009)
19. Zhang, X., Changzhi, W., Li, J., Wang, X., Yang, Z., Lee, J.-M., Jung, K.-H.: Binary artificial algae algorithm for multidimensional knapsack problems. Appl. Soft Comput. **43**, 583–595 (2016)

# SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments

Hichem Rahab[1](✉), Abdelhafid Zitouni[2], and Mahieddine Djoudi[3]

[1] ICOSI Labs, University of Khenchela,
BP 1252 El Houria, 40004 Khenchela, Algeria
rahab.hichem@univ-khenchela.dz

[2] Lire Labs, Abdelhamid Mehri Constantine 2 University,
Ali Mendjli, 25000 Constanine, Algeria
abdelhafid.zitouni@univ-constantine2.dz

[3] TECHNE Labs, University of Poitiers,
1 rue Raymond Cantel, 86073 Poitiers Cedex 9, France
mahieddine.djoudi@univ-poitiers.fr

**Abstract.** It is a challenging task to identify sentiment polarity in Arabic journals comments. Algerian daily newspapers interest more and more people in Algeria, and due to this fact they interact with it by comments they post on articles in their websites. In this paper we propose our approach to classify Arabic comments from Algerian Newspapers into positive and negative classes. Publicly-available Arabic datasets are very rare on the Web, which make it very hard to carring out studies in Arabic sentiment analysis. To reduce this gap we have created SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) a corpus dedicated for this work. Comments are collected from website of well-known Algerian newspaper Echorouk. For experiments two well known supervised learning classifiers Support Vector Machines (SVM) and Naïve Bayes (NB) were used, with a set of different parameters for each one. Recall, Precision and F_measure are computed for each classifier. Best results are obtained in term of precision in both SVM and NB, also the use of bigram increase the results in the two models. Compared with OCA, a well know corpus for Arabic, SIAAC give a competitive results. Obtained results encourage us to continue with others Algerian newspaper to generalize our model.

**Keywords:** Opinion mining · Sentiment analysis · Arabic comments · Machine learning · Natural Language Processing · Newspaper · Support Vector Machines · Naïve Bayes

## 1 Introduction

The proliferation of Internet use and application in our daily life we offer a large amount of data of several forms and about all domains, this treasury need a powerful means to take benefit from. A lot of available data in the web is constituted by user generated content (UGC) like product reviews, and comments submitted by users of Web sites

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_12

such as Epinions.com and Amazon.com [1], also in websites of Algerian newspaper such Echorouk[1], elkhabar[2],… etc., or television channels like Aljazeera[3] [2].

Sentiment analysis or opinion mining [3, 4], is a new field in the cross road of data mining and Natural Language Processing/Natural Language Understanding (NLP/NLU) [5] which the purpose was to extract and analyze opinionated documents and classify it into positive and negative classes [6, 7], or in more classes such as in [8] and [9]. Unlike data mining, where the work is to track meaningful knowledge from structured data, in sentiment analysis it is subject to find structured knowledge from unstructured amounts of data [10].

A challenging task in opinion mining is comments classification to their positive and negative sentiment toward article subject. This problem will be increased in the case of Arabic language due to the morphologic complexity and the nature of comments, for instance the behavior of the reviewers could be affected by the culture in Arabic countries [7].

We have organized this paper as fellows: In the second section related works are presented, Third section of the paper deals with the approach we carried out for sentiment analysis, and the different steps we fellow are detailed. Experimental evaluation techniques and metrics are explained in the section four. The fifth is dedicated to present and discuss obtained results. We conclude the work in the sixth section with giving perspectives to future works.

## 2 Related Works

An important baseline to conduct studies in opinion mining is the language resources, in [7] the OCA, a publicly available corpus, is designed to implement sentiment analysis applications for Arabic language. The authors collect 500 movie reviews from different web pages and blogs in Arabic, and they take benefit from the rating system of websites to annotate them as positive or negative. For experiments two machine learning algorithms, SVM and NB were used, and their performances are compared. A 10-fold cross validation method was implemented. The best results were obtained with SVM, and the use of trigram and bigram overcome the use of unigram model.

In their next work [11] the authors of OCA [7], using a Machine Translation (MT) tool, have translated the OCA corpus into English, generating the EVOCA corpus (English Version of OCA) contains the same number of positive and negative reviews. Following the work in [7] Support Vector Machines (SVM) and Naïve Bayes (NB) were applied for classification task. Obtained results are worse than OCA (90.07% of F_measure).

The authors in [6], in the goal of improving accuracy of opinion mining in Arabic language, they investigate the available OCA corpus [7], and they use the two well know machine learning algorithms SVM and NB with different parameters of SVM.

[1] http://www.echoroukonline.com.

[2] http://www.elkhabar.com.

[3] http://www.aljazeera.net.

Then 10, 15, and 20 fold cross validation were used. For SVM method the highest performance was obtained with Dot, Polynomial, and ANOVA kernels. For NB, its highest accuracy was achieved with BTO (Binary Term Accuracy).

In [8] and [9] a classification in five classes with SVM method was used. In [8] Arabic reviews and comments on hotels are collected from Trip Advisor website and classified into five categories: "ممتاز" (excellent); "جيد جدا" (very good); "متوسط" (middling); "ضعيف" (weak) and"مروع" (horrible), the modeling approach combined SVM with kNN provides the best result (F_measure of 97%). The authors in [9] proposed a three steps system consists of: corpus preprocessing, features extraction, and classification. The corpus used is obtained from Algerian Arabic daily Newspapers. So they focus on the second step where 20 features were used, and a combination of many SVM was used to classify comments into five classes.

The work in [12] investigate in the sentence and document levels. For the feature selection they start by the basic known feature model, the bag-of-words (BOW), where the feature model contain only the available words as attributes. A second type of feature model is created by adding the polarity score as attribute, using SentiWordNet via a machine translation step.

The work in [13] focus on Arabic tweets to study the effect of stemming and n-gram techniques to the classification process. Also the impact of feature selection on the performance of the classifier is studied. Support Vector Machines (SVM), Naïve Bayes, (NB), and K-nearest neighbor (KNN) are the used classifiers. We remark that the authors don't study the effect of changing parameters of different classifiers. In the results the authors mention that the use of feature selection technique improves significantly the accuracy of the three classifiers, and the SVM outperforms the other classifiers.

In [2] the authors used two available Arabic Corpora OCA created in [7] and ACOM which is collected from the web site of Aljazeera channel[4]. For the classification task, Naïve Bayes, Support Vector Machines and k-Nearest Neighbor were used. Stemming is investigated and the work concludes that the use of light-stemming is better than stemming. Obtained results show that the classification performance is influenced by documents length rather than the data sets size.

It is clear from this study of related works that publicly available resources for sentiment analysis in Arabic language are seldom. And those available are generally collected from movie reviews, which limit their domain of use. This fact makes it very important, for us, to create our proper corpus to be adequate with the purpose of our study.

## 3 Our Approach

In this step we will present our proposed approach for sentiment analysis in Algerian Arabic Newspaper. In our work we use RapidMiner tool kit[5], which is free for educational purpose. And we collect comments from the Algerian daily echorouk[6]. Figure 1 show the general approach we adopt.

[4] www.aljazeera.net.

[5] https://rapidminer.com/.

[6] www.echoroukonline.com.

### 3.1 Corpus Generation

We have created our corpus for Algerian Arabic SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) which mean in arabic (سياق, Context).

So we construct our corpus principally from the web site of echorouk newspaper[7]. The articles cover several topics (News, politic, sport, culture).

Compared to the important visitors of Algerian Newspaper websites, except Echorouk web site the number of comments is very low. And a lot of them are out of the main topic of the article, these comments are considered as neutral, and in the remained comments we found that negative ones largely outnumber the positives, which make for us a challenge to have equilibrium in our corpus. Table 1 present the statistics of comments collected from Echorouk web site for different categories in SIAAC before the equilibrium between different classes. It is clear that negative ones (91) largely outnumber other categories.

**Table 1.** Number of comments in SIAAC before equili-brium

| Positive | Negative | Neutral | **Total** |
|---|---|---|---|
| 32 | 91 | 24 | **147** |

Another difficulty is lied to the rating system, unlike review web sites such Amazon[8] for example where user can give points in scale e.g. from 0 to 5, to the article. In our case with web site Echorouk (and in the other Algerian newspaper websites) the user rather than giving points to the article, he can give one positive or negative point to other comments, this make annotation task very difficult, because having a certain number of points for a given comment have no meaning for positive or negative sentiment of the comments. So we must read carefully each comment and understand if it present a positive or negative sentiment, or even is off topic comment [14].

To generate our corpus, we take 92 comments, 60 comments from the negative category and all the 32 from the positive category (this is due to the nature of available comments -as mentioned above-) and this to have equilibrium in our corpus. The neutral comments are removed from our corpus and altered to a future work. Table 2 show the statics of our corpus.

**Table 2.** Number of comments from different categories in SIAAC after equilibrium

| Positive | Negative | **Total** |
|---|---|---|
| 32 | 60 | **92** |

[7] https://www.echoroukonline.com.

[8] https://www.amazon.com.

Despite the law number of comments in our generated corpus, we continue our study, due to the fact the important number of tokens per comment is 36.45, and is the documents length that influence in the classification performance more than the data set size [2].

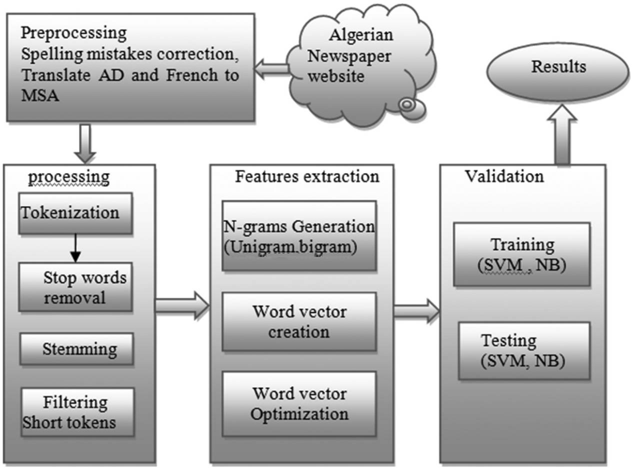


**Fig. 1.** Our approach general process

### 3.2 Pre Processing

To reduce the comment's vector size, some pre processing steps were conducted. Spelling mistakes were corrected manually, and words written in Algerian Dialect (AD) and in French are translated into Modern Standard Arabic (MSA), that stemming algorithms do not perform well with dialectical words and this dialectical words need an extended set of stopwords [15]. Characters encoding are resolved on UTF-8 (Table 3).

**Table 3.** Sample of comment manual correction

| Extraction from Original comment | Comment after correction |
|---|---|
| عندما يقول هذا السؤول بان البنزين غير مغشوش فلماذا اصبحنا نرانه ينفذ بسرعةهذا السوال موجه اليك لانك انت بنفسك لاتراعي لهذا لانك تعبء سيارتك باطل كل شيء من عند البايلك ولاتصرف عليه مليم واحدمن جيبك اطلب من بخابر نفطال ان تفسر لنا هذا مادمت انك تقول لاوجود للغش فبماذا تفسر ذالك ياسي ايزو | عندما يقول هذا المسؤول بان البنزين غير مغشوش فلماذا اصبحنا نرانه ينفذ بسرعة هذا السؤال موجه اليك لانك انت بنفسك لاتهتم لهذا لانك تعبيء سيارتك مجانا كل شيء من الدولة ولاتصرف عليه مليم واحد من جيبك اطلب من مخابر نفطال ان تفسر لنا هذا مادمت انك تقول لاوجود للغش فبماذا تفسر ذالك ياسيد ايزو |

### 3.3 Comments Processing

Before feature extraction, a sequence of processing steps were carried out that each comment goes through.

**Tokenization**
Each token representing a word, in this process we use simply spaces between words.

**Stop Words removal**
We note that some authors such in [2] recommend that these lists should be hand crafted as it is domain and language-specific.

**Stemming**
In stemming the words are reduced to their roots known as the base form or stem [16]. There are two different stemming techniques; generally stemming simply called stemming and light-stemming. For our work we use the basic Arabic stemmer.

**Filtering Short tokens**
Tokens with less than two letters were removed because of their low significance in opinion mining task.

### 3.4 Feature Selection

Feature selection is a process that selects a subset of original features to be used in the classification task [17]. The optimality of a feature subset is measured by some evaluation criterions [18]. We have used several feature selection parameters and different results are computed and compared.

**N-grams Generation**
In this work two n-grams were generated: Unigram and bigram.

**Word vector creation**
To create vector representing all comments, we use four different parameters:

1. **Term Frequency (TF):** A ratio representing the number of term occurrences over the total number of words.
2. **Term Frequency Inverse Document Frequency (TF-IDF):** Define the weight of a term in the context of a document [8].
3. **Term Occurrences (TO):** Each element represent the number the word occur in the comment (0 to n).
4. **Binary Term Occurrences (BTO):** The element takes 1 if the word appears at least once in the comment and 0 otherwise.

**Word vector optimization**
Very occur words and very rare ones have a low significance in opinion mining [17]. So we eliminate words appear less than 3% in corpus, and those appear more than 30%.

# 4 Experimental Evaluations

In this section, the proposed system is evaluated. Several experiments have been accomplished. We have used cross-validation to compare the performance of two of the most widely used learning algorithms: SVM and NB.

In our experiments, the 10-fold cross-validation has been used to evaluate the classifiers.

## 4.1 Classifiers

For the classification task, two well known supervised learning methods were utilized: support vector machine (SVM) and naïve Bayes (NB).

**Support vector machines**

The Support vector machines (SVM) are a widely used classifier in different disciplines, due to its ability to modeling diverse sources of data, their flexibility in handling data of high-dimensionality, and the high obtained accuracy.

**Naïve Bayes**

The Naive Bayes classifier is a well known algorithm used in text classification. In the "Naive Bayes assumption" all attributes of the examples are independent of each other given the context of the class [19]. If document belongs to different classes with different probabilities, it is classified in the class that have the highest posterior probability [13].

## 4.2 Performance Measures

Performance measures are defined and computed from this table; three parameters were used, precision, recall, and $F_1$_measure (or too simply F_mesure).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Precision and recall are complementary one to the other, we combine the two using the $F_1$ measure called generally $F_1$, given as:

$$\mathrm{F}_1\_\mathrm{measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

# 5 Results and Discussion

## 5.1 Evaluation with Support Vector Machines

In the scope of our work, we have applied the SVM algorithm with two different kernel types, Anova and polynomial. And both Unigram and Bigram models were tested. For the comments vector creation four variants was implemented, Term occurrence (TO), Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), and Binary Term Frequency (BTO).

Best results are obtained in precision when TF and TF-IDF are suited for vector creation and this for both SVM implemented kernels. The use of bigram increase results in most of cases (Table 4).

**Table 4.** Results with SVM

| | | Unigram | | | Bigram | | |
|---|---|---|---|---|---|---|---|
| | Kernel | F_measure | Precision | Recall | F_measure | Precision | Recall |
| Term occurrence | Anova | 79.64% | 88.33% | 72.50% | 79.45% | 93.33% | 69.17% |
| | Polynomial | 61.19% | 47.80% | 85.00% | 62.93% | 48.87% | 88.33% |
| Term frequency | Anova | 82.86% | 96.67% | 72.50% | 83.50% | **100%** | 71.67% |
| | Polynomial | 67.16% | 88.33% | 54.17% | 75.00% | **100%** | 60.00% |
| TF-IDF | Anova | 81.78% | 100% | 69.17% | 79.39% | **100%** | 65.83% |
| | Polynomial | 70.11% | 96.67% | 55.00% | 63.64% | **100%** | 46.67% |
| BTO | Anova | 83.00% | 91.67% | 75.83% | 84.78% | **97.50%** | 75.00% |
| | Polynomial | 69.58% | 57.40% | 88.33% | 68.31% | 55.69% | 88.33% |

For comparing the results of our SIAAC corpus with OCA used in [7], we choose the term frequency vector and the anova kernel for SVM, and this because the work with OCA in [7] does not testing several SVM parameters (Table 5).

**Table 5.** Comparison between SIAAC And OCA in SVM classification

| | N-gram model | Precision | Recall | Other metrics |
|---|---|---|---|---|
| SIAAC | Unigram | **96.67%** | 72.50% | F_measure = 82.86% |
| | Bigram | **100%** | 71.67% | F_measure = 83.50% |
| OCA in [7] | Unigram | 86.99% | **95.20%** | ACC = 90.20% |
| | Bigram | 87.38% | **95.20%** | ACC = 90.60% |

The results show that in term of precision SIAAC outperform OCA, which mean that our predictive results are more important. In the other hand the recall of OCA is better than SIAAC which indicate that OCA documents are well classified than SIAAC ones.

### 5.2 Evaluation with Naïve Bayes

As with SVM best results are found in precision, with the different vector creation methods. And the use of bigram also increases the obtained results (Table 6).

**Table 6.** Results with Naïve Bayes

| | Uni-gram | | | Bi-gram | | |
|---|---|---|---|---|---|---|
| | F_measure | Precision | Recall | F_measure | Precision | Recall |
| Term occurrence | 80.05(%) | 95.00(%) | 69.17(%) | 80.93(%) | 97.50(%) | 69.17(%) |
| Term frequency | 80.05(%) | 95.00(%) | 69.17(%) | 80.93(%) | 97.50(%) | 69.17(%) |
| TF-IDF | 72.15(%) | 90.48(%) | 60.00(%) | 73.55(%) | 95.00(%) | 60.00(%) |
| BTO | 80.93(%) | 97.50(%) | 69.17(%) | 81.78(%) | 100(%) | 69.17(%) |

As for the SVM, we will compare our the results of our corpus in classification with Naïve Bayes classifier with OCA corpus [7] (Table 7).

**Table 7.** Comparison between SIAAC And OCA in NB Classification

| | n-gram Model | Precision | Recall | Other metrics |
|---|---|---|---|---|
| SIAAC | Unigram | **97.50**(%) | 69.17(%) | F_measure = 80.93% |
| | Bigram | 81.78(%) | **100**(%) | F_measure = 69.17% |
| OCA in [7] | Unigram | 79.99(%) | **85.60**(%) | Acc = 81.80% |
| | Bigram | **82.75**(%) | 88.80(%) | Acc = 84.60% |

In this case our system SIAAC outperforms OCA in term of precision when using Uigram model, and in term of recall when using the bigram model.

## 6 Conclusion and Future Works

Our exploration of obvious achieved works in sentiment analysis, especially in Arabic language, show the lack in publicly available resources dedicated for carried out Arabic sentiment analysis studies. And in these rarely available ones, we found that the most are interested by movie reviews; perhaps because of the important number of web sites inciting people to review films and serials. Such available resources are not adapted to use in other domains such as newspaper comments sentiment analysis which cover several topics like politics, culture, sports, medicine, etc.

In this work we present our approach of sentiment analysis in Algerian Arabic daily newspapers. The approach starts with the corpus creation where 92 comments are collected from echorouk newspaper web site. And SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) was created to be used in the scope of this study. Some processing operations were conducted.

Comments are represented in different vector models, TO, TF, BTO and TF IDF, also with unigram and bigram models. For classification, two well known methods are used, support vector machines SVM and naïve bayes NB. And different parameters for each classifier were tested. In the validation process 10-fold cross validation method was conducted to train and test the models.

Obtained results are very promising, in term of precision and recall. But in term of F_measure as a compromise between precision and recall the results remain modest which need more work to improve this rate.

Compared with the well know corpus for Arabic OCA, our approach SIAAC give a competitive results, for the SVM model SIAAC outperform OCA in classification precision. These results encourage us to continue in this issue in future works.

As perspective to this work we would to add the neutral class which allow us using all available comments (or at least an important part of them), so other methods can be implemented or a combination of existing methods can be used to resolve the multi classes problem. Also the corpus must be enriched by comments from other Algerian newspaper to generalize the model.

Another point is to dealing with comments written in Algerian dialect and French language directly without the need of the manual translation. In this point machine translation may be envisaged as an automated process.

Also it is very important to take in consideration in future works, the article topic when searching the sentiment orientation of comments which can allow us considering more comments in the classification step, that a lot of comments are removed from our corpus due to the fact that their semantic orientation is strongly related to the article topic.

## References

1. Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W.: Sentiment analysis of Chinese documents: from sentence to document level. J. Am. Soc. Inf. Sci. Technol. **60**(12), 2474–2487 (2009)
2. Mountassir, A., Benbrahim, H., Berraba, I.: Sentiment classification on arabic corpora. A preliminary cross-study. Doc numérique **16**(1), 73–96 (2013)
3. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
4. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing (2010)
5. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, vol. 5. John Benjamins Publishing Company, Amsterdam (2002)
6. Atia, S., Shaalan, K.: Increasing the accuracy of opinion mining in Arabic. In: Proceedings —1st International Conference on Arabic Computational Linguistics: Advances in Arabic Computational Linguistics ACLing 2015, pp. 106–113 (2015)
7. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M.: OCA: opinion corpus for Arabic. J. Am. Soc. Inf. Sci. Technol. **62**(10), 2045–2054 (2011)
8. Cherif, W., Madani, A., Kissi, M.: Towards an efficient opinion measurement in Arabic comments. Procedia Comput. Sci. **73**(Awict), 122–129 (2015)
9. Ziani, A., Tlili Guaissa, Y., Nabiha, A.: Détection de polarité d'opinion dans les forums en langues arabe par fusion de plusieurs SVMs. **7**, 17–21 (2013)

10. Salloum, S.A., Al-emran, M., Monem, A.A., Shaalan, K.: A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. **2**(1), 127–133 (2017)
11. Rushdi-Saleh, M., Martín-Valdivia, M.: Bilingual experiments with an Arabic–English corpus for opinion mining. In: Proceedings on International Conference on Recent Advances in Natural Language Processing 2011, pp. 740–745, September 2011
12. Alotaibi, S.S., Anderson, C.W.: Extending the knowledge of the Arabic sentiment classification using a foreign external lexical source. Int. J. Nat. Lang. Comput. **5**(3), 1–11 (2016)
13. Brahimi, B., Touahria, M., Tari, A.: Data and text mining techniques for classifying arabic tweet polarity. J. Digit. Inf. Manag. **14**(1), 15–25 (2016)
14. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. O'Reilly Media, Inc. (2012)
15. Duwairi, R.M.: Sentiment analysis for dialectical Arabic. In: 6th International Conference on Information and Communication Systems (ICICS), 2015, pp. 166–170, February 2015
16. El-defrawy, M.: Enhancing root extractors using light stemmers. In: 29th Pacific Asia Conference on Language, Information and Computation, pp. 157–166 (2015)
17. Agarwal, B., Mittal, N.: Prominent feature extraction for sentiment analysis (2016)
18. Huan, L., Lei, Y.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**(4), 491–502 (2005)
19. McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: AAAI/ICML-98 Work Learning Text Category, pp. 41–48 (1998)

# Enrichment Ontology Instance by Using Data Mining Techniques

## A Case of Thai Tourist Interest in Culture Tourism

Kunyanuth Kularbphettong(✉)

Computer Science Program, Suan Sunandha Rajabhat University, Bangkok 10300, Thailand
kunyanuth.ku@ssru.ac.th

**Abstract.** Ontology is an agreement about a shared conceptualization, which includes frameworks for modeling domain knowledge and agreements about the representation of particular domain theories, often captured in some form of a semantic web formally. However, building ontology is a time consuming task. however, the paper was presented an approach to enrich instances into the exiting ontology and this research presented the technique to extract information from the unstructured text from websites. Support vector machine was used to create model. The results showed that feature reduction and SVM techniques presented the highest precision than SVM approach.

**Keywords:** Cultural tourism · Support vector machine · Ontology enrichment

## 1 Introduction

Cultural tourism plays as an important role of the tourism industry because there are diverse and versatile forms to supply tourists to learn and understand different cultures. Cultural tourism is defined as the activity tourist enables to experience the different ways of life of other people and gains at first hand an understanding of their customs, traditions, and the physical environment [1].

Cultural tourism is an educational tour of knowledge in areas of historical and cultural importance that tells the story of social development through history as a result of cultural relevance. This knowledge is valuable and can be reflected the natural environment, living conditions and well-being of people in each era as well. Nowadays, there are many websites that offer information related to tourists who are searching for information on the Internet. However, it was found that there is a huge of information related to travel on the Internet and it makes difficult and waste time to find the corrected and relevant information.

With the advance of technology, ontology is an agreement about a shared conceptualization, which includes frameworks for modeling domain knowledge and agreements about the representation of particular domain theories, often captured in some form of a semantic web formally. Also, data mining techniques is the promising methodology to extract valuable information in this objective and it can analyze relevant information results and produce different perspectives to understand more about

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_13

extracted knowledge. Therefore, this research aims to enrich the concept of developing knowledge related to cultural tourism by utilizing ontological theory to categorize tourist information and using data mining techniques to help categorize information and to classify tourists' interest in cultural tourism faster.

And the rest of this paper is organized as follows. Section 2 reviews about the related literatures and the related methodologies used in this work. Section 3 presents the implementation based on the purposed data mining techniques. In Sect. 4 the result is presented. Finally, Sect. 5 was shown the conclusion and the future research issues.

## 2 Literature Reviews

A literature search shows that most of the related researches have deployed data mining techniques to classify the concept of developing knowledge related to cultural tourism by following this: According to Hiep Luong et al. [2], the research was shown how to implement and validate an ontology learning framework to enrich the vocabulary of the domain ontology from Web documents related to the domain and from the WordNet semantic lexicon. Also, feature extraction based on TF-IDF and classification based on SVM was applied to classify two BBC datasets and five groups of 20 Newsgroup datasets [3]. The effectiveness of word2vec and tf-idf can outperform because word2vec provides complementary features [4]. Similarity measure based on the size of the intersection of the sets of variables is important for the class separation of the instances in both input ontologies by using a VC dimension variable selection criterion elaborated for support vector machines [5]. Also there are many researches that have been investigated in variable selection improved by SVM learning in practice [6, 7].

## 3 The Methodologies

Data Mining is the data analyzing process from different perspectives also summarizing the useful information results. The data mining process uses many principles as machine learning, statistics and visualization techniques to discover and present knowledge in an easily comprehensible form. There is another definition as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [8, 9].

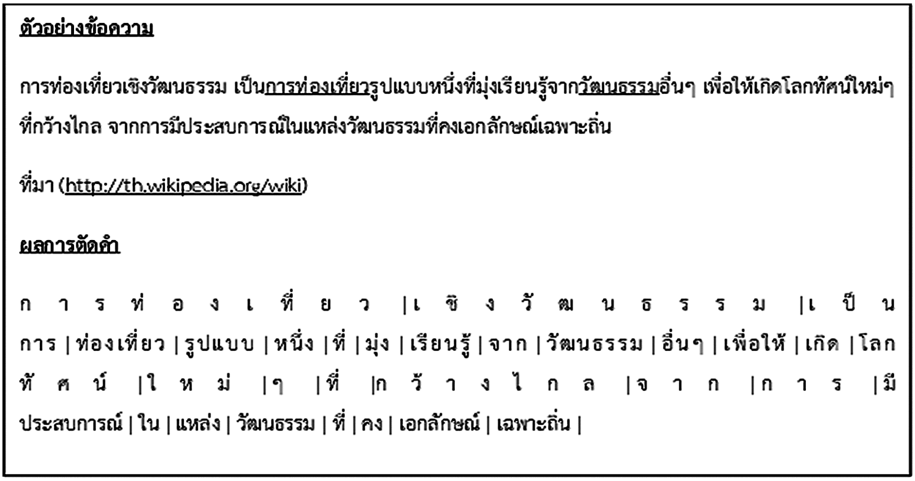
SVMs (Support Vector Machines) are a useful technique for data classification technique. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes [10]. Also, Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [11, 12].

Ontology is a popular feature in many appliances in order to provide a semantic framework for knowledge management. Ontology refers to a content representing specific knowledge primitives (classes, relations, functions and constants). Ontology stands for the hierarchical knowledge structure about things by subcategorizing them by their essential qualities.

## 4 Experimental Setup

Building ontologies is a time consuming and complex task, requiring a high degree of human supervision. To develop the knowledge of cultural tourism, ontology acts as a basis for the development of cultural tourists' interest and data mining techniques was used to enrich the concept of ontology. The domain cultural tourism ontology, an existing small, manually-created ontology, is used to capture the semantic information among different terms in the documents. Data was collected from Thai tourism websites related to cultural tourism information. The research was divides in to three phrases as following: preprocessing phrase, feature extraction, and performance testing.

In preprocessing phrase, data used in the research was online and html tags and tables were included. Therefore, data was eliminated the irrelevant and non-trivial data and removed html tags. Thai language may have more restrictions on natural language processing than English because Thai does not have a punctuation mark and there is no represents ending sentences. Thai word segmentation is different from English and this research used longest matching technique to word segmentation as shown in Fig. 1.



ตัวอย่างข้อความ

การท่องเที่ยวเชิงวัฒนธรรม เป็นการท่องเที่ยวรูปแบบหนึ่งที่มุ่งเรียนรู้จากวัฒนธรรมอื่นๆ เพื่อให้เกิดโลกทัศน์ใหม่ๆ
ที่กว้างไกล จากการมีประสบการณ์ในแหล่งวัฒนธรรมที่คงเอกลักษณ์เฉพาะถิ่น

ที่มา (http://th.wikipedia.org/wiki)

ผลการตัดคำ

การท่องเที่ยว |เชิงวัฒนธรรม |เป็น
การ | ท่องเที่ยว | รูปแบบ | หนึ่ง | ที่ | มุ่ง | เรียนรู้ | จาก | วัฒนธรรม | อื่นๆ | เพื่อให้ | เกิด | โลก
ทัศน์ |ใหม่ |ๆ |ที่ |กว้างไกล |จาก |การ |มี
ประสบการณ์ | ใน | แหล่ง | วัฒนธรรม | ที่ | คง | เอกลักษณ์ | เฉพาะถิ่น |

**Fig. 1.** Example of Thai word segmentation

In the feature extraction phrase, the keywords based on the frequency of the occurrence were indexed to consider the word appeared in text related to cultural tourism. TF-IDF (Term Frequency-Inverse Document Frequency) method was used to determine the weight of keywords in the document [13]. This is a method for evaluating the importance of words in a document by calculating TF-IDF as shown in equations.

$$W_{(\mathrm{f,d})} = TF_{(\mathrm{f,d})} \times IDF_{(\mathrm{f})} \tag{1}$$

$$IDF_{(\mathrm{f})} = log \frac{|D|}{\left|DF_{(\mathrm{f})}\right|} \tag{2}$$

Where $W_{(\mathrm{f,d})}$ is the weight of the characteristic (f) in the document (d).

$TF_{(\mathrm{f,d})}$ is the frequency of the characteristic (f) in the document (d).

$|D|$ is the total number of documents in the training set.

$|DF_{(\mathrm{f})}|$ is the number of all documents with the attribute (f) listed.

The results from the calculation were taken into account the attributes that give the TF-IDF a high value in each category of key words. The number of features was chosen from ascending and then converted to the form of a matrix of documents (Document-Terms Matrix) as shown in Table 1.

**Table 1.** Document-terms matrix

| **Document** | **Features** | | | | **Class** |
|---|---|---|---|---|---|
| | **Feature$_1$** | **Feature$_2$** | **…** | **Feature$_j$** | |
| Doc$_1$ | F$_{11}$ | F$_{12}$ | … | F$_{1j}$ | วัด |
| Doc$_2$ | F$_{21}$ | F$_{22}$ | … | F$_{2j}$ | อาคาร |
| Doc$_3$ | F$_{31}$ | F$_{32}$ | … | F$_{3j}$ | พระราชวัง |
| Doc$_i$ | F$_{i1}$ | F$_{i2}$ | … | F$_{ij}$ | พิพิธภัณฑ์ |

Performance testing is measured with precision, recall, and F-measure.

$$precision = \frac{\text{number of correct positive predictions}}{\text{number of positive predictions}} \times 100 \tag{3}$$

$$recall = \frac{\text{number of correct positive prediction}}{\text{number of positive example}} \times 100 \tag{4}$$

$$F - measure = \frac{2 * \text{precision} * \text{Recall}}{\text{precision} + \text{Recall}} \tag{5}$$

## 5 Results of Experimental

The data used in the experiments was corrected from 70 Thai cultural tourism websites (20,000 words approximately) and the preprocessing applied to handle with these documents. F-measure was used to evaluate the performance of the classification of attraction category as shown in Table 2.

**Table 2.** Classification performance

| Class | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|
| | SVM | Feature reduction and SVM | SVM | Feature reduction and SVM | SVM | Feature reduction and SVM |
| 1 | 67.67 | 75.46 | 53.33 | 73.45 | 56.34 | 71.96 |
| 2 | 65.81 | 77.58 | 62.46 | 77.45 | 62.65 | 77.65 |
| 3 | 53.33 | 79.12 | 67.45 | 77.33 | 65.32 | 75.24 |
| 4 | 63.62 | 79.84 | 64.23 | 77.69 | 64.34 | 75.78 |

The results were shown that feature reduction and SVM techniques presented the highest precision than SVM approach. However, some words in Thai are still provided less accuracy to classify because of ambiguous of words.

## 6 Conclusion

This paper presented the method of developing knowledge related to cultural tourism by utilizing ontological theory to categorize tourist information and using data mining techniques to help categorize information and to classify tourists' interest in cultural tourism. The results showed that feature reduction and SVM techniques presented the highest precision than SVM approach. However, in term of the future experiments, other text mining techniques will be investigate to research about to enhance this project and also apply the tool to extend concepts in the experiment.

**Acknowledgment.** The authors gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

## References

1. ICOMOS Charter on Cultural Tourism (1976). http://www.icomos.org/tourism/tourism_charter.html
2. Luong, H., Wang, Q., Gauch, S.: Ontology Learning Using Word Net Lexical Expansion and Text Mining. INTECH Open Access Publisher, Rijeka (2012)
3. Lilleberg, J., Zhu, Y., Zhang, Y.: Support vector machines and word2vec for text classification with semantic features. In: Proceedings of IEEE ICCI*CC, pp. 136–140, July 2015
4. Todorov, K., Geibel, P.: Variable selection as an instancebased ontology mapping strategy. In: Proceedings of the 2009 International Conference on Semantic Web and Web Services, pp. 3–9. CSREA Press (2009)
5. Joachims, T.: Text categorization with support vector machines: learning with many relevant features (1998). www.cs.comell.edu

6. Polpinij, J., Ghose, A.K.: An ontology-based sentiment classification methodology for online consumer reviews. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 518–524. IEEE Computer Society (2008)
7. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
8. Dadgar, S.M.H., Araghi, M.S., Farahani, M.M.: A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH), pp. 112–116 (2016)
9. Fayyad, U.M., Pitatesky-Shapiro, G., Smyth, P., Uthurasamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Cambridge (1996)
10. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge discovery in databases: an overview. AI Mag., 213–228 (1992). Fall
11. Jakkula, V.: Tutorial on Support Vector Machine. School of EECS, Washington State University (2006)
12. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. J. Mach. Learn. Res. **2**, 125–137 (2002)
13. Tf-idf. en.wikipedia.org

# Solving the Set Covering Problem Using Cat Swarm Optimization Algorithm with a Variable Mixture Rate and Population Restart

Broderick Crawford, Ricardo Soto, and Hugo Caballero(✉)

Pontificia Universidad Católica de Valparaíso,
Avenida Brasil 2950, 2374631 Valparaśo, Chile
{broderick.crawford,ricardo.soto}@pucv.cl, hcaballec@gmail.com

**Abstract.** Cat swarm optimization (CSO) is a novel metaheuristic based on swarm intelligence, presented in 2006 has demonstrated great potential generating good results and excellent performances simulating the behavior of domestic cats using two behavior: seeking and tracing mode, this mode are classified using a mixture rate (MR), this parameter finally defines the number of individuals who work by exploring and exploiting. This work presents an improvement structure of a binary cat swarm optimization using a total reboot of the population when loss diversity it is detected.

**Keywords:** Metaheuristics · Combinatorial Optimization · Diversity loss

## 1 Introduction

The Set Covering Problem (SCP) is a NP-hard combinatorial optimization problem which it has multiple practical applications, for example: air and maritime ports, factories, warehouses, retail outlets, schools, hospitals, bus stops, subway stations, electronic switching centers, Daskin introduces the term Facility to identify public and private services centers, to find where to install these facility satisfying certain restrictions. The main objective of the SCP consists is to find these places with a minimun cost and distributing efficiently these centers.

The SCP because is a classical question in combinatorics, computer science and complexity theory, many works have been written to solve it using different techniques such as: Genetic algorithm [1], Particle Swarm Optimization [2] and algorithm based in the nature [3], in this work we will use Binary Cat Swarm Optimization [4] a novel metaheuristic based on swarm intelligence presented in 2006 modifying one of the main parameters that defines how many cats will be exploiting or exploring into spaces search.

## 2 Set Covering Problem

The set covering problem (SCP) objective is to locate the minimum number of facilities required to satisfy all of the demand [5]. It is one of Karp's 21

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_14

NP-complete problems shown to be NP-complete in 1972 [6]. There are 2 ways to solve it, using exact methods and approximate algorithm, Heuristics or metaheuristics. SCP is formally defined as fallows: Let A $= (a_{ij})$ it is zero-one matrix with *m-rows n-columns*. We say that a column $j$ cover a row $i$ if $a_{ij} = 1$. Each column $j$ is associated with a non-negative real cost $c_j$. Let $I = 1, ..., m$ and $J = 1, .., n$ be the row set and column set, respectively. The SCP calls a minimal cost to a subset $S \subseteq J$ such that each row $i \in I$ is covered by at least one column $j \in J$. A mathematical model for the SCP is

$$Minimize\ Z = \sum_{j=1}^{n} c_j x_j \qquad j \in \{1, 2, 3, ..., n\} \tag{1}$$

Subject to:

$$\sum_{j=1}^{n} a_{ij} x_j \geq 1 \qquad i \in \{1, 2, 3, ..., m\} \tag{2}$$

$$x_j \in \{0, 1\} \tag{3}$$

The objective function (1) minimize the total fixed cost of the siting configuration rather than the number of facilities sited. Constraint set (2) ensures that each demand node is covered by at least one facility. Constraint set (3) enforces the yes or no nature of the siting decision.

## 3 Cat Swarm Optimization

Chu and Tsai [7] presented CSO algorithm for solving optimization problems simulating the conduct of domestic cats. This algorithm is based on a swarm intelligent [8] of $N$ individuals(cats) working in $M$ dimensional space and over the cat behavior defined by two states: seeking mode and tracing mode that define the movement of the population. To ensure the balance between of these states, the population is divided using a mixture rate (MR) and represent a percentage of the population, $N$. Chu and Tsai propose a low value of MR [7] favoring the exploration, however this value will depend on the problem that needs to be solved.

- Seeking Mode, exploitation. Resting but being alert - looking around its environment for the next move improving the position, $C_{sm}$ will represent the cats in mode seeking
$$C_{sm} = MR \cdot N \tag{4}$$
- Tracing Mode, exploration. Running to the prey, $C_{tm}$ will represent the cats in mode tracing
$$C_{tm} = N - C_{sm} \tag{5}$$

The CSO was originally developed for continuous spaces, however there are a lot optimization problems, like as SCP, in which is necessary work in discrete space. Sharafi et al. proposed the discrete version of this technique [4], Binary Cat Swarm Optimization (BCSO). The main difference between CSO and BCSO is the representation of the vector position of each cat which composed only of ones and zeros. The cat behaviors are described below.

### 3.1 Seeking Mode (SM: Resting and Observing).

The SM corresponds to a global search technique in the search space of the optimization problem. In this way SMP copies of the cat is done. There are four main parameters [4]:

- SMP Seeking memory pool.
- SRD Seeking Range of the selected dimension
- CDC Counts Dimension Change
- SPC Self-Position Consideration

It's steps are:

1. Make $(SMP - 1)$ copies $cat_k$ actual position. If the value of SPC is true, let $j = (SMP - 1)$, then retain the present position as one of the candidates.
2. For each $(SMP - 1)$ copies, and using CDC, randomly plus or minus SRD percent the present value and replace the old ones.
3. To obtain the fitness values (FS) of all candidate.
4. If all FS are not exactly equal, calculate the selecting probability of each candidate point by (4), else set the selecting probability in 1 for all copies.

$$P_i = \frac{|SSE_i - SSE_{max}|}{SSE_{max} - SSE_{min}} \tag{6}$$

5. Randomly pick the point to move to from the candidate points, and replace the position of $cat_k$. If fitness function is a minimum solution, $FSb = FSmax$, otherwise $FSb = FSmin$

### 3.2 Tracing Mode (TM): Running After a Target

In this mode, the cats are moving following the targets spending high energy. Define position and velocity of $i_{th}$ cat in the *D-dimensional* space as $X_i$=($X_{i1}$,$X_{i2}$,$X_{i3}$ $\ldots X_{iD}$) and $V_i$=($V_{i1}$,$V_{i2}$,$V_{i3}$ $\ldots V_{iD}$) where $(1 \leq d \leq D)$ is the dimension. The global best position of the cat swarm is represented as $X_g$=($X_{g1}$,$X_{g2}$,$X_{g3}$ $\ldots X_{gD}$). The steps involved in tracing mode are:

1. Compute the new velocity of *i th* cat using:

$$V_{id} = w \ V_{id} + c \ r \ (X_{gd} - X_{id}) \tag{7}$$

where w is the inertia weight, c is the acceleration constan and r is a random number uniformly distributed in the range [0,1]

2. Compute the new position of *i th* cat using

$$V_{id} = X_{gd} - X_{id} \tag{8}$$

3. If the new position of *i th* cat corresponding to any dimension goes beyond the search space, then the corresponding boundary value is assigned to that dimension and the velocity corresponding to that dimension is multiplied by -1 to continue the search in the opposite direction.

The main process of CSO is described in Algorithm 1.

**Algorithm 1.** BCSO MAIN ALGORITHM

```
Input: $MR, SMP, SRD, CDC, SPC, NumGen, SizePob$
Output: The best Fittness
$Cat_{SizePob} \leftarrow CreateBinaryCatPopulation(SizePob, MR)$
$Iteration_i \leftarrow 1$
while $Iteration_i \leq NumGen$ do
    for $i \leftarrow 1$ to $SizePob$ do
        if $Cat[i]$ is $SeekingMode$ then
            $SolutionCandidate \leftarrow applySeekingMode(Cat[i])$
        else
            $SolutionCandidate \leftarrow applyTracingMode(Cat[i])$
        $Fitness_{best} \leftarrow EvaluateBestSolution(Cat_k, Fitness_{bestbefore})$
    $Cat_{SizePob} \leftarrow RepickCats(Cat_{SizePob}, MR)$
    $Iteration_i \leftarrow Iteration_i + 1$
return $Fitness_k$
```

Where

- NumGen: Iterations number
- CreateBinaryCatPopulation(SizePob, MR): Randomly create the binary cats population(velocity and position) and distribute it according to $MR$
- EvaluateBestSolution( Cat[i],$Fitness_k$ ): Compute the fitness, and keep the best fitness
- applySeekingMode( Cat[i] ): Execute Seeking Mode in according to MR
- applyTracingMode( Cat[i] ): Execute Tracing Mode in according to MR
- RepickCats(SizePob, MR): Re-pick number of cats and set them into tracing mode according to MR, and set the other into seeking mode

## 4 Algorithm Proposed. Binary Cat Swarm Optimaztion Using Restart Population: BCSO-RP

Our purpose is to introduce a mechanism to improve the results obtained for SCP in previous studies. The basic idea is, to use original BCSO to determine the best MR that resolves the SCP instance, after that, run a new BCSO using the best MR and apply population restart when the premature convergence is detected. The premature convergence happens in all virtual population working with search methods, this behavior it is related when the population loses the capacity to generate new a better solutions, this process is known as "diversity loss". Our strategy will be to regenerate [9,10] the population to explore unvisited regions in the search spaces when the solution does not improve. The effect of this mechanism it is showed below Figs. 1 and 2. Therefore our proposal will contain 2 phases: Phase one; Run BCSO varying MR from 0.0 to 1.0, we seek the best rate in a minimal the iterations, and Phase two; Determined MR(phase 1), we execute BCSO controlling premature convergence and restarting the population when it's detected (Table 1) (Fig. 3).

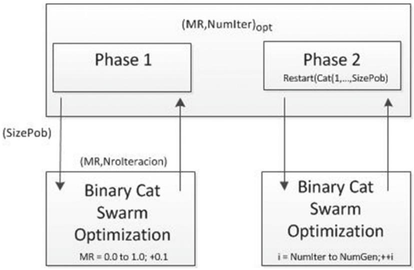

(MR,NumIter)$_{opt}$
Phase 1
Phase 2
Restart(Cat(1,...,SizePob)
(SizePob)
(MR,NroIteracion)
Binary Cat Swarm Optimization
MR = 0.0 to 1.0; +0.1
Binary Cat Swarm Optimization
i = NumIter to NumGen;++i

**Fig. 1.** BCSO-RP, MR tunning and restart population

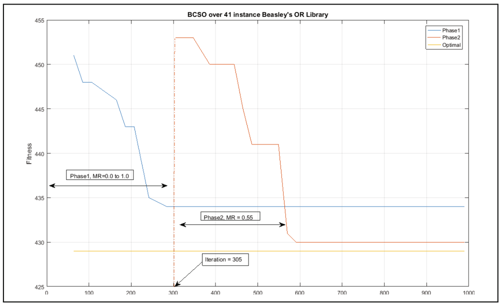


**Fig. 2.** Effect of proposal( Phase 1 and 2) over instance 41. We improved the objective function

**Algorithm 2.** BCSO-RP: BCSO WITH RESTART POPULATION

```
Input: $SMP, SRD, CDC, SPC, NroGen, SizePob$
Output: Best Fittness
$Iteration_i \leftarrow 1$
$(MR_{best}, MR_{iter}) \leftarrow DetermineBestMR(SizePob)$  /* Best MR using a minimal iteration */
$Iteration_i \leftarrow MR_{iter}$
$Cat_{SizePob} \leftarrow CreateCatPopulation(SizePob, MR_{best})$
while $Iteration_i \leq NroGen$ do
    for $i \leftarrow 1$ to $SizePob$ do
        if Cat[i] is SeekingMode then
            $SolutionCandidate \leftarrow applySeekingMode(Cat[i])$
        else
            $SolutionCandidate \leftarrow applyTracingMode(Cat[i])$
        $Fitness_{best} \leftarrow EvaluateBestSolution(SolutionCandidate, Fitness_{bestBefore})$
    if $Converge(Fitness_k, Iteration_i)$ Detected then
        $Cat_{SizePob} \leftarrow CreateCatPopulation(SizePob, MR_{best})$
    else
        $Cat_{SizePob} \leftarrow RepickCats(Cat_{SizePob})$
    $Iteration_i \leftarrow Iteration_i + 1$
return $Fitness_k$
```

**Table 1.** Analysis for SCP4.x,$max(\omega) = 46.6$ and $min(\omega) = 14.7$ with both value the phase2 reached the optimal

| Instance | MR | Iteration | Fitness<br>Phase1 | Optimal<br>Known | Fitness<br>Phase2 | omega<br>(%) |
|---|---|---|---|---|---|---|
| SCP41 | 0,55 | 371 | 431 | 429 | 430 | 37.10 |
| SCP410 | 0,4 | 219 | 516 | 514 | 514 | 21.90 |
| SCP42 | 0,9 | 466 | 515 | 512 | 512 | 46.60 |
| SCP43 | 0,45 | 250 | 525 | 516 | 517 | 25.00 |
| SCP44 | 0,65 | 362 | 499 | 494 | 496 | 36.20 |
| SCP45 | 0,41 | 160 | 514 | 512 | 514 | 16.00 |
| SCP46 | 0,25 | 154 | 560 | 560 | 560 | 15.40 |
| SCP47 | 0,45 | 147 | 430 | 430 | 430 | 14.70 |
| SCP48 | 0,6 | 405 | 494 | 492 | 492 | 40.50 |
| SCP49 | 0,5 | 262 | 656 | 641 | 652 | 26.20 |

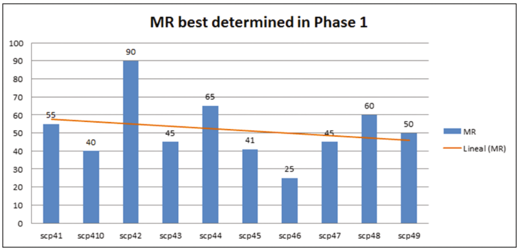


**Fig. 3.** MR tends to be close to 0.5

Where

- $DetermineBestMR(SizePob, MR_{best})$ : Determines the best MR for SCP instance, is defined by the Algorithm 3
- $Converge(Fitness_k, Iteration_i)$ : Detects if the solution does not progress
- $CreateCatPopulation(SizePob, MR_{best})$ : Create a new population and it is distributed according to $MR_{best}$

## 5 Experimental Results

The BCSO-RP was implemented in Java Programming language. The execution was made using a laptop with OS X Yosemite 10.10.5 operating system, Intel Core i5 2.50 GHz with 16 GB of RAM. The instances for SCP, obtained from OR-Library [11]. The results are compared using the relative percentage deviation ($RPD$). The $RPD$ value quantifies the deviation of the objective value $Z_{min}$

**Algorithm 3.** Phase I:Search for the best MR

```
Input: BCSO parameters : $SMP, SRD, CDC, SPC$
Input: Population parameters: $NroGen, SizePob$
Output: The best Fittness
$MR_i \leftarrow 0.0$
$Cat_{SizePob} \leftarrow CreateCatPopulation(SizePob, MR_i)$
while $MR_i \leq 1.0 \wedge$ ( Not $Converge(Fitness_k, Iteration_i)$) do
    for $i \leftarrow 1$ to $SizePob$ do
        if $Cat[i]$ is SeekingMode then
            $SolutionCandidate \leftarrow applySeekingMode(Cat[i])$
        else
            $SolutionCandidate \leftarrow applyTracingMode(Cat[i])$
        $Fitness_{best} \leftarrow$
          $EvaluateBesSolution(SolutionCandidate], Fitness_{bestBefore})$
    $Cat_{SizePob} \leftarrow RepickCats(SizePob, MR)$
    $Iteration_i \leftarrow Iteration_i + 1$
    $MR_i \leftarrow MR_i + 0.1$
return $MR, NroIter$
```

from $Z_{opt}$ that in our experiment is the minimal best known value for each instance and it is calculated as follows (Table 2)

$$RPD = \left(\frac{Z_{min} - Z_{opt}}{Z_{opt}}\right) \times 100 \tag{9}$$

In all experiments the BCSO-RP was executed using 1000 iterations and 31 times each instance. The best MR in all instances was close 0.5, coincides with MR used in [12], however our results were better by the population reset introduced when the population loses the diversity, it allows to explore other areas of the solution space and greatly improves results. The Tables 3, 4, and 5 shows these results. We can observed that Standard Desviation(SD) is very small, except in the instance 4.2 and 4.3. From Fig. 2 it can observed are $RPD_{min}$ and $RPD_{avg}$ are worst in the A.x ans B.x, however these same instances have good standard deviation. Finally we reachead 11 optimal, these results are promising compared to the studies done in [12] and [13] (Table 6).

The results BCSO-RP were compared with BCSO algorithm for the non-unicost and BCSO with Different Binarization Methods for Solving Set Covering Problems [12,13], Binary Firefly Optimization (BFO) [14]; and Binary Artificial Bee Colony (BABC) [15]. The column $Z_{opt}$, for instance executed, the columns $Z_{min}$ and $Z_{avg}$ represent the minimum and average objetive functions values respectively and finally, $RPD$, the minimal relative percentage deviation calculated.

**Table 2.** Experimental results with 4.x and 5.x Beasley's OR Library instances

| Inst | Opt | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD | Inst | Op | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.1 | 429 | 430 | 438 | 0.233 | 433.65 | 1.083 | 1.70 | 5.1(*) | 253 | 253 | 263 | 0.000 | 258.26 | 2.078 | 2.65 |
| 4.2(*) | 512 | 512 | 545 | 0.000 | 525.10 | 2.558 | 8.36 | 5.2 | 302 | 303 | 322 | 0.331 | 309.94 | 2.628 | 5.08 |
| 4.3 | 516 | 517 | 573 | 0.194 | 533.81 | 3.451 | 12.05 | 5.3(*) | 226 | 226 | 233 | 0.000 | 229.87 | 1.713 | 1.52 |
| 4.4 | 494 | 496 | 515 | 0.405 | 502.74 | 1.770 | 4.52 | 5.4(*) | 242 | 242 | 248 | 0.000 | 244.81 | 1.160 | 1.60 |
| 4.5 | 512 | 514 | 518 | 0.391 | 515.58 | 0.699 | 1.57 | 5.5 | 211 | 214 | 222 | 1.422 | 217.52 | 3.088 | 1.98 |
| 4.6(*) | 560 | 560 | 565 | 0.000 | 561.16 | 0.207 | 1.85 | 5.6(*) | 213 | 213 | 224 | 0.000 | 217.42 | 2.075 | 3.85 |
| 4.7(*) | 430 | 430 | 445 | 0.000 | 433.68 | 0.855 | 3.17 | 5.7 | 293 | 294 | 308 | 0.341 | 299.71 | 2.290 | 3.36 |
| 4.8(*) | 492 | 492 | 512 | 0.000 | 496.97 | 1.010 | 5.08 | 5.8(*) | 288 | 288 | 314 | 0.000 | 299.03 | 3.831 | 6.59 |
| 4.9 | 641 | 652 | 678 | 1.716 | 661.77 | 3.241 | 6.43 | 5.9 | 279 | 280 | 286 | 0.358 | 280.52 | 0.543 | 1.12 |
| 4.10(*) | 514 | 514 | 524 | 0.000 | 518.10 | 0.797 | 2.10 | 5.10 | 265 | 268 | 278 | 1.132 | 272.29 | 2.751 | 2.27 |

**Table 3.** Experimental results with 6.x and A.x, B.x, C.x, D.x and D.x Beasley's OR Library instances

| Inst | Opt | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD | Inst | Opt | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6.1 | 138 | 141 | 149 | 2.174 | 144.03 | 4.371 | 2.18 | C.1 | 227 | 229 | 241 | 0.881 | 235.03 | 3.538 | 2.36 |
| 6.2(*) | 146 | 146 | 151 | 0.000 | 148.97 | 2.033 | 1.80 | C.2 | 219 | 221 | 233 | 0.913 | 226.74 | 3.535 | 2.65 |
| 6.3(*) | 145 | 145 | 154 | 0.000 | 148.77 | 2.603 | 1.96 | C.3 | 243 | 251 | 271 | 3.292 | 258.68 | 6.452 | 4.69 |
| 6.4(*) | 131 | 131 | 135 | 0.000 | 133.13 | 1.625 | 0.96 | C.4 | 219 | 225 | 240 | 2.740 | 231.39 | 5.656 | 4.32 |
| 6.5 | 161 | 164 | 170 | 1.863 | 166.52 | 3.426 | 1.96 | C.5 | 215 | 220 | 234 | 2.326 | 226.13 | 5.176 | 3.84 |
| A.1 | 253 | 254 | 264 | 0.395 | 257.87 | 1.925 | 2.47 | D.1(*) | 60 | 60 | 65 | 0.000 | 62.77 | 4.624 | 1.73 |
| A.2 | 252 | 259 | 266 | 2.778 | 262.77 | 4.275 | 2.00 | D.2 | 66 | 69 | 70 | 4.545 | 69.55 | 5.376 | 0.51 |
| A.3 | 232 | 234 | 248 | 0.862 | 241.81 | 4.227 | 3.70 | D.3 | 72 | 75 | 79 | 4.167 | 77.48 | 7.616 | 1.09 |
| A.4 | 234 | 238 | 249 | 1.709 | 243.00 | 3.846 | 2.00 | D.4 | 62 | 63 | 66 | 1.613 | 64.48 | 4.006 | 1.00 |
| A.5(*) | 236 | 236 | 240 | 0.000 | 237.81 | 0.765 | 1.01 | D.5 | 61 | 62 | 65 | 1.639 | 63.74 | 4.495 | 0.82 |
| B.1 | 69 | 70 | 75 | 1.449 | 73.16 | 6.031 | 1.16 | E.1(*) | 29 | 29 | 30 | 0.000 | 29.90 | 3.115 | 0.30 |
| B.2 | 76 | 80 | 86 | 5.263 | 83.00 | 9.211 | 1.83 | E.2 | 30 | 32 | 34 | 6.667 | 33.23 | 10.753 | 0.96 |
| B.3(*) | 80 | 80 | 84 | 0.000 | 81.68 | 2.097 | 1.25 | E.3 | 27 | 29 | 32 | 7.407 | 30.87 | 14.337 | 1.02 |
| B.4 | 79 | 81 | 85 | 2.532 | 83.19 | 5.308 | 1.49 | E.4 | 28 | 29 | 33 | 3.571 | 31.58 | 12.788 | 1.15 |
| B.5(*) | 72 | 72 | 73 | 0.000 | 72.97 | 1.344 | 0.18 | E.5(*) | 28 | 28 | 30 | 0.000 | 29.74 | 6,221 | 0,51 |

**Table 4.** Experimental results with F.x, G.x and H.x Beasley's OR Library instances

| Inst | Opt | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD | Inst | Opt | Min | Max | RPD min(%) | Prom | RPD prom(%) | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F.1 | 14 | 15 | 17 | 7.143 | 16.61 | 18.664 | 0.56 | G.4 | 168 | 175 | 183 | 4.167 | 180.29 | 7.316 | 2.08 |
| F.2 | 15 | 16 | 18 | 6.667 | 17.45 | 16.344 | 0.68 | G.5 | 168.1 | 172 | 184 | 2.320 | 180.65 | 7.463 | 2.60 |
| F.3 | 14 | 16 | 17 | 14.286 | 16.84 | 20.276 | 0.37 | H.1 | 63 | 67 | 71 | 6.349 | 69.81 | 10.804 | 0.75 |
| F.4 | 14 | 15 | 17 | 7.143 | 15.94 | 13.825 | 0.44 | H.2 | 63 | 65 | 67 | 3.175 | 66.87 | 6.144 | 0.43 |
| F.5 | 13 | 15 | 16 | 15.385 | 15.90 | 22.333 | 0.30 | H.3 | 59.2 | 65 | 69 | 9.797 | 67.48 | 13.993 | 1.26 |
| G.1 | 176.4 | 182 | 193 | 3.175 | 190.00 | 7.710 | 2.68 | H.4 | 58.1 | 63 | 67 | 8.434 | 64.81 | 11.543 | 0.98 |
| G.2 | 154.1 | 160 | 167 | 3.829 | 165.23 | 7.220 | 1.50 | H.5 | 55 | 59 | 61 | 7.273 | 60.32 | 9.677 | 0.54 |
| G.3 | 166.2 | 178 | 182 | 7.100 | 180.06 | 8.342 | 1.44 | | | | | | | | |

**Table 5.** Comparison of BCSO-RP against BCSO original, BEE, FIREFLY ANT metaheristics

| **Tech/SCP** | 4.1 | 4.2 | 4.3 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 4.10 | 5.1 | 5.2 | 5.3 | 5.4 | 5.5 | 5.6 | 5.7 | 5.8 | 5.9 | 5.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Best Know** | 429 | 512 | 516 | 494 | 512 | 560 | 430 | 492 | 641 | 514 | 253 | 302 | 226 | 242 | 211 | 213 | 293 | 288 | 279 | 265 |
| **BCSO-RP** | 430 | 512 | 517 | 496 | 514 | 560 | 430 | 492 | 652 | 514 | 253 | 303 | 226 | 242 | 214 | 213 | 294 | 288 | 280 | 268 |
| **BCSO-BI** | 432 | 517 | 531 | 496 | 514 | 560 | 434 | 494 | 660 | 518 | 258 | 306 | 229 | 242 | 216 | 217 | 294 | 294 | 280 | 271 |
| **BCSO-ORI** | 429 | 517 | 519 | 495 | 514 | 563 | 430 | 497 | 655 | 519 | 279 | 339 | 247 | 251 | 230 | 232 | 332 | 320 | 295 | 285 |
| **BEE** | 430 | 512 | 516 | 494 | 512 | 561 | 430 | 493 | 643 | 514 | 254 | 309 | 228 | 242 | 211 | 213 | 296 | 288 | 280 | 266 |
| **FIREFLY** | 481 | 580 | 619 | 537 | 609 | 653 | 491 | 565 | 749 | 550 | 296 | 372 | 250 | 277 | 253 | 264 | 337 | 326 | 350 | 321 |
| **ANT** | 429 | 512 | 516 | 494 | 512 | 560 | 430 | 492 | 641 | 514 | 253 | 302 | 228 | 242 | 211 | 213 | 293 | 288 | 279 | 265 |

| **Tech/SCP** | **6.1** | **6.2** | **6.3** | **6.4** | **6.5** | **A.1** | **A.2** | **A.3** | **A.4** | **A.5** | **B.1** | **B.2** | **B.3** | **B.4** | **B.5** | **C.1** | **C.2** | **C.3** | **C.4** | **C.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Best Know** | 138 | 146 | 145 | 131 | 161 | 253 | 252 | 232 | 234 | 236 | 69 | 76 | 80 | 79 | 72 | 227 | 219 | 243 | 219 | 215 |
| **BCSO-RP** | 141 | 146 | 145 | 131 | 164 | 254 | 259 | 234 | 238 | 236 | 70 | 80 | 80 | 81 | 72 | 229 | 221 | 251 | 225 | 220 |
| **BCSO-BI** | 143 | 146 | 148 | 133 | 165 | 271 | 259 | 238 | 241 | 237 | 70 | 80 | 80 | 81 | 73 | 232 | 225 | 251 | 231 | 222 |
| **BCSO-ORI** | 151 | 152 | 160 | 138 | 169 | 286 | 274 | 257 | 248 | 244 | 79 | 86 | 85 | 89 | 73 | 242 | 240 | 277 | 250 | 243 |
| **BEE** | 140 | 146 | 145 | 131 | 161 | 254 | 254 | 234 | 234 | 237 | 69 | 76 | 80 | 79 | 72 | 230 | 219 | 244 | 220 | 215 |
| **FIREFLY** | 173 | 180 | 160 | 161 | 186 | 285 | 285 | 272 | 297 | 262 | 80 | 92 | 93 | 98 | 87 | 279 | 272 | 288 | 262 | 262 |
| **ANT** | 138 | 146 | 145 | 131 | 161 | 253 | 252 | 232 | 234 | 236 | 69 | 76 | 80 | 79 | 72 | 227 | 219 | 243 | 219 | 215 |

| **Tech/SCP** | **D.1** | **D.2** | **D.3** | **D.4** | **D.5** |
|---|---|---|---|---|---|
| **Best Know** | 60 | 66 | 72 | 62 | 61 |
| **BCSO-RP** | 60 | 69 | 75 | 63 | 62 |
| **BCSO-BI** | 60 | 69 | 76 | 63 | 64 |
| **BCSO-ORI** | 65 | 70 | 79 | 64 | 65 |
| **BEE** | 60 | 67 | 73 | 63 | 62 |
| **FIREFLY** | 71 | 75 | 88 | 71 | 71 |
| **ANT** | 60 | 66 | 72 | 62 | 61 |

**Table 6.** Comparison of BCSO-RP against BCSO original, BEE, FIREFLY ANT metaheristics

| Tech/SCP | NRE1 | NRE2 | NRE3 | NRE4 | NRE5 | NRF1 | NRF2 | NRF3 | NRF4 | NRF5 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Best Know** | 29 | 30 | 27 | 28 | 28 | 14 | 15 | 14 | 14 | 13 |
| **BCSO-RP** | 29 | 32 | 29 | 29 | 28 | 15 | 16 | 16 | 15 | 15 |
| **BCSO-BI** | 30 | 34 | 29 | 32 | 30 | 17 | 16 | 17 | 15 | 16 |
| **BCSO-ORI** | 29 | 34 | 31 | 32 | 30 | 17 | 18 | 17 | 17 | 15 |
| **BEE** | 29 | 30 | 27 | 28 | 28 | 14 | 15 | 14 | 14 | 13 |
| **FIREFLY** | 32 | 36 | 35 | 34 | 34 | 17 | 17 | 21 | 19 | 16 |
| **ANT** | 29 | 30 | 27 | 28 | 28 | 14 | 15 | 14 | 14 | 13 |
| Tech/SCP | NRG1 | NRG2 | NRG3 | NRG4 | NRG5 | NRH1 | NRH2 | NRH3 | NRH4 | NRH5 |
| **Best Know** | 176 | 154 | 166 | 168 | 168 | 63 | 63 | 59 | 58 | 55 |
| **BCSO-RP** | 182 | 160 | 178 | 175 | 172 | 67 | 65 | 65 | 63 | 59 |
| **BCSO-BI** | 191 | 165 | 182 | 180 | 183 | 69 | 67 | 69 | 64 | 61 |
| **BCSO-ORI** | 190 | 165 | 187 | 179 | 181 | 70 | 67 | 68 | 66 | 61 |
| **BEE** | 176 | 154 | 166 | 168 | 168 | 63 | 63 | 59 | 58 | 55 |
| **FIREFLY** | 230 | 191 | 198 | 214 | 223 | 85 | 81 | 76 | 75 | 68 |
| **ANT** | 176 | 155 | 166 | 168 | 168 | 64 | 64 | 59 | 58 | 55 |

## 6 Conclusion and Future Work

There are not published results on using BCSO modifying the rate mixture and using restart the population when premature convergence it is detected, for this reason we think that our results are quite promising. The motivation to choose MR and work with this parameter is based to study the effects of population changes and their consequences on the quality of solutions compared to other metaheuristics. In this preliminary investigation we obtained good results and as future works we think improve them considering:

- Optimize phase 1 to determine a MR set rather than a single value and use it in the SCP solution
- Improve the mechanism of restart, considering properties of elitism, Genetic Mutation and transferring experience from one restart to another
- Improve the mechanism used in phase 1 to include other parameters of the technique and experiment with them. The idea is to achieve a balance characteristics of exploration and exploitation defined by MR (Mixture Rate) and the others parameters BCSO techniques.

**Acknowledgements.** The author Broderick Crawford is supported by grant CONICYT/FONDECYT/REGULAR/1171243 and Ricardo Soto is supported by Grant CONICYT/FONDECYT/REGULAR/1160455

## References

1. Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. Mach. Learn. **3**(2), 95–99 (1988)
2. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, MHS 1995, pp. 39–43. IEEE (1995)
3. Fister, I., Strnad, D., Yang, X.-S., Fister Jr., I.: Adaptation and hybridization in nature-inspired algorithms. In: Adaptation and Hybridization in Computational Intelligence, pp. 3–50. Springer (2015)
4. Sharafi, Y., Khanesar, M.A., Teshnehlab, M.: Discrete binary cat swarm optimization algorithm. In: 3rd International Conference on Computer, Control & Communication (IC4), 2013, pp. 1–6. IEEE (2013)
5. Current, J., Daskin, M., Schilling, D., et al.: Discrete network location models. Facility Locat. Appl. Theor. **1**, 81–118 (2002)
6. Beasley, J.E.: An algorithm for set covering problem. Eur. J. Oper. Res. **31**(1), 85–93 (1987)
7. Chu, S.-C., Tsai, P.-W., Pan, J.-S.: Cat swarm optimization. Pacific Rim International Conference on Artificial Intelligence, pp. 854–858. Springer (2006)
8. Fister Jr, I., Yang, X.-S., Fister, I., Brest, J., Fister, D.: A brief review of nature-inspired algorithms for optimization, arXiv preprint arXiv:1307.4186 (2013)
9. Auger, A., Hansen, N.: A restart cma evolution strategy with increasing population size. In: The 2005 IEEE Congress on Evolutionary Computation, 2005, vol. 2, pp. 1769–1776. IEEE (2005)

10. Moscato, P., Cotta, C.: Una introducción a los algoritmos memeticos. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial **7**(19), 131–148 (2003)
11. Beasley, J.E.: Or-library: distributing test problems by electronic mail, Operations Research (OR) problems. Brunel University London, OR-Library (2016)
12. Crawford, B., Soto, R., Berríos, N., Johnson, F., Paredes, F., Castro, C., Norero, E.: A binary cat swarm optimization algorithm for the non-unicost set covering problem. Math. Probl. Eng. vol. 2015 (2015)
13. Crawford, B., Soto, R., Berrios, N., Olguín, E.: Cat swarm optimization with different binarization methods for solving set covering problems. In: Artificial Intelligence Perspectives in Intelligent Systems, pp. 511–524. Springer (2016)
14. Crawford, B., Soto, R., Olivares-Suárez, M., Paredes, F.: A binary firefly algorithm for the set covering problem. In: Modern Trends and Techniques in Computer Science, pp. 65–73. Springer (2014)
15. Crawford, B., Soto, R., Cuesta, R., Paredes, F.: Application of the artificial bee colony algorithm for solving the set covering problem. Sci. World J. **2014**, 8 pages (2014)

# Analysis of Students' Behavior Based on Educational Data Mining

Kunyanuth Kularbphettong(✉)

Computer Science Program, Suan Sunandha Rajabhat University,
Bangkok 10300, Thailand
kunyanuth.ku@ssru.ac.th

**Abstract.** This research aims to develop a model for analysis of student behavior through e-Learning based on data mining technique in case of Suan Sunandha Rajabhat University. The student data set was composed of 5392 personal records and, to compare the effective of algorithm, the model was created under decision tree and Bayesian networks techniques. The result found that showed that Bayesian networks technique showed higher performance and the percentage of prediction is accurate 91.32%.

**Keywords:** Educational data mining · Student's behavior · Decision tree · Bayesian networks

## 1 Introduction

With no longer barrier by space and time, information technology systems play the important rule to support human and social life. Web based and mobile based Learning systems have become more and more used in teaching and learning to enhance the ability of both students and teachers. e-Learning is an educational system that involves learning and teaching through the Internet. With plenty of information readily available, students can study their interested courses though a web-based class so as to enhance their knowledge at any time and any place and teachers can easily manage their online classes and monitor student's performance as well. However, most web-based educational courses can rarely support interactive and adaptive abilities. Hence, educational data mining (EDM) is used to analyze student's behavior so as to find the patterns of system usage by teachers and students and to discover the students 'learning behavior patterns.

Educational data mining is the process of converting raw data from educational systems to useful information and it can analyze relevant information results and produce different perspectives to understand more about the users' behavior. Data Mining can extract knowledge through the analysis of the information available in the form of data generated by their users. Therefore, the purpose of this research aims to develop a model for analysis of student behavior through e-Learning based on educational data mining technique in case of Suan Sunandha Rajabhat University.

The remainder of this paper is organized as follows. Section 2 presents the research methodologies used in this work. Section 3 presents the experimental results based on the purposed model based on educational data mining technique. This project

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_15

demonstrates how to analyze student behavior from log file from e-Learning and how to evaluate purposed methodology. Finally, in Sect. 4 conclude the paper with future research.

## 2 Literature Reviews

A literature search shows that most of the related researches have deployed data mining techniques to analyze student's learning behaviors by following this: According to Romero et al. [1], the research was shown the usefulness of the data mining techniques in course management system and the rules can help to classify students and to detect the sources of any incongruous values received from student activities. Kularbphettong [2] used data mining techniques with data log file provided by Learning Management Systems (LMSs) in relation to visits and times, resources viewed, assessments, activities and etc. [3–6]. Also there are many researches that have been investigated in the on-line learning environment. For example, Efrati et al. presented a case study of a data mining approach based on cluster analysis to support the detection of learning styles in a community of learners [7]. Minaei-Bidgoli et al. [8] presented an approach to classify students in order to predict their final grade based on features extracted from logged data in an education web-based system.

## 3 The Methodologies

This section described the specified methodologies used in this project. Educational data mining is the data analyzing process from different perspectives also summarizing the useful information results.

A decision tree is one of the most well-known classification approaches that are commonly used to examine data and induce the tree in order to make predictions [9]. J48 classifier is a simple C4.5 decision tree to create a binary tree developed by Quinlan [10] and it is an open source Java implementation of the C4.5 algorithm under WEKA data mining platform. C4.5 is an algorithm used to generate a decision tree. C4.5 builds decision trees from a set of training data by using the concept of information entropy.

Bayesian classification is a supervised learning algorithm for classification and It is simple probabilistic classifier based on Bayesian theorem with strong independence assumption. The conditional independence assumptions in the graph are estimated by statistical and computational models [11, 12].

## 4 Experimental Setup

This research was adopted classification approaches like decision tree and Bayesian Networks techniques to create model for prediction the pattern of student's learning behavior through e-Learning. The data of this experiment was collected from information technology subject, Suan Sunandha Rajabhat University, during the period of 2014–2015. The student data set was composed of 5392 personal records and the data was consisted of personal records, course (face-to-face) records, score tests, project scores and students' log file from e-Learning system.

In this class, student should be required to attend classroom and they were assigned to take post exams in class and they must propose and present project they must work together in group. In addition, student needs to take post quizzes online, to review and use materials on e-Learning system and to participate in exercises. The data is preprocessed, and transformed to be appropriated format in order to apply data mining techniques as shown in Fig. 1.

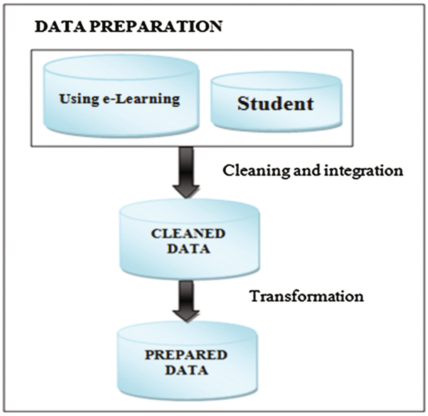


**Fig. 1.** The data preparation process phrase

In data preparation process phrase, to develop a model for analyzing the learning behavior of learners on the e-Learning system, attributes were defined from the learner's intended learning behavior like number of reading material defined how much student accessed to the course materials, number of downloads shown the behavior of the students regularly using the course materials and number of attendance class presented the intention and responsibility of learner. Also, all continuous attributes have been transformed to nominal attributes so as to conveniently be processed and understand the meaning. To transform numerical attributes to discrete attributes, equal width method was used to partition the value of continuous attributes into five nominal values: VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH. Learning set and testing set were divided to create the model.

After preparation data, decision tree and bayesian networks algorithms were applied to discover valuable patterns. Data was analyzed by WEKA. WEKA, the Waikato Environment for Knowledge Analysis, is a collection of machine learning algorithm to analyze data set for data mining tasks [15]. Decision tree and Bayesian Networks algorithms were used to estimate and evaluate for creating the model. To take measure the result, the K-fold cross validation method was provided to validate the result. And to create the effectiveness model, the results of each algorithm were evaluated by the percentage correct, precision, recall and F measure (Table 1).

**Table 1.** The listed of important attributes

| Name | Description |
|---|---|
| Result-PostTest_number | Mark obtained from posttest |
| Time-of-PostTest_number | Time spent on posttest |
| SumofPostTest | Total mark obtained from all posttest |
| Time-of-viewing–material | Total time spent on viewing material |
| score-of-attendance | Scores of attendance in class |
| Project_score | Mark obtained from project |
| Midterm-score | Mark obtained from Midterm |
| Final-score | Mark obtained from Final |
| Grade | Final mark obtain from this class |

## 5 Experimental Results

In this research, the data was analyzed by using J48 and Bayesian Networks algorithms and using the appropriated result was created the model to analyze learners' learning behavior using educational data mining techniques. The results of this experiment were shown in Table 2, compared the effectiveness of each algorithm.

**Table 2.** Comparison of model performance measurements.

| | Precision | Recall | F-measure | Accuracy |
|---|---|---|---|---|
| J48 | 90.2 | 95.33 | 92.75 | 92.16 |
| Bayesian networks | 91.32 | 96.12 | 93.64 | 92.73 |

The results of the experiment show that the Bayesian Networks algorithm is more efficient than J48 algorithm. The percentage of prediction is accurate 91.32% and precision and Fig. 2 was compared the effectiveness of each algorithm.

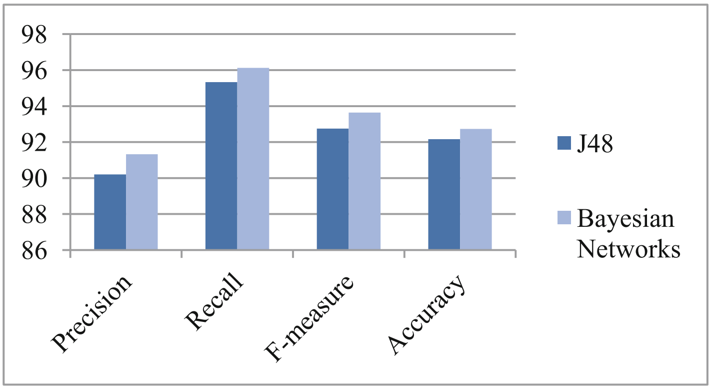


**Fig. 2.** The effectiveness of each algorithm

## 6 Conclusion

This paper presented the prototypes model for the ongoing improvement project to build a model for analyzing learners' learning behavior using educational data mining techniques. The research was divided the study and developed the model into 2 phases: the development of models for analyzing the learning behavior of learners on the e-learning system in the first part was to prepare the log files of the e-learning and e-learning data and student attendance data and transformed the data into a suitable format for further study and development and to develop the model using decision tree and bayesian Networks techniques. This model can be beneficial to discover the pattern of learning behavior. However, in term of the future experiments, other data mining techniques are be applied to conduct a research to enhance this project and also apply the tool to analyze the learning behavior of learners on the e-learning system.

**Acknowledgments.** The authors gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

## References

1. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Comput. Educ. **51**(1), 368–384 (2008)
2. Kularbphettong, K., Tongsiri, C., Waraporn, P.: Analysis of student motivation behavior on e-learning based on association rule mining. In: Proceedings of International Conference on Education and Information Technology, Paris, France, 27–28 July 2012
3. Kularbphettong, K., Tongsiri, C.: Student motivation behavior on e-learning based on data mining techniques. In: Proceedings of International Conference on Data Analysis and Decision Making, Prague, Czech Republic, 8–9 July 2013
4. García, E., Romero, C., Ventura, S., Castro, C.: Using rules discovery for the continuous improvement of e-learning courses. Lecture Notes in Computer Science, LNCS, vol. 4224/2006, pp. 887–895 (2006)
5. Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., Heiner, C.: An educational data mining tool to browse tutor–student interactions: time will tell. In: Proceedings of the Workshop on Educational Data Mining, Pittsburgh, USA, pp. 15–22 (2005)
6. Zorrilla, M.E., Menasalvas, E., Marin, D., Mora, E., Segovia, J.: Web usage mining project for improving web-based learning sites. In: Web Mining Workshop, Cataluna, pp. 1–22 (2005)
7. Efrati, V., Limongelli, C., Sciarrone, F.: A data mining approach to the analysis of students' learning styles in an e-learning community: a case study. In: Universal Access in Human–Computer Interaction. Universal Access to Information and Knowledge, pp. 289–300. Springer International Publishing, Berlin (2014)
8. Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G., Punch, W.F.: Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA. In: Proceedings of ASEE/IEEE Frontiers in Education Conference. IEEE, Boulder, CO (2003)
9. Edelstein, H.: Introduction to Data Mining and Knowledge Discovery, 3rd edn. Two Crows Corporation, Potomac (1999)
10. Quinlan, J.R.: Induction of Decision Trees. Mach. Learn. **1**(1), 81–106 (1986)

11. Korb, K.B., Nicholson, A.E.: Introduction to Bayesian networks. In: Bayesian Artificial Intelligence, pp 29–54. CRC Press, Boca Raton (2010)
12. Singh, M., Valtorta, M.: Construction of Bayesian network structures from data: a brief survey and an efficient algorithm. Int. J. Approx. Reason. **12**, 111–131 (1995)

# Inference Algorithms in Latent Dirichlet Allocation for Semantic Classification

Wan Mohammad Aflah Mohammad Zubir(✉), Izzatdin Abdul Aziz,
Jafreezal Jaafar, and Mohd Hilmi Hasan

Department of Computer and Information Sciences,
Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia
wanmohammadaflah@gmail.com,
{izzatdin,jafreez,mhilmi_hasan}@utp.edu.my

**Abstract.** There are existing implementations of Latent Dirichlet Allocation (LDA) algorithm as a semantic classifier to arrange the data for efficient retrieval. However, the problem of learning or inferencing the posterior distribution of the algorithm is trivial. Inferencing directly the prior distribution could lead to time taken to increase exponentially. It is due to the coupling of the hyperparameters. Several inference algorithms have been implemented together with LDA to solve this issue. The inference algorithm used in this research work is Gibbs sampling. Research using Gibbs sampling shows promising results in comparison to other inference algorithms, especially in the performance of the algorithm. It still takes a long time to compute the topic distribution of the data. There are still room for improvement in the time taken for the algorithm to complete the topic distribution. Using two datasets, an evaluation of the performance of the algorithm has been conducted. Results show that Gibbs sampling as the inference algorithm provides a better prediction on the optimal number of topic of the data in comparison to Variational Expectation Maximization (VEM).

**Keywords:** Latent Dirichlet Allocation · Inference engine · Topic models · Semantic · Text classification · Information retrieval

## 1 Introduction

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic modelling algorithm. It is used as an unsupervised learning classifier for unstructured data. As it is a generative probabilistic algorithm, it will randomly generate observable values based on hidden parameters. LDA uses two hidden hyper parameters, alpha and beta. Based on the generated observable values, we could learn the distributions and the pattern of the words occurrence. However, to infer completely the distributions are almost impossible as it would lead to the coupling of the hyper parameters. Several approximations inference algorithms have been used in LDA such as the Variation Expectation Maximization (VEM) and Gibbs sampling.

In most research regarding inferring the LDA, Gibbs sampling emerged as one of the best algorithm to solve the issue. However, it will still take a long time for the

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_16

algorithm to compute the topic distribution of the data. It is because the inferring process still involve a huge amount of iterations or cycles through all the words in the dataset. This shows that there is still room for improvement in the time taken for the algorithm to complete the topic distribution. In this paper, we discuss on the capability of the inference algorithms through tests with real world dataset.

The paper focuses to summarizing popular inference algorithms of LDA algorithm. The algorithm will be tested against two datasets including one real world industrial dataset obtained from an oil and gas company. The existing implementation of Gibbs sampling and Variational Expectation Maximization (VEM) as the inference algorithm for LDA are tested in the experiment. The experimentation is also limited to the evaluation of the harmonic means of both Gibbs sampling and Variational Expectation Maximization (VEM). Through harmonic means evaluation, an estimation of the ability of the inference algorithm to discover the optimal number of topics could be done. The inference algorithm should achieve high performance. High performance is defined as the ability to infer the topic distribution with shortest time taken possible. This is important as the inference algorithm should not hinder the performance of the overall algorithm.

# 2 Literature Review

This section is split to 4 sections in discussing the comparison between different methods in semantic classifier. Section 2.1 elaborates on the selection of semantic classifier algorithm, Sect. 2.2 discusses LDA, Sect. 2.3 deliberates on existing work of LDA and Sect. 2.4 is a critical analysis on the inference algorithms of LDA.

## 2.1 Selection of Semantic Classifier Algorithm

There are existing semantic classifier algorithms that had been researched and utilized. Table 1 discusses the comparison of the existing semantic classifier algorithms. A critical analysis on the advantages and disadvantages of existing semantic search algorithm was carried out prior to selecting a suitable algorithm for this research work.

Table 1 describes the existing method in semantic classifying data. For this research, LDA is selected to be the classifier. The main reason LDA is chosen is due to the feature that it does not require prior knowledge on the corpus or document. It is important that the algorithm does not require prior knowledge because it will not be confined to certain predefined words and can adapt to new corpus. LDA is able to define its own knowledge base which can be append when needed. This will produce a more accurate result as the algorithm will continue to improve its own capability. Even though Latent Semantic Analysis allows this but it still lacks the ability to distinguish polysemy. This will cause an issue if we encounter the same word but of different context. Probabilistic Latent Semantic Indexing overcame this problem but LDA is still better in terms of having the consistency in generating the topic distribution. As LDA assumed that the multivariate probabilities are generated by Dirichlet distributions, it prevents overfitting issue found with PLSI. The next section discusses in brief on LDA.

**Table 1.** Selection of semantic classifier algorithm

| Name | Features | Advantages | Disadvantages |
|---|---|---|---|
| Latent Dirichlet Allocation (Blei et al. 2003) | • A model that explains sets of observations by unobserved groups through deducing why some parts of the data are similar (Blei et al. 2003)<br>• Automatically discover topics without prior knowledge on the document (Chen 2011) | • Does not require priori, or prior knowledge on relations the document<br>• Results are close to Google's Page Rank (Fishkin 2010)<br>• Requires no labelled data (Hu 2009) | • Ignore the words order (Chen 2011) |
| Latent Semantic Analysis (Landauer et al. 2013) | • A model that analyses relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms | • Able to create relationship between the context of the document and keyword. This will return a more meaningful result | • Harder to handle polysemy (Hofmann 2001). Polysemy is defined as a word that has many meaning |
| Probabilistic Latent Semantic Indexing (PLSI) (Girolami and Kabán 2003) | • Developed in improving latent semantic indexing in term of precision and recall | • A low perplexity language algorithm. Having a low perplexity has been proved to increase the precision-recall (Azzopardi et al. 2003) | • The generation of semantic is not consistent<br>• The inconsistency leads to difficulty in assigning the probabilities to the words to their topics |

## 2.2 Latent Dirichlet Allocation (LDA)

LDA is a topic modelling model. It is a generative probabilistic model that is able to find and cluster the keywords in the document into their respective topic (Blei et al. 2003). The method selected is able to classify into their respective meaningful themes (Eight to Late 2015). It is generated based on an idea that documents are consisting of mixtures of latent topics and the topics are characterized by the distribution of words (Blei et al. 2003). Latent is defined as something which presence but not visible. Figure 1 illustrates how LDA works.

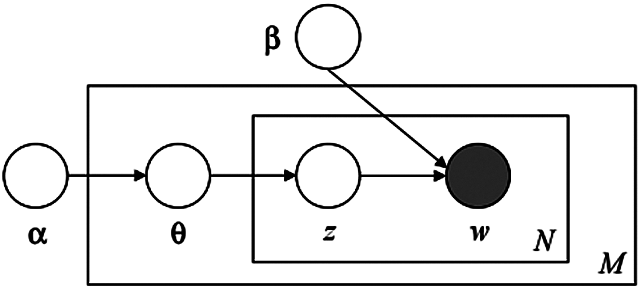


**Fig. 1.** Graphical model of LDA

In Fig. 1, M denotes the documents, N represents the set of words in the document, w represents the denoted words in the document, z represents the topic for the denoted word, θ represents the topic distribution for the document, α denotes the parameters set for the topic distribution per-document and β denotes the parameters set for the topic distribution per-topic word (Knispelis 2016).

In LDA, the user needs to specify the number of topics that the algorithm will classify. The algorithm will then find the latent relationship between the words and the topic associated with the words. LDA implements the bag-of-words model which ignores the order of the words (Wu et al. 2010). Bag-of-words model is a model that tokenize every word regardless of the grammar and their word order. However, LDA will view the words independently and the occurrence of each word is utilized in training a classifier (Maas et al. 2011). The next section discusses existing works that utilize LDA.

### 2.3 Existing Works in Latent Dirichlet Allocation

There are several existing works that implements LDA to semantically analyse datasets. Table 2 exhibits the comparison between existing work in LDA and the research gaps found.

**Table 2.** Existing work in Latent Dirichlet Allocation

| Name | Type of data | Findings | Way forward |
|---|---|---|---|
| Kim et al. (2011) | Document | • Eliminated the manual labelling but still lacks the same granularity<br>• The application of LDA can scale up to large number of documents | • Huge potential in analysing unstructured data |
| Arora and Ravindran (2008) | Document | • LDA-based model outperformed the performance of cluster-based approached | • Improve the estimation of LDA by applying Multi-Document Summarization Algorithm in Inferencing part |
| Wei and Croft (2006) | Web database | • Consistently outperformed the performance of cluster-based approach | • Improve efficiency in information retrieval through approximation in the inferencing part. The approximation is done through tweaking the hyper parameters |
| Yu et al. (2014) | E-commerce sites | • LDA able to determine meaningful user intents accurately<br>• LDA outperforms the eBay ranking on user satisfaction | • Include diversified retrieval approach to improve accuracy |

The content in Table 2 explains the comparison between works implementing LDA in semantically analysing the data. Based on all existing works, all of them proved that LDA could perform semantic search accurately. Either it is searching through documents or multimedia data, LDA can determine the semantic meaning of the words.

However, most of the existing works explained that there is minimal work in improving the performance of the algorithm. Since LDA is a generative probabilistic algorithm, it tends to take a longer time to compute a fairly distribution of topic assignments. This is because LDA has an intractable coupling issue (Blei et al. 2003; Hoffman et al. 2010). It normally occurs in a Bayesian Inference problem such as LDA. In LDA, we have two hyper parameters which are observable but not entirely independent. These will create a lot of possible latent variables when we want to sample the posterior distributions. Even though it is possible to integrate them out or marginalize the observable variables, there will be a likelihood that there is a problem in computation. Hence, inference methods or algorithms are used to infer the posterior distributions (Blei et al. 2003).

## 2.4 Inference Algorithms for Latent Dirichlet Allocation

There are several inference algorithms that is used with LDA. Table 3 exhibits the comparison between inference algorithms in LDA.

**Table 3.** Inference algorithms of Latent Dirichlet Allocation

| Name | Type | Findings | Way forward |
|---|---|---|---|
| Gibbs Sampling (Yildirim 2012) | Markov chain Monte Carlo | Posterior samples are generated by going through each variable to sample from its conditional distribution with the remaining variables with their current values unchanged (Yildirim 2012) | Not as fast as Variational Expectation Maximization when it comes to large sets of data |
| Variational Expectation Maximization (Meila 2014) | An extension of the Expectation-Maximization (EM) | It uses deterministic approximation instead of direct sampling from the posterior which resulting in faster computation than Gibbs Sampling | Although it is faster than Gibbs Sampling in computing the topic distribution, it will also limit the possible posterior distribution |

As it is best to avoid the coupling of the hyper parameters, these two inference algorithms had been widely used in LDA. Variational Expectation Maximization (VEM) has been used by the original researcher as it produce a fast result given the sample test data that they used (Blei et al. 2003). The time taken to infer the topic distribution is faster than Gibbs Sampling when a large dataset is used (Meila 2014). However, as it does not sample from the true posteriors, it will not be able to match the accuracy of Gibbs Sampling. Besides that, it also uses a lot more computational resources when larger dataset is used in comparison to Gibbs sampling. This will be a major drawback as we are dealing with large dataset. Thus, in this research, Gibbs Sampling is chosen as the inference for the LDA.

Gibbs sampling has several room of improvements such as in terms of the time taken to compute the topic distribution. This creates a gap in improving the performance of the algorithm in terms of the time taken to compute the total topic distribution.

## 3 Methodology

In this research, there parameters that is studied are the performance and accuracy. The parameter is selected based on the research gaps found from the existing implementations of LDA algorithm. Performance in this context is defined as having fast processing speed of any jobs given. Hence, performance of the algorithm is measured as the time taken to complete the tasks assigned. However, it is also important to ensure that the result obtained from the inference algorithm is accurate. Based on the critical analysis performed, LDA is selected as the search algorithm for this research.

The algorithm is written in Java. Most of the programming parts are referred from (Blei et al. 2003; McCallum 2002; Nguyen 2007) (Fig. 2).

The main part of the program is to determine the topics of words in a collection of documents. Through identification of the topic, the relation between the words and the document can be established which will bring a more meaningful result for a search request. The first step of the program will be to initialize a sampling method for the algorithm.

The sampling method will generate the prior distribution of the words in the documents using the value of $\alpha$ and $\beta$ are entered. Value of $\alpha$ and $\beta$ represents the hyper parameters of the algorithm. If the value of $\alpha$ is high, each document is likely to have a mix of most of the topics assigned. If the value of $\beta$ is high, each topic is likely to contain a mixture of most of the words. Both value of $\alpha$ and $\beta$ basically deals with the sparsity or the uniformity in the collection of documents. In this research, we are initializing the value of $\alpha$ of 0.1 and value of $\beta$ of 0.01. The values are decided based on experiments conducted by (Griffiths and Steyvers 2004).

After all the prior distribution have been calculated, the values are passed to the inference algorithm integrated in the LDA. Firstly, an initialization of Z, the number of topics in the document is done. In the proof of concept, the value of 20 is used as initialization. Gibbs sampling, as the inference algorithm, will update Z once it learns the document through the drawing of $Z_n^{(d)}$. After the first update has been completed, the value of the hyper parameters are also updated to reflect the learned distribution

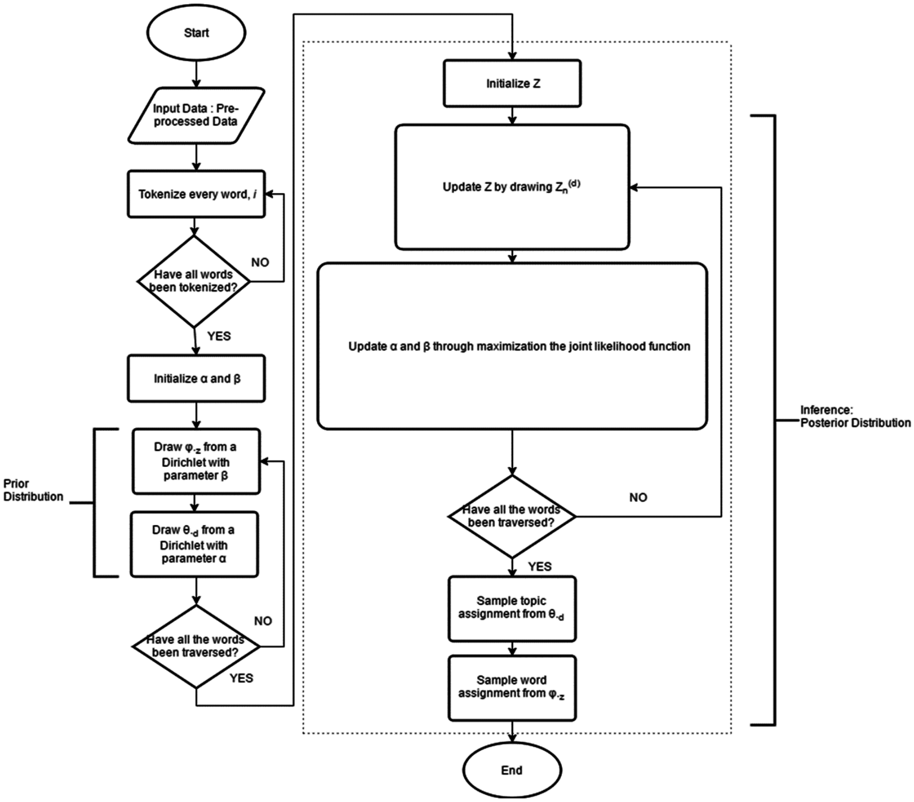

Start
Input Data : Pre-processed Data
Tokenize every word, *i*
Have all words been tokenized?
NO
YES
Initialize α and β
Prior Distribution
Draw $\varphi_{\cdot z}$ from a Dirichlet with parameter β
Draw $\theta_{\cdot d}$ from a Dirichlet with parameter α
Have all the words been traversed?
NO
YES
Initialize Z
Update Z by drawing $Z_n^{(d)}$
Update α and β through maximization the joint likelihood function
Have all the words been traversed?
NO
YES
Sample topic assignment from $\theta_{\cdot d}$
Sample word assignment from $\varphi_{\cdot z}$
Inference: Posterior Distribution
End

**Fig. 2.** High-level graphical model of LDA with Gibbs sampling as the inference algorithm

through the maximization of the joint likelihood function. After all the updates have been completed, the algorithm will then sample both the topic assignment and word assignment from the documents. Through this sampling process, we can retrieve the posterior distribution which is the topic distribution of the documents.

## 4 Experimentation

The aim for the experiment is to prove that Gibbs sampling of LDA able to discover the optimal number of topic with better than Variational Expectation Maximization (VEM), in terms of accuracy and performance.

### 4.1 Data Preparation

The data are in the form of unstructured documents. The dataset will undergo pre-processing to ensure that it will be suitable for further analysed by the chosen method. Details about the documents are as in Table 4.

**Table 4.** Parameters of dataset used in the experiments

| | Dataset A | Dataset B |
|---|---|---|
| Source | News articles in NYTIMES (Blei et al. 2003) | Hydrocarbon Well Reports (PETRONAS 2001) |
| Number of words, N | 90,000 | 35,127 |
| Vocabulary, W | 6525 | 1922 |
| Number of topic | 13 | 3 |

### 4.2 Experimentation Setup

This evaluation is done through computing the harmonic means of the inference algorithm when learning the optimal number of topics for a dataset. LDA algorithm is applied to each of the dataset with either Gibbs sampling or Variational Expectation Maximization (VEM) as the inference algorithm. On each of the dataset, the algorithm will compute the harmonic means for different number of topics. The highest point of the harmonic mean reflects the optimal number of topic for the dataset (Wallach et al. 2009).

$\alpha = 0.1$, $\beta = 0.01$ are used as initialization. It is important that all experiments are given the same value of the hyper parameters to ensure a fair comparison.

The value of burn-in is set to 1000. It is a widely accepted value as to ensure that the algorithm has properly learn the latent structure of the documents (Yildirim 2012).

## 5 Results and Discussions

In our experiments, we are aiming to evaluate the accuracy and performance of LDA to discover optimal number of topics with both the inference algorithm without any improvements or alterations.

Firstly, the algorithm will use the initialized the number of topics before analysing the data to learn the structure of the data. Then, the trained algorithm will analyse the data to compute the distribution of topics. Both evaluation will then be repeated for a different number of topics. To find the optimum number of topics for each of the datasets used, we analysed the harmonic mean of the algorithm when we choose different number of topics. Several research have proved that the highest value of harmonic means corresponds to optimal number of topics (Wallach et al. 2009). As the problem is intractable, the heuristic approach (Harmonic mean method) is sufficient in determining the number of topics of the dataset (Zhao et al. 2015). The formula used in calculating the number of topics are

$$\frac{1}{P(w)} = \sum_{z} \frac{P(z|w)}{P(w|z)} \simeq \frac{1}{S} \sum_{s} \frac{1}{P(w|z^{(s)})}, \tag{1}$$

where $w$ means the particular word that we are, $z$ is the instance of the topic, S is the total number of topics, $P(w|z), P(z|w)\ and\, P(w|z^{(s)})$ are the Bayes probabilities of each of the symbols. $Z^{(s)}$ is drawn from P(z|w). Through this calculation it gives an estimator for P(w| $\varphi$, $\alpha$ m):

$$P(\boldsymbol{w}|\Phi,\alpha\boldsymbol{m}) \simeq \frac{1}{\frac{1}{S}\sum_{s}\frac{1}{P(\boldsymbol{w}|\boldsymbol{z}^{(s)},\Phi)}} = \mathrm{HM}\left(\left\{P(\boldsymbol{w}|\boldsymbol{z}^{(s)},\Phi)\right\}_{s=1}^{S}\right), \tag{2}$$

$\phi$ represents the probability distribution of the topic in the corpus, $\alpha$ is the hyper parameter of the Dirichlet distributions, HM denotes the harmonic means calculation from the resulting equation.

Based on Figs. 3 and 4, we could analyse the harmonic means corresponds to the number of topics tested against the dataset. All figures are summarized in Table 5.

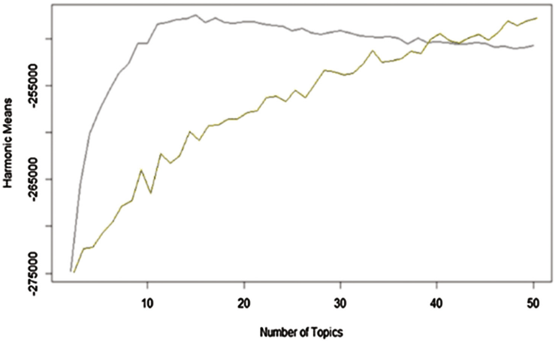


**Fig. 3.** Harmonic means for Dataset A (Gibbs sampling (black line) versus variational expectation maximization (yellow line))

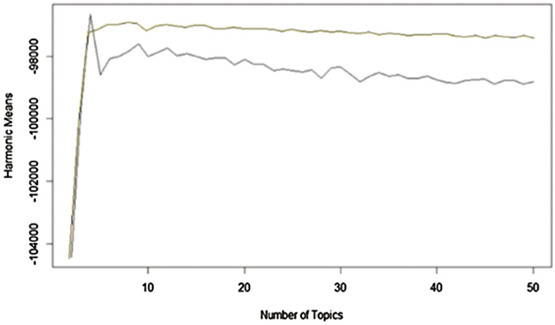


**Fig. 4.** Harmonic means for Dataset B (Gibbs sampling (black line) versus variational expectation maximization (yellow line))

**Table 5.** Time taken for calculating the harmonic means and optimal number of topics (Gibbs sampling and Variational Expectation Maximization)

| | Gibbs sampling | | Variational Expectation Maximization (VEM) | |
|---|---|---|---|---|
| | Dataset A | Dataset B | Dataset A | Dataset B |
| Time taken (minutes) | 19.31865 | 8.50527 | 18.45791 | 6.735076 |
| Optimal number of topics | 14 | 4 | 48 | 8 |
| Actual number of topics | 13 | 3 | 13 | 3 |
| Deviation | 1 | 1 | 35 | 5 |

Based on Table 5, a comparison was done between the result obtained from Gibbs sampling and Variational Expectation Maximization (VEM). The time taken for Gibbs sampling to discover the optimal number of topics are longer on average than Variational Expectation Maximization (VEM). However, the difference is minimal as the average difference of time taken is about 3–4 min.

Gibbs sampling performed better as the average difference of the number of topics computed is 1.67. Variational Expectation Maximization (VEM) overestimates the optimal number of topics as the average difference of number of topics is 15.67. One of the possible reason is the Variational Expectation Maximization (VEM) would require more iterations to arrive in the accurate optimal number of topics.

The results show that Gibbs sampling is better in discovering the optimal number of topics of any given data. Even though there is still a difference of the computed number of topics in comparison to the actual number of topics, it is still minimal when compared to Variational Expectation Maximization (VEM).

## 6 Conclusion

An analysis conducted towards the LDA algorithm. Through this analysis, it was found that the algorithm depends on inference algorithm to compute the topic distribution. Further analysis on the existing implementations of inference algorithm were conducted. Two inference algorithms, Gibbs sampling and Variational Expectation Maximization (VEM) were analyzed. Gibbs sampling was selected as it shows promising result in comparison to Variational Expectation Maximization (VEM), especially in the accuracy of the topic distribution. However, there is a gap of performance in the implementation of Gibbs sampling in LDA. It takes a longer time to reach to a topic distribution in comparison to Variational Expectation Maximization (VEM). Through experimentation, the results confirm to the earlier mentioned studies, which depict Gibbs sampling as a more robust inference algorithm.

## References

Arora, R., Ravindran, B.: Latent Dirichlet allocation based multi-document summarization. In: Paper Presented at the Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data (2008)

Azzopardi, L., Girolami, M., van Risjbergen, K.: Investigating the relationship between language model perplexity and IR precision-recall measures. In: Paper Presented at the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada (2003)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Chen, E.: Introduction to latent Dirichlet allocation (2011). http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Fishkin, R.: Latent Dirichlet allocation (LDA) and Google's rankings are remarkably well correlated (2010). https://moz.com/blog/lda-and-googles-rankings-well-correlated

Girolami, M., Kabán, A.: On an equivalence between PLSI and LDA. In: Paper Presented at the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2003)

Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(Suppl. 1), 5228–5235 (2004)

Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent Dirichlet allocation. In: Paper Presented at the Advances in Neural Information Processing Systems (2010)

Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42** (1–2), 177–196 (2001)

Hu, D.J.: Latent Dirichlet allocation for text, images, and music. University of California, San Diego (2009). Accessed 26 Apr. 2013

Kim, D.-K., Motoyama, M., Voelker, G.M., Saul, L.K.: Topic modeling of freelance job postings to monitor web service abuse. In: Paper Presented at the Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (2011)

Knispelis, A. (Producer): LDA topic models (2016). https://www.youtube.com/watch?v=3mHy4OSyRf0

Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Psychology Press, Hove (2013)

Eight to Late: A gentle introduction to topic modeling using R (2015). https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Paper Presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Portland, OR (2011)

McCallum, A.K.: MALLET: a machine learning for language toolkit (2002). http://mallet.cs.umass.edu

Meila, M.: Variational methods and variational EM (2014). http://www.stat.washington.edu/courses/stat539/spring14/Handouts/l7-variational.pdf

Nguyen, C.-T., Phan, X.-H.: GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) (2007). http://gibbslda.sourceforge.net/

PETRONAS: Final well report ANGSI (2001)

Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Paper Presented at the Proceedings of the 26th Annual International Conference on Machine Learning (2009)

Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Paper Presented at the Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2006)

Wu, L., Hoi, S.C., Yu, N.: Semantics-preserving bag-of-words models and applications. IEEE Trans. Image Process. **19**(7), 1908–1920 (2010)

Yildirim, I.: Bayesian inference: Gibbs sampling. Technical note, University of Rochester (2012)

Yu, J., Mohan, S., Putthividhya, D.P., Wong, W.-K.: Latent Dirichlet allocation based diversified retrieval for e-commerce search. In: Paper Presented at the Proceedings of the 7th ACM International Conference on Web Search and Data Mining (2014)

Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W.: A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinform. **16**(13), S8 (2015)

# Between Data Mining and Predictive Analytics Techniques to Cybersecurity Protection on eLearning Environments

José Manuel, Raul Cordeiro, and Carla Silva(✉)

ULHT, Lisboa, Lisbon, Portugal
jmf@uninova.pt, cinel.raul@gmail.com, carla.silva@ulusofona.pt

**Abstract.** This paper aims to present a hypothetic theory of intelligent security system. In society the threat of cyber-attacks is getting louder and the use of computers, criminal activity has also changed from physical to cybernetic intrusion. There had been many cyber security solutions used to counteract these attacks, however we highlight the importance of self-protected systems in defense and in a correct analysis of cyber attacks. The internet is vulnerable to cyber-attacks as well as the information found in data systems and through a form of recognition and extraction of relevant information, we can represent data as shared data and integrated to intelligent system. What was used us a static firewall is now intended to be dynamic and self-critical. By techniques of data analysis, statistics, machine learning, data mining, the cybersecurity and privacy challenges are within our reach. This paper examines data mining techniques in order to predict pathways of Internet security and which considerations are involved in the theoretical solutions presented for the privacy systems such as the e-Learning environments.

**Keywords:** Data mining · Predictive models · Cybersecurity · Intelligent firewall · Intrusion detection systems

## 1 Introduction

In our current society, the threat of cyber intrusion is increasingly high and harmful. With the rise of usage in computers, criminal activity has also shifted from physical intrusion into cyber intrusion. Many well-known cybersecurity solutions are in place to counteract these attacks.

The Web is typically our first source of information about new software vulnerabilities and cyber-attacks. Information is found in semi-structured vulnerability databases as well as in text from security bulletins, news reports, cyber security blogs and Internet chat rooms. It can be useful to cyber security systems if there is a way to recognize and extract relevant information and represent it as easily shared and integrated semantic data.

Increasingly, techniques from data analytics fields of statistics, machine learning, data mining, and natural language processing are being employed for challenges in cyber-security and privacy. This paper examines which techniques from these fields are

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_17

essential for current and future cyber-security actioners and which considerations involved in successfully solving security and privacy challenges of the future.

Intrusion detection systems provide the ability to identify security breaches in a system by any action which is unauthorized. Current methods used for these systems are anomaly detection or a trusted signature database.

Cyber security is the efficient and effective security of computer systems to ensure the security of the data and security of communication processes between them. This means blocking reading and unauthorized access especially by write actions in media. Its important to define very clearly what we want to protect, we can know what are the means indicated and actions more recommended. In general we can say that the function of an effective cyber security system will be to protect vital, confidential and important information, as well as to avoid manipulation and unauthorized modification of the vital system parameters to the safety and survival of the system.

It is therefore necessary to identify correctly in each case, what are the real security needs and what procedures applied. These depends on each threat in concrete, by conducting a careful and comprehensive analysis of the types of attacks, the data analysis by Date Mining techniques, make it possible to build directed defense systems precisely against such attacks, specifically build optimized and focussed firewall systems.

Our solution provides knowledge to understand which data mining tools identified a log file, detecting patterns that may be considered an unauthorized activity. The tool gains additional patterns and grows more effective which allowed us to detect password cracking and Denial-of-Service (DoS) attacks, between other kinds of intrusion in the firewall systems. Such behaviour is introduced to enhance all traditional security techniques.

In education systems the most vulnerable platforms are the e-learning and b-learning systems. If in those laboratories harmful actions are pursued, safety and integrity can be attacked because some includes access to remote labs (RCMS), and it can destroy or damage some of its components.

E-learning systems or b-learning allow student to access the materials and exercises that can be performed any day or time according to convenience, helping to create an educational and training system, centralized on students. However, there are several administrative information (notes, statements of work and other assessment elements) that usually become a target for hackers.

It is therefore extremely important ensure proper protection of these systems, and to do this it's necessary identify what threats and common vulnerabilities exists and implement effective defense systems, such as "Firewalls and Honey pot" that are more effective and resistant to attacks. The objective is to enhance data mining analysis to build effective defense systems that can predict these kinds of attacks.

## 2 What Is the Relation Between Data Mining and Predictive Analysis?

As the role of information reaches a dimension that goes beyond what everyone thought, this dimension becomes excessive. Data mining is a field of computer science, which involves discovering patterns from large data sets through methods of artificial intelligence, machine learning, statistics, and database systems. The main aim of the data mining process is to extract information from a data set and transform it into an understandable format for future use. Apart from basic analysis, the data mining process covers database and data management aspects, data pre-processing, inference considerations, complexity considerations, post-processing of discovered structures, and online updating [1]. New data technology has been characterized by relational technology and increased research and development activities in new and powerful data systems, like promoting advanced data models such as extended, object-oriented, object-relational, and deductive relationships models. Application-oriented database systems, including spatial, temporal, multimedia, active, flow and sensor, scientific and engineering data, knowledge bases, and office information bases, have flourished. Issues related to the distribution, diversification and sharing of data have been extensively studied. Heterogeneous database and global information systems based on the Internet, such as the World Wide Web, play a vital role in the information industry. Data streams that include the World Wide Web flow in and out as in applications such as video surveillance, telecommunications, and sensor networks. Effective and efficient analysis of data in such different ways becomes a challenging task. Without powerful tools the increasing amount of data collected and stored in large repositories (*New arrival data* and *Network data*) are far of our human capacity to understanding [2].

It is through these precious tools that we can mine and supervise the data and find patterns. Therefore, an increasingly important task in data mining is to explore complex types of data. In addition, many data mining applications need to undermine standards including sequential patterns, subgraph patterns, and features in interconnected networks [3].

- Clustering – It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – It is the task of generalizing known structure which can be applied to new data. For example, an email programming attempts to classify an email as genuine or spam. Regular algorithms are decision tree learning, Naïve Bayesian classification, neural networks (soft computing) and support vector machines.
- Regression - Attempts to find a function which models the data with the least error. Regression analysis is a statistical methodology that is most often used for numeric prediction hence the two terms are often used synonymously. Classification and numeric prediction are the two major types of prediction problems [4].
- Association Rule Learning - Searches for relationships between variables.

In literature, a number of data mining based algorithms have been proposed to deal with the information security and privacy problems, by using approaches like classification, frequent pattern mining, and clustering methods to do intrusion detection, anomaly detection, and privacy preserving [5]. Application of these data mining methods have resulted in stimulating results that has concerned many researchers in both data mining and information security areas.

## 2.1 Data Mining Predictive Analysis

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can provide us with a better understanding of ***big data***. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data [6].

### Algorithms for Building the Predictive Model

The collected historical/log data from the network are learned by the classification algorithms for building the predictive model to identify the hackers and attackers. In this section the most popular classifiers also called as supervised learners namely probabilistic algorithm Nave Bayes (NB), tree based C4.5 (J48) and Instance based IB1 (Instance-based) like described [7].

NaiveBayes (NB): Bayesian classifiers are statistical classifiers. They can predict class members' hip probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classifiers are statistical classifiers. They can predict class member ship probabilities, such as the probability that a given tuple belongs to a particular class. This classification algorithm uses Bayes' theorem and the features of the training dataset are assumed as independent to the given class labels for building the predictive model. This classifier relies on discriminant function as seen in Eq. (1):

$$\int_{i}(\boldsymbol{X}) = \prod\nolimits_{j=1}^{N} \boldsymbol{P}(\boldsymbol{x}_i|\boldsymbol{C}_i)\boldsymbol{P}(\boldsymbol{C}_i) \tag{1}$$

This algorithm computes the conditional probabilities $P(x_j|c_i)$ and prior probabilities $P(c_i)$ on given training dataset to build the predictive model. $P(c_i)$ are computed by counting data which present in the Class label $C_i$ divides the resultant count based on the number of the training data. The same way is followed to compute the probabilities through observed frequency of feature distribution in $x_j$ within the training dataset which is labelled. The posterior probability is computed on each class to predict the unknown labelled data.

C4.5 (J48): This algorithm uses the decision tree[1] to build the predictive model (Fig. 1). The decision tree is constructed in numerous methods. All these methods

[1] A decision tree is a tree data structure consisting of decision nodes and leaves. A leaf species is a class value. A decision node species a test over one of the attributes, which is called the attribute selected at the node. For each possible outcome of the test, a child node is present. In particular, the test on a discrete attribute A has h possible outcomes $A = d_1, \ldots, A = dh$, where d1;:::dh are the known values for attribute A. The test on a continuous attribute has two possible outcomes, A t and A > t, where t is a value determined at the node, and called the threshold [8].

convert the given dataset in to a tree structure. The nodes of the tree represent the features and the edges represent the association between the features by value of features the lowest level of the node represents the class label. Recursively the value of the features are calculated by the information gain or entropy measure to convert the training datasets in tree structure. The low entropy and high information gain value, the feature is selected as repetitive node, split and convert the dataset in a tree structure. The tree structure is used as a rule to predict the unlabeled data in prediction.

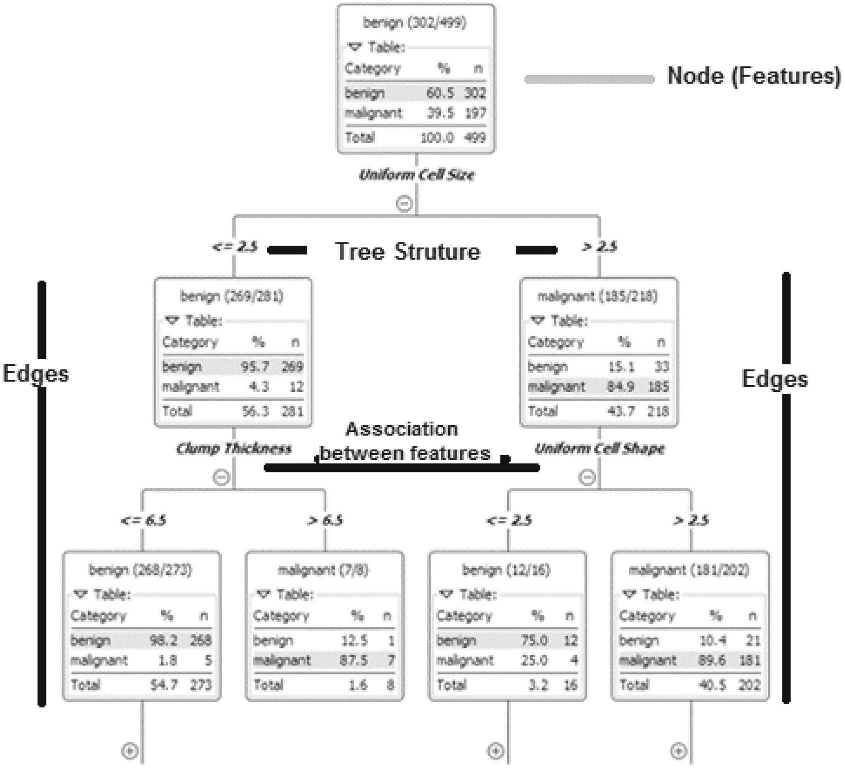


**Fig. 1.** Decision tree as a predictive model [8]

IB1: This algorithm uses the nearest neighbor principle to construct the predictive model. In this approach, the distance between the training instance and the given test instance are calculated by the Euclidean distance measure. If more than one instance has the smallest distance to the test, the first found is used. Nearest neighbor is one of the most significant learning algorithms; it can be adapted to solving wider problems. Let a dataset D has X instances ($X_1$, $X_2$, $X_{3,}$ …, $X_n$) and F feature ($F_1$, $F_2$, $F_3$, …, $F_{m)}$ with the class label $C_j$ where j = 1, 2, …, K.

This algorithm ranks the distance value of the neighboring instances to predict the unlabeled data X with the Class label. In this way the unlabeled data is predicted by the weight to calculate the nearest neighbors of the particular class to predict the unknown data. The algorithm operates on a set of d-dimensional vectors, D = {$x_i$|i = 1, …, N}, where $x_i \in R^d$ denotes the $i^{th}$ data point. The algorithm is initialized by picking *k* points in $R^d$ as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is the expectations (weighted mean) of the data partitions. The algorithm converges when the assignments (and hence the $c_j$ values) no longer change [9]. The number of iterations required for convergence varies and may depend on N, but as a first cut, this algorithm can be considered linear in the dataset size. One issue to resolve is how to quantify "closest" in the assignment step. In Eq. 2

$$\sum\nolimits_{i=1}^{N}\left(\boldsymbol{argmin}\left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|_2^2\right) \tag{2}$$

will decrease whenever there is a change in the assignment or the relocation steps, and hence convergence is guaranteed in a finite number of iterations. The greedy-descent nature of k-means on a non-convex cost also implies that the convergence is only to a local optimum, and indeed the algorithm is typically quite sensitive to the initial centroid locations.

Generally, volumes of tremendous and potentially infinite data flows are generated by real-time surveillance systems, communication networks, Internet traffic, and online transactions in the financial market or retail industry, electrical networks, manufacturing processes in the industry, Scientific and engineering experiments, Sensors and other dynamic environments. Unlike traditional datasets, data flow moves in and out of a computer system and with different refresh rates. It may be impossible to store an entire data stream or scan it multiple times because of its tremendous volume. To discover knowledge or patterns from data flows, it is necessary to develop methods of processing and analysis of single-scan, online, multilevel and multidimensional flow [9]. The knowledge about these attacks is acquired from the huge volume of network data with data mining tools. This knowledge facilitates the security system to identify the attackers or hackers based on their behavior in a network. Since the internet access is getting cheaper, people are always connected by computer or mobile phones. Therefore to protect the information exchanged over internet, cyber security standards are required, which enable organizations to practice safe security techniques to minimize the number of successful cyber security attacks. In the current scenario cyber-attacks and digital spying are identified as the biggest threat to any nation [5].

## 2.2 Data Mining Analysis on Cyber – Attacks in LMS and Remote Laboratories

There are several types of attacks to the informatics systems and networks [10]. Each time they are more common and frequent and it's necessary to find more effective prevention measures and techniques that avoid or at least minimize the effect of these attacks. In Fig. 2 the data is captured from the data communication traffic and they are classified in “package” section. By a filter process with the application of algorithms, the patterns of cybernetic trends (virus, worms, Trojans, other more) is identified and stored in “Detect know attacks” database that will be used as a data repository to the deep analysis process. With this analysis the attack and hacker profile is built by the system and confirmed by the human analyst. Through this identification a general summary of attacks is produced and cataloged identifying the most important and serious kinds of attacks.

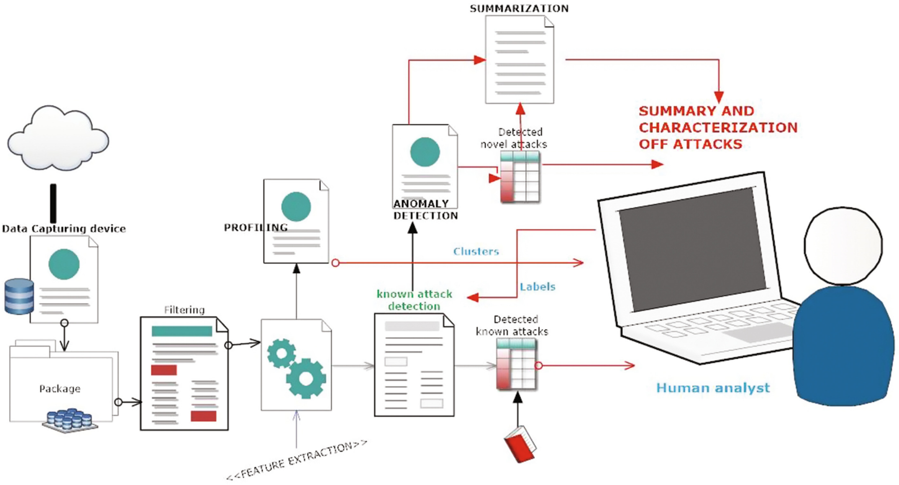


**Fig. 2.** Data mining analysis of cyber attacks

### Anomaly Detection

Anomaly detection approaches and builds models of normal data and detects deviations from normal model in observed data. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning. Anomaly detection algorithms have the advantage that they can detect emerging threats and attacks (which do not have signatures or labeled data corresponding to them) as deviations from normal usage. Moreover, unlike misused detection schemes (which build classification models using labeled data and then classify an observation as normal or attack), anomaly detection algorithms do not require an explicitly labeled training data set, which is very desirable, as labeled data is difficult to obtain in a real network setting [11].

### Profiling Network Traffic Using Clustering

Clustering is a widely used data mining technique which groups similar items, to obtain meaningful groups/clusters of data items in a data set [12]. These clusters represent the dominant modes of behavior of the data objects determined using a similarity measure. A data analyst can get a high level understanding of the characteristics of data set by analyzing the clusters. Clustering provides an effective solution to discover the expected and unexpected modes of behavior and to obtain a high level understanding of the network traffic [11].

## 3 Application of Predictive Analysis to LMS, RCMS Cyber Protection Systems

The browsers features define behaviours, than it is essential for students to understand the functionalities and features of the browser they use. Understanding what features can enable/disable the system, will help them to determine how they affect their privacy and security of their computers. This indicates specially, what features does browsers have to help protect the students from dangerous downloads and to help secure the connection between them and browsed sites, helping to defend them against browser attacks.

All browser vendors propose to their users through the graphical user interface, options panels allowing them to enabling or disabling manually some features setting. The application of this predictive technique intends that with the data obtained through this technique of analysis, it is possible to predict more easily the computer attacks that may occur and design more effectively the protection systems, especially the firewall applications. In Fig. 3 the firewall is installed in the router [13].

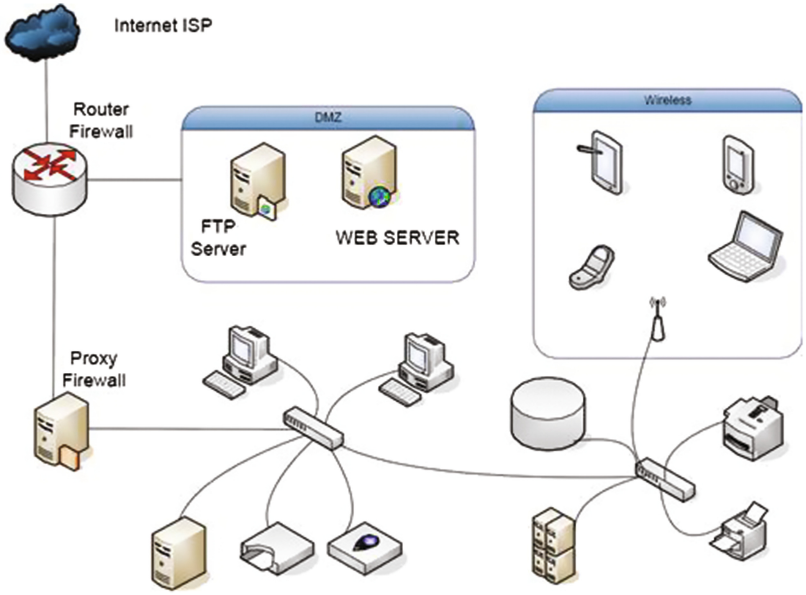


**Fig. 3.** Firewall between tree area networks: Internet, DMZ an internal network

The main function of the router with a firewall is to protect the internal network of external attacks. In this case the router protects the network and also makes the interface between the internet, the Demilitarized Zone (DMZ) and the internal network. The firewall is strategically located between the 3 networks and specially protects all the computers and devices on the internal network. Some firewalls today are already included in the operating systems, protecting this way the computer where they are installed. In Server operating systems is more common to use specific firewalls developed by software houses.

The functions of a perimeter firewall are much deeper that the simple protection of the computers against malicious code. They can also control all data traffic between the networks, identifying the protocols and codes used allowing ones and blocking other ones.

The perimeter firewalls, allow interconnect and secures several areas of a network or interconnect different networks with distinct levels of trust and security.

In e-learning systems and remote laboratory systems the computers that provide the physical access to the laboratories will be always on the internal network, and the LMS and the Service Broker Server of the remote laboratory system, will be in the DMZ to allow access from the internet. The Service Broker Server is a server that supports all the schedule software to make an agenda of the experiences which users try, and guarantees a secondary support to some databases (users, experiments, time scheduling), which communicates with the remote lab controller computer.

To increase the level of security all the calls to the remote lab system should be done through the LMS and the Server broker Service resident in DMZ. By this way all these calls should cross the firewall twice (with different security levels). This firewall is daily updated with Data Mining analysis techniques to define and constantly redefine their access lists.

## 4 Development and Future Research

In the actual cybernetic world, every day appears new trends and new types of attacks on systems. The traditional static Intrusion and Detection System and firewall it's not sufficient and efficient to properly protect a system. The solution is to create a complete new protection system that learns from input data provided by trends databases produced by software houses dedicated to security systems, which works as an intelligent system learning with data mining analysis.

With this strategy the Intrusion Data System will be continuously improved and the level of security and reliability of the system also increased. To reach this purpose is necessary to develop two fronts of work: Increase the quality of the algorithms to classify the trends patterns and develop continuously a major database and records of all types of attacks suffered buy the system. In intelligent network philosophy of protection we should aim for a system that, in addition to including intrusion detection systems and dynamic firewalls, also shares security information (algorithms and databases with properly classified threats) among all networks that should be protected to continuously increase levels of security and reliability.

## References

1. Lin, T., Hinke, T., Marks, D., Thuraisingham, B.: Security and data mining. In: Database Security, pp. 391–399 (1996)
2. Phridvi Raj, M.S.B., Guru Rao, C.V.: Data mining—past, present and future—a typical survey on data streams. Proc. Technol. **12**, 255–263 (2014)
3. Silva, J., Fonseca, C.: Educational data mining: a literature review. In: Advances in Intelligent Systems and Computing. Europe and MENA Cooperation Advances in Information and Communication Technologies (2016)
4. Verma, R., Kantarcioglu, M., Marchette, D., Leiss, E., Solorio, T.: Security analytics: essential data analytics knowledge for cybersecurity professionals and students. IEEE Secur. Priv. **13**(6), 60–65 (2015)
5. Woody, C., Ellison, R., Nichols, W.: Predicting cybersecurity using quality data. In: 2015 IEEE International Symposium on Technologies for Homeland Security, HST 2015 (2015)
6. Tobergte, D.R., Curtis, S.: Data manning **53**(9) (2013)
7. Ruggieri, S.: Efficient C4. 5 [classification algorithm]. Knowl. Data Eng. IEEE Trans. **14**(2), 438–444 (2002)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Elsevier, San Francisco (2006)
9. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
10. Scambray, J., Mcclure, S., Kurtz, G.: Hackers Expostos. Pearson Education, London (2002)
11. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. **1**(2), 12–23 (2000)
12. Varghese, B.M., Jose Tomy, J., Unnikrishnan, A., Poulose, K.: Clustering student data to characterize performance patterns. Int. J. Adv. Comput. Sci. Appl. **2**, 138–140 (2010)
13. Saliah-Hassane, H., Correia, R.C., Fonseca, J.M.: A network and repository for online laboratory, based on ontology. In: IEEE—2013 IEEE Global Engineering Education Conference, pp. 1177–1189, March 2013

# A Performance Evaluation of Chi-Square Pruning Techniques in Class Association Rules Optimization

Han Chern-Tong(✉) and Izzatdin Abdul Aziz

Computer Information Science Department, Universiti Teknologi PETRONAS,
UTP, Tronoh, Malaysia
jathniel.tong@gmail.com, izzatdin@utp.edu.my

**Abstract.** Associative classification is recognized by its high accuracy and strong flexibility in managing unstructured data. However, the performance is still induced by low quality dataset which comprises of noised and distorted data during data collection. The noisy data affected support value of an itemset and so it influenced the performance of an associative classification. The performance of associative classification is relied on the classification where the classification is worked based on the class association rules which generated from frequent rule mining process. To optimize the frequent itemsets based on the support value, in this research, we proposed a new optimization pruning technique to prune decision tree according to the correlation of each decision tree branches using genetic algorithm.

**Keywords:** Data mining · Associative classification · Genetic algorithm · Association rules mining · Pruning · Decision tree

## 1 Introduction

When it comes to select a technique for a particular problem, the choice will be very crucial since one single data mining technique could work well for a task and poor elsewhere. There are many factors that must be considered before taking such decision, such as the size of the data set, the attribute types, the number of attributes in the data set, and the goal of the task [1].

Association rule mining, ARM is one of the fundamental parts in data mining which it encounters rules that pass certain user constraints in a data set [2, 3]. ARM is a solid tool to capture the relationships among items inside a transaction database [4–7]. The association rules illustrate how robust the relation between each attribute-value pairs (or item) that grow repetitively in a specified data set [8]. The finding of association rules is based on frequent item-set mining or frequent pattern mining. Association rules must satisfy certain criteria regarding their accuracy (confidence) and the proportion of the data set that they actually represent [4], thus, it is capable to achieve global optimality.

Once the association rules are obtained, the classification took place and the prediction process could be started by using the provided association rules.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_18

Classification is another fundamental task in data mining. Given a collection of records in a data set, each record consists of a group of attributes and one of the attributes is a class label. The classification task involves constructing a model from the classified objects, in order to categorize earlier unseen objects as precisely as possible [9]. This process involves prediction of future class labels, whereas ARM involves only the description of the relationship among items in a database. In addition, there is one and only one pre-specified target class in classification; however, the target classes for association rule are not pre-specified.

By integrating the association rule mining and classification, a new approach is introduced called Associative Classification. A few successful classifier based on associative classification have been presented in last few years, such as CBA [10], CMAR [11], CPAR [12].

The integration is done by focused on a special subset of association rules whose right hand side is restricted to the classification class attribute. There are three basic steps in associative classification: discretizing continuous attributes, generating all the class association rules (CARs) and building a classifier based on the generated CARs.

A class association rule is generally expressed as IF-THEN rule, i.e., IF [term1 AND term2 AND …] THEN [class]. Each term of the antecedent is a pair of [attribute, value]. The consequent is the result of classification that is the class value of the attributes. However, the association rules that are generated are various and some of them are inefficient. Therefore pruning techniques must be applied to prune out redundant or shabby rules which will affect the classifier accuracy.

# 2 Problem Statements

## 2.1 Scoping Area of Problems

Associative Classification (AC) derive a large set of rules [11] because the classification data sets are highly correlated and the ARM methods that are used for rules discovery. As a result, there have been many attempts to reduce the size of classifiers produced by AC approaches, mainly focused on preventing rules that are either redundant or misleading from taking any role in the prediction process of test data objects. The removal of such rules can make the classification process more effective and accurate.

Several pruning methods have been used effectively to reduce the size of associative classifier, some of which have been adopted from decision trees, such as pessimistic estimation, others from statistics such as chi-square testing. These pruning techniques are utilized during the construction of the classifier; for instance, a very early pruning step, which eliminates association rules that do not passed the support threshold, may occur in the process of finding frequent itemsets. Another pruning approach such as chi-square testing may involve when generating the CARs, and late pruning methods, such as database coverage, could also carried out while building the classifier.

In ARM, a transaction or training object can be used to generate many rules; therefore, there are tremendous numbers of potential rules. Without adding support value threshold on the rule discovery and generation phases or imposing appropriate pruning, the very large number of rules will decrease the understandability and the maintainability of the outcome. Therefore, pruned out noisy and redundant rules becomes an important task.

### 2.2 Narrowed Down to Statements

The problem attached to the current pruning technique is that the CARs generated from the ARM are tending to overfit to the training data [13, 14]. The over fitting of rules is mean by the CARs that has been generated is too specific to a case with the result that it perform well on the training data but weak on unseen instances.

Consequently, the CARs generated from the training data must be general to all cases and must allow a small loss of accuracy [11]. So that the CARs obtained from it could accept new unseen data. The over fitting problem could be derived from the factors such as the dataset is too small which could only provide limited association rules to be mined out or the dataset contained too much of noise data.

The timing to prune a decision tree also plays an important role as if it is too early the pruning process took place the decision tree is not yet matures and so it will killed a lot of interesting rules. Therefore, the decision tree needed to be structured first before pruning process begins.

The pruning technique proposed by [11] comes with a great loss of information as the technique itself cut out the whole branch of nodes from the decision tree. With this method, the decision tree is hard to achieve its optimal level as some of the nodes inside a branch may be crucial. Therefore, it is needed to consider each tree node contribution to a decision tree instead of blindly taken a branch away from decision tree.

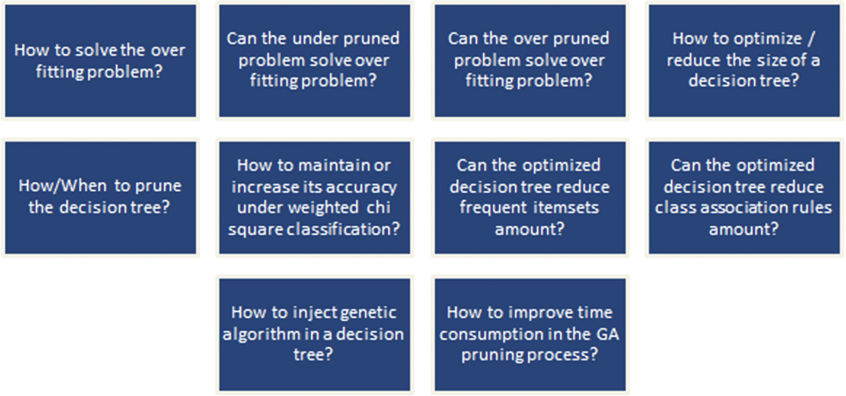
When a decision tree is built, it consists of noise and misclassified data. Therefore, it is needed to test each support value of a node to check its influence to the decision tree. Besides that because an association rule is built up by a few nodes from the decision tree which represented as a branch, all forms of combination should be tested.

Since a class is not decided just by using only one association rules, the correlation between each branch also should take into account. To handle the complexity of the evaluation process an optimization technique must be applied. Therefore, genetic algorithm has been chosen to simplify the process.

However, the traditional genetic algorithm consumes quite an amount of time. The calculation of the fitness value is also feed on the resources of an experiment process such as the time, CPU bandwidth, and computer memory. Thus, an advanced genetic algorithm is introduced to our proposed pruning technique to avoid resources being dried out.

## 3 Research Questions

In order to produce a good pruning technique, there are some aspects that are needed to focus on as in Fig. 1. In a decision tree, each node represents an itemset and a branch from it is meanwhile part of an association rule [15]. Pruned out any part of the decision tree will affect how the association rule being produced. If a pruning technique focuses only on the accuracy, it will produce association rules that have small error rate.



How to solve the over fitting problem?

Can the under pruned problem solve over fitting problem?

Can the over pruned problem solve over fitting problem?

How to optimize / reduce the size of a decision tree?

How/When to prune the decision tree?

How to maintain or increase its accuracy under weighted chi square classification?

Can the optimized decision tree reduce frequent itemsets amount?

Can the optimized decision tree reduce class association rules amount?

How to inject genetic algorithm in a decision tree?

How to improve time consumption in the GA pruning process?

**Fig. 1.** Questioning about targeted problems

However, this kind of pruning technique tends to over prune the decision tree. On the other hand, under pruned problem occurred if the decision tree does not clean deep enough and left out those misleading branches.

Overfitting problem [16], the rules are created from an associative classification are said over fitting when the rules are fit only to the training data set. When rules are pruned by refer to the training set and make it perfectly in shape only to that particular condition, it will has high classification accuracy on the training data but decreasing when test against test data [11].

To produce an optimized decision tree, it must not only focus on the accuracy but also the organization of the decision tree. A small error rate decision tree or a small size decision tree is not always the most optimized one. As the nodes represent the frequent itemsets and the class association rules are the branches, cut out any part of the decision will distort the decision tree. In order to check a node effect in a decision tree, genetic algorithm is used.

## 4 Objectives to Solve

Based on the problem background we discussed above, there are some research objectives derived in order to enhance current AC pruning technique. Firstly, this research we carried out is aimed to fine-tune a decision tree. By optimizing the decision tree using this newly pruning technique, we able to subsequently solving the projected problems: over fitting problem, under pruned problem, and over pruned problem.

Secondly, the research is sought to construct a genetic algorithm based pruning algorithm in associative classification. This objective is response to the research question questioning on how and when to pruning the decision tree and what is the optimized size of a decision tree. In order to achieve this idea, the algorithm is focusing in reducing the size of decision tree (to produce less frequent itemsets, less class association rules, and less complexity of the classifiers).

In order to decide the significance of class association rules, there are a few criteria that the proposed algorithm abided in doing the judgment. Therefore, to embed weighted Chi-Square in proposed GA based pruning algorithm is the crucial part of this algorithm. By integrating weighted Chi-Square and GA, the proposed pruning algorithm tends to handle statistical irregularity data, misclassified data, and dataset with few records.

## 5 Experiment Setting

CMAR has been selected as AC technique to be enhanced of its pruning technique. With support level of 1% and confidence level of 50%, the decision tree has been generated using Apriori-TFP. After the total support tree is successfully created, genetic algorithm will take place to optimize it by using 5% of random mutation and 80% of Roulette crossover as common GA operators.

## 6 Experiment Result and Analysis

The result in Table 1 shown that the number of association rules produced from the corrosion dataset had successfully reduced from 44 rules to 8 rules, while improving the accuracy of the prediction from 93.3333% to 98.6667%. By using the proposed genetic algorithm enhanced pruning techniques, associative classification able to produce less but general rules in order to predict a corrosion outcome for oil and gas pipeline.

**Table 1.** Experiments result of CMARPGA

| | Before optimize | After optimize |
|---|---|---|
| Number of frequent/large (supported) sets in T-tree | 123 | 46 |
| Number of generated CMAR classification rules | 44 | 8 |
| Accuracy | 93.3333% | 98.6667% |
| Time consumption | 0.15 ms | 27.979 ms |

While having the pros from accuracy, the proposed model is though suffered from the increase of time from 0.15 ms to 27.979 ms. This increment is due to the extra pruning algorithm injected into the proposed prediction model. However, the time consumption in training and producing classifiers is insignificant as the training is run once a time where in real time the generated classifiers are used.

From the Table 2, the average testing accuracy of CBA is increased compare to previous experiment where in tenth fold cross validation CBA achieved 61.7% testing accuracy. For the high standard deviation in CBA, 19.7993, it is proven that CBA suffered from overfitting problem as discussed in Sect. 4. However, the proposed technique managed to maintain the testing accuracy of 87.5% with 1.6499 low standard deviation values. From the validation done, it is verified that the first objective of this research achieved by consistently in successfully predict the corrosion severity level.

Furthermore, from Table 2 we could find that the number of frequent itemset generated from proposed technique is the lowest comparing to other techniques. With the low amount of frequent itemset, the proposed techniques is managed to produce low number of high quality association rules. Hence, the proposed technique is achieved the third objective of maintaining or improving the testing accuracy while optimizing the decision tree.

**Table 2.** Tenth fold cross validation on 100 records

| Dataset | Accuracy | | | Frequent itemset | | | Association rules | | |
|---|---|---|---|---|---|---|---|---|---|
| | CBA | CMAR | CMARPGA | CBA | CMAR | CMARPGA | CBA | CMAR | CMARPGA |
| N100#1 | 71 | 84 | 88 | 1331.9 | 1331.9 | 244.3 | 25.6 | 73.9 | 17.6 |
| N100#2 | 77 | 63 | 86 | 1219.8 | 1148.8 | 205 | 26.8 | 55.9 | 17.5 |
| N100#3 | 79 | 82 | 87 | 1148.8 | 1148.6 | 199.1 | 25.2 | 69.6 | 13.3 |
| N100#4 | 49 | 81 | 87 | 1148.6 | 1265.4 | 215.5 | 6.5 | 69.8 | 18.1 |
| N100#5 | 36 | 79 | 90 | 1265.4 | 1061.4 | 172.9 | 8.1 | 45.6 | 15.6 |
| N100#6 | 85 | 78 | 86 | 1061.4 | 977.9 | 148 | 24.8 | 53.1 | 16.3 |
| N100#7 | 74 | 81 | 90 | 977.9 | 1121 | 191.2 | 21.9 | 83.2 | 17.3 |
| N100#8 | 70 | 81 | 89 | 1121 | 1202.6 | 220 | 22.1 | 74.7 | 17.2 |
| N100#9 | 28 | 77 | 86 | 1202.6 | 1212.6 | 166.6 | 2.8 | 73.4 | 16.9 |
| N100#10 | 48 | 65 | 86 | 1212.6 | 1219.8 | 224.1 | 6.1 | 65.3 | 18.4 |

From Fig. 2, the line chart shown the stability of proposed algorithm which having 1.6499 low standard deviation values when running through 10 sets of 100 records dataset. The low standard deviation is crucial to avoid high error rate in predicting the severity of corrosion in oil industry as huge profit loss will caused when miscalculation happened in shutting down the whole oil platform.

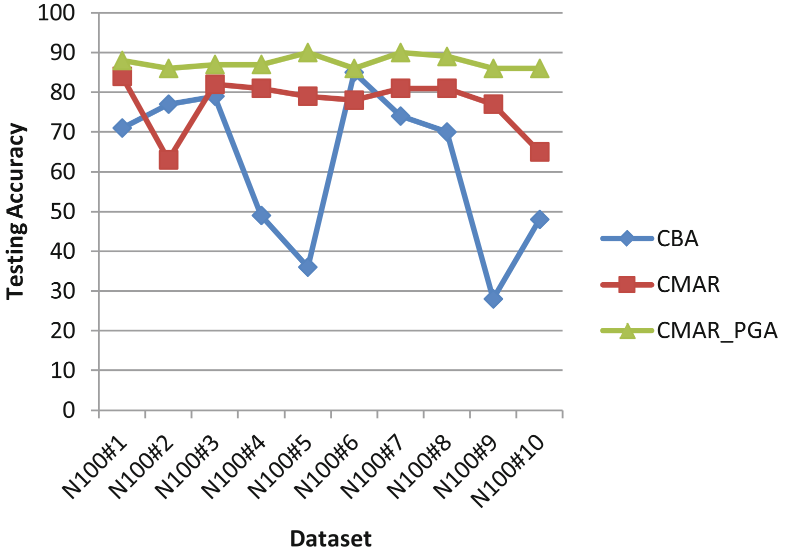


**Fig. 2.** Line chart of tenth fold cross validation on testing accuracy (100 records)

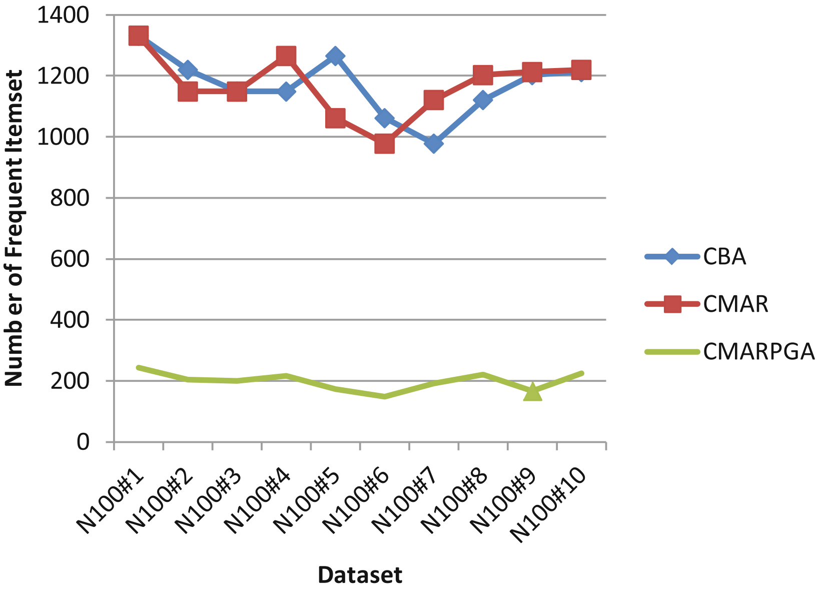


**Fig. 3.** Line chart of tenth fold cross validation on number of frequent itemset (100 records)

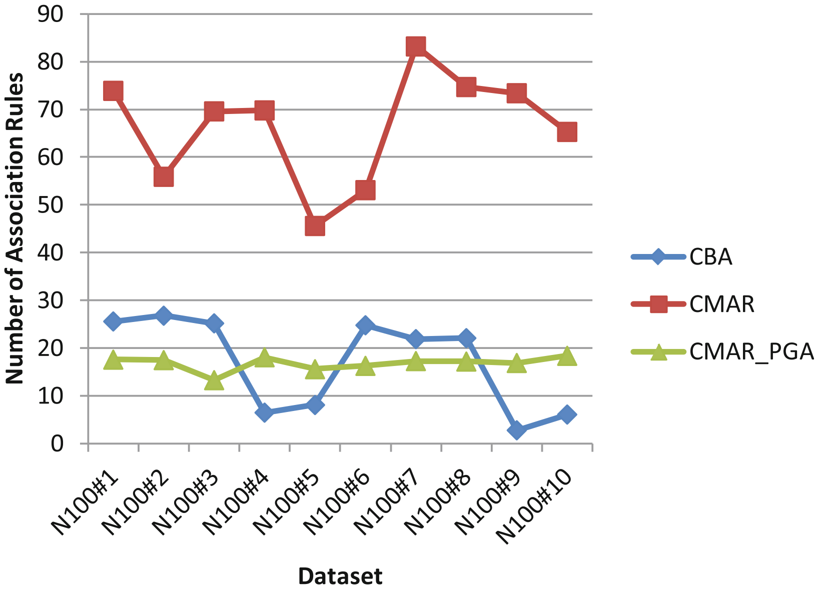


**Fig. 4.** Line chart of tenth fold cross validation on number of association rules (100 records)

As proposed technique able in providing stability and consistency in low records quantity environment, it is proven that the proposed technique had generalized the association rules by optimizing the decision tree. The generalization results we could view it in Figs. 3 and 4 which will be further discussed by looking at the number of frequent itemset and number of association rules from the tenth fold cross validation results.

Figure 3 shows that the number of frequent itemset generated by the proposed technique (green line) is the lowest amount among all other techniques which had verified that the decision tree had been pruned down. With the decision tree had been optimized, pruning out low quality nodes, the proposed technique able to produce a more general association rules without over fitting to the training data.

The low number of association rules shown in Fig. 4 is a result from tenth fold cross validation done toward 10 different set of 100 records dataset. The green line is a result of proposed technique, while, blue and red line are CBA and CMAR. With tenth fold cross validation, the proposed algorithm besides successfully in providing high testing accuracy consistently. It is as well managed to produce low amount of association rules. By using fewer amounts of association rules to predict an outcome and successfully in having high testing accuracy, it is verified that the second research objective: avoiding over fitting problem in training phase is achieved.

## References

1. Thabtah, F., Cowling, P., Peng, Y.: Real performance of categorization-based association rule techniques (2005)
2. Agrawal, S., Pandey, N.K.: A comparison between two association rule mining techniques. Curr. Trends Inf. Technol. **1** (2012)
3. Tran, A., Truong, T., Le, B.: Structures of association rule set. In: Intelligent Information and Database Systems, pp. 361–370 (2012)
4. Agrawal, R., Imielinski, T.: Mining association rules between sets of items in large databases (1993)
5. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules, pp. 487–499 (1994)
6. Agrawal, A., Thakar, U., Soni, R., Chaurasia, B.K.: Efficiency enhanced association rule mining technique. In: Advances in Parallel Distributed Computing, pp. 375–384 (2011)
7. Vedula, V.R., Thatavarti, S.: Binary association rule mining using Bayesian network (2011)
8. Tran, A., Truong, T., Le, B.: Structures of association rule set. In: Intelligent Information and Database Systems, pp. 361–370 (2012)
9. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Rec. **31**, 76–77 (2002)
10. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. Appeared in KDD-98, New York (1998)
11. Wenmin, L., Jiawei, H., Jian, P.: CMAR: accurate and efficient classification based on multiple class-association rules, pp. 369–376 (2001)
12. Yin, X., Han, J.: CPAR: classification based on predictive association rules. Society for Industrial & Applied Mathematics, p. 331 (2003)
13. Chen, J., Wang, X., Zhai, J: Pruning decision tree using genetic algorithms, pp. 244–248. IEEE (2009)

14. Fürnkranz, J., Gamberger, D., Lavrač, N.: Pruning of rules and rule sets. In: Foundations of Rule Learning, pp. 187–216 (2012)
15. Coenen, F., Leng, P., Ahmed, S.: Data structure for association rule mining: T-trees and P-trees. IEEE Trans. Knowl. Data Eng. **16**, 774–778 (2004)
16. Hawkins, D.M.: The problem of overfitting. J. Chem. Inf. Comput. Sci. **44**, 1–12 (2004)

# A Hybrid Method Based on Intelligent Water Drop Algorithm and Simulated Annealing for Solving Multi-depot Vehicle Routing Problem

Absalom E. Ezugwu(✉), Micheal O. Olusanya,
and Aderemi O. Adewumi

School of Mathematics, Statistics and Computer Science,
University of Kwazulu-Natal, Westville Campus, Private Bag X54001,
Durban 4000, South Africa
{ezugwua, olusanyam, adewumia}@ukzn.ac.za

**Abstract.** The vehicle routing problem and its variants such as the multi-depot vehicle routing problem are well-known NP-hard combinatorial optimization problems with wide engineering and theoretical background. In this paper a new hybrid technique based on intelligent water drop algorithm and simulated annealing is proposed to solve the multi-depot vehicle routing problem. The intelligent water drop algorithm is a stochastic population based metaheuristic optimization algorithm that uses a constructive approach to find optimal solutions of a given problem. Simulated annealing is a popular local search meta-heuristic approach with the key features of being able to provide a means to escape local optima by allowing hill-climbing moves with the hope of finding a global optimum. The performance of the hybrid algorithm is evaluated on a set of 23 benchmark instances and the results obtained compared with the best known solutions. The computational results show that the proposed method can produce good solutions, indicating that it is a good alternative algorithm for solving the multi-depot vehicle routing problem.

**Keywords:** Multi-depot vehicle routing problem · Metaheuristics · Intelligent water drops · Simulated annealing

## 1 Introduction

Managing fleet of vehicles which are outsourced for the distribution of specific number of products to a set of customers with specific supply and demand is considered an important challenge in distribution problems. The challenge here is not only restricted to making decision on the number of vehicles to be dispatched on the road, but in deciding how a customer receives a service and which customer receives services first based on assigned priority [1]. This type of problem can be presented as a vehicle routing problem (VRP) and modelled using graph based metaheuristic algorithms. The expected performance metrics involve determining the optimal sequence of customers to be visited by each vehicle, which satisfies the criteria such as travel time, the length

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_19

of route, and the cost involved in the operation [2]. The VRP optimization problem is widely-studied with several attractive solutions and different implementation techniques proposed in the literatures [3–6]. Similarly, several variants of the VRP which design concepts are based on the operational mechanism and mathematical modelling of the problem's diverse conditions in real-world applications have been studied recently in the literatures [7–10]. Some of these variants include [10, 11], capacitated vehicle routing problem (CVRP), heterogeneous fleet vehicle routing problem (HFVRP), multi-depot vehicle routing problem (MDVRP), periodic vehicle routing problem (PVRP), stochastic vehicle routing problem (SVRP), and vehicle routing problem with time windows (VRPTW).

The intelligent water drop (IWD) algorithm is a graph-based metaheuristic algorithm, which makes it suitable for solving VRP and its variants including the MDVRP. The IWD algorithm is a very simple and effective population-based optimization technique, which uses a constructive solution approach in finding optimal solution to a given problem [12, 13]. The IWD is inspired by natural phenomena, which is based on the idea of water drops and their interactions with the soils in the river beds. The process is such that each water drop would construct a solution by traversing in the problem search space and at the same time modifying its environment. The IWD algorithm has found applications in a wide range of optimization problems among which include the well-known travelling salesman problem [14], VRP and its variants [15], software quality assurance testing [16], and so on. Results from different literature show that the IWD algorithm compete favorably well with other state-of-the-arts metaheuristic algorithms [15].

In this paper, a new hybrid algorithm that comprises of IWD algorithm and simulated annealing (SA) is proposed. The reason for the hybridization is basically to optimize parameters that affect performance of the IWD algorithm using simulated annealing local search characteristics. Since SA has been applied to solve a record number of optimization problems and with fairly good results in most cases [17], it was selected on the basis of its high and better objective values, and its ability to move from the current solution to the neighborhood solutions. This invariably helps the search process escape from local minima in its search for the global optima by using the specified acceptance probability criteria to either accept or reject solutions with worse objective values.

The proposed hybrid algorithm referred to as IWD-SA in this paper is enhanced by the capability of the improved IWD and SA to explore and exploit the solution search space of the MDVRP in more efficient and effective way. Therefore, this paper presents a new metaheuristic hybrid algorithm based on the intelligent water drop algorithm and simulated annealing to solve the MDVRP, while the main purpose of the paper is to find even better search strategy and optimal set of performance parameters that will achieve high quality solution and faster convergence speed, especially with MDVRP benchmark problems with graphs ranging from 50 up to 360 nodes. In order to demonstrate the effectiveness of the proposed algorithm, we chose the Cordeau's MDVRP benchmark instances taken from [18], to test the performance of the new method. The experimental results obtained demonstrate the efficiency of our method when compared to the best known solution.

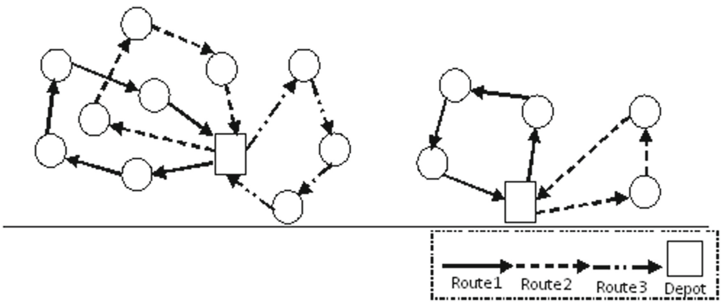


**Fig. 1.** MDVRP model with 2 depots and 15 customers

## 2 Multi-depot VRP Model

In [17], the MDVRP was presented as a least cost problem, with the objective of finding routes with the least cost from each designated depots to a set of geographically located customers. In modelling the MDVRP, certain assumptions are made regarding the vehicle routing plan, capacity, and customers. First, each route begins and ends at the same depot. Second, each customer is served exactly once by a vehicle. Third, the total demand on each route is less than or equal to the capacity of the vehicle assigned to that route. Finally, customer's demand can be met. An example illustration of the MDVRP with 2 depots and 15 customers is shown in Fig. 1. In solving the MDVRP, three decision making processes are involved [19], first is clustering, which deals with the grouping of the set of customers to be served by the same depot based on the distance of each customer to the servicing depots, second is routing, which is the assignment of customers of the same depot to several routs such that the capacity constraint of the vehicle is not violated and lastly, the scheduling, which handles the delivery sequence of each route in every depot.

The main objective of the MDVRP is to minimize the total delivery distance or time spent in attending to each customer. As shown in Fig. 1, the two rectangular boxes represent the actual depots, while the circles represent the actual customers to be visited. If we defined Fig. 1 in terms of a complete undirected graph $G = (V, E)$, where the set $V = \{1, 2, \ldots, n\}$ is the node set and $E = \{(i, i+1) : i, i+1 \in V, i < i+1\}$ is the edge set. A cost matrix $D = \left\{d_{i,i+1}, v_i, v_{i+1} \in V\right\}$ corresponding to the distance is defined on $E$. The cost matrix satisfies the triangle inequality whenever $d_{i,k} + d_{i+1,l} \leq d_{i,l} + d_{i+1,k}$ for all $1 \leq i < i+1 \leq n, 1 \leq k < l \leq n$ or $d_{i,i+1} \leq d_{i,k} + d_{k,i+1}$, for all $i, i+1, k$. In particular, this is the case of planer problems for which the nodes are points $p_i = (x_i, y_i)$ in the plane, and $d_{i,i+1} = \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}$ is the Euclidean distance. The triangle inequality is also satisfied if $d_{i,i+1}$ is the length of a shortest path from $i$ to $i+1$ on $G$. For the given MDVRP, the distance between customer $i$ and depot $h$, can likewise be represented as $d(i, h) = \sqrt{(x_i - x_h)^2 + (y_i - y_h)^2}$. The calculated Euclidean distance between customers and depots can then be used to make grouping decision of assigning customers to specific depots and routing decision of assigning customers on the same link to several routes.

## 3 Intelligent Water Drops Algorithm

The IWD algorithm problem solving approach is modelled in the form of a graph $G = (V, E)$, where $V$ and $E$ denote sets of nodes and edges. The structure of the graph depends on the problem representation. The orientation of the problem for which the IWD algorithm is supposed to optimize and find solution is usually viewed based on the assumption that there exists a node source from which the water drop moves through a selected path to the next unvisited node. The paths through which the water drops traverses have some loads of soil, therefore, the choice of selecting a specific path by the IWD is dependent on the amount of soil present on the unvisited path and the path with less soil is usually selected. However, during this process the velocity of the water drop may change depending on the quantity of soil is able to offset or accumulate along the selected path of movement. The whole process is repeated iteratively and the best path updated periodically until the best solution or global solution is found and updated subsequently, after which the algorithm is terminated. The main algorithm procedure is presented in Algorithm listing 1.

```
Algorithm listing 1: Basic IWD algorithm steps
Input: Graph $G = (V, E)$ where $V$ and $E$ are set of nodes and edges
Output: global best solution
Parameter initialization:
Set static parameters: population size, maximum iteration, initial soil, soil update parameters and velocity update parameters
do
    Initialize: IWDs, list of visited nodes, initial velocity, and the initial amount of soils load on water drop
    Construct solutions by IWDs
    Search for the current best solution
    Update the soils path that forms the current best solution
    Update the best solution
While (termination condition is not met)
Return the best solution
```

There are two type of parameter settings defined for the IWD algorithm; these include static and dynamic parameter settings. Examples of the static parameters are, termination criteria, which determine when the algorithm should be terminated, the initial soil paths and velocity update parameters, which are constant throughout the iterative execution of the algorithm. On the other hand, examples of the dynamic parameters includes list of visited nodes and initial amount of soil load on water drop, these type of parameters changes their values with every increment in the iteration steps.

## 4 IWD Framework for Solving MDVRP

In accordance with the IWD algorithm working procedures presented in algorithm listing 1, we adopt and configure the following steps to solve the MDVRP.

***Solution Construction***

The steps presented in this section help minimize the total dispatching cost of each vehicle assigned to a customer. First, a graph with $n$ nodes and $n(n-1)/2$ directed edges is constructed, which is used by the IWDs as input to construct solutions for the optimization process. In this case, a node denotes a customer, while an edge denotes a route to a customer. Every IWDs begins its journey starting from the first or initial node and terminate by visiting the end node on the graph. We therefore, discuss some of the important factors and parameters employed by the IWD algorithm to solve the optimization process of the MDVRP.

***Parameter Initialization***

The first step in the solution construction is the initialization of all the static and dynamic parameters. In each step of the solution construction process by the IWD algorithm, an empty set of list of nodes called tabu list, which the IWD is not allowed to visit is created. Let denote this set by $C = \{\ \}$.

***Probability Distribution Function***

In the proposed optimization method, each IWD utilizes a probability distribution function assigned along each edge starting from the current node through to all other nodes, which do not violate the assigned constraints of the problem under consideration. The selection mechanism for choosing an edge connected to the next node by the IWD is expressed as follows. Let an IWD be at the start node $i$, then the probability distribution denoted by $P_i^{IWD}$ which is required to select an edge that would allow IWD to be able to move from node $i$ to node $j$ is calculated using the fitness function given in Eq. (1) as follows.

$$P_i^{IWD}(i+1) = \frac{f(soil(i,i+1))}{\sum_{n \notin C} f(soil(i,n))} \tag{1}$$

$$f(soil(i,i+1)) = \frac{SM(i,i+1)}{\varepsilon + g(soil(i,i+1))} \tag{2}$$

$$SM(i,i+1) = d(h,i) + d(h,i+1) - d(i,i+1) \tag{3}$$

$$d(i,h) = \sqrt{(x_i - x_h)^2 + (y_i - y_h)^2} \tag{4}$$

where the function $SM(.)$ is the saving matrix proposed by Clarke and Wright [20], which is used to compute the distance travelled by the IWDs along the edges to visit the nodes or as in our case the distance travelled by the vehicles for serving the customers. Here the saving matrix is constructed for every two customers $i$ and $i+1$ on the same link path to the given depot $h$. The parameter $\varepsilon$ is a very small positive number assigned to prevent singularity or possible division by zero. The function $g(soil(i,i+1))$ serves as a shift function that moves the soil through an edge joining any two nodes $i$ and $j$ toward a positive value. The function is given as follows:

$$g(soil(i,i+1)) = \begin{cases} soil(i,i+1) & if \ \min\limits_{n \notin C}(soil(i,n)) \geq 0 \\ soil(i,i+1) - \min\limits_{n \notin C}(soil(i,n)) & otherwise \end{cases} \tag{5}$$

where $C$ denotes the set of nodes that IWD is not allowed to visit.

Therefore the edge selection procedure between two nodes $i+1$ and $k$ from node $i$ can be summarized based on the following conditions:

If $P_i^{IWD}(i+1) < P_i^{IWD}(k)$, then select edge $k$ to visit the connected node
If $P_i^{IWD}(i+1) > P_i^{IWD}(k)$, then select edge $i+1$ to visit the connected node
If $P_i^{IWD}(i+1) = P_i^{IWD}(k)$, then select edge arbitrarily to visit any of the connected node

***Set of Visited Nodes***

An updated list of the set of visited paths $C$ created earlier is maintained by the IWD as a means of keeping track of all the nodes already visited. Iteratively each node is first evaluated on the basis of whether it has been visited or not, before a decision is taken and if it is confirmed that the node has been visited previously, the node is deleted. The check is performed so as to prevent the IWD from traversing the same path twice. Relating this techniques with the MDVRP for example, having customer $i$ on the route of vehicle $k$ from depot $h$, and after $i$ has been visited by $k$, it is removed completely from that route and any subsequently visit by $k$ to $i$ is declared tabu for a specific number of iteration. This condition holds for as long as $k$ is only allowed to serve $i$ just once. Otherwise, the tabu status of $i$ can be revoked if the new solution is better than the current solution on the same route as $i$.

***Local Soil and Velocity Update***

As IWD move from node $i$ to node $i+1$, its velocity needs to be updated subsequently as follows:

$$vel^{IWD}(t+1) = vel^{IWD}(t) + \frac{a_v}{b_v + c_v.soil(i,i+1)} \tag{6}$$

where $a_v$, $b_v$, and $c_v$ are the IWD velocity updating parameters and $vel^{IWD}(t)$ is the previous velocity of the IWD. The time taken by IWD to move from node $i$ to node $i+1$ is computed as follows:

$$Time\left(i, i+1; vel^{IWD}\right) = \frac{d(i, i+1)}{\max(\varepsilon_v, vel^{IWD})} \tag{7}$$

$$d(i, i+1) = \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \tag{8}$$

The $max(.,.)$ returns the maximum value between its arguments'. This value is then used to threshold the negative velocities to a very small positive number $\varepsilon_v$. The function $d(.)$ represents the distance taken by IWD to move from node $i$ to node $i+1$.

The amount of soil on the link will reduce, since IWDs carries some soil as it traverses from node $i$ to node $i+1$. Therefore, the soil carried by the IWD as it moves along the link from node $i$ to node $i+1$ is updated as follows:

$$Soil(i, i+1) = (1-\rho).soil(i, i+1) - \rho.\Delta soil(i, i+1) \tag{9}$$

$$soil^{IWD} = soil^{IWD} + \Delta Soil(i, i+1) \tag{10}$$

$$\Delta Soil(i, i+1) = \frac{a_s}{b_s + c_s.Time(i, i+1; vel^{IWD})} \tag{11}$$

where $a_s$, $b_s$, and $c_s$ are the IWD soil updating parameters, $\rho$ is a small positive number $(0 < \rho < 1)$.

***Fitness Function***

The main reason for determining the fitness function is to increase the chances of finding a global best solution and to also improve the convergence speed of the IWD algorithm. The fitness function determines the ranking of the individual solution obtained by calculating the total length of the constructed route traversed by each IWD in the iterations. The solution with the minimum route length among all the IWD constructed routes is then taken as the best solution. Since this is a minimization case, then the route length denoted by $T^{IWD}$ can be expressed as follows:

$$T^{IWD} = \sum_{i=1}^{n-1} d\left(T_i^{IWD}, T_{i+1}^{IWD}\right) + d\left(T_n^{IWD}, T_1^{IWD}\right) \tag{12}$$

Therefore, the fitness function can be defined as follows:

$$f = min\left(\sum_{\forall T^{IWD}} \left(T^{IWD}\right)\right) \tag{13}$$

where $n$ is the total number of nodes or customers and the function $d(.)$ is the Euclidean distance between customer $i$ and customer $i+1$.

***Global Update***

To prevent IWD from plunging into local minima, the amount of soil on each of the current iteration's best solution with the minimum route length $T_M$ is updated subsequently as follows.

$$Soil(i, i+1) = (1-\rho).soil(i, i+1) + \rho.\frac{2.soil^{IWD}}{n(n-1)} \forall(i, i+1) \in T_M \tag{14}$$

If at the end of each iteration process, $T_M$ is found to be shorter than the best solution found so far denoted by $T_B$, then the best route is updated as follows:

$$T_B = \begin{cases} T_M & if\, f(T_B) \geq f(T_M) \\ T_B & otherwise \end{cases} \tag{15}$$

***Termination Condition***

The program is terminated once there is no further improvement on the global soil updating, that is after a number of successive iterations have been performed and this would also correspond to the value of the constant parameter referred to as the maximum number of iteration, which in our case is set to 100.

## 5 Simulated Annealing

Since its introduction as a solution method into the field of optimization techniques, SA algorithm has been used to solve several optimization problems, either on the basis of classical algorithm or as part of a hybrid algorithm with fairly good results in comparisons with other heuristic based algorithms [21–23]. The SA algorithm is deeply studied in the literatures and in some cases specifically applied to solve the VRP and its variants problems [24].

The introduction of SA into the IWD was on the basis of developing an improved local search method that prevents the IWD from getting stuck at local minima. Since SA can be viewed as a search process that can always attempt to move from one current solution to another solution in its neighborhood solutions, it therefore has the potential of providing better objective values. Unlike the hill climbing, SA is able to escapes from being trapped into local minima by allowing worse moves (lesser quality) or uphill steps to be taken at random some of the time. The SA choice of selecting best solution, is based on its movement procedure, which is such that, if the anticipated move is better than its current position then SA will always take it and if the move is worse, then it will be accepted based on some probability. For the MDVRP, the SA procedure begins by considering the solution $T_i^{IWD}|i = 1, 2, \ldots, n$, obtained by the IWD through the set of given customers with an update solutions $T_{i+1}^{IWD}$ created by randomly switching the orders of two customers. The cost function or fitness function, which represents the quality of the solution $T_i^{IWD}$, is denoted by $f\left(T_i^{IWD}\right)$. The relative change in cost $\Delta f$ between $T_i^{IWD}$ and $T_{i+1}^{IWD}$ is expressed as follows:

$$\Delta f = f\left(T_{i+1}^{IWD}\right) - f\left(T_i^{IWD}\right) \tag{16}$$

Beginning with the initial solution, only the solution which results in smaller fitness value than the previous solution is accepted by the algorithm, in other words, a solution is only accepted when the fitness value of $f\left(T_{i+1}^{IWD}\right) < f\left(T_i^{IWD}\right)$. However, accepting or rejecting a new solution with higher fitness values for $T_{i+1}^{IWD}$ can be based on the probability acceptance function given in Eq. 17 as follows:

$$p(f, T_k) = exp\left(-\frac{f\left(T_{i+1}^{IWD}\right) - f\left(T_i^{IWD}\right)}{kT_k}\right) \tag{17}$$

However, for large problem sizes, instead of using Eqs. 17 and 18 can be employed to improve the performance of the SA [25].

$$p(f, T_k) = exp\left(-\frac{f\left(T_{i+1}^{IWD}\right) - f\left(T_i^{IWD}\right)}{T_k}\right) \tag{18}$$

where $T_k$ is the temperature at the $k^{th}$ instance of accepting a new solution. The probability of accepting a new solution is a function of both the temperature of the system and the difference in the fitness value. It has been noted that the probability of accepting a worse solution decreases as the temperature deteriorates, which means that as the temperature reduces to zero, then only better solution will be accepted. In this paper the following cooling schedule (Eq. 19) is adopted.

$$T_{k+1} = \alpha T_k \tag{19}$$

where, $\alpha$ denotes the rate at which the temperature is lowered each time a new solution $T_{i+1}^{IWD}$ is discovered.

The new IWD-SA algorithm introduces the SA probability of acceptance criteria in determining between the current best cost and the new solution cost, which to choose as a better solution to be updated. The algorithm is able to perform this process by computing and comparing iteratively the quality of solution obtained by both the old and new IWDs, which revolves around $T_M$ and $T_B$. Therefore, the acceptance rule is evaluated based on two conditions namely, the fitness function or quality of solution and the environmental temperature. The acceptance criteria enable the IWD process to easily avoid entrapment in local optima and thereby increasing the rate of the algorithm's convergence and exploitation and exploration capability (also referred to as intensification and diversification). The detailed explanation of how the IWD and SA methods work together to achieve the aforementioned steps is presented in Algorithm listing 2, while Fig. 2 shows the flowchart of IWD-SA procedure.

**Algorithm listing 2**: IWD-SA algorithm steps

**Input**: Graph $G = (V,E)$ where $V$ and $E$ are set of nodes and edges

**Output**: global best solution

```
Parameter initialization:
Set static parameters: population size, maximum iteration, initial soil, soil update parameters and velocity update parameters
Do
Initialize IWDs, list of visited nodes, initial velocity of water drop, and initial amount of soils load on water drop
Construct solutions by IWDs
Do
  For $k$ = 1 to $N_C$ where $N_C$ is the number of nodes on the graph
Search for the current best solution by spreading the IWDs randomly on the problem graph
Update all the dynamic parameters to include: list of visited nodes, initial velocity of the water drop and the initial amount of soil loaded onto water drop
End for
While (not end of graph)
Calculating the fitness functions $f( T_i^{IWD})$ and $f(T_{i+1}^{IWD})$
  If $f(T_{i+1}^{IWD}) \leq f( T_i^{IWD})$ then
        $T_i^{IWD} \leftarrow T_{i+1}^{IWD}$
  Else
    If  $p > r(0,1]$ then
        $T_i^{IWD} \leftarrow T_i^{IWD}$
  End if
Update soil value for paths visited by the $IWD \in T_M$
Update the global best cost $T_B$ by comparing the fitness values of  $T_M$ and  $T_B$
If  $(f(T_B) \geq f(T_M))$ then
      $T_B = T_M$
Else
  $T_B = T_B$
End if
Update temperature $t_{k+1} = \alpha t_k$
While (termination condition is not met)
Return the best cost $T_B$
```

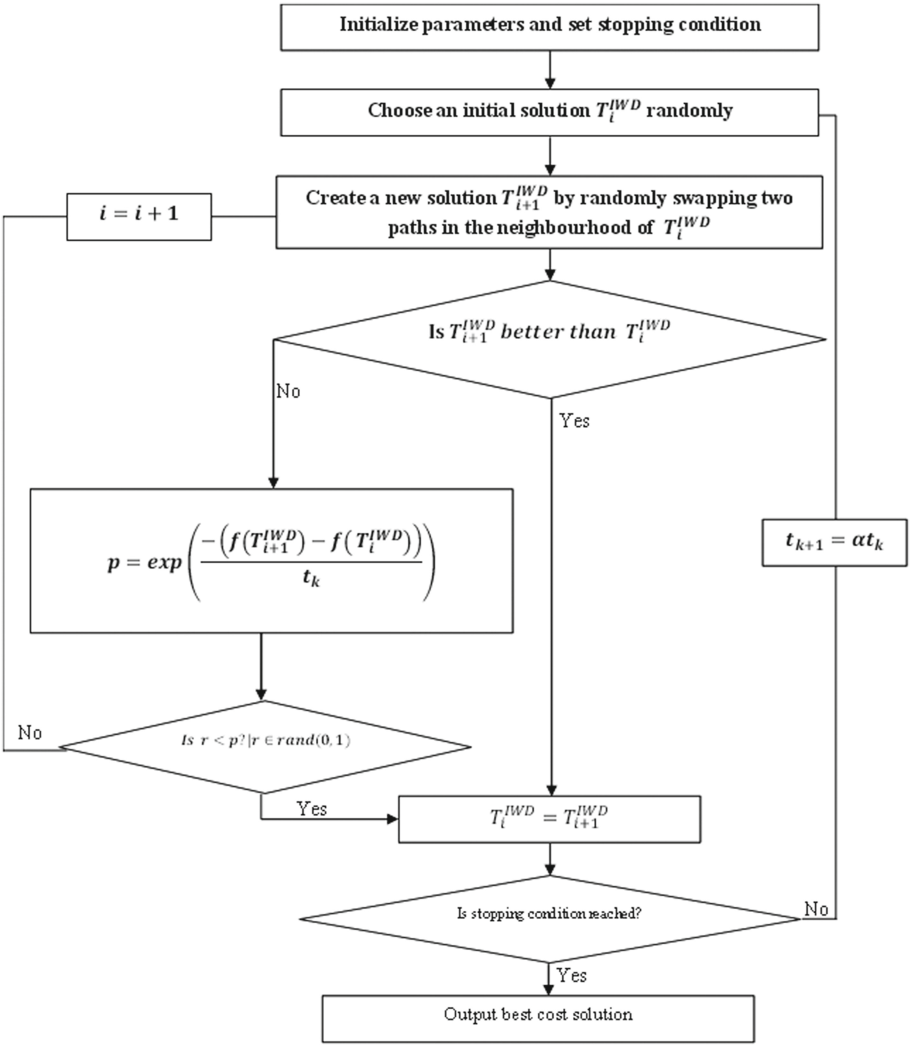

Initialize parameters and set stopping condition
Choose an initial solution $T_i^{IWD}$ randomly
Create a new solution $T_{i+1}^{IWD}$ by randomly swapping two paths in the neighbourhood of $T_i^{IWD}$
$i = i + 1$
Is $T_{i+1}^{IWD}$ better than $T_i^{IWD}$
No
Yes
$p = exp\left(\frac{-\left(f(T_{i+1}^{IWD}) - f(T_i^{IWD})\right)}{t_k}\right)$
$t_{k+1} = \alpha t_k$
Is $r < p? | r \in rand(0,1)$
No
Yes
$T_i^{IWD} = T_{i+1}^{IWD}$
Is stopping condition reached?
No
Yes
Output best cost solution

**Fig. 2.** Flow diagram of the proposed IWD

## 6 Numerical Results

This section presents computational results of the hybrid algorithm, which are compared with the best known solution from the Cordeau's instances taken from [18]. The IWD-SA algorithms were executed on Windows 7 OS, Intel Core i7-2600 CPU@3.40 GHz 3.40 GHz with 4 GB of RAM. MATLAB R2015a was used as the programming language. The MDVRP instances results described in [18] was benchmarked to evaluate the performance of the proposed method. As mentioned earlier that IWD has two types of parameters namely, static and dynamic parameters, the static parameters relative to the velocity update are set using similar theoretical values from the work presented in [12] as follows: $a_v = 1000$, $b_v = 0.01$, $c_v = 1$. For the soil updating parameters, we initialize the following variables $a_s = 1000$, $b_s = 0.01$, $c_s = 1$, while we initialize the amount of soil on each path to be *initial_Soil* = 1000, and the IWD velocity to be *initial_Velocity* = 100. Finally, for the SA parameters, we initialize the following variables: initial temperature $t_0 = 200$ and the temperature cooling rate $\alpha = 0.99$. Also, each instance of the problem was run 10 times for a maximum number of 100 iterations.

Table 1 show that the solution quality of IWD-SA is better than that of the standard IWD and in some cases competed favourably with the best known solutions (BKS). It is obvious that the IWD is outperformed by the IWD-SA based on the compared average values of the two techniques. This significant difference in the performance of the proposed method can be attributed to the local search and hill climbing characteristics of the SA introduced into the standard IWD. These two features of the SA are efficient mechanisms which assist the IWD to escape from being trapped into local minima and it increases also the explorative and exploitative power of the IWD within

**Table 1.** Computational results obtained for 23 Cordeau's MDVRP benchmark instances for improved IWD and IWD-SA

| Instance | | | | | IWD | | | | IWD-SA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S. No. | Inst. | n | h | BKS [18] | AVE | Best | Std. Dev | Gap (%) | AVE | Best | Std. Dev | Gap (%) |
| 1 | P01 | 50 | 4 | 576.87 | 576.87 | 576.87 | 0.00 | 0.00 | 576.87 | 576.87 | 0.01 | 0.00 |
| 2 | P02 | 50 | 4 | 473.53 | 491.17 | 473.53 | 0.48 | 0.00 | 479.03 | 473.53 | 0.07 | 0.00 |
| 3 | P03 | 75 | 5 | 641.19 | 648.19 | 643.58 | 1.58 | 0.37 | 641.19 | 641.19 | 0.03 | 0.00 |
| 4 | P04 | 100 | 2 | 1001.59 | 1010.23 | 1006.19 | 1.05 | 0.46 | 1002.23 | 1001.59 | 0.68 | 0.00 |
| 5 | P05 | 100 | 2 | 750.03 | 758.13 | 754.84 | 1.01 | 0.64 | 752.13 | 750.03 | 0.52 | 0.00 |
| 6 | P06 | 100 | 3 | 876.50 | 888.49 | 879.71 | 2.45 | 0.37 | 881.49 | 876.50 | 1.05 | 0.00 |
| 7 | P07 | 100 | 4 | 885.80 | 901.83 | 885.82 | 4.80 | 0.00 | 889.12 | 885.80 | 2.04 | 0.00 |
| 8 | P08 | 249 | 2 | 4437.68 | 4572.56 | 4492.39 | 24.56 | 1.23 | 4430.48 | 4430.48 | 0.66 | −0.16 |
| 9 | P09 | 249 | 3 | 3900.22 | 3942.81 | 3910.62 | 11.24 | 0.27 | 3912.81 | 3900.10 | 4.69 | 0.00 |
| 10 | P10 | 249 | 4 | 3663.02 | 3689.54 | 3663.29 | 7.44 | 0.01 | 3662.54 | 3659.84 | 0.88 | −0.09 |
| 11 | P11 | 249 | 5 | 3554.18 | 3695.27 | 3565.17 | 4.72 | 0.31 | 3601.72 | 3552.34 | 7.89 | −0.05 |
| 12 | P12 | 80 | 2 | 1318.95 | 1376.03 | 1324.34 | 16.63 | 0.41 | 1321.41 | 1318.95 | 1.51 | 0.00 |
| 13 | P13 | 80 | 2 | 1318.95 | 1353.95 | 1324.05 | 8.81 | 0.39 | 1318.95 | 1318.95 | 0.52 | 0.00 |
| 14 | P14 | 80 | 2 | 1360.12 | 1371.91 | 1369.38 | 0.65 | 0.68 | 1360.12 | 1360.12 | 0.44 | 0.00 |
| 15 | P15 | 160 | 4 | 2505.42 | 2565.02 | 2539.25 | 8.67 | 1.35 | 2515.02 | 2505.42 | 3.06 | 0.00 |
| 16 | P16 | 160 | 4 | 2572.23 | 2596.83 | 2580.91 | 4.16 | 0.34 | 2580.18 | 2572.23 | 1.84 | 0.00 |
| 17 | P17 | 160 | 4 | 2709.09 | 2729.23 | 2721.28 | 2.05 | 0.45 | 2709.09 | 2709.09 | 0.44 | 0.00 |
| 18 | P18 | 240 | 6 | 3702.85 | 3807.22 | 3743.12 | 11.95 | 1.09 | 3713.92 | 3702.85 | 3.58 | 0.00 |
| 19 | P19 | 240 | 6 | 3827.06 | 3951.21 | 3946.61 | 1.29 | 3.12 | 3839.21 | 3827.06 | 3.03 | 0.00 |
| 20 | P20 | 240 | 6 | 4058.07 | 4168.37 | 4109.06 | 15.09 | 1.26 | 4060.37 | 4058.07 | 0.53 | 0.00 |
| 21 | P21 | 360 | 9 | 5474.84 | 6101.68 | 5543.29 | 145.51 | 1.25 | 5531.48 | 5474.84 | 17.75 | 0.00 |
| 22 | P22 | 360 | 9 | 5702.16 | 5984.87 | 5736.01 | 63.39 | 0.59 | 5741.03 | 5702.16 | 15.58 | 0.00 |
| 23 | P23 | 360 | 9 | 6095.46 | 6145.58 | 6134.91 | 3.32 | 0.65 | 6091.43 | 6088.96 | 0.63 | −0.11 |

the given solution search space. To conclude our comparison, we now evaluate the gap between the two algorithms.

The gap between the computed results for the two algorithms relative to the best known solution is identified, the percentage gap being calculated as follows:

$$Gap = \frac{Best\ result - best\ known\ solution}{best\ known\ solution} \times 100\% \qquad (20)$$

The computed percentage gap results presented in Table 1 and illustrated in Fig. 3 shows that the IWD-SA has solved all the instances efficiently compared to the IWD algorithm and the deviations also never exceed 1% in all the instances considered. The result thus shows that our proposed method (IWD-SA) is more stable than IWD, since it's computed average solution is very close to the best known solution. Figure 4 shows that there is no significant difference between the run-time of the two methods, despite the exponential calculation of the acceptance probability function that often incurs additional computational cost for the IWD-SA approach. Therefore, we conclude that the IWD-SA is more stable and very suitable approach for solving large scale problems.

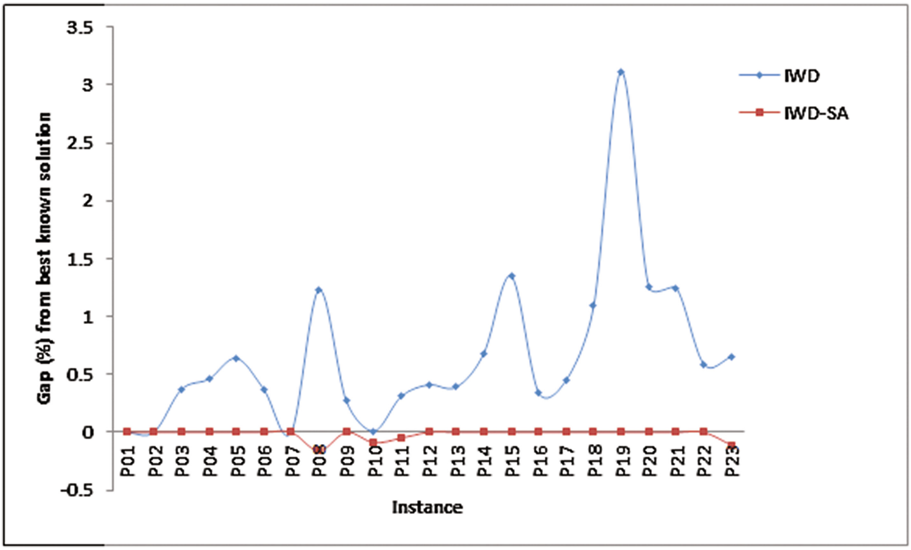


**Fig. 3.** Percentage gap of each algorithm best solution to the best known solution for 23 instance (over 10 runs)

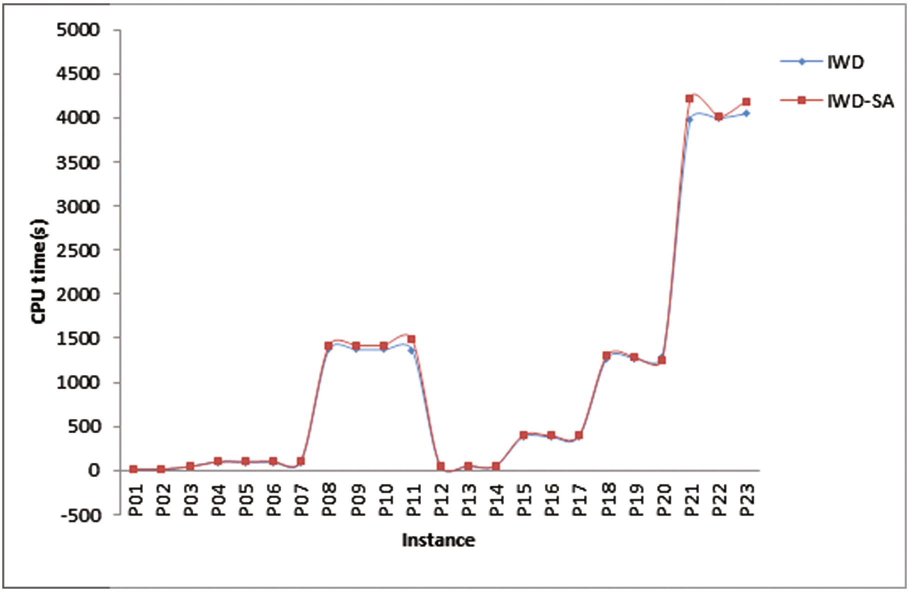


**Fig. 4.** Average running times for IWD and IWD-SA algorithms

## 7 Conclusion

In this paper the multi-depot vehicle routing problem has been studied. A hybrid algorithm that incorporates intelligent water drops algorithm and simulated annealing based local search for the MDVRP is presented. This study utilizes both the IWD algorithm and SA solution implantations to develop a better alternative algorithm for solving MDVRP NP-hard problem. The two algorithms presented namely, basic IWD and hybrid IWD-SA were tested using the Cordeau et al. [18] benchmark data, covering the instances P01-P23. Descriptive statistical analysis was also conducted to verify the performances of the proposed methods. The computational results show that the hybrid IWD-SA solution approach is effective and efficient in finding good and promising results. Future research direction may consider analyzing the impact of the exponential calculation of SA acceptance probability on the performance of the hybrid algorithm.

**Acknowledgements.** This paper was supported in part by the National Research Foundation, NRF CSUR, under the Research Grant CSUR160427163266. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. The authors are also grateful to the anonymous referees for their constructive comments that have greatly improved the presentation of this paper.

## References

1. Ai, T.J., Kachitvichyanukul, V.: Particle swarm optimization and two solution representations for solving the capacitated vehicle routing problem. Comput. Ind. Eng. **56**(1), 380–387 (2009)
2. Kachitvichyanukul, V., Sombuntham, P., Kunnapapdeelert, S.: Two solution representations for solving multi-depot vehicle routing problem with multiple pickup and delivery requests via PSO. Comput. Ind. Eng. **89**, 125–136 (2015)
3. Parragh, S.N., Doerner, K.F., Hartl, R.F.: A survey on pickup and delivery problems. J. für Betriebswirtschaft **58**(1), 21–51 (2008)
4. Giosa, I.D., Tansini, I.L., Viera, I.O.: New assignment algorithms for the multi-depot vehicle routing problem. J. Oper. Res. Soc. **53**(9), 977–984 (2002)
5. Anbuudayasankar, S.P., Ganesh, K., Mohapatra, S.: Survey of methodologies for TSP and VRP. In: Models for Practical Routing Problems in Logistics, pp. 11–42. Springer (2014)
6. Archetti, C., Speranza, M. G.: The split delivery vehicle routing problem: a survey. In: The Vehicle Routing Problem: Latest Advances and New Challenges, pp. 103–122. Springer (2008)
7. Zirour, M.: Vehicle routing problem: models and solutions. J. Qual. Meas. Anal. JQMA **4** (1), 205–218 (2008)
8. Çatay, B.: A new saving-based ant algorithm for the vehicle routing problem with simultaneous pickup and delivery. Expert Syst. Appl. **37**(10), 6809–6817 (2010)
9. Caceres-Cruz, J., Arias, P., Guimarans, D., Riera, D., Juan, A.A.: Rich vehicle routing problem: survey. ACM Comput. Surv. (CSUR) **47**(2), 32 (2015)
10. Toth, P., Vigo, D. (eds.).: Vehicle Routing: Problems, Methods, and Applications. Society for Industrial and Applied Mathematics (2014)
11. Daneshzand, F.: The vehicle-routing problem. Logist. Oper. Manag. **8**, 127–153 (2011)
12. Shah-Hosseini, H.: The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm. Int. J. Bio-Inspired Comput. **1**(1–2), 71–79 (2009)
13. Shah-Hosseini, H.: An approach to continuous optimization by the intelligent water drops algorithm. Proc. Soc. Behav. Sci. **32**, 224–229 (2012)
14. Hosseini, H. S.: Problem solving by intelligent water drops. In: 2007 IEEE Congress on Evolutionary Computation, CEC 2007, pp. 3226–3231. IEEE (2007)
15. Teymourian, E., Kayvanfar, V., Komaki, G.M., Zandieh, M.: Enhanced intelligent water drops and cuckoo search algorithms for solving the capacitated vehicle routing problem. Inf. Sci. **334**, 354–378 (2016)
16. Agarwal, K., Goyal, M., Srivastava, P.R.: Code coverage using intelligent water drop (IWD). Int. J. Bio-Inspired Comput. **4**(6), 392–402 (2012)
17. Wu, T.H., Low, C., Bai, J.W.: Heuristic solutions to multi-depot location-routing problems. Comput. Oper. Res. **29**(10), 1393–1415 (2002)
18. Cordeau, J.F., Gendreau, M., Laporte, G.: A tabu search heuristic for periodic and multi-depot vehicle routing problems. Networks **30**(2), 105–119 (1997)
19. Ho, W., Ho, G.T., Ji, P., Lau, H.C.: A hybrid genetic algorithm for the multi-depot vehicle routing problem. Eng. Appl. Artif. Intell. **21**(4), 548–557 (2008)
20. Clarke, G., Wright, J.W.: Scheduling of vehicles from a central depot to a number of delivery points. Oper. Res. **12**(4), 568–581 (1964)
21. Dowsland, K.A.: Simulated annealing, modern heuristic techniques for combinatorial problems. In: Reeves, C.R. (ed.) (1993)
22. Metroplis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1086–1092 (1953)

23. Johnson, D.S., Aragon, C.R., McGeoch, L.A., Schevon, C.: Optimization by simulated annealing: an experimental evaluation; part II, graph coloring and number partitioning. Oper. Res. **39**(3), 378–406 (1991)
24. Vincent, F.Y., Lin, S.W., Lee, W., Ting, C.J.: A simulated annealing heuristic for the capacitated location routing problem. Comput. Ind. Eng. **58**(2), 288–299 (2010)
25. Seshadri, A.: Traveling Salesman Problem (TSP) using simulated annealing. http://www.mathworks.com/matlabcentral/fileexchange/9612-traveling-salesman-problem-tsp-using-simulated-annealing. Accessed 31 Jan 2017

# Information Retrieval Based on the Extracted Social Network

Mahyuddin K.M. Nasution(✉), Rahmad Syah, and Maria Elfida

Information Technology Department, Fakultas Ilmu Komputer dan Teknologi Informasi (Fasilkom-TI) and Information System Centre,
Universitas Sumatera Utara, Medan 1500 USU, Medan, Sumatera Utara, Indonesia
mahyuddin@usu.ac.id

**Abstract.** It is possible that a technology affects other technologies. In this paper, we explored the possibility to reveal the performance of improved information retrieval through the extraction method of social network. Any extracted social network structurally is not a complete graph so it is possible to build the star social networks as the optimal form of graph, which guides to model the implication of information retrieval: the formulation of recall and precision, by using a sample, it show better performance on average over 90% and 58%, respectively.

**Keywords:** Information space · Webpages · Recall · Precision · Graph · Degree

## 1 Introduction

The social network as a resultant of extraction method from Web is a representation of relationship between web pages through any search engine [1]. The resultant based on occurrence and co-occurrence [2]. It depends not only on the logically relevance between the query and the webpage but also the similarity between the query and the information available [4], i.e. the answer to the required information as accurately as possible [3]. The latter case has become the concentration of information retrieval that is knowledge technology that focuses on the effectiveness and efficiency for retrieving information from information space like Web [5].

The method of social network extraction from the Web not only gives birth to social structure, but reveals the need for that information to be trusty [6,7]. Therefore, as a technology an extraction method of social networks must be equipped with evaluation tools [8]. In other words, when the method produces a resultant, another methods needs to be expressed from that result. This paper aims to reveal an IR model based on analysis of the extracted social network.

## 2 Problem Definition

Completeness of social networks is generally expressed in graph $G(V, E)$. The set of vertices $V = \{v_i | i = 1, \ldots, I\}$ as visual representations of a set of social

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_20

actors $A = \{a_i | i = 1, \dots, I\}$, $a_i$ are the names of possible social actors in/from information space $\Omega$, hit count $|\Omega_{a_i}|$ as the cardinality of $a_i$ [10]. The set of edges $E = \{e_j | j = 1, \dots, J\}$ as a visual representation of relation between two social actors $a_i, a_j \in A$ in the information space $\Omega$, hit counts $|\Omega_{a_i} \cap \Omega_{a_j}|$ as the cardinality of $a_i$ AND $a_j$ [11]. Hit count of two occurrences and co-occurrence computationally within similarity distance [12], e.g. using

$$sim_{our}(a_i, a_j) = \frac{2|a_i||a_j|}{|a_i|^2 + |a_j|^2} = \frac{2|\Omega_{a_i} \cap \Omega_{a_j}|}{|\Omega_{a_i}|^2 + |\Omega_{a_j}|^2} \tag{1}$$

is to give weight to each relationship. The weight not only determines the rank of any relation between social actors, but encourages the growth of social networks. Moreover, semantically the weights indicate the proximity of webpages that are not actually interlinked with each other [13]. Thus, social structures indirectly form another structures of the documents scattered within the information space, the different structures than links built into webpages or within the server in which the webpage resides.

In general, if a webpage is accessed then another webpage that is interlinked to it will indirectly impacted by the click event. Likewise, if one of social actor's name becomes the keyword of another social actor's name in the co-occurrence form, through the extracted social network, then the different webpages, although having one of the social actor names, will be clustered by search engine into appropriate cluster.

**Proposition 1.** *If a vertex is representation of a social actor in the extracted social network, then a cluster of webpages is representation of a social actor in Web.*

*Proof.* Based on Eq. (1) to get information about a social actor $a \in A$ from the Web is assigned a query $q$, and it generates a collection of webpages $\Omega_a = \{\omega_k | k = 1, \dots, K_a\}$ as a composition of information: a hit count $|\Omega_a|$ and list of snippets $L_{sa}$. Formally, $|\Omega_a| \leftarrow q = a$ or $L_{sa} \leftarrow q = a$. In other words, for all $v_i \in V$ we have

$$v_i = q(a_i) = \Omega_{a_i} = \{\omega_j | j = 1, \dots, m_i\} \tag{2}$$

where $\Omega_{a_i}$ is a cluster of webpages.

**Proposition 2.** *If an edge is representation of the relation between two social actors in the extracted social network, then a cluster of webpages is representation of the relation in Web.*

*Proof.* Equation (1) have shown that for getting information of the relation between two social actors $a_i a_j$ $\forall a_i, a_j \in A$ from the Web is assigned a query $q$, and it produces a collection of webpages $\Omega_{a_i a_j} = \{\omega_k | k = 1, \dots, K_a\}_i \cap \{\omega_k | k = 1, \dots, K_a\}_j$ as a composition of information: a hit count $|\Omega_{a_i} \cap \Omega_{a_j}|$ and list of snippets $L_{sa_i sa_j}$. In other words, for all $e_l \in E$ we have

$$e_l = q(a_i a_j) = \Omega_{a_i a_j} = \Omega_{a_i} \cap \Omega_{a_j} = \{\omega_k | k = 1, \dots, m_l\} \tag{3}$$

where $\Omega_{a_i a_j}$ is a cluster of webpages.

The cluster of webpages in either occurrence or co-occurrence consists of webpages in general conical to intersection between a cluster based on $a_i$ and a cluster based on $a_j$. Therefore, to obtain a reliable technology for extracting social network it is necessary to assess the performance of social network extraction methods. Assessment is done on the side of getting trusty information and getting side of technology associated with it, i.e. information retrieval. Assume $D_r$ is the set of documents relevant to the query and $D_c$ is an achieved documents, then the commonly used the assessment measures are recall $Rec$ and precision $Prec$ as follows [14].

1. To measure the permissibility (approximate ability) of reaching approach to all relevant documents based on query,

$$Rec = \frac{|D_r \cap D_c|}{|D_r|} \tag{4}$$

2. To measure the ability of the approach by achieving only documents relevant to the query and by rejecting the irrelevant documents,

$$Prec = \frac{|D_r \cap D_c|}{|D_c|} \tag{5}$$

**Theorem 1.** *If a social network is representation of social structure, then resultant of method for extracting social network from Web is representation of webpages structure.*

## 3 An Approach

If the social network extracted from the information source consists of $n$ social actors, then in general the social structure is formed can be expressed through the degree of social actors. In this case, the social network theoretically consist of two categories as defined below.

**Definition 1.** *A social network is completely shaped or abbreviated a* complete social network *consists $n$ vertices and $n(n-1)$ edges. Hereinafter it denoted as $SN(n, n(n-1))$.*

**Definition 2.** *A star-shaped social network or abbreviated a* star social network *consists $n$ vertices and $n-1$ edges. Hereinafter it denoted as $SN(n, n-1)$.*

The comparison between the complete social network and the star social network based on degree of all vertices is to reveal that the extracted social network will be between two categories, see Fig. 1. Thus, the predictably extracted social network may consist of parts, i.e. the subsets of social network are composed of social actors as the center and the social actors who become the leaf.

**Lemma 1.** *If social actor has the highest degree in social network, then the social actor become candidate of center in social network.*

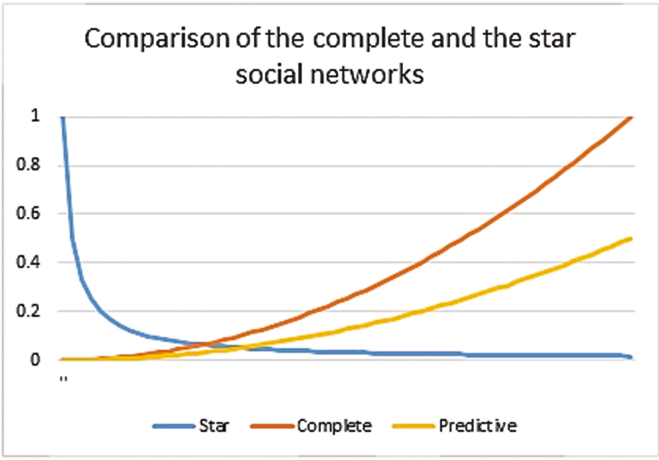


**Fig. 1.** Star and complete social network in comparison

*Proof.* Suppose there are $m$ social actors within social networks within each degree can be sorted as follow $d(v_1) > d(v_2) \geq d(v_3) \geq \cdots \geq d(v_m)$. It is clear that the social actor with the highest degree $d(v_1)$ is a candidate of center in the social network.

**Proposition 3.** *If there is more than one social actor is of the highest degree and likely to have multiple leaves, the social actors are the candidates of center in subs of social network.*

*Proof.* Based on Lemma 1, if edges between the highest degree vertices in social network are eliminated but each vertex still has leaves, it is clear that the social actors represented by the vertices become the candidates of center in sub of social network.

In graph theory, a tree is the optimal-shaped of graph, and the star is one of tree forms.

**Lemma 2.** *If there are vertices with the highest degree in social network, then the sub of social network with social actors as the center are the optimal form of social networks.*

*Proof.* Based on Proposition 3, the sub of social networks have a center of candidate that makes them be a sub social network independently. Suppose that on each sub social network there is $m_i$ vertices, by eliminating the edges that does not cover the center and leaf vertices, or eliminating the leaf-linking edges, it produces $m_i - 1$ edges within the sub social network. A sub social network is with a center has a degree is $m_i - 1$ while other vertices are of degree 1. This social network is denoted by $SN_s(m_i, m_i - 1)$ be sub of star-shaped social network or the star social networks.

**Proposition 4.** *If there are the star social networks in a social network, then the collection of information spaces about social actors is conjoined into the social actors' information space as the center of star social networks.*

*Proof.* In formally, $SN_s(m_i, m_i - 1) \subset SN(n, n(n-1))$ are star social networks, and $m_i \leq n$, $i = 1, \ldots, I$. Based on Lemma 2 and Proposition 2, for each star social networks there are $m_i - 1$ $(\Omega_a \cap \Omega_{a_{j-1}})$, where $\Omega_a \cap \Omega_{a_{j-1}} = \{\omega_k | k = 1, \ldots, m_l\}_{i-1}$. In other words, for vertex as center of the star social network, we have $m_i - 1$ edges or number of $|\Omega_a \cap \Omega_{a_{j-1}}|$ is $m_i - 1$, and

$$\sum_{j=2}^{m_i-1} |\Omega_a \cap \Omega_{a_{j-1}}| = \left| \bigcup_{j=2}^{m_i-1} \Omega_a \cap \Omega_{a_{j-1}} \right| \approx \left| \bigcup_{j=2}^{m_i-1} \Omega_a \right| = |\Omega_a| \qquad (6)$$

or

$$\left| \bigcup_{j=1}^{m_i-1} \Omega_a \cap \Omega_{a_{j-1}} \right| \approx |\Omega_a| \qquad (7)$$

So based on Eq. 1, whereby Proposition 1, 2 and 4 take the appropriate role to structure the webpages, then Theorem 1 is proved.

**Table 1.** Statistic of dataset $D_r$

| Personal name | Position | Number of documents |
|---|---|---|
| Abdul Razak Hamdan | Professor (A1) | 103 |
| Abdullah Mohd Zin | Professor (A2) | 105 |
| Shahrul Azman Mohd Noah | Professor (A3) | 160 |
| Tengku Mohd Tengku Sembok | Professor (A4) | 210 |
| Md Jan Nordin | Professor (A5) | 70 |

## 4 Information Retrieval

As the application of Theorem 1, we modelled the structure of webpages for information retrieval in social network extraction perspective based on Eqs. 4 and 5. The documents set $D_r$ is modeled as a collection of all the collected webpages based on the academic actor (professor), in which each document has the unique identity that is URL address. From a collection of academic actors, there are the star social networks in which each professor as a center and another academic actors (who is not a professor) as a leaf, and each query is built from a pair of such social actors. The documents $D_c$ was obtained based on the query. For evaluation of the approaches, in this case we use the sample based on the test concept [15,16]. This sample is used as a representation of the population to prove some of the above theories [17–19]. However, further testing is required

**Table 2.** Comparison between disambiguation method and based on social network for recall and precision

| Personal name | Disambiguation | | Star social network | |
|---|---|---|---|---|
| (Professor) | Recall | Precision | Recall | Precision |
| A1 | 47/103 (45.63%) | 47/154(30.52%) | 99/103 (96.12%) | 99/194(51.03%) |
| A2 | 49/105 (46.67%) | 49/158(31.01%) | 101/105 (96.19%) | 101/188(33.72%) |
| A3 | 72/160 (45.00%) | 72/152(47.37%) | 155/160 (96.88%) | 155/222(69.82%) |
| A4 | 92/210 (43.81%) | 92/148(62.16%) | 200/210 (95.24%) | 200/258(77.52%) |
| A5 | 32/70 (45.71%) | 32/154(20.78%) | 68/70 (97.14%) | 68/164(41.46%) |

by involving the large data including using the available dataset [20]. Therefore, we have gathered and labeled a dataset of 648 webpages like Table 1. By using concept the name disambiguation, Eqs. 4 and 5 produce Table 2.

In general, there is an increase in the performance of the access process by means of involving social networks whereby the webpages of influential social actors will be dug up by other social actors as keywords rather than keywords that are not the names of social actors. Thus, the formulation of recall and precision based on Proposition 4 are as follows:

$$Rec_a = \frac{|D_r \cap \bigcup_{i=1}^{m} D_c|}{|D_r|} \tag{8}$$

and

$$Prec_a = \frac{|D_r \cap \bigcup_{i=1}^{m} D_c|}{|\bigcup_{i=1}^{m} D_c|} \tag{9}$$

where $m$ is degree of center in star social network.

## 5 Conclusion and Future Work

By studying the principle of social network extraction, from the extraction method, there is a change in formulation on the implication of information retrieval that is recall and precision, which in this sample it has better performance based on computation. Furthermore, to test this formula will be built larger datasets.

## References

1. Nasution, M.K.M., Noah, S.A.M.: Superficial method for extracting social network for academic using web snippets. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNAI, vol. 6401, pp. 483–490 (2010). doi:10.1007/978-3-642-16248-0_68
2. Nasution, M.K.M.: Social network mining (SNM): a definition of relation between the resources and SNA. Int. J. Adv. Sci. Eng. Inf. Technol. **6**(6), 975–981 (2016)

3. Nasution, M.K.M., Noah, S.A.: Information retrieval model: a social network extraction perspective. In: Proceedings of 2012 International Conference on Information Retrieval and Knowledge Management, CAMP 2012, pp. 322–326 (2012). doi:10.1109/InfRKM.2012.6204999
4. Nasution, M.K.M., Sitompul, O.S.: Enhancing extraction method for aggregating strength relation between social actors. Adv. Intell. Syst. Comput. **573**, 312–321 (2017). doi:10.1007/978-3-319-57261-1_31
5. Nasution, M.K.M., Noah, S.A.M., Saad, S.: Social network extraction: superficial method and information retrieval. In: Proceeding of International Conference on Informatics for Development (ICID 2011), pp. c2-110–c2-115 (2011)
6. Nasution, M.K.M., Sitompul, O.S., Sinulingga, E.P., Noah, S.A.: An extracted social network mining. In: Proceedings of 2016 SAI Computing Conference, SAI 2016, pp. 1168–1172 (2016). doi:10.1109/SAI.2016.7556125
7. Žižka, J., Dařena, F.: The comparison of effects of relevant-feature selection algorithms on certain social-network text-mining viewpoints. Adv. Intell. Syst. Comput. **573**, 354–363 (2017). doi:10.1007/978-3-319-57261-1_35
8. Nasution, M.K.M.: Modelling and simulation of search engine. J. Phys. Conf. Ser. **801**(1), 012078 (2016). doi:10.1088/1742-6596/801/1/012078
9. Nasution, M.K.M.: New method for extracting keyword for the social actor. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNAI, vol. 8397 (Part 1), pp. 83–92 (2014). doi:10.1007/978-3-319-05476-6_9
10. Nasution, M.K.M.: Simple search engine model: adaptive properties. Cornell University Library arXiv:1212.3906v1 (2012)
11. Nasution, M.K.M.: Simple search engine model: adaptive properties for doubleton. Cornell University Library arXiv:1212.4702v1 (2012)
12. Mahyuddin, K.M.N., Sitompul, O.S., Nasution, S., Ambarita, H.: New similarity. IOP Conf. Ser. Mater. Sci. Eng. **180**(1), 012297 (2016). doi:10.1088/1757-899X/180/1/012297
13. Nasution, M.K.M., Noah, S.A.: Extraction of academic social network from online database. In: 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011, pp. 64–69 (2011). doi:10.1109/STAIR.2011.5995766
14. Croft, W.B., Metzler, D., Strohman, T.: Search Engines Information Retrieval in Practice. Addison Wesley, New York (2010)
15. Bekkerman, R., McCallum, A.: Disambiguating web appearances of people in a social network. In: WWW 2005, Chiba, Japan, 10–14 May 2005
16. Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., Fleck, M.: Using social network analysis to enhance information retrieval systems. In: Social Networks Applications Conference (2008)
17. Nasution, M.K.M., Elveny, M., Syah, R., Noah, S.A.: Behavior of the resources in the growth of social network. In: Proceedings of 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015, pp. 496–499 (2015). doi:10.1109/ICEEI.2015.7352551
18. Nasution, M.K.M., Hardi, M., Syah, R.: Mining of the social network extraction. J. Phys. Conf. Ser. **801**(1), 012020 (2017). doi:10.1088/1742-6596/801/1/012020
19. Nasution, M.K.M., Syah, R., Elveny, M.: Studies on behaviour of information to extract the meaning behind the behaviour. J. Phys. Conf. Ser. **801**(1), 012022 (2017). doi:10.1077/1742-6596/801/1/012022
20. Bollegala, D., Noman, N., Iba, H.: RankDE: learning a ranking function for Information Retrieval using differential evolution. In: GEOCCO 2011, 12–16 July 2011, Dublin, Ireland (2011)

# Strategic Decision Method Structured in SWOT Analysis and Postures Based in the MAGIQ Multicriteria Analysis

Sergio Alexandre Barreira Forte(✉),
Sergio Henrique Arruda Cavalcante Forte,
and Placido Rogério Pinheiro

Universidade de Fortaleza (UNIFOR), Fortaleza, Ceará 60811-905, Brazil
sergio.alexandre92@gmail.com, sergioforte@unifor.br,
placidrp@gmail.com

**Abstract.** Strategic decisions are those that have a far reaching effect on the environment and on the organization itself, however, they need a strategic diagnosis. Decision makers, for lack of knowledge of the literature and without a more efficient methodology, sometimes delay and/or make decisions that might not have been the best. This work aims to mitigate this problem, proposing a method that can assist the decision maker. The solution is based on two tools in the field of business strategy: the SWOT matrix (strengths and weaknesses in internal analysis and opportunities and threats in external analysis) and the map of strategic postures (survival, maintenance, growth and development), with support of the Multi-Attribute Global Inference of Quality (MAGIQ) multicriteria analysis tool [1]. Thus, depending on the situation of the strategic diagnosis in which the organization is located, the solution would indicate what possible strategic decisions the organization should adopt. A methodology for feeding, filtering, calculating, positioning and selecting strategies, including inbound and outbound reports, is proposed.

**Keywords:** SWOT Matrix · Strategic decision · MAGIQ

## 1 Introduction

The strategic process is based on four dimensions: diagnosis, decision, implementation and monitoring. In the diagnosis, the internal analysis and external analysis of the organization is carried out. Then, for the decision-making phase comes the positioning and the strategic decision [2].

Among the known strategic diagnosis tools, the SWOT matrix (Strength, Weakness, Opportunity and Threat) is the most widely used [3–5] and aims to integrate and manage information in relation to four categories: strengths and weaknesses on the internal side of the organization and threats and opportunities on the external side of the organization [6], aiming for a strategic positioning of the organization [7, 8].

On the other hand, the literature on business strategy is abundant in strategic typologies, one of the most detailed being the strategic posture matrix, with a link

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_21

between the SWOT matrix [7] and the strategic posture maps, indicating four possible situations in which an organization can be (Survival Strategic Posture = Weaknesses and Threats; Maintenance Strategic Posture = Strengths and Threats; Growth Strategic Posture = Weaknesses and Opportunities; and Development Strategic Posture = Strengths and Opportunities). Each strategic posture is composed of a set of business strategies [9].

There are some methodologies for linking the SWOT matrix with business strategy [3, 10–13] but a methodology was not found to select the strategies according to the SWOT analysis and the strategic postures using the Multi-Attribute Global Inference of Quality (MAGIQ) [1].

On the other hand, the links between the SWOT matrix and the map of strategic posture have not been properly studied and therefore this research aims to contribute to this theoretical and methodological gap.

Inserting the expression SWOT Matrix on Google Scholar in the title, there are 107 papers, in the position of August 2016. However, in studies dealing with the links between SWOT matrix and business strategy, by any method, the studies do not adopt theoretical reference for the establishment of pre-defined strategies [14].

Inserting the expression “Multi-Attribute Global Inference of Quality” in “any part of the article” in Google Scholar’s “advanced search” link [14] in the September 2016 position, 47 papers were shown, and from 2000 to 2009 there were 10 papers published and from 2010 to 2016, 37 papers were published, demonstrating that this technique has already been used in the international scope and in the last five years the number of papers practically quadrupled compared to the previous decade.

So, the research problem is how to make a connection between strategic positions and business strategies with the SWOT matrix through the MAGIQ methodology?

The main objective of the research will be to build a method structured in multi-criteria analysis that can support the strategic decision makers of organizations through the SWOT matrix and strategic maps through the MAGIQ methodology.

The relevance of the work is methodological since it contributes to fill the gaps in the field of the interconnection between diagnosis and strategic decision, and empirical or practical, because it aims to support strategic decision makers in organizations.

The proposed method does not aim to operationalize how strategic decisions should be implemented, but to indicate what possible paths the organization could adopt.

## 2 Process and Strategic Diagnosis

### 2.1 Strategic Diagnosis Process and Techniques

Strategic management is defined as the art and science of formulating, implementing and evaluating some decisions to achieve the organization goals [15].

Many flows and frameworks on the strategic process are presented, but they basically fit in cartesian decisions since the diagnosis phase until the prescription or decisive, evaluation, monitoring, and review phases [2, 16].

As techniques of external analysis of the general environment there is the [16] model of six dimensions (Legal and Political Conditions, Economic Climate, Demographic

Trends, Cultural Changes, Technological Changes and Specific International Events), and another model also diffused, The PESTEL model, which is divided into Political, Economic, Social/Cultural/Demographic, Technological, Ecological (Environment) and Legal variables [17].

The most used techniques of external analysis for industry analysis, i.e. the analysis of the sector in which the company is inserted, also called specific environment, industrial operational or simply task, are: the technique of analysis of the industry structure (Five strengths of Porter) [16, 18, 19], the technique of Strategic Groups (distinct strategic groups within the same industry) [19] and the Boston Consulting Group (BCG) and General Electric/McKinsey matrices [20], used to evaluate the portfolios of units and business areas of companies of the same industry.

In the internal analysis, the VRIO (Value, Rarity, Impersonability and Organization) model of VBR (Resource Based View) by [16] and the Balanced Scorecard [21] are used as techniques.

However, there is a technique that acts both in the external evaluation and in the external evaluation, called SWOT Analysis, discussed below.

## 2.2 Multicriteria Analysis Between SWOT Matrix and Strategy

In a strategic planning process, the SWOT Matrix is one of the most widespread tools used as a method of evaluating internal and external analysis of an organization [3, 4, 6, 8].

[22] recommends listing three to seven variables for each of the four factors or dimensions: strengths, weaknesses, opportunities, and threats. [14] suggest from 10 to 20 factors or variables by internal dimensions (strengths and weaknesses) and external dimensions (opportunities and threats). Therefore, there is no sizing or consensus.

The variables for analysis can be divided basically into three groups: (1) general environment; (2) task environment; and (3) internal environment [19].

Based on these issues, a SWOT matrix can be constructed [7]. The strategic posture of Survival is attributed when there is Weakness and Threat (Mini-Mini); The strategic posture of Maintenance is attributed when there is Strength and Threat (Maxi-Mini); The strategic posture of Growth is when there is Weakness and Opportunity (Mini-Maxi), and finally the strategic posture of Development occurs when there is Strength and Opportunity (Maxi-Maxi).

The junction of the AHP method with the SWOT matrix is also called A’WOT [8]. The prioritization process was developed based on the [23] scale, which varies from 1 to 9 according to the relative contribution of one event/variable over the other in relation to the problem studied [23]. According to the amplitude of the criteria, the comparisons generate a matrix A (n × n) where the matrices are normalized and the relative weights are found. In order to find the importance of all the factors of the SWOT matrix, the importance degrees of the second level are multiplied with those of the third level [23].

# 3 Research Methodology

As for epistemology, this research is classified as positivist [24] and the descriptive type and by means of secondary data [25], using applicable approaches [26].

## 3.1 Research Process

The research was carried out from June 2016 to March 2017. For the assembly of the SWOT matrix, it was based on the general, industry and enterprise environments as a set of features. For the list of strategies, the Strategic Postures map was used as reference.

As a qualitative analysis technique, the content analysis [27] was used to structure SWOT variables and Strategic Postures. As a support technique for the proposed method, the [28] participant rule, Quartis [29], the Delphi technique [30] and the MAGIQ methodology were used as well as the use of the Likert scale.

## 3.2 Multi-Attribute Global Inference of Quality (MAGIQ) Method

The MAGIQ is a multicriteria decision analysis method that supports the determination, by means of comparable weights, of the relative importance of factors and subfactors related to a basic question or decision [1]. This method has similarities to the Analytical Hierarchy Process - AHP method [31, 32].

Because it does not use a paired analysis methodology of each variable for another, which demands a great effort on the part of the respondents and a probability of inconsistencies, the MAGIQ was among those cited by [33], the most feasible for SWOT analysis and for the Strategies in this research.

For [34], MAGIQ is essentially a variation of the AHP technique, which explains a large number of AHP-focused research. However, the AHP method has been the subject of much criticism, including the use of an arbitrary scale inducing a non-existent order [35] and changing rankings due to the addition of different criteria [34]. Finally, according to [34], the AHP technique is highly correlated to MAGIQ ($R^2 > 0{,}9$), which further assures its use.

The first step of MAGIQ analysis is to determine the relative weights of factors at the highest hierarchy levels for each “j” respondent using the “Rank Order Centroids” (ROC) concept, which allows the conversion of sort orders (e.g., 1st, 2nd, 3rd etc.) into numeric values (Formula 1):

$$\left[\sum_{i=k}^{N}(1/i)\right]\Big/N \quad \text{for } i = 1, 2, 3, \ldots N \tag{1}$$

After sorting and assigning the weights relative to each of the levels of the hierarchy, the next step in the MAGIQ analysis is to calculate the quality global goal value. The global value is given by the simple weighted sum of all weights of the comparison attributes, the final sum of all vectors being equal to 1.0.

# 4 Proposition of an Integration Method Between SWOT Matrix and the Strategic Posture Map

## 4.1 Method Process

**Phase (1) Development of the SWOT Matrix.** Strategic external and internal decision makers are invited to take part on the Strategic Planning. It is suggested from 15 to 30 participants, according to Godet's methodology for the construction of strategic prospective scenarios [28].

Decision makers select a set of external analysis and internal analysis variables. The external analysis variables are divided into the following categories: general environment and sector environment.

The External Variables - General Environment, incorporate the PESTEL model [8, 17] and the General Environment model [16]. The external variables of the Sector according to [19] are formed by the industry; Customer-Market; Suppliers and Competition (Rivalry, new entrants, substitute products).

According to [16, 21], the company variables are Financial, Strategy/Management, Marketing/Commercialization, Process/Production/Engineering and Learning/Skills.

The tool presents a main table with variables to choose from, a database with several other variables and an option to append a variable not available in the main list and database.

This phase is divided into two rounds. In the first round the variables for judgment will be chosen. Each decision maker chooses whether he/she prefers it to be an Opportunity or Threat for external analysis or Strength or Weakness for internal analysis. If a variable is not chosen, it will not be part of the SWOT matrix.

In case there is a divergence among the amount of opportunity or threat, or of strength or weakness, the majority wins. Thus, the tool would only compute the winning variables in frequency. If there is a tie, the process manager decides, or a Committee decides, or a new voting round after a coordinated discussion (Delphi round).

If the study is carried out for a specific sector, the external variables to be considered will be those referring to the General Environment and the internal variables to the Environment of the sector chosen for the study.

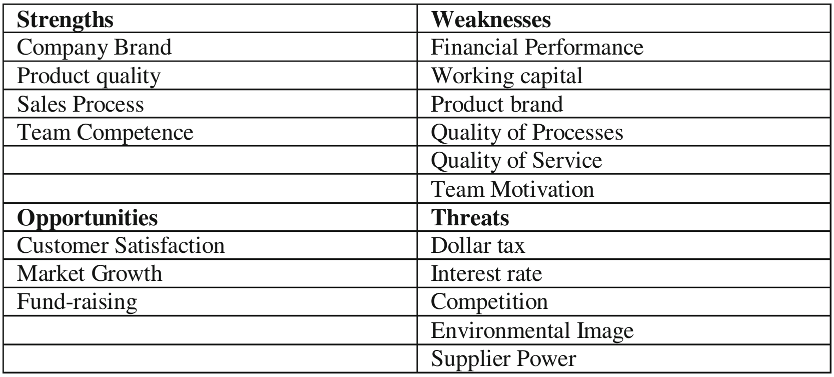
In the second round, the quartis technique [29] is used to reduce the variables for later use of the MAGIQ technique. A Likert scale is applied from 0 to 4. The variables that would be from the 4th quartile will be chosen, that is, where the average would be in the closed range of 3 to 4.

For strengths the scale will be: 0 (negligible); 1 (slightly strong); 2 (moderately strong); 3 (strong); 4 (very strong). For weaknesses it will be: 0 (negligible); 1 (slightly weak); 2 (moderately weak); 3 (weak); 4 (very weak). For Opportunities the scale will be: 0 (negligible); 1 (slightly opportune); 2 (moderately opportune); 3 (opportunity); 4 (great opportunity). For the Threats the scale will be: 0 (negligible); 1 (slightly threatening); 2 (moderately threatening); 3 (Threat); 4 (great Threat).

If any variable gets an average score below 3, but one participant wants to defend it to include it in the SWOT matrix, it is suggested a group discussion, or a new round

(Delphi round), or the higher hierarchy manager decides whether it should be included or not.

The Delphi technique (up to two new rounds) would serve to improve the consensus of the score judgement, presenting to each decision-maker the average of the group and the note that he applied. Then a report is issued showing only the variables selected for the MAGIQ round (see Fig. 1 below).



| **Strengths** | **Weaknesses** |
|---|---|
| Company Brand | Financial Performance |
| Product quality | Working capital |
| Sales Process | Product brand |
| Team Competence | Quality of Processes |
|  | Quality of Service |
|  | Team Motivation |
| **Opportunities** | **Threats** |
| Customer Satisfaction | Dollar tax |
| Market Growth | Interest rate |
| Fund-raising | Competition |
|  | Environmental Image |
|  | Supplier Power |

**Fig. 1.** SWOT matrix - example. Source: Made by the authors (2017).

Figure 1 already suggests that there are more weaknesses than strengths and more threats than opportunities.

With the selected variables, a new round will be applied to the participants of the strategic decision, this time using the Multi-Attibute Global Inference of Quality (MAGIQ) technique.

To obtain the factor indices (S, W, O, T), the tool asks: Which factor represents the company's situation? Each participant chooses one of the four. The exercise repeats itself now with three factors up to n − 1, that is, up to the third factor. Of course the fourth factor would not be chosen. The tool establishes the relative weight of each factor.

Then, the same procedure is started for each sub-factor (variable) of each factor (SWOT dimension). In the case of strong points, it is asked (which one is stronger?). In the case of weaknesses (which is the weakest?). In the case of opportunities, it is asked (which one favors our company/business?). In the case of threats, it is asked (which is the one that most disadvantages our company/business)? Such an approach would greatly facilitate the choice of priorities, rather than the respondent having to order the priorities at once as the original MAGIQ methodology suggests. So, there would only be one standard question (which one?).

As the exercise assumes several judges, the weight of each factor is the average MAGIQ weights of each variable, that is, we add the MAGIQ weights of each judge and then divide it by the amount of judges, a technique based on [36].

For all variables selected within the four dimensions of the Matrix, relative weights are assigned so that the total is 1. The weights are calculated by the formula presented in Sect. 3.2.

The global (final) weights of each variable are calculated within each SWOT factor by multiplying the weights of the variables and by the weight of each SWOT factor to which it belongs. The Table 1 below shows the distribution of the ordinations (example of 4 participants):

**Table 1.** *Ranking* and final weight of SWOT factors and variables (example)

| Factors/Variables | *Ranking* by participants | | | | Average | Final weight | # |
|---|---|---|---|---|---|---|---|
| | P1 | P2 | P3 | P4 | | | |
| *Strengths* | 0,156225 | | | | | | |
| S1. Company brand | 0,5208 | 0,1458 | 0,1458 | 0,5208 | 0,3333 | 0,0520 | 2 |
| S2. Product quality | 0,1458 | 0,0625 | 0,2708 | 0,2708 | 0,1875 | 0,0293 | 3 |
| S3. Sales process | 0,0625 | 0,2708 | 0,0625 | 0,1458 | 0,1354 | 0,0212 | 4 |
| S4. Team competence | 0,2708 | 0,5208 | 0,5208 | 0,0625 | 0,3437 | 0,0536 | 1 |
| *Sum final weights* | | | | | | **0,1562** | |
| *Weaknesses* | 0,42705 | | | | | | |
| W1. Financial performance | 0,2566 | 0,0400 | 0,4566 | 0,0900 | 0,2108 | 0,0900 | 2 |
| W2. Working capital | 0,1566 | 0,0900 | 0,1566 | 0,1566 | 0,1399 | 0,0598 | 4 |
| W3. Quality of processes | 0,0400 | 0,2566 | 0,0400 | 0,0400 | 0,0941 | 0,0402 | 5 |
| W4. Quality of service | 0,4566 | 0,4566 | 0,2566 | 0,4566 | 0,406 | 0,1736 | 1 |
| W5. Team motivation | 0,0900 | 0,1566 | 0,0900 | 0,2566 | 0,1483 | 0,0633 | 3 |
| *Sum final weights* | | | | | | **0,4269** | |
| *Opportunities* | 0,187475 | | | | | | |
| O1. Customer satisfaction | 0,2777 | 0,2777 | 0,2777 | 0,1111 | 0,2360 | 0,0442 | 2 |
| O2. Market growth | 0,6111 | 0,6111 | 0,6111 | 0,6111 | 0,6111 | 0,1146 | 1 |
| O3. Fund-raising | 0,1111 | 0,1111 | 0,1111 | 0,2777 | 0,1527 | 0,0286 | 3 |
| *Sum final weights* | | | | | | **0,1874** | |
| *Threats* | 0,2292 | | | | | | |
| T1. Dollar tax | 0,2566 | 0,4566 | 0,0900 | 0,2566 | 0,2649 | 0,0607 | 2 |
| T2. Interest rate | 0,4566 | 0,2566 | 0,0400 | 0,4566 | 0,3024 | 0,0693 | 1 |
| T3. Competition | 0,1566 | 0,1566 | 0,4566 | 0,1566 | 0,2316 | 0,05307 | 3 |
| T4. Environmental image | 0,0900 | 0,0900 | 0,2566 | 0,0900 | 0,1316 | 0,0302 | 4 |
| T5. Supplier power | 0,0400 | 0,0400 | 0,1566 | 0,0400 | 0,0692 | 0,0158 | 5 |
| *Sum final weights* | | | | | | **0,2292** | |

Source: Made by the authors (2017).

**Phase (2) Positioning in Strategic Posture.** There are four Cartesian quadrants. The right side of the X axis are the markings for the Opportunities weights and the left side for the Threats. Each side ranges from 0 to 1 (right) and 0 to −1 (left). The Y axis ranges from 0 to 1 (up) and 0 to −1 (down), with the top part being allocated to the markings of the Strengths weights and bottom part to the Weaknesses weights.

The MAXI MAXI quadrant corresponds to the Development posture (opportunities and strengths). The MAXI MINI quadrant corresponds to the Growth posture (weaknesses and opportunities). The MINI MAXI quadrant corresponds to the Maintenance posture (threats and strengths). The MINI MINI quadrant corresponds to the Survival posture (weaknesses and threats).

According to the [3] technique, the positioning is calculated by the difference between the offensive capacity (Forces minus Weaknesses) and defensive capacity (Opportunities minus Threats), and then the posture in which the company is located is found (see Fig. 1).

In this methodology, an alternative path of Ingaldi's methodology [3] is chosen, since we also want to find the second, third and fourth postures to go through. The ordering of the other positions will be given according to the diagonal result of the other quadrants, when plotting the factor values (final weights) of the SW (Y axis) and OT (X axis) axes.

Thus, there would be a link between strategy and the diagnosis variables, through positioning (posture) and combinations of the variables SO, WO, TS and TW.

Thus, there is a sequence of postures from the most valued to the least valued.

In the example, it is: Offensive Capacity = Strengths - Weaknesses = 0.156225 - .42705 = −0.2708. Defensive Capacity - Opportunities - Threats = 0,187475 - 0,22915 = −0,0417.

Figure 2 shows the positioning graph.

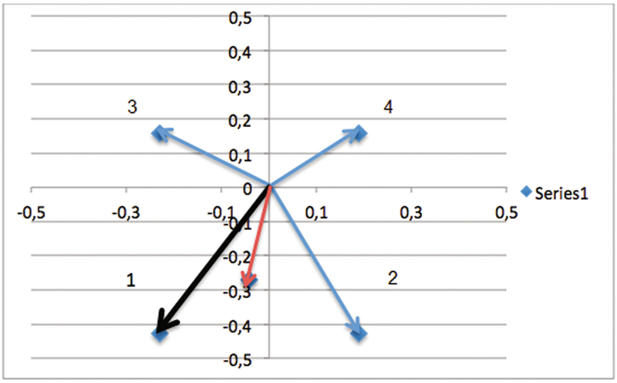


**Fig. 2.** Strategic positioning chart. Source: Made by the authors (2017).

It is verified that the Offensive Capacity was positioned in the Mini-Mini quadrant, that is, in the Survival Posture. There was also an alignment of the position with the largest vector of the graph in the same Strategic Posture.

The second posture was the Mini-Maxi, or the Growth Posture. The third was the Maxi-Mini, that is, the Maintenance or Competitive Posture. The fourth and last was the Maxi-Maxi, that is, the Development Posture.

**Phase (3) Strategies Selection.** Once each posture is selected in an order sequence, the literature strategies for each posture will be chosen using the same initial selection method used with the SWOT variables.

In the first place, strategies will be chosen by posture using the frequency technique. In this way, the strategies of higher frequency of the decision group will enter.

In the same way as in the previous phase, a Likert scale with the following configuration is applied to each variable (strategy) of each posture: 0 (will not be applied); 1 (slightly applicable); 2 (moderately applicable); 3 (applicable); 4 (strongly applicable).

In the same way as the SWOT phase, the strategies which averages are below 3 are excluded, and only the ones of the fourth quartile are included.

Also like in the SWOT phase, if the participants have discordance about the excluded variables, there may be a discussion and once approved, a new round, using the Delphi technique [30], should be applied, where each participant can adjust their score in relation to the average of the group or consensus; or a larger decision.

Finally, the tool presents the framework for the chosen strategies, according to the multicriteria analysis judgement of the MAGIQ tool (see Table 2).

**Table 2.** Ranking e final weight of strategies (example).

| Postures/Strategies | *Ranking* by participants | | | | Average | Final weight | # | TOWS matrix |
|---|---|---|---|---|---|---|---|---|
| | GG | GC | GP | GA | | | | |
| *Survival* | 0,48455 | | | | | | | |
| Cost reduction | 0,25 | 0,25 | 0,75 | 0,75 | 0,5 | 0,24227 | 1 | W1; W2 |
| Debts renegotiation | 0,75 | 0,75 | 0,25 | 0,25 | 0,5 | 0,24227 | 2 | W1, T1, T2. O3 |
| *Growth* | 0,466304 | | | | | | | |
| Market share increase | 0,25 | 0,25 | 0,75 | 0,75 | 0,5 | 0,23315 | 2 | O1, 02 |
| Quality certification program | 0,75 | 0,75 | 0,25 | 0,25 | 0,75 | 0,34972 | 1 | W3, W4 |
| *Maintenance* | 0,27729 | | | | | | | |
| Product specialization | 0,25 | 0,75 | 0,75 | 0,75 | 0,625 | 0,17330 | 1 | O2, S2, S4 |
| Outsourcing | 0,75 | 0,25 | 0,25 | 0,25 | 0,375 | 0,10398 | 2 | S4, W3 |
| *Development* | 0,24401 | | | | | | | |
| Develop. of new markets | 0,25 | 0,25 | 0,75 | 0,75 | 0,5 | 0,12200 | 2 | S1, O2 |
| Vertical backward integration | 0,75 | 0,75 | 0,25 | 0,25 | 0,5 | 0,12200 | 1 | S1, S2 |

Source: Made by the authors (2017).

As can be seen in Table 2, according to the final weights of the strategies, there is a sequence of strategic decisions to be adopted in descending weight order.

Finally, participants will fill in a column for each taken strategy, indicating the SWOT variables that will support those strategies, i.e., the TOWS matrix [13].

## 5 Conclusion

The literature presents methodologies for interconnecting SWOT variables with strategy through the MAGIQ method, which is easier to operationalize, implying in error reduction, besides being strongly correlated with the AHP.

It is proposed that, based on the method presented in this paper, called *Strong Decisions*, a tool is made based on a web environment that would have the following modules: (1) registration of SWOT variables; company; decision makers. (2) selection, classification and ranking of variables and strategies. (3) Strategic posture chart (positioning); (4) Array of selected strategies and postures; (5) Queries and position and evolution Reports.

It is suggested to operationalize the web environment tool in the ASP.NET MVC platform using the C# programming language and also to expand the tool including a study of the selected strategies, integrating it with the scorecard indicators methodology (monitoring by indicators).

## References

1. Mccaffrey, D.J., Koski, N.: Competitive analysis using MAGIQ. MSDN Mag. **21**(11), 35–39 (2006). http://msdn2.microsoft.com/enus/magazine/cc300812.aspx
2. Paludo, A.V., Procopiuck, M.: Planejamento governamental: Referencial teórico. Conceitual e Prático. Atlas, São Paulo (2011)
3. Ingaldi, M.: Use of the SWOT analysis the 3 × 3 matrix to determine the technological position of the chosen metal company. In: Acta Metallurgica Slovaca – Conference, Czestoshowa, vol. 4, pp. 207–21 (2014)
4. Motefaker, H., et al.: Formulating gas company strategy of Lorestanbasek on SWOT analysis and ANP process. In: International Conference of Management, Innovation and National Production. Shahivar, Qom, Iran (2013)
5. Öztürk, S., Tönük, G.: Stakeholder participation as a means for river basin management plan. J. Environ. Prot. Ecol. **14**(3), 1097–1106 (2013). ISSN: 1311-5065
6. Akbulak, C., Cengiz, T.: Determining ecotourism strategies using A'WOT hybrid method: case study of Troia Historical National Park, Çanakkale, Turkey. Int. J. Sustain. Dev. World Ecol. **21**(4), 380–388 (2014)
7. Andrews, K.R.: The Concept of Corporate Strategy. Richard D. Irwin, New York (1980)
8. Stankovic, J., Stankovic, J., Jankovic-Milic, V.: Integrated approach of SWOT and multi-criteria analysis in strategic decision making. In: SYMORG 2012. Innovative Management & Business Performance, pp. 1224–1232 (2012)
9. Forte, S.H.A.C.: Uma Contribuição para a Tipologia no Campo da Estratégia Empresarial. In: Costa, B.K., Almeida, M.R. (Org.). Modelos e Inovações em Estratégia. Universidade Metodista de São Paulo, São Paulo, pp. 101–122 (2007)
10. Al Khassabi, M.H.: A suggested framework for evaluating the status of design by using the concepts of (Prioritization Matrix) and (SWOT). Int. Des. J. **5**(1), 99–111 (2015). Egypt
11. Harisudin, M., Setyowati, N., Utami, B.: Formulating and choosing strategy of processed catfish product development using the SWOT matrix and QSPM: a case study in Bouolali Regency. World Appl. J. **30**, 56–61 (2014). (Innovation Challenges in Multidiciplinary Research & Practice)
12. Salehi, M., Askari, J., Behrouzi, S.: Strategy formulation by SWOT and QSPM Matrix (Case study: Sanitary Ware Company of Golsar Fars). Int. SAMANM J. Mark. Manag. **2**(2) (2014). Fars Iran

13. Gupta, M., Shri, C., Agrawal, A.: Strategy formulation for performance improvement of Indian Corrugated Industry: An application of SWOT matrix and QSPM matrix. J. Appl. Packag. Res. Indian **7**(3), 3 (2015)
14. GoogleScholar (2016). https://schollar.google.com.br
15. David, F.R.: Strategic Management: A Competitive Advantage Approach, Concepts and Cases. Prentice Hall, Upper Saddle River (2014)
16. Barney, J.G., Hesterly, W.S.: Administração estratégica e vantagem competitiva: conceitos e casos. Pearson, São Paulo (2011)
17. Johnson, G., Whittington, R., Scholes, K.: Exploring Strategy. Pearson Education Limited, Essex (2011)
18. Mintzberg, H., et al.: O Processo da Estratégia: conceitos, contextos e casos selecionados, 4th edn. Bookman, Porto Alegre (2011)
19. Porter, M.E.: Estratégia competitiva: Técnicas de Análise de Indústrias e de Concorrência: Campus, Rio de Janeiro (2005)
20. Kolbina, O.: SWOT analysis as a strategic planning tool for companies in the food industry. Probl. Econ. Transit. **57**(9), 74–83 (2015)
21. Kaplan, R.S., Norton, D.: Mapas Estratégicos. Campus, Rio de Janeiro (2004)
22. Quevedo, M.C.: AHP-Enhanced SWOT Matrix Teaching Strategy
23. Saaty, T.L., Vargas, L.: Decision Making in Economic, Political, Social and Technological Environments with the Analytic Hierarchy Process. RWS Publications, Pittsburgh (1994)
24. Saccol, A.Z.: Um Retorno ao Básico: Compreendendo os paradigmas de pesquisa e sua aplicação na pesquisa em administração. Revista de Administração. **2**(2), 250–269 (2009). UFSM, Santa Maria, maio/ago
25. Thomas, J.R., Nelson, J.K., Silverman, S.J.: Métodos de pesquisa em atividade física, 5th edn. Artmed, Porto Alegre (2007)
26. Creswell, J.W.: Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: Projeto de pesquisa métodos qualitativo, quantitativo e misto. Artmed, Porto Alegre (2010)
27. Bardin, L.: Análise de conteúdo. Edições 70, São Paulo (2011)
28. Godet, M.: A caixa de ferramentas da prospectiva estratégica: problemas e métodos. Caderno do Centro de Estudos de Prospectiva e Estratégia, Lisboa, n. 5 (2000)
29. Larson, R., Farber, B.: E: rank reversals in multicriteria decision analysis with statistical modeling of ratio scale pairwise comparisons. J. Oper. Res. Soc. **56**(7), 855–861 (2005)
30. Landeta, J.: Current validity of the Delphi method in social sciences. Technol. Forecast. Soc. Change **73**(5), 467–482 (2006). doi:10.1016/j.techfore
31. Mota, M.O., Nogueira, C.A.G., Ogasavara, M.H.: The internationalization strategies of information technology firms from Brazil: an AHP analysis of Ivia's case. Internext – Revista Eletrônica de Negócios Internacionais da ESPM **6**(1), 21–41 (2011)
32. Saaty, T.L.: Decision making with the analytic hierarchy process. Int. J. Serv. Sci. **1**(1), 83–98 (2008)
33. Saaty, T.L., Ergu, D.: When is a decision-making method trustworthy? Criteria for evaluating multi-criteria decision-making methods. Int. J. Inf. Decis. Mak. **14**, 1–17 (2015)
34. Mccaffrey, D.J.: Using the multi-attribute global inference of quality (MAGIQ) technique for software testing. In: Sixth International Conference Information Technology, New Generations, ITNG 2009, pp. 738–742. IEEE (2009)
35. Schenkerman, S.: Inducement of nonexistent order by the analytic hierarchy process. Decis. Sci., 1–6 (1997). Spring
36. Lipovetsky, S.: Comparison of a dozen AHP techniques for global vectors in multiperson decision making and complex hierarchy. In: International Symposium on the Analytic Hierarchy Process, vol. 10. ISAHP, Pittsburgh (2009)

# Automatic Structuring of Arabic Normative Texts

Ines Berrazega[1](✉) and Rim Faiz[2]

[1] LARODEC, University of Tunis – ISG, 2000 Bardo, Tunisia
ines_berrazega@yahoo.fr
[2] LARODEC, University of Carthage – IHEC,
2016 Carthage Presidency, Tunisia
rim.faiz@ihec.rnu.tn

**Abstract.** The amount of unstructured documents daily produced has dramatically increased in the last few years. As a result, automatic structuring of these contents has become an urgent need: it constitutes a prerequisite to any further automatic processing in term of annotation, indexing, information retrieval, etc. Nevertheless, a lack of automatic structuring methods for the Arabic normative texts is perceived. In this context, a method for automatic structuring of Arabic normative texts is presented in this paper. A standardized structure of Arabic normative texts is defined: two levels of granularity are identified: thematic and logic. A semantic annotation rule base is also developed to automatically structure documents according to these levels of granularity. Obtained results are very promising: the overall performance reached 94.53% for Precision, 91.21% for Recall and 92.84% for F-score.

**Keywords:** Automatic structures · Normative texts · Thematic structuring · Legal XML · DTD · Arabic natural language processing

## 1 Introduction

Nowadays, huge amounts of textual documents are daily produced in all areas. These documents are produced in most cases in an unstructured way, which constitutes a brake on their effective exploitation. Thus, the automatic structuring of textual documents is a determinant task. It enables a better document modeling, a better content analysis and thus a better knowledge management. It also enables a better access to the relevant information hidden in these texts. The structural properties extracted from these documents could also be exploited to index them and to enhance the information retrieval process. In this regard, proposing computational methods and developing effective systems for automatic structuring of textual documents constitute an interesting task.

The need of automatically structuring these contents was increasingly felt in the legal domain, which is characterized by a variety of legal documents categories: codes, normative texts, jurisprudence, contracts etc. Each of which admits a different structure. For instance, normative texts are characterized by a deep hierarchical logical structure: articles constitute elementary elements, and are grouped under more general subdivisions like sections, chapters, books, etc. In this concern, we argue that automatic

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_22

identification and semantic annotation of these structural properties would facilitate exploiting the wealth of information conveyed by these structures. In this way, it would be easier to effectively index these contents and to enhance the performance of legal information retrieval system.

Though the importance of this task, we have noticed that normative texts written in Arabic language have not been yet processed. Thus, it would be wise to set up structuring methods able to take into account the specificities of Arabic normative texts. These methods should also be able to overcome the ambiguities raised by the Arabic language in text segmentation such as the lack of capital letters and regular punctuation, which makes conventional segmentation methods non appropriate to Arabic [1].

In the scope of this work, an automatic structuring method is proposed and evaluated over a corpus of Arabic normative texts collected from the Official Gazette of the Republic of Tunisia[1]. These documents do not meet any standard structuring or markup process. They are diffused in a raw unstructured format: structural portions constituting the normative texts (articles, sections, etc.) are not tagged. Consequently, accessing to the formal and semantic contents of texts could not be effectively done. In this concern, we argue that proposing a structuring method of Arabic normative texts is a primordial condition to enhance access of relevant information contained in these contents. The remainder of this paper is structured as follows. Background and related work are presented in Sect. 2. The proposed structuring method is detailed in Sect. 3. Evaluation and obtained results are presented and discussed in Sect. 4. Conclusion and future work are presented in Sect. 5.

## 2 Background and Related Work

The automatic processing of structural information in legal textual contents has been processed over two major kinds of legal documents: juridical judgments and normative texts. For juridical judgments, several approaches have been proposed. For instance, [2] proposed an approach for automatic summarization of legal judgments written in English. The authors extracted relevant units in the source judgments by identifying the discourse structures constituting these judgments. To do so, the authors defined the following thematic structures: Decision, Data, Introduction, Context, Juridicial Analysis and Conclusion. Then, the authors used the predefined structures to determine the semantic roles of these segments in order to thematically structure documents. For this purpose, the authors conducted a filtering step to eliminate unimportant quotations and noises. Then, they selected candidate units to generate judgments summaries Preliminary evaluation results reaches 90% of F-measure for thematic segmentation.

[3] proposed a system for document segmentation and automatic summarization of legal judgments. To perform document segmentation, the authors made use of the Conditional Random Field, a machine learning technique, to identify the rhetorical roles and extract structured head notes from sentences. For automatic summarization, the authors applied a set of probabilistic models to automatically extract key sentences

[1] http://www.legislation.tn/.

and compose the relevant chunks in the form of a headnote. The authors argue that the determination of basic structures and distinct segments helps improving the final presentation of the summary.

As argued by [2, 3], the automatic structuring of legal documents is very useful for further Natural Language Processing applications like information extraction and text summarization. Though the importance of this task, we have noticed a lack of computational methods for the automatic structuring of Arabic legal documents expect the work of [4, 5] who proposed a linguistic method for Arabic jurisprudence decisions' structuring. In this work, the authors constructed a training corpus of Arabic jurisprudence decisions and asked legal experts to manually annotate it. Then, they conducted a linguistic analysis over the corpus in order to extract linguistic markers expressing structural borders of the thematic segments constituting a decision. The extracted markers were used in a further step to develop a set of linguistic patterns to automatically structure the decisions.

On the other hand, automatic structuring of normative texts has been addressed in the scope of the DEFT challenge [6]. Participants were called to propose methods to segment an extract of a European Union law written in French into a series of articles. Two types of approaches have been proposed: statistical approaches based on the lexical cohesion concept and on statistical and distributional algorithms including the TextTiling algorithm [7] and the C99 algorithm [8]; and linguistic approaches.

Statistical approaches consist on calculating terms repetition based on similarity scores between sentences or on terms densities in the terms space of the text. Statistical approaches have the advantage of being domain independent and they do not require a learning phase. However, they give low accuracy levels. In the other hand, linguistic approaches are based on cohesion linguistic markers. Linguistic approaches are less used because they require a considerable learning effort over large corpora, and they are domain dependent. Nevertheless, they provide more accurate results, compared with statistical ones.

As shown along this section, almost all related work in the automatic structuring of legal documents domain processed jurisprudence texts. There is a lack of work dealing with normative texts, which admits different structures and employs different vocabulary when compared with jurisprudence decisions. For this reason, we assume that proposing a method dealing with the structural properties of normative texts written in Arabic would be very interesting. As we are working on a domain dependant corpus, we opt for linguistic approaches. We therefore propose a linguistic structuring method for Arabic normative texts, which constitutes to our knowledge, the first method proposed to automatically structure Arabic normative texts.

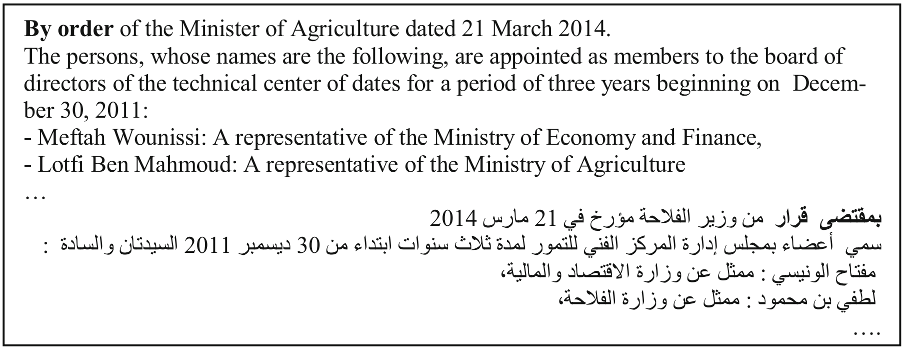
## 3 Proposed Automatic Structuring Method

To automatically structure texts, a corpus analysis was carried on as a first step to identify structural regularities of Arabic normative texts. Based on the observed regularities, a Document Type Definition (DTD) was built. In a second step, a structuring rule base composed of two parts was developed: a part grouping thematic rules, and a part grouping organizational rules. Finally, an automation process was conducted to translate the performance of these rules on a computational algorithm. The enhanced steps are detailed in what follows.

### 3.1 Definition of a Standardized Document Structure

The definition of a standardized structure of Arabic normative texts was done by analyzing a corpus of 100 Arabic normative texts collected from the Official Gazette of the Republic of Tunisia. At this stage it's worth noting that an in-house corpus composed from 200 texts was constructed: 100 texts were used as a training dataset and 100 texts were used as a test data set to evaluate the performance of the proposed method. The corpus analysis allowed us to study the structural regularities of texts. Two different structures were identified as follows:

**First Structure:** Texts having this structure are short, generally constituted of a title followed by a single paragraph or by a sentence followed by a series of indents. This is the case for example of orders appointing persons in a given position. The title starts with an expression indicating the type in the form: "By decree/بمقتضى أمر" or "By Order/بمقتضى قرار". An example of this first structure is shown in Fig. 1.



**By order** of the Minister of Agriculture dated 21 March 2014.
The persons, whose names are the following, are appointed as members to the board of directors of the technical center of dates for a period of three years beginning on December 30, 2011:
- Meftah Wounissi: A representative of the Ministry of Economy and Finance,
- Lotfi Ben Mahmoud: A representative of the Ministry of Agriculture
…

**بمقتضى قرار** من وزير الفلاحة مؤرخ في 21 مارس 2014
سمي أعضاء بمجلس إدارة المركز الفني للتمور لمدة ثلاث سنوات ابتداء من 30 ديسمبر 2011 السيدتان والسادة :
مفتاح الونيسي : ممثل عن وزارة الاقتصاد والمالية،
لطفي بن محمود : ممثل عن وزارة الفلاحة،
….

**Fig. 1.** First possible structure of texts

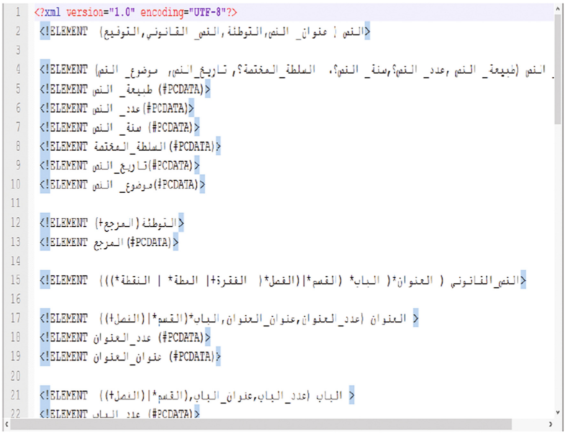
**Second Structure:** Texts having this structure are composed of four informational blocks namely a title (عنوان النص), a preamble (التوطئة), a body (النص القانوني) and a signature (التوقيع). Each segment is composed in turn of a set of information.

For texts belonging to this second structure, examples belonging to different possible types (laws, decrees and orders) were studied. We noticed that the elements constituting each segment differ slightly depending on the nature of the text. The constituents of each segment are detailed in Table 1.

The segment "text body/النص القانوني" follows a hierarchical structure composed of different subdivision levels. The presence/absence of these levels varies from one text to another depending on their volumes. The following subdivisions could be found in a normative text (from the broadest level to the lowest level): "Book/كتاب"; "title/عنوان"; "Chapter/باب"; "Section/قسم"; "Article/فصل"; "Paragraph/فقرة"; "Indent/مطة". This sequence of subdivisions is the most complete for a normative text. However, in most

**Table 1.** Second possible structure of texts

| Segment | Structure |
|---|---|
| Title<br>“ عنوان النص ” | • Text nature “طبيعة النص”<br>• Competent authority "السلطة المختصة"<br>• Date of text “ تاريخ النص ”<br>• Object of text “ موضوع النص ” |
| Preamble<br>“التوطئة ” | Composed of a set of legal references (visas) that correspond to the legal basis of the text. These references are listed according to the usual order in the hierarchy of legal standards and in a chronological order. |
| Text Body<br>“ النص القانوني ” | Either composed of a single article or a series of articles that could be grouped in titles "عناوين", chapters "أبواب" and sections "أقسام".<br>An article could be composed of pargraphs "فقرات" and / or indents "مطات" |
| Signature<br>“التوقيع ” | • Place and date of signature " تاريخ و مكان التوقيع"<br>• The autority who signed the text "الجهة المعنية بالتوقيع"<br>• The autority who approved it "الجهة المعنية بالموافقة" |



```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT النص ( عنوان_النص,التوطئة,النص_القانوني,التوقيع)>

<!ELEMENT _النص (طبيعة_النص ,عدد_النص?,سنة_النص?، السلطة_المختصة?, تاريخ_النص, موضوع_النص)
<!ELEMENT طبيعة_النص (#PCDATA)>
<!ELEMENT عدد_النص (#PCDATA)>
<!ELEMENT سنة_النص (#PCDATA)>
<!ELEMENT السلطة_المختصة (#PCDATA)>
<!ELEMENT تاريخ_النص (#PCDATA)>
<!ELEMENT موضوع_النص (#PCDATA)>

<!ELEMENT التوطئة (المرجع+)>
<!ELEMENT المرجع (#PCDATA)>

<!ELEMENT النص_القانوني ( العنوان*( الباب* (القسم*|(الفصل*( الفقرة+| المطة* | النقطة*)))>

<!ELEMENT العنوان (عدد_العنوان,عنوان_العنوان,الباب*(القسم*|(الفصل+))>
<!ELEMENT عدد_العنوان (#PCDATA)>
<!ELEMENT عنوان_العنوان (#PCDATA)>

<!ELEMENT الباب (عدد_الباب,عنوان_الباب,(القسم*|(الفصل+))>
<!ELEMENT عدد_الباب (#PCDATA)>
```

**Fig. 2.** Document type definition of Arabic normative texts

cases, the largest used level is the Article. Other levels are only used in case of long texts that need to group Articles into Sections, Chapters and Titles according to their topics. Note also that the Chapter subdivision may contain several sections (including articles). In other cases, a Chapter may begin with one or more Articles followed by one or more Sections. The complete architecture covering all possible cases was translated in a DTD as shown in Fig. 2.

## 3.2 Automatic Structuring of Texts

Automatically structuring texts amount to develop a set of structuring rules able to identify all segments borders and attribute tags delimiting them. Two structuring levels were conducted: a thematic level and a logic/organizational level. This mechanism was traduced throw the development of a tool which takes as input a raw unstructured text file and produces an XML document thematically and logically structured (as shown in Fig. 5). Texts were structured using a tree based modeling. For this purpose, the DOM standard[2] (Document Object Model) is used. The principle of structuring text by a node tree is to create a root node which is the access point to all the segments constituting the text. Thereafter, each text segment is assigned to a child node. In the scope of our work, the root of a text has 4 children nodes corresponding to the 4 thematic segments. Then, the textual content of each thematic segment is stored in its corresponding node. Thus, access to a segment could be done by directly identifying the node in question. This method greatly facilitates access to portions of text that meet users' requirements.

**Preprocessing: Sentence Splitting**

As previously mentioned, the corpus study is published as plain unstructured texts, constituted from a series of lines delimited by the sentence delimiter "\ n". Thus, a preprocessing step was carried on to segment this text into a set of sentences. For this purpose, a sentence splitting method was developped as shown in Fig 3. It takes as input an unstructured plain text and produces a tree nodes XML document by storing the content of each line in a distinct node. The root node of the XML document is labeled "Text النص". Each sentence is stored in a child node with an identifier and the label "Sentence/جملة". We deemed this step essential to guarantee text clarity at the thematic and logical structuring steps. Indeed, DOM tree modeling considers all the textual content of one node as a single sentence. Thus, the organization of the text into distinct sentences within 4 thematic nodes would be lost. This solution was adopted to keep a certain organization of the text. If not, all sentences belonging to a child node in the DOM tree would be considered in following steps as a single sentence, which would further complicate the structuring process.

```
Let C be a raw corpus constituted of a series of lines Li;
C={Li}
Create the text root TRx « النص »
Initialize the sentence counter Idi←0
For each line Li in the text Do
   Create a new node Ni « جملة »
   Assign Idi to the node Ni as an Identifier
   Add the content of Li to Ni
   Increment Idi ; Idi←Idi+1
   Add Ni to the text root RTx
End For
```

**Fig. 3.** Our sentence splitting method

**Contextual Exploration Based Thematic structuring**

Thematic structuring was achieved based on the Contextual Exploration (CE) method [9]. CE is a linguistic method of textual analysis. It consists on identifying, for a given notion, a set of linguistic markers having contextual dependencies. In a first step, a set

[2] https://www.w3.org/TR/WD-DOM/introduction.html.

of *main indicators* that could be used to express the sought notion are searched in text. Once identified, *complementary clues* are then searched to confirm or infirm that the pre-identified indicators do express the sought notion. Thus, the complementary clues could be respectively categorized as *positive/negative* clues, once identified they help to confirm/infirm the presence of searched notion in text. This method provides access to the semantic content of a text without conducting morphological and syntactic analysis or using external knowledge. Thus, we were able to overcome the lack of Arabic natural language processing tools like deep syntactic analyzers and semantic role labelers to determine syntactic dependencies between the relevant linguistic markers and their corresponding semantic roles.

Several computational approaches were based on the Contextual Exploration method to tackle different Natural Language Processing issues: Arabic Text segmentation [10], Events extraction from News articles [11], Indexing and retrieval of learning objects [12], Image and Text Mining [13], etc. Hence, we would investigate the effectiveness of applying this method to process Arabic legal corpora. Thus, the CE-based mechanism was used to propose a thematic structuring method to identify discourse markers used by legislators to express the boundaries of thematic segments constituting normative texts. The proposed method couples a thematic structuring rule base and an automatic structuring algorithm.

***Definition of the thematic structuring rule base.***

Let *BRST* be a CE-based thematic structuring rule base. BRST admits a node tree structure and is defined as follows:

$$BRST \; = \cup_1^4 RST_j \; with \, j = \{Title, \; Preamble, \; Text\,Body, \; Signature\}$$

With:

- BRST: the tree root node
- $\text{RST}_{\text{Title}}$, $\text{RST}_{\text{Preamble}}$, $\text{RST}_{\text{Text Body}}$ and $\text{RST}_{\text{Signature}}$ four child nodes, grouping respectively the structuring rules of the four thematic segments that constitute a normative text.

***Definition of a thematic structuring rule***

Let $\text{R}_\text{i}$ be a CE-based thematic structuring rule with:

$$BRST = \cup_1^n R_i \tag{1}$$

$$\begin{aligned} R_i \; &= \; \{Id, \; Icnd, \; Icpg_j, \; Icpd_j, \; TS_k\} \; with\, j = \{1, 2\}; \\ TS_k \; &= \; \{Title, \; Preamble, \; Text\,Body, \; Signature\} \end{aligned} \tag{2}$$

Each rule $R_i$ is composed from:

- Id: a set of trigger indicators (main indicators)
- Icnd: a set of negative complementary clues searched in CnD
- $\text{Icpd}_1$: a first set of positive complementary clues searched in CnD
- $\text{Icpd}_2$: a second set of positive complementary clues searched in CnD

- $Icpg_1$: a first set of positive complementary clues searched in CnG
- $Icpg_2$: a second set of positive complementary clues searched in CnG
- TSk: the type of the thematic segment in question
- CnD and CnG represent respectively the right and left contexts of the indicator Id in the rule $R_i$

```
Thematic Structuring (BRST, CSP){
Create the Root Node of the thematically structured corpus NR
« الرائد_الرسمي »
   Sentences_List= {Phi; $\bigcup_1^n$Phi =CSP}
   For each sentence $Ph_i$ ∈ Sentences_List Do
     Create a list Candidates_Rules_List
           For each Rule $R_i$ ∈ BRST Do
              If($t_i$∈$Ph_i$)and ∃($Id_i$∈$R_i$); $t_i$=$Id_i$ Then
                 Candidates_Rules_List=Add($R_i$;$t_i$=$Id_i$and $Id_i$∈$R_i$)
           End For
      Let ErD and ErG be respectively the right and left
      search spaces of $Id_i$ in $Ph_i$
      Let CnD and CnG be respectively the right and left con-
      texts of $Id_i$ in $R_i$
      If Candidates_Rules_List is not empty Then
          Found=false
          While (Found=false)
                If for each($Icnd_j$ ∈ CnD) ∄ ($t^1_i$ ∈ ErD);
                ($t^1_i$=Icnd)Then
                      If for each($Icpd_j$ ∈ CnD ∃ ($t^2_i$ ∈ ErD;
                      $t^2_i$=$Icpd_j$)Then
                         If for each($Icpg_j$ ∈ CnG) ∃ ($t^3_i$ ∈
                         ErG);$t^3_i$=$Icpg_j$) Then
                            (Found=true)
                                If $TS_k$= Title then
                                    Create_node($TN_i$,NR)
                                    Create_node($ST_{titre}$,$TN_i$)
                                    Insert($Ph_i$,$ST_{titre}$)
                                Else
                                    Create_node ($ST_j$, $TN_i$)
                                    Insert($Ph_i$,$ST_j$)
                                End If
                         End If
                      End If
                End If
          End while
      Else
          Insert($Ph_i$,$ST_j$)
   End For
```

**Fig. 4.** Proposed thematic structuring algorithm

(1) The set of all the rules $R_i$, whatever the segment to which they are applied, constitute our base BRST.
(2) All rules admit a unified structure. This structure is defined based on the formalism of the Contextual Exploration method.

***Proposed CE-based thematic structuring algorithm***

Let *BRST* be a CE-based thematic structuring rule base; $BRST = \cup_1^n R_i$

Let $R_i$ a thematic structuring rule; $R_i \in \left\{R_{Title}, R_{Preamble}, R_{Text\,Body}, R_{Signature}\right\}$

From these two definitions, we can draw the following remarks:

1. From a structural point of view, the base BRST is constituted in the form of a node tree having RST as a root, with $RST = \bigcup_1^4 RST_j$ (def1).
2. From a content point of view, BRST is constituted from a set of rules $R_i$ (def2).

Thus, it is possible to say that: $BRST = \bigcup_1^4 RST_j = \bigcup_1^n RS_i$

Let *CSP* be a corpus segmented into sentences $Ph_i$; $CSP = \bigcup_1^n Ph_i$. CSP is constituted from a set of normatifs texts $TN_i$ with $CSP = \bigcup_1^n TN_i$

Let be $TN_i = \left\{\bigcup_1^4 ST_j\right\}$ with $ST_j = \{Title, Preamble, Text\,Body, Signature\}$.

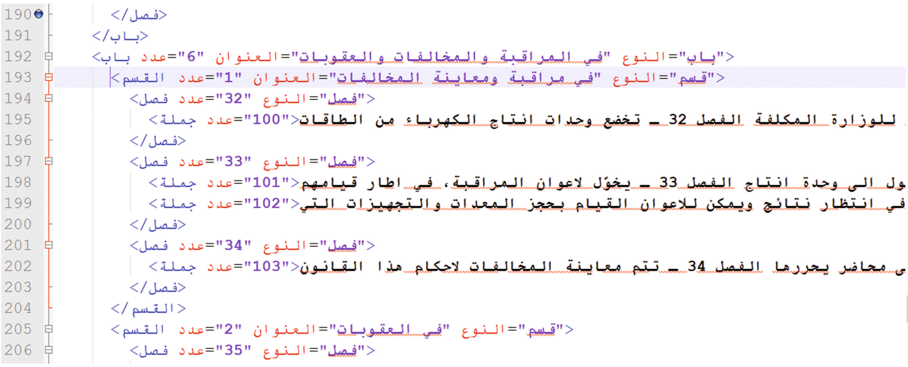
Each normative text $TN_i$ is composed from four thematic segments $ST_j$. The proposed thematic structuring algorithm is detailed in Fig. 4.

**Organizational Texts Structuring**

Once the corpus was thematically structured, an organizational structuring step was conducted. For each text, the Preamble segment was decomposed into a series of Visas and the Text Body segment was decomposed into a series of Articles, Sections, Chapters, etc. (as previously detailed in Sect. 3.1). Organizational structuring was conducted by developing a set of organizational structuring rules based on regular expression patterns.

For the visas, the identification of the beginning of a new segment was based on the identification of expressions like "After seeing the Constitutive Act number… الاطلاع على القانون التأسيسي عدد بعد" or "And the Decision of… وعلى القرار المؤرخ في", etc. within the *Preamble* segment. Thus, the organizational structuring of the Preamble into Visas consists on crossing the sentences nodes constituting the Preamble, until spotting one of the pre-mentioned expressions in the head of a sentence.

For the *Text Body* segment, identifying the beginning of a new subdivision is based on the identification of a term such as "Article الفصل, Section القسم, Chapter الباب, Title العنوان, etc." at the head of the current sentence node, followed by a sequential number "2, 3, 4, etc." in the case of Articles or by a term such as "First الأول, Two الثاني, Three الثالت, etc." in the case of subdivisions superior to Articles. The organizational segmentation of the *Text Body* segment is conducted as follows: the sentence nodes constituting the segment are crossed one by one. For each sentence, the presence of an



```
    </فصل>
  </باب>
  <باب عدد="6" العنوان="في المراقبة والمخالفات والعقوبات" النوع="باب">
    <القسم عدد="1" العنوان="المخالفات ومعاينة في مراقبة" النوع="قسم">
      <فصل عدد="32" النوع="فصل">
        <جملة عدد="100">للوزارة المكلفة الفصل 32 ـ تخضع وحدات انتاج الكهرباء من الطاقات
      </فصل>
      <فصل عدد="33" النوع="فصل">
        <جملة عدد="101">ول الى وحدة انتاج الفصل 33 ـ يخوّل لاعوان المراقبة، في اطار قيامهم
        <جملة عدد="102">في انتظار نتائج ويمكن للاعوان القيام بحجز المعدات والتجهيزات التي
      </فصل>
      <فصل عدد="34" النوع="فصل">
        <جملة عدد="103">ى محاضر يحررها الفصل 34 ـ تتم معاينة المخالفات لاحكام هذا القانون
      </فصل>
    </القسم>
    <القسم عدد="2" العنوان="في العقوبات" النوع="قسم">
      <فصل عدد="35" النوع="فصل">
```

**Fig. 5.** Example of the identification of a section segment

expression denoting the beginning of a new subdivision is checked (while respecting the hierarchical order defined in the DTD, namely: Book كتاب/Title عنوان/Chapter باب/Section قسم/Article فصل/Paragraph فقرة/Indent مطة). If one of these expressions is found, then a new node is created under the Text Body segment of the current normative text, and three attributes are assigned to it: "number عدد", "title العنوان" and "type النوع". Figure 5 shows an example of the identification of a *Section* having *1* as sequence number and "*On sanctions في مراقبة ومعاينة المخالفات*" as title.

# 4 Evaluation and Obtained Results

To evaluate the performance of the proposed method, a test dataset composed of 100 texts was used. The choice of texts was based on a set of criteria to ensure better representation from structural and content point of views. Selected texts admit different types (28 laws, 42 orders and 30 decrees); address various topics; are published in different dates (between 2000 and 2016); represent the two possible structures (16 texts having the first structure and 84 texts having the second structure); have different lengths (ranging from texts composed of one paragraph to texts composed of more than 80 articles) with a variation in terms of presence/absence of subdivisions. The total number of Visas and Articles in the test dataset reached respectively 987 and 1274. The total number of sentences in the corpus reached 7460: 3415 constitute the beginning of a new segment and 4045 constitute normal sentences.

The performance of the structuring process was evaluated by calculating the Precision, Recall and F-score values. Precision and Recall are used to better assess the degree of noise and silence of our structuring rules. The evaluation was performed for the 4 types of thematic segments (Title, Preamble, Text Body and Signature); for all Visas constituting the Preamble of each text; as well as for all Articles constituting the body of each text. For the other organizational subdivisions (Books, Chapters, Sections, etc.), Precision, Recall and F-score values were calculated one time for all subdivisions to simplify calculations. The overall obtained results are shown in Table 2.

The Precision, Recall and F-score values for all the thematic and logic segments were calculated as follows:

$$Precision = \frac{NCIS}{NIS}$$

$$Recall = \frac{NCIS}{TNS}$$

$$\mathrm{F\text{ - }score} = \frac{2 * \mathrm{NCIS}}{\mathrm{NIS} \ + \mathrm{TNS}}$$

With:

**NCIS:** Number of sentences Correctly Identified as a beginning of a Segment
**NIS:** Number of sentences Identified as a beginning of a Segment by the system
**TNS:** Total Number of sentences representing a beginning of a Segment in the corpus

The overall performance of the proposed method reached 94.53% for Precision, 91.21% for Recall and 92.84% for F-score. The Precision values range from 79.54% for the identification of the Visas to 100% for the other types of segments. The Recall values range from 70.92% for the identification of the Visas to 100% for some of the other types of segments.

Errors occurring in the identification of Visas are explained by the observation of some new structural regularities in the test data set that have not been observed in the training dataset. By consequence, their patterns do not exist in the rules base. This silence could be explained by the changes occurring in the legislative power in Tunisia, starting from 2011 until today in accordance with the establishment of new institutional bodies.

**Table 2.** Obtained results

| Nom du segment | NCIS | NIS | TNS | P | R | F-S |
|---|---|---|---|---|---|---|
| Title | 100 | 100 | 100 | 100 | 100 | 100 |
| Preamble | 75 | 75 | 84 | 100 | 89.28 | 94.33 |
| Text body | 100 | 100 | 100 | 100 | 100 | 100 |
| Signature | 80 | 80 | 84 | 100 | 95.23 | 97.56 |
| Visas | 700 | 880 | 987 | 79.54 | 70.92 | 74.98 |
| Articles | 1274 | 1274 | 1274 | 100 | 100 | 100 |
| Other subdivisions | 786 | 786 | 786 | 100 | 100 | 100 |
| Total | 3115 | 3295 | 3415 | 94.53 | 91.21 | 92.84 |

## 5 Conclusion

This paper presents an automatic structuring method for Arabic normative texts. This method consists in a first step on defining a standard structure for Arabic normative texts and building a legal DTD describing their structural content. The second step of this method consists on developing a structuring rule base allowing to thematically and logically structure texts. The proposed method was trained and tested over a corpus of Arabic normative texts collected from the Official Gazette of the Republic of Tunisia. Obtained results are very encouraging. The overall performance of the proposed method reached 94.53% for Precision, 91.21% for Recall and 92.84% for F-score.

We assume that this work would provide a considerable assistance in the production of structured Arabic legal documents and to ensure the interoperability of data represented in such content in order to facilitate their access and management by users.

## References

1. Aloulou, C., Belguith Hadrich, L., Ben Hamadou, A.: MASPAR: Multiagent System for Parsing Arabic, IEEE International Conference on Systems, Man and Cybernetics, vol. 7, pp. 6–9, Hammamet-Tunisie (2002)
2. Farzindar, A., Lapalme, G.: Legal text summarization by exploration of the thematic structures and argumentative roles. In: Text Summarization Branches Out Conference held in conjunction with ACL 2004, pp. 27–38 (2004)
3. Saravanan, M., Ravindran, B.: Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artif. Intell. Law **18**(1), 45–76 (2010)
4. Dhouib, K., Gargouri, F.: An applied legal ontology in Arabic for the jurisprudence decision-structuring. IJKSR **6**(1), 43–54 (2015)
5. Dhouib, K., Gargouri, F.: A textual jurisprudence decision structuring methodology based on extraction patterns and Arabic legal ontology. J. Decis. Syst. **23**(1), 69–81 (2014)
6. Azé, J., Heitz, T., Mezaour, A.D., Peinl, P., Roche, M., Mela, A.: Présentation de DEFT 2006 (DÉfi Fouille de Textes). In: Proceedings of DEFT 2006, vol. 1, pp. 3–12 (2006)
7. Hearst, M.A.: TextTiling: segmenting text into multiparagraph subtopic passages. Comput. Linguist. **23**, 33–64 (1997)
8. Choi, F.: Advances in domain independent linear text segmentation. Presented at the First Conference on North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, Washington (2000)
9. Desclés, J.-P. Contextual exploration processing for discourse automatic annotations of texts. In: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2006. AAI Press, California (2006)
10. Belguith, L., Baccour, L., Mourad, G.: Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles TALN 2005, pp. 451–456 (2005)
11. Elkhlifi, A., Faiz, R.: French-Written Event Extraction Based on Contextual Exploration, Dans Proc, FLAIRS. AAAI Press, Palo Alto (2010)
12. Smine, B., Faiz, R., Desclés, J.P.: A semantic annotation model for indexing and retrieving learning objects. J. Dig. Inf. Manag. (JDIM) **9**(4), 159–166 (2011)
13. Le Pirol, F.: Image and text mining based on contextual exploration from multiple points of view. In: Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference FLAIRS 2011. AAAI Press, Palo Alto (2001)

# Computer Aided Analysis of the Mobile Crane Handling System Using Computational Intelligence Methods

Wojciech Kacalak, Zbigniew Budniak, and Maciej Majewski(✉)

Department of Technical and IT Systems Engineering,
Faculty of Mechanical Engineering, Koszalin University of Technology,
Raclawicka 15-17, 75-620 Koszalin, Poland
{wojciech.kacalak,zbigniew.budniak,maciej.majewski}@tu.koszalin.pl

**Abstract.** This article describes computer aided analysis using computational intelligence methods for analysis and simulation research of a crane system during sequential movements. A parametric solid model has been specified, designed with a CAD/CAE environment, which allows to evaluate its stability for selected configurations and conditions of operation. Neural-network-supported analysis of varying contact forces exerted by the outriggers onto the ground, stabilizing and overturning torques, mass centre during handling allowed to specify trajectories ensuring stability of the crane. The results of the simulation research have been presented as changes of stability conditions depending on: angular position of the column with its telescopic arms and booms, the position of the telescopic arms themselves, the mass of individual components of the load system, as well as the load value applied onto it.

**Keywords:** Machine control · Mobile crane · Interactive system · Computer aided analysis · CAD/CAE · Neural networks · Artificial intelligence · Computational Intelligence Methods

## 1 Introduction

This study presents computer aided analysis of the mobile crane handling system using computational intelligence methods based on the methodology developed with the use of a simulation model built in the integrated CAD/CAE environment. The model proposed consists of the main crane assemblies coupled together: the truck with outrigger system and the base, the slewing column, the inner and outer arms, the six-member telescopic boom, the hook with lifting sling and the transported load. In the modelling of the crane system, the masses of the majority of the equipment and the assemblies that load the system were taken into account. An example of the application of the method proposed to determine an optimal trajectory of the handling assignment being performed was presented as the results of simulation testing.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_23

## 2 Analysis and Simulation Research Methodology

In the computer aided simulation of the handling task of the mobile crane, the methodology presented in Fig. 1 was used.

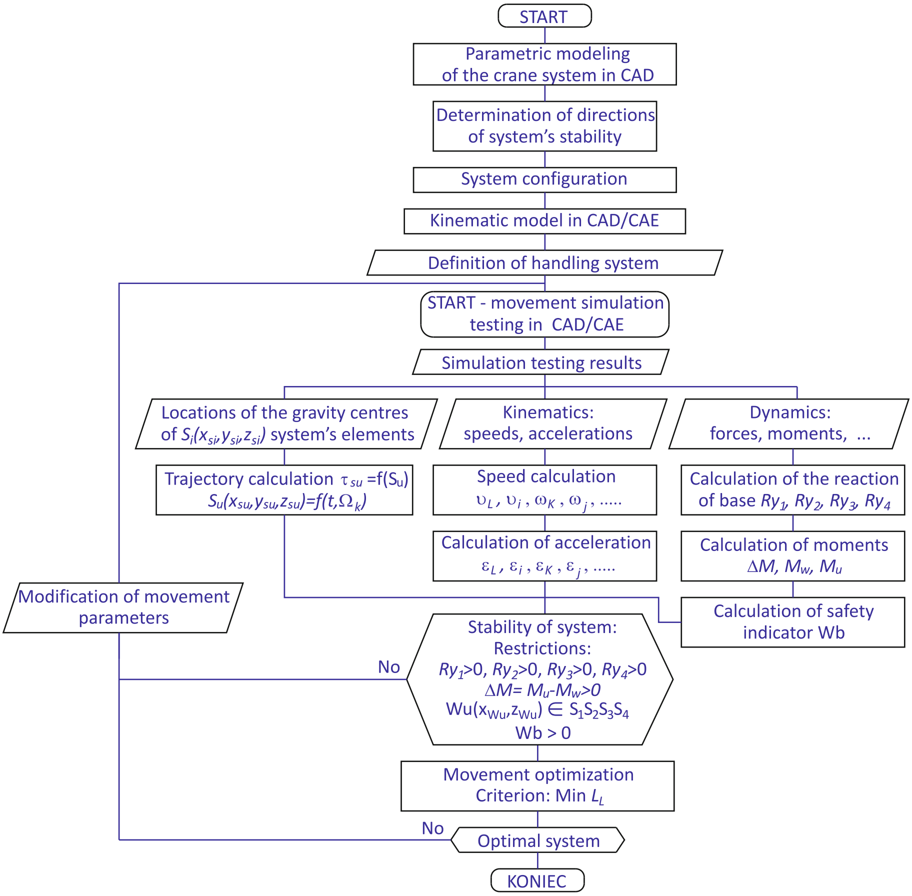

START
Parametric modeling of the crane system in CAD
Determination of directions of system's stability
System configuration
Kinematic model in CAD/CAE
Definition of handling system
START - movement simulation testing in CAD/CAE
Simulation testing results
Locations of the gravity centres of $S_i(x_{si}, y_{si}, z_{si})$ system's elements
Kinematics: speeds, accelerations
Dynamics: forces, moments, ...
Trajectory calculation $\tau_{su}$ =f($S_u$) $S_u(x_{su}, y_{su}, z_{su})=f(t,\Omega_k)$
Speed calculation $\upsilon_L, \upsilon_i, \omega_K, \omega_j$, .....
Calculation of the reaction of base $Ry_1, Ry_2, Ry_3, Ry_4$
Calculation of acceleration $\varepsilon_L, \varepsilon_i, \varepsilon_K, \varepsilon_j$, .....
Calculation of moments $\Delta M, M_w, M_u$
Calculation of safety indicator Wb
Modification of movement parameters
Stability of system: Restrictions: $Ry_1>0, Ry_2>0, Ry_3>0, Ry_4>0$ $\Delta M = M_u - M_w > 0$ $Wu(x_{Wu}, z_{Wu}) \in S_1S_2S_3S_4$ Wb > 0
No
Movement optimization Criterion: Min $L_L$
No
Optimal system
KONIEC

**Fig. 1.** Computer aided analysis and simulation of the mobile crane handling system.

The following are the basic elements of the method implemented:

- parametric modelling of the elements and the entire crane system in the CAD system for the defined configuration;
- determination of the system stability conditions (a notation of equations that constitute a mathematical model to calculate the following: the trajectory of the mass centres of the elements of the crane system, the reaction of the base on the crane outrigger system, the stabilizing torque $Mu$ and the overturning torque $Mw$ as well as the safety indicator);

- building of a kinematic model of the crane and carrying out simulation testing in the integrated CAD/CAE system;
- an analysis of the kinematic and dynamic quantities of the crane system during handling in connection with maintaining constant balance (stability);
- optimization of the trajectory of the displacements of the crane working elements for specified assignments taking limiting conditions into account.

Integrated CAD - SolidWorks software as well as the module for computations and engineering analyses: CAE - SolidWorks Motion was used for the purpose of the modelling and numeric tests of the crane handling system.

## 3 Model of the Handling Crane

The assemblies of the truck crane type HDS Hiab XS (Cargotec Poland) in relation to which the voice control system was proposed [1,2], include the design of all the main parts and sub-assemblies as well as other important elements of the construction [3,4]. The model of the support system (Fig. 2A) is composed of the following crane assemblies that are coupled together: the truck frame, the outrigger system placed in the crane base frame connected with the frame of the truck chassis, the crane base, the slewing column, the inner and outer boom with the installed six-member telescopic boom and the hook including lifting slings loaded with the transported cargo. A method for computer aided analysis for the mobile crane handling system's configuration during operation is proposed (Fig. 2B), which consists of mapping of the position of the crane's working elements using self-organizing networks [5].

Configuration of the mobile crane's cargo handling system as a combination of connected elements was analyzed as sets of local coordinate systems connected with the crane's components. The cargo's position vector $\overrightarrow{q}_l$, in the absolute coordinate system $Oxyz$, is given with the following formula:

$$\begin{aligned}\overrightarrow{q}_l &= L(x_L, y_L, z_L) = [x_L, y_L, z_L]^T \\ &= \overrightarrow{r}_f + \overrightarrow{r}_b + \overrightarrow{r}_k + \overrightarrow{r}_{W_w} + \overrightarrow{r}_{W_z} + \overrightarrow{r}_t + \overrightarrow{r}_h + \overrightarrow{r}_z + \overrightarrow{r}_l\end{aligned} \tag{1}$$

where: $\overrightarrow{r}_t = \overrightarrow{r}_{t_1} + \overrightarrow{r}_{t_2} + \overrightarrow{r}_{t_3} + \overrightarrow{r}_{t_4} + \overrightarrow{r}_{t_5} + \overrightarrow{r}_{t_6}$
$\overrightarrow{r}_f$, $\overrightarrow{r}_b$, $\overrightarrow{r}_k$, $\overrightarrow{r}_{W_w}$, $\overrightarrow{r}_{W_z}$, $\overrightarrow{r}_t$, $\overrightarrow{r}_h$, $\overrightarrow{r}_z$, $\overrightarrow{r}_l$ - vectors defining local coordinate systems origins' positions located at points $F$, $B$, $K$, $Ww$, $Wz$, $T$, $H$, $Z$, $L$, which belong to the truck $f$, the crane's base $b$, slewing column $k$, outer $W_w$ and inner $W_z$ arms, six-member telescopic boom $t$, hook $h$, lifting sling $z$, and the cargo $l$.

A change to the configuration of the crane system is connected with its working movements [3,4]. An analytical description of the configuration of the crane kinematic system involves strenuous conversions of vector-matrix equations [3], until explicit dependences have been obtained that determine the variable angular and linear quantities. Knowledge of these dependences is very desirable.

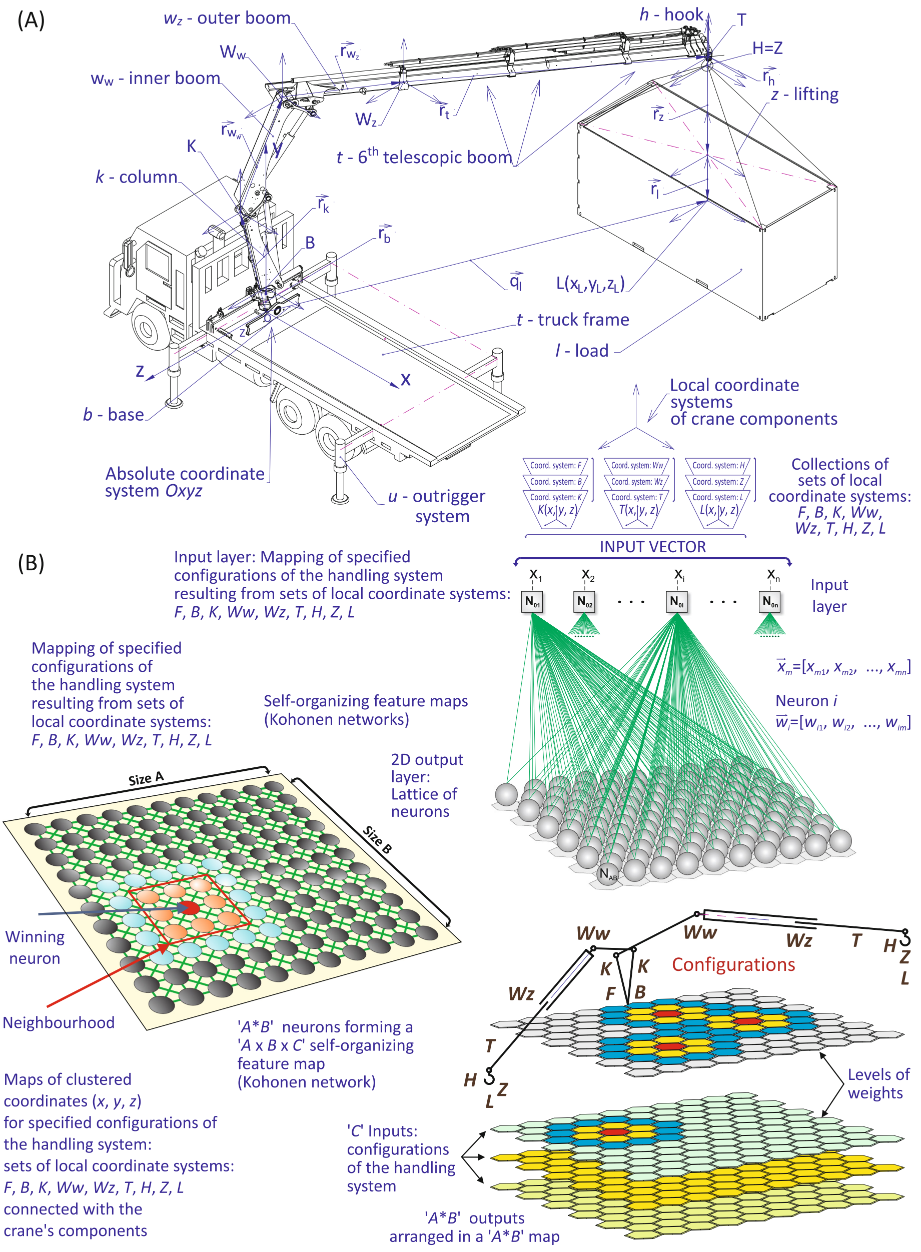

(A)
$w_z$ - outer boom
$W_w$
$w_w$ - inner boom
$\vec{r}_{w_z}$
$\vec{r}_{w_w}$
K
$k$ - column
$W_z$
$\vec{r}_t$
$t$ - 6th telescopic boom
$h$ - hook
T
H=Z
$\vec{r}_h$
$z$ - lifting
$\vec{r}_z$
$\vec{r}_l$
$\vec{r}_k$
$\vec{r}_b$
B
y
$\vec{q}_l$
$L(x_L,y_L,z_L)$
$t$ - truck frame
$l$ - load
Z
X
$b$ - base
Absolute coordinate system $Oxyz$
$u$ - outrigger system
Local coordinate systems of crane components
Coord. system: F
Coord. system: B
Coord. system: K
K(x, y, z)
Coord. system: Ww
Coord. system: Wz
Coord. system: T
T(x, y, z)
Coord. system: H
Coord. system: Z
Coord. system: L
L(x, y, z)
Collections of sets of local coordinate systems: $F, B, K, Ww, Wz, T, H, Z, L$
(B)
Input layer: Mapping of specified configurations of the handling system resulting from sets of local coordinate systems: $F, B, K, Ww, Wz, T, H, Z, L$
INPUT VECTOR
$x_1$
$x_2$
$x_i$
$x_n$
$N_{01}$
$N_{02}$
$N_{0i}$
$N_{0n}$
Input layer
$\bar{x}_m=[x_{m1}, x_{m2}, ..., x_{mn}]$
Neuron $i$
$\bar{w}_i=[w_{i1}, w_{i2}, ..., w_{im}]$
Mapping of specified configurations of the handling system resulting from sets of local coordinate systems: $F, B, K, Ww, Wz, T, H, Z, L$
Self-organizing feature maps (Kohonen networks)
2D output layer: Lattice of neurons
Size A
Size B
$N_{AB}$
Winning neuron
Neighbourhood
'$A*B$' neurons forming a '$A$ x $B$ x $C$' self-organizing feature map (Kohonen network)
Ww
Ww
Wz
T
H
Z
L
K
K
F
B
Wz
T
H
Z
L
Configurations
Levels of weights
Maps of clustered coordinates ($x, y, z$) for specified configurations of the handling system: sets of local coordinate systems: $F, B, K, Ww, Wz, T, H, Z, L$ connected with the crane's components
'$C$' Inputs: configurations of the handling system
'$A*B$' outputs arranged in a '$A*B$' map

**Fig. 2.** (A) Mobile crane handling system's configuration during operation; (B) mapping of the position of the crane's working elements using self-organizing networks.

## 4 Kinematic Model of the Handling System

In simulation testing, a kinematic model was used of the mobile crane handling system with four degrees of freedom. Considering a large number of elements and their construction characteristics [3,4], a number of necessary simplifications were used in the kinematic model (Fig. 3).

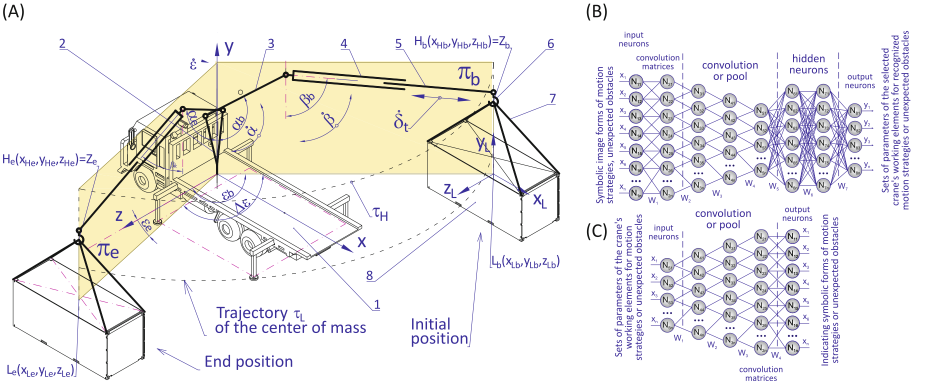

(A)
2
3
4
5
6
7
8
1
$\dot{\varepsilon}$
y
$H_b(x_{Hb}, y_{Hb}, z_{Hb})=Z_b$
$\pi_b$
$\beta_b$
$\dot{\beta}$
$\dot{\delta}_t$
$\alpha_b$
$\dot{\alpha}$
$\alpha_e$
$y_L$
$z_L$
$x_L$
$H_e(x_{He}, y_{He}, z_{He})=Z_e$
$\varepsilon_b$
$\Delta\varepsilon$
$\varepsilon_e$
z
x
$\tau_H$
$\pi_e$
$L_b(x_{Lb}, y_{Lb}, z_{Lb})$
Trajectory $\tau_L$ of the center of mass
Initial position
End position
$L_e(x_{Le}, y_{Le}, z_{Le})$
(B)
input neurons
convolution matrices
convolution or pool
hidden neurons
output neurons
Symbolic image forms of motion strategies, unexpected obstacles
Sets of parameters of the selected crane's working elements for recognized motion strategies or unexpected obstacles
$W_1$ $W_2$ $W_3$ $W_4$ $W_5$ $W_6$ $W_7$
(C)
input neurons
convolution or pool
output neurons
convolution matrices
Sets of parameters of the crane's working elements for motion strategies or unexpected obstacles
Indicating symbolic forms of motion strategies or unexpected obstacles
$W_1$ $W_2$ $W_3$ $W_4$

**Fig. 3.** (A) Simplified kinematic model of the crane handling system, where: 1 - truck including outrigger system, 2 - slewing column, 3 - inner arm, 4 - outer arm, 5 - telescopic boom, 6 - hook, 7 - lifting slings, 8 - cargo, $\tau_L$ - trajectory of cargo gravity centre, $\tau_H$ - trajectory of point $H = Z$ (hang point: lifting slings and hook); (B) Deep convolutional neural networks for recognition of motion strategies or unexpected obstacles and corresponding sets of parameters of the selected crane's working elements; (C) Deconvolutional neural networks for recognition of indicating symbolic forms of motion strategies or unexpected obstacles.

In order to determine dependences between the configuration coordinates ($\varepsilon$, $\varepsilon_b$, $\varepsilon_e$, $\alpha$, $\alpha_b$, $\alpha_e$, $\beta$, $\beta_b$, $\beta_e$, $\delta t$, $\delta t_b$, $\delta t_e$, where: $b$ and $e$ - indices for the initial and end positions) and the base coordinates of the location of the cargo, temporary $3D$ bonds were introduced into the simulation model, which determine the location of the handling system and its elements. In the model developed, drives were defined that perform the rotary motion of the crane column with velocity $\dot{\varepsilon}$ and linear drives that force the rotary motion of the inner and outer arms with velocities $\dot{\alpha}$ and $\dot{\beta}$ as well as sliding out of the six-member telescopic boom with velocity $\dot{\delta t}$. The computer aided analysis includes a method (Fig. 3B) based on deep convolutional neural networks [6] for recognition of motion strategies or unexpected obstacles and corresponding sets of parameters of the selected crane's working elements, and a method (Fig. 3C) consisting of deconvolutional neural networks [6] for recognition of indicating symbolic forms of motion strategies or unexpected obstacles.

## 5 Handling System's Stability

Owing to the performance of simulation testing in line with the methodology proposed in Fig. 1, it is possible to determine the optimum trajectory of cargo displacements for a selected handling assignment. The value of the moment required to maintain balance in relation to the tip-over axis may constitute the measure of the risk of the crane tipping over. Loading with the moment from the mass of the crane elements and the loads is additionally summed up with the moments that originate from inertia forces (caused by the movement of the cargo and its parts) and from the load with wind. The overturning torque $M_w$ is counteracted by the stabilizing torque $M_u$ with an opposite direction that is dependent from the mass and the location of the mass centre of the crane elements (Fig. 4A). The computer aided analysis methods (Fig. 4B) include probabilistic neural networks [7] for classification of vertical reactions and weights of the crane system to the classes: stabilizing torque, overturning torque, stability indicator.

According to international standards, it is accepted that the crane is stable if at any position of the boom loaded with lifting capacity with an adequate extension, the stabilizing torque $M_u$ is greater than the overturning torque $M_w$ by the value of $\Delta M$.

$$\Delta M = M_u - M_w > 0 \tag{2}$$

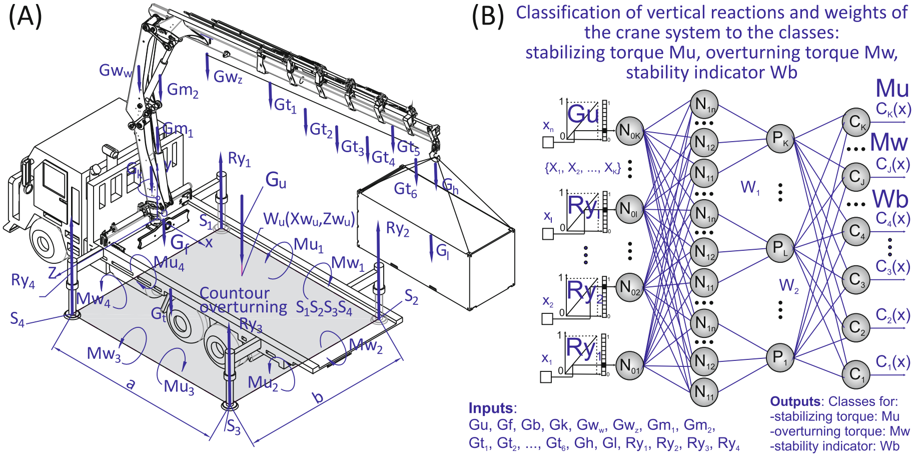


**Fig. 4.** (A) Diagram of forces and torques that act on the crane outrigger system: where: $Gu$ - total weight of the crane system; $Gf$ - weight of the truck including the outrigger system; $Gb$ - crane base weight; $Gk$ - weight of the slewing column; $Gw_w$ - weight of the inner arm, $Gw_z$ - weight of the outer arm; $Gm_1$, $Gm_2$ - weights of hydraulic cylinders; $Gt_1$, $Gt_2$,.., $Gt_6$ - weights of the arms of the six-member crane boom; $Gh$ - hook weight, $Gl$ - cargo weight; $Ry_1$, $Ry_2$, $Ry_3$, $Ry_4$ - vertical reactions of the base; a&b - spacing of the crane outriggers; (B) Probabilistic neural networks for classification of vertical reactions and weights of the crane system to the classes: stabilizing torque, overturning torque, stability indicator.

The following may also constitute the measure of the crane stability:

1. The value of the pressure on the base of the least loaded crane support and the value of the changes of this force in time;
2. The location of the symmetric mass centre of the handling system of the crane in relation to the support points. The system is stable if, in the projection on the horizontal plane, the mass centre is located inside the quadrangle that is established by the support points of the crane outrigger system;
3. Safety indicator $Wb$. The authors presented a new effectiveness assessment method of the handling assignment that permits the determination of the value of the safety indicator $Wb$ as a criterion of the stability of the crane system. The indicator $Wb$ accepts values from 0 to 1. The value of the indicator of $Wb = 0$ constitutes the lower limit of safe operation.

The stability indicator $Wb$ was defined as follows:

$$Wb = min \in \left\{ \frac{min(Ry_i)_t}{G_u{\cdot}k_1{\cdot}(1-k_2)} - \frac{k_2}{1-k_2} \right\}_t \tag{3}$$

where:

$$t = t_e - t_b \qquad t = \sum \mathit{\Delta} t_j \tag{4}$$

- $i = 1$ - 4 - number of the outrigger,
- $j =$ - number of the elementary fragment of the trajectory,
- $min(Ry_i)$, kN - the smallest of the vertical reactions of the base on the outrigger $i$,
- $Gu$, kN - total weight of the crane system,
- $k_1$ - index of the maximum load of the crane outrigger, $Ry_{max} = Gu{\cdot}k_1$, where: $k_1 \leq 0.25$ - for a crane with four outriggers,
- $k_2$ - index that determines the minimum load of the crane outrigger, $Ry_{min} = Gu{\cdot}k_2$,
- $t$, s - time of the working cycle of the handling assignment,
- $t_b = 0$, s - start of the crane working cycle,
- $t_e$, s - end of the crane working cycle.

In order to guarantee the stability of the crane system, the value of the indicator $Wb$ should be greater than zero when $min(Ry_i) > k_1{\cdot}k_2$. The value of the indicator $k_2$ is determined considering safety on the level that depends from the crane working conditions.

## 6 Handling Task

Simulation of the handling task was carried out for an example of a mobile crane of the HDS Hiab XS-111 type. The configuration of the movement of the working mechanisms of the crane during the execution of the four variants of the handling assignment is presented in Table 1, where the denotations of

**Table 1.** Parameters of sequential movements for four variants of handling assignment

| Movement sequence | I | II | III | IV |
|---|---|---|---|---|
| 1 | $\Delta\beta = 15°$ | $\Delta\beta = 15°$ | $\Delta\beta = 15°$ | $\Delta\beta = 15°$ |
| 2 | $\Delta\varepsilon = 108°$ | $\Delta\delta t = 6\,\mathrm{m}$ | $\Delta\delta t = -2.1\,\mathrm{m}$ | $\Delta\delta t = 6\,\mathrm{m}$ |
| 3 | $\Delta\delta t = 6\,\mathrm{m}$ | $\Delta\varepsilon = 108°$ | $\Delta\varepsilon = 108°$ | $\Delta\varepsilon = -252°$ |
| 4 | $\Delta\alpha = -9.3°$ | $\Delta\alpha = -9.3°$ | $\Delta\delta t = 8.1\,\mathrm{m}$ | $\Delta\alpha = -9.3°$ |
| 5 | - | - | $\Delta\alpha = -9.3°$ | - |

the location parameters were accepted according to Fig. 3. The cargo located in position $A$ was to be transported and positioned in location $B$ (Fig. 5A). Proposed methods using computational intelligence for analysis of the crane handling system and process to support handling assignment presents (Fig. 5B).

For simulation purposes, the following assumptions were accepted in simulation testing:

- the crane is positioned on a stable horizontal base (inclination up to 1%);
- the rotation angle of the crane column was measured from the plane $Ozy$ and its range $\varepsilon = 0 \div 360°$ was determined; the lifting angle of the outer boom, measured in vertically, changed in the range of $\alpha = 34.13° \div 185.16°$; the value of the angle of rotation of the inner arm was $\beta = 9.9° \div 182.5°$,
- in the simulation testing, the following propeller speeds were accepted: $\dot{\varepsilon} = 18$ deg/s, $\dot{\alpha} = 2.5$ deg/s, $\dot{\beta} = 5$ deg/s, $\dot{\delta t} = 0{,}3\,\mathrm{m/s}$.
- the value of the extension of the telescopic boom was in the range of $\delta t = 0 \div 11.85$ m,
- the value of the safety indicator was $k_1 = 0.25$. This means that in the most favourable position of the centre of the mass $Wu(x_{Wu}, z_{Wu})$ of the crane system, in a projection on the horizontal plane, all the vertical reactions in the outriggers are identical and they constitute 25% of the total load $Gu$.
- the value of the safety indicator $k_2 = 0.05$,
- it was accepted in the simulation testing that the crane is not subject to the wind pressure force (the wind speed is smaller than $v_w < 8.3$ m and it is neglected),
- the working movements of the crane are smoothly controlled, hence it was accepted that inertia forces can be neglected,
- the values of the masses of the crane elements and the locations of their gravity centres were taken into account in modelling,
- the mass of the cargo carried is $m_l = 560\,\mathrm{kg}$,
- in the initial position, the configuration coordinates of the individual working mechanisms are as follows: $\varepsilon_p = 175°$, $\alpha_p = 112°$, $\beta_p = 139.8°$, $\delta t_p = 2.68$ m.

The integrated CAD/CAE system with an additional computational application was used in simulation testing, which permitted the following among others:

- an accurate determination of the coordinates of any point of the crane system based on the mathematical model that describes its configuration [3,4],

- establishing the trajectory of the gravity centre of the crane $Wu(x_{Wu}, z_{Wu})$,
- calculation of the reaction in the outriggers $Ry_1$, $Ry_2$, $Ry_3$, $Ry_4 = f\{G_l$, $Wu$ $(x_{Wu}$, $z_{Wu})$, $t\}$,
- calculation of the difference of the torques $\Delta M = Mu - Mw = f\{G_l$, $Wu$ $(x_{Wu}$, $z_{Wu})$, $t\}$,
- calculation of the safety indicator $Wb = f\{G_l, Wu(x_{Wu}, z_{Wu}), t\}$,
- determination of the values of the working loads and the crane lift curves,
- determination of the crane stability conditions in the function of its working load and extension,
- evaluation of the stability of the performance of the entire crane handling cycle,
- optimization of the movement trajectory of the crane working elements for the accepted optimization criterion $min\ Wb$ or $min\ L_l$.

# 7 Simulation Research Results

The trajectories presented in Fig. 6A that are determined by the gravity centres $Wu(x_{Wu}, z_{Wu})$ of the crane system are located inside the tip-over outline $S_1S_2S_3S_4$. This is confirmed by the courses of the formation of the value of the

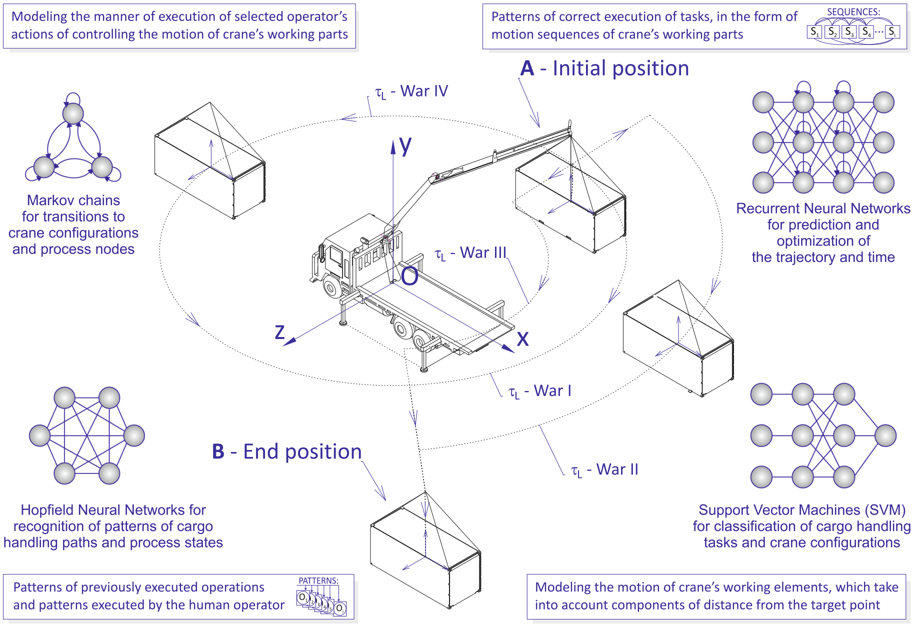

Modeling the manner of execution of selected operator's actions of controlling the motion of crane's working parts
Patterns of correct execution of tasks, in the form of motion sequences of crane's working parts
SEQUENCES:
S1 S2 S3 S4 ... Si
A - Initial position
τL - War IV
y
Markov chains for transitions to crane configurations and process nodes
Recurrent Neural Networks for prediction and optimization of the trajectory and time
τL - War III
O
z
x
τL - War I
B - End position
τL - War II
Hopfield Neural Networks for recognition of patterns of cargo handling paths and process states
Support Vector Machines (SVM) for classification of cargo handling tasks and crane configurations
Patterns of previously executed operations and patterns executed by the human operator
PATTERNS:
O1 ... Oi
Modeling the motion of crane's working elements, which take into account components of distance from the target point

**Fig. 5.** Computational intelligence methods for analysis of the crane handling system and process to support handling assignment consisting in carrying the cargo from its initial position $A$ to position in point $B$, for four displacement variants.

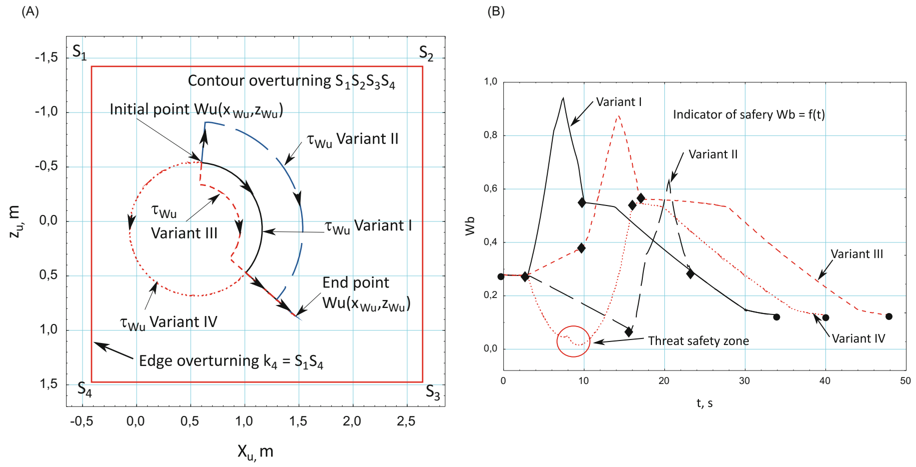


**Fig. 6.** (A) Projection of the mass centre's $\tau_{Wu} = Wu(x_{Wu}, z_{Wu})$ carrying trajectory of the crane onto the $Oxz$ horizontal plane for four handling assignment variants; (B) Courses of the value of the safety indicator $Wb$ for the four variants of the handling assignment, where: ● - start and end of movement, ◊ - start and end of circular motion.

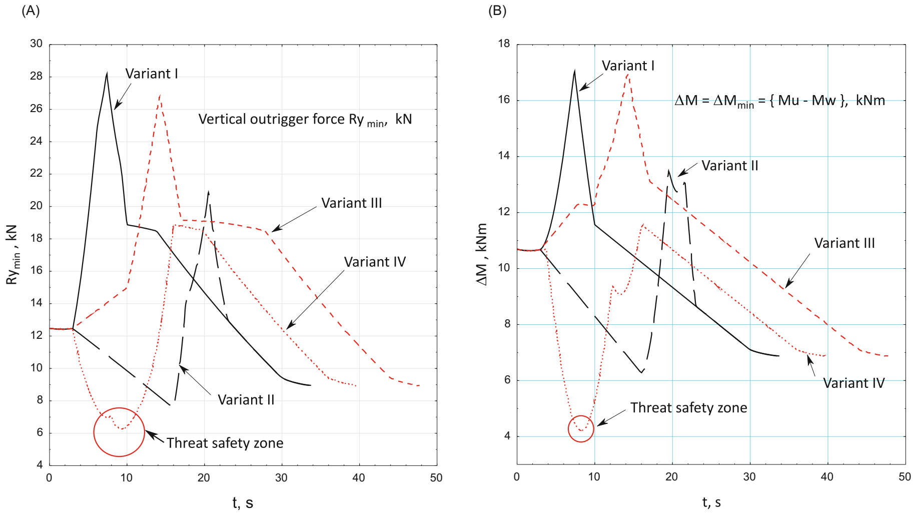


**Fig. 7.** Changes to the value of the base vertical reaction forces $min(Ry_i)$: (A) and values $\Delta M_{min}$, (B) during displacement of the cargo for the four variants of handling assignment.

safety indicator $Wb$ that are presented in Fig. 6B. It is evident for the handling assignment example presented that the minimum value of the safety indicator for all of the three cases is greater than 0; hence, the crane system is stable over the whole range. For the fourth variant of the handling assignment, however, the value of this indicator $Wb = 0.002$ is very small. This means that for the trajectory of the load carried $\tau_H$ the working conditions are the least favourable as there is a risk of a loss of the crane system stability. This is confirmed with the diagrams from Figs. 7 and 8.

By analysing the courses presented it can be found that in spite of ensuring the crane's static stability, there may occur a risk to its operation (Fig. 7: variant IV). In the time interval between the 9th and 11th second, the values of the horizontal reaction force $Ry_{min}$ (Fig. 7A) and the torque differences $\varDelta M$ (Fig. 7B) are the lowest. The gravity centre $Wu$ (Fig. 6B) is located too close o the tip-over axis $S_1S_4$: as little as in the distance of $d = 0.34$ m. This is also confirmed by the diagrams from Fig. 8 for the fourth case of the handling assignment. The values of the vertical reactions $Ry_2$ and $Ry_3$ and the torque difference $\varDelta M$ are the lowest for this variant.

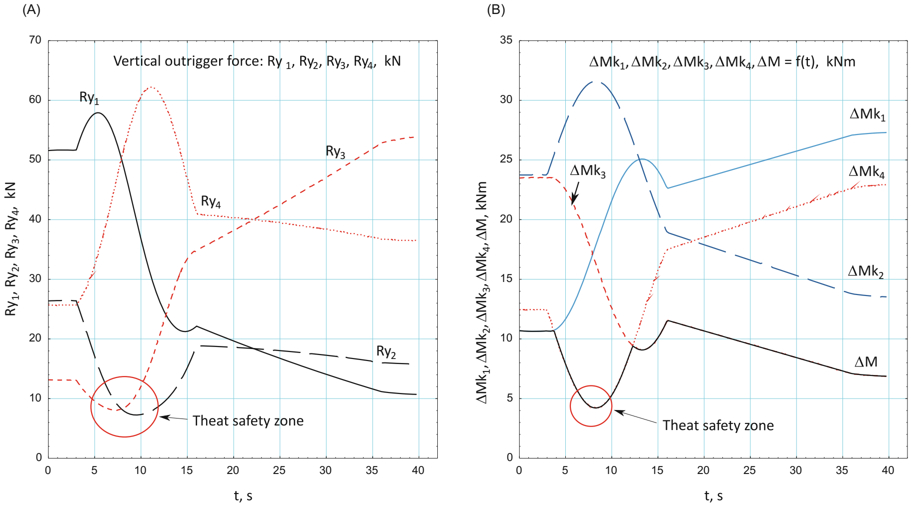


**Fig. 8.** Courses of changes to the values of the vertical reactions of the base $Ry_i (i = 1, .., 4)$ for all the outriggers (A) and the difference of torques $\varDelta M_k (k = 1, 2, .., 4)$ (B) in relation to all the tip-over axes of the crane outrigger system for the four variant of the handling cargo.

## 8 Conclusion

In this article a crane simulation model allowing to analyze the crane's stability during sequential movements, that is rotation of its column, rotation of its inner and outer arms as well as the extension pistons of its six-member telescopic boom. The model designed with a CAD/CAE environment allows to determine: crane's variable configuration setups in the Cartesian space, positions of mass centres of the crane system, reactions and moments interacting with the outrigger system, as well as values of the safety indicator. The obtained results of numerical simulations, which meet the stability criteria, allow to determine the optimal trajectory of carrying the cargo for a given handling assignment. Implementation of trajectory correction of crane's moving components may prevent the outriggers from losing contact with the ground, which provides safe operation in all conditions.

**Acknowledgements.** This project is financed by the National Centre for Research and Development, Poland (NCBiR), under the Applied Research Programme - Grant agreement No. PBS3/A6/28/2015.

## References

1. Majewski, M., Kacalak, W.: Innovative intelligent interaction systems of loader cranes and their human operators. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Prokopova, Z., Silhavy, P. (eds.) Artificial Intelligence Trends in Intelligent Systems, CSOC 2017. AISC, vol. 573, pp. 474–485. Springer, Switzerland (2017)
2. Majewski, M., Kacalak, W.: Smart control of lifting devices using patterns and antipatterns. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Prokopova, Z., Silhavy, P. (eds.) Artificial Intelligence Trends in Intelligent Systems, CSOC 2017. AISC, vol. 573, pp. 486–493. Springer, Switzerland (2017)
3. Kacalak, W., Budniak, Z., Majewski, M.: Crane stability for various load conditions and trajectories of load translocation. Mechanik **2016**(12), 1820–1823 (2016)
4. Kacalak, W., Budniak, Z., Majewski, M.: Simulation model of a mobile crane with ensuring its stability. Model. Eng. **29**(60), 35–43 (2016). PTMTS
5. Kohonen, T.: Self-Organization and Associative Memory. Springer, Heidelberg (1984)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
7. Specht, D.F.: Probabilistic neural networks. Neural Netw. **3**(1), 109–118 (1990). Elsevier

# The Study of the Impact of Technological Innovation Network on Dual Innovation

Cheng Song[1], Longying Hu[1], and Haiyan Yuan[2](✉)

[1] School of Management, Harbin Institute of Technology, Harbin 150001, China
[2] Department of Mathematics, Heilongjiang Institute of Technology,
Harbin 150050, China
yhysc82_47@163.com

**Abstract.** According to the adaptive behavior of organizational innovation, this paper explores the impact of organizational inertia on incremental innovation and radical innovation from the dimensions of structural inertia and cognitive inertia by analyzing the root of technological innovation network inertia, and taking the network embeddedness as adjustment variable. The research is carried out with the industry with high R & D intensity as the object, and the multiple regression model is used to do the empirical test, it is proved that organizational inertia has significant positive effect on incremental innovation, adaptive behavior helps to reveal the organization of technological innovation network, to enhance organizational innovation ability, and has important significance to maintain the network stability.

**Keywords:** Technological innovation · Organizational network · Organizational inertia

## 1 Introduction

In the fast changing market environment, the process of technological innovation becomes a kind of network process [1]. The multi-agent collaborative innovation network can not only share the information and complementary technology, but also share the wind and improve the operational efficiency [2]. However, in the turbulence of the technical and market environment, innovation network presents instability and high failure rate, this constraints the organizational innovation capability [3]. It is mainly because the organization is lack of effect adaptive behavior to the environment and ignores its effect in the process of promoting innovation network. In the present, many studies the organizational behavior from the two mutually exclusive perspective of the “rational” and “irrational”, under the “environment-behavior” model, they think that the organizational behavior will change with the environment and appear “delay”, “slow”, then promote or inhibit the innovation performance [4]. However, these two perspectives provide an incomplete view of the network organizational behavior and the innovation, and ignore the promote and hinder effect of the inertia characteristics of adaptive behavior to innovation. So, it is necessary to study the innovation efficiency of the technology innovation network organizational inertia.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_24

The organizational inertia which has typical adaptive behavior characteristics is generally recognized as the direct cause of Organizational failure which is similar to the organizational rigidities [5], this could leads to the "self closing" behavior and then inhibit innovation performance [6]. However, with the development, more and morescholars believe that organizational inertia is the cognitive and behavioral orientation of the organization, which is an adaptive outcome of flexible organizational behavior, and it could be strengthen as the accumulation of environment and time, the strong organizational inertia is conducive to the organization's cooperative innovation [7]. However, this verification ignores the internal structure of organizational inertia and its different influence on different types of innovation.

For one thing, the organizational inertia has many forms which including structural inertia, cognitive inertia and knowledge inertia, the different types of organizational inertia have the corresponding rational or irrational act, they can directly impact the the local or remote search methods, and then influence the innovation performance. For the other thing, there are incremental innovation and breakthrough innovation, there are differences in needs of the exploration learning and the use of learning, it needs a basis to adapt to the learning methods. The two aspects show that the dual innovation needs to match the organizational inertia so as to play its effectiveness. At the same time, as the organization is embedded in the multi-agent cooperation innovation network, the embeddedness acts on the organization inertia directly or indirectly, and promotes the differentiated innovation strategy choice. Therefore, it is necessary to explore the impact of organizational inertia on dual innovation from the perspective of network embeddedness.

This paper combine the organizational inertia, the dual innovation and the network embeddedness into the same framework, study the source performance of organizational inertia in technological innovation network, reveal the influence of different organizational inertia on breakthrough and incremental innovation. It could also explore the regulating effect of the network embeddedness on the dual innovation. So, it has the theoretical and practical significance in revealing the adaptive behavior characteristics of the technological innovation network organization and guiding the smooth operation of the network.

## 2 Preparatory Knowledge

### 2.1 Theoretical Basis

1. The organizational inertia

The organizational inertia and the organizational flexibility are the opposite concepts, the organizational flexibility is the performance of the rational allocation and organic integration of the organizational resources. On the contrary, the organizational inertia is the performance of maintaining the original state, curing and internal inertia, it has the negative impact [8]. But with the in-depth study, scholars find that in the multi-agent cooperation innovation network, organizational inertia of the network members present as they can strengthen the cognition of environment, adapt to the changes of external environment and form a flexible organization.

The structural inertia of the network members performance present as the organization is vulnerable to other members, produce endogenous dependence, has maintained stable state and is not sensitive to the environment change, and these characteristics will strengthen with the increasing of the diversification [9]. The Structural inertia includes organizational structure, organizational practices, culture and institutions, etc. Hou et al. [10] believed that in the stable environment, the higher the structural inertia and the stronger the homogeneity of network organization. Although the adaptation ability of the organizational dynamic is poor, it can maintain the organizational gene and avoid the occurrence of mutation. The cognitive inertia is the strategic characteristics of the network members, it is the organization of cognitive activity practice, it depends on the historical experience, the memory and the learning knowledge in the process of cooperation from other members and guide the organizational activity practice in the new environment, this feature will strengthen with the increase of knowledge diversity.

The cognitive inertia includes organizational experience, environmental awareness and imprinting. Argot et al. [11] explain the cognitive inertia from the perspective of organizational experience, they believed that the process of changing the experience into knowledge is the main content of organizational learning, one can improve his learning efficiency with the successful experience. We believe that the structural inertia is the basis of the evolution of the organizational structure of the network, it need continuity of practice to maintain. The cognitive inertia is the motivation of innovation network organization evolution, it is the strategic basis for the organizational change. Therefore, this study will explore the mechanism of organizational inertia.

2. The dual innovation

The study of organizational duality focuses on the stability and continuity of the organization, it divide the dual innovation into breakthrough innovation and incremental innovation according to the novelty of the innovation [12].

The breakthrough innovation will enters a new technical field, by through the exploratory learning, it will subvert the existing technological knowledge, and reflect the forward market orientation, it could guarantee the long-term strategic development of organization [13]. The incremental innovation will enter the existing technology fields, by through the use of learning, it will improve the existing technology knowledge, and reflect the existing market orientation, it could guarantee the organization's robust development [14].

The dual innovation of technological innovation network is different from that of non network members, the dual innovation of non network members originates in the organizational development strategy. The dependence of the dual innovation of network members on the structure of the whole network, leads to the paradox of in compact networks and sparse networks [15], the compact networks which strengthened by the organizational relationship and the knowledge, will promote the incremental innovation, but, the excessive embedding and homogeneity will inhibit the breakthrough innovation; because of the diversity and openness of the sparse network, it can promote the breakthrough innovation, but the lack of trust and deep exchange will inhibit the incremental innovation. It can be seen that the formation of network structure depends on the dynamic change of organization cooperation behavior. For the network members, the two dimensions of innovation is closely related to the adaptive behavior among network members, which is based on the difference between the breakthrough and incremental innovation, in this paper, we think that the organization will take different inertia behavior, for one thing, it can promote the differentiation of the network structure, for the other thing, it can realize the innovation strategy which will adapt to the organization's development. Therefore, it is necessary to further clarify the influence of the network members' inertia on the dual innovation.

3. The network embeddedness

The cooperative innovation organization is embedded in the complex network which has multiple relationships, the network embeddedness characterizes the relationships between the organizational network location and the cooperation, determines the amount of resources allocation and integration, and affects the decision-making behavior of the organization [9]. In the present study, the network embeddedness is divided into relational embeddedness and structural embeddedness [16, 17].

The relational embeddedness which takes Granovetter as its scholar represent emphasizes the social bonding relationship between organizations, it believes that the strong and weak links between organizations can promote cooperation and promote the organization to obtain heterogeneous knowledge. The relationship embeddedness can not only maintain the stability in cooperative relations, influence the efficiency of the knowledge transfer enhance the reputation of the organization, but also can inhibit opportunism, coordinate the conflicts between groups and promote the norms of reciprocity and cooperation consensus [7]. The structural embeddedness which takes Burt as its scholar represent, emphasizes the structure of the organization network and the location characteristics of the organization in the network, it believes that the network position of the organization has a direct factor affecting on innovation performance. On the one hand, the organization who occupies the center location is the distributed knowledge and control node, the edge organization is willing to establish cooperative relations with them, this will force the network structure core/edge formation, and result in the differentiation of "strong stronger and weak weaker". On the other hand, the advantage of this kind of structure can make the organization contact the heterogeneous resources in the network, make the resource internalization, combination, externalization and socialization, and then use the knowledge to regulate and coordinate the relationship between organizations.

Based on the above analysis, the innovative organization as a member of the network inevitably is affected by external effects of complex network embedded in the process of organizational inertia effect on the dual innovation, so we need further explore the role of network embeddedness.

## 2.2 Some Assumptions

**Hypothesis H1:** Compared with the cognitive inertia, the structural inertia has the stronger positive effect on the incremental innovation.

The incremental innovation needs to expand and integrate the existing capacity and technology development trajectory, which is based on the stable and appropriate knowledge reservation. Whether the structural inertia or cognitive inertia and gradual progress: Firstly, all of the two types of organizational inertia will adjust the learning behavior to match organization and environment, ensure the organization's resource base with the learning ability, obtain some knowledge and technology and promote the integration and improvement of the organization. These conform to the knowledge requirements of incremental innovation [1]. Secondly, in the process of development, the organization has accumulated many successful experiences, and has stronger ability to identify the opportunities. It can guide the organization's decision making by updating the experience database. Thirdly, organizational inertia improves the stability of the cooperative relationship, ensures the organization to carry out the use of learning, support the implementation of incremental innovation.

Although both the structural inertia and the cognitive inertia can promote the incremental innovation, but due to the differences of the root causes, the impact on incremental innovation are different: Firstly, compared with the cognitive inertia, the structural inertia depends on the stable state of the organizational structure too much, which can promote the application ability, but limit the generation of the organization's exploratory learning behavior [10]. Secondly, the existence of cognitive inertia, means that the organization will consolidate the previous knowledge reserves at the same time, carry out a modest exploratory learning to expand the knowledge flow, update the appropriate cutting-edge knowledge innovation in the meantime. Thirdly, because of the path dependence characteristics of the structural inertia and the cognitive inertia of the network organization, and the path dependence intensity of the structural inertia is larger than that of the cognitive inertia, therefore, the uncertainty risk of organizational learning is reduced.

Above all, the organizational inertia can help promote the sustainable development of the network organization. The structural inertia and the cognitive inertia enhance the knowledge' utilization efficiency, and structural inertia can inhibit more uncertainty in innovation than the cognitive inertia, and thus enhance the incremental innovation then promote incremental innovation.

**Hypothesis H2:** Compared with the structural inertia, the cognitive inertia impacts stronger U on the breakthrough innovation.

The breakthrough innovation needs to subvert and destroy the existing ability and technology development track to carry out exploratory learning, which is based on a variety of frontier knowledge [2]. Moderate organizational inertia will support the breakthrough innovation in an effective way of learning, but it may produce the opposite inhibitory effect when it exceeds a certain threshold value. Whether the structural inertia or cognitive inertia and breakthrough innovation: Firstly, the excessive organizational inertia can lead to over dependence on the path, the organization is too conservative to resist external environmental changes, increase the probability of conflict between internal and external knowledge, and it is affected by the distance between organizations, so it is likely to produce the relationship slicle and reduce the possibility of breakthrough innovation [4]. Secondly, the organizational relationships maintenance costs increase. The excessive inertia reduces the organization flexibility, produce rigid "isolation" mode, limit the incremental range, result in the rejection of outside study, and then inhibit the organization's ability to explore and use. Thirdly, the excessive organizational inertia causes a limited "rational adaptation" behavior, curbs the adaptive change of organizational behavior, causes the emergence of innovation delay.

The moderate organizational inertia can promote breakthrough innovation, but because of the internal differences, the impact on the breakthrough innovation are different: Firstly, compared with the structural inertia, the cognitive inertia emphasizes to repeat the previous success strategic experience, not only the accumulation of existing knowledge, but also the integration of external knowledge, which will cause stronger exploratory learning ability. Secondly, the structural inertia has a higher effect on the neighborhood of organizational innovation activities than the cognitive inertia, which leads to the improvement of the knowledge of the organizational network, and limits the breakthrough innovation to some extent. Thirdly, the effect of cognitive inertia on the knowledge expansion is higher than that of structural inertia, and the accumulation of knowledge is becoming more and more obvious, the impact on the breakthrough innovation increases gradually.

Above all, moderate organizational inertia helps to promote the sustainable development of network organization, on the one hand, the organizational inertia needs to contact the appropriate diversified knowledge so as to carry out breakthrough innovation, on the other hand, the impact of the cognitive inertia on the exploratory learning effect is enhanced.

**Hypothesis H3a:** The influence of positive regulation on organizational inertia of the relationship embeddedness on on dual innovation

**Hypothesis H3b:** The influence of positive regulation on organizational inertia of the structural embeddedness on on dual innovation

The positive effects of cooperative innovation between organizations is rooted in knowledge sharing and learning, it has a strong dependency on the organizational network structure, so in the process of acquiring sustainable resources the organizational inertia the is influenced by the embedded external network [18].

Under the effect of the relationship embeddedness, firstly, the higher relationship embeddedness provide knowledge transfer channels for the heterogeneous resources, promote the internalization, combination, externalization and socialization of the organizational knowledge, and improve the efficiency of resource exchange. Secondly, the higher relational embeddedness is an important component of the network governance mechanism, which can guide the cooperation behavior of organizational inertia, strengthen inter organizational trust, and reduce the probability of opportunistic behavior. Thirdly, as the cooperative relationship becomes stable, the higher relational embeddedness can reduce transaction costs and risks, and ensure the smooth operation of the organization's inertia orientation, reduce the environmental uncertainty in routine search, improve the organization's ability to use and explore, and promote the organization's dual innovation performance.

Under the effect of structural embeddedness, firstly, the higher structure embedding indicates that the organization is in the core position of the network, and its knowledge power is high, and it has the ability to coordinate the knowledge transfer and knowledge spillover among organizations. So, under the inertia action, its self strengthening becomes more and more obvious, and the cooperative relationship between the organizations tends to be stable, the ability of exploration and learning becomes stronger. Secondly, the diversity of the structural embeddedness will enhance the organization's knowledge reserves, promote the scope and the efficiency of transmission spread, then, guarantee the continued impact of inertia behavior. Thirdly, the diversity of the structural embeddedness means that network organization may occupy the structural holes, have non redundant knowledge resources. This competitive advantage is not only the basis of the structural inertia to maintain the stability of the organization, but also the external source of successful experience, learning style and environmental perception.

In brief, the higher the network structure embeddedness and the relational embeddedness, the stronger the adaptability of the organizational innovation network. It will not only guarantee the promotion of the structural inertia on the continuity and stability, but also reduce the internal knowledge "lock" effect, maintain cognitive inertia's ability of adapting to the environment, and enhance the dual innovation performance.

# 3 The Study Method

## 3.1 The Collection of the Data and Samples

The first source of data was surveyed in 4 industries from computer, pharmaceuticals, communication and automobile body and parts manufacture. The survey were carried out in four steps. Firstly, select the 4 leading enterprises in the industries (Lenovo, Sunflower pharmaceutical, Huawei and Volvo), by way of snowball sampling, the enterprises are required to fill major cooperative innovation partner enterprises recently in the questionnaire; Secondly, according to the questionnaires, send the partner

**Table 1.** Sample distribution

| | The industries | The quantity | The proportion |
|---|---|---|---|
| Industry characteristics | The computer | 106 | 35.81 |
| | The pharmaceuticals | 74 | 25.00 |
| | The mobile communication | 78 | 26.35 |
| | The parts manufacturing | 38 | 12.83 |
| Enterprise type | Chinese | 152 | 51.35 |
| | Joint venture | 48 | 16.21 |
| | Foreign investment | 96 | 32.43 |
| Enterprise scale | <300 | 95 | 32.09 |
| | 200–500 | 163 | 55.06 |
| | >500 | 38 | 12.83 |
| Total assets | <1500 Millions | 152 | 51.35 |
| | 1500–5000 Millions | 107 | 36.15 |
| | >5000 Millions | 37 | 12.50 |
| Founding time | <5 years | 108 | 36.48 |
| | 5–10 years | 155 | 52.36 |
| | >10 years | 33 | 11.15 |

**Table 2.** Information statistics

| | The category | The quantity | The proportion |
|---|---|---|---|
| Gender | Male | 190 | 64.19 |
| | Female | 106 | 35.81 |
| Age | <25 | 75 | 25.34 |
| | 25–35 | 102 | 34.45 |
| | 35–50 | 84 | 28.38 |
| | >50 | 35 | 11.82 |
| Education | Junior college | 103 | 34.79 |
| | Bachelor | 126 | 42.56 |
| | Master | 48 | 16.21 |
| | Doctor | 19 | 6.41 |
| Post | Grassroots workers | 47 | 15.87 |
| | Artisan | 100 | 33.78 |
| | Middle managers | 125 | 42.22 |
| | Top managers | 24 | 8.11 |

companies new questionnaires; Thirdly, relying on the Harbin's Pingfang industrial park, Da Qing high tech industrial development zone, select four representative enterprises in the development zone of the four major industries, a total of 472 questionnaires were collected, and a total of 176 valid questionnaires were deleted, and the effective recovery rate was 296, and the effective recovery rate was 61.157%, and the sample distribution was shown in the Tables 1 and 2.

## 3.2 Measuring the Variables

In order to ensure the reliability and validity of the variables involved in the research, on the basis of combing the related literature at home and abroad, we select the measurement structure for the relevant variables, and transform the characteristics of the technological innovation network. Use the Likert 7 scale measurement, 1 denotes that does not comply, and so on, 7 the full compliance with.

(1) The organizational inertia

In Lu [4] and Karim [17], the measure of organizational inertia was studied, the structural inertia and cognitive inertia were measured respectively. The structure inertia is measured by 4 items: the companies in the field of innovation will not compete with the co-operative enterprise, the company will reject the new methods in the choice of the work content because of the habit, the companies prefer to transform existing products rather than design new ones, the company's production method is complex but the staff department link closely; The cognitive inertia is measured by 4 items: the attention of companies paid influences the environmental dynamic changes in innovation, companies change their the strategy faster than that of the partners as the turn of market, the companies' decision making is based on historical experience and current system, the companies have a long research and development cycle.

(2) Dual innovation

In Cao [18] and Dang [19], the dual innovation measure both the breakthrough innovation and incremental innovation. The breakthrough innovation sets 5 items to measure: develops new products and new technology, introduces of new technology, eliminates the mature products and technologies, and introduces more new products in the new market areas, be in the leading position in the industry; The incremental innovation sets 4 items to measure: dedicates to the improvement of existing products and technologies, improves product production equipment and production process, dedicates to the application of current technology innovation, introduces more innovative products which fit in with the development needs of enterprises.

(3) Network embeddedness

In Zhang [20], the network embeddedness measures both the relational embeddedness and the structural embeddedness. The relational embeddedness sets 4 items to measure: the frequency of cooperation between partners, the scope of cooperation between the partners, the duration of the relationship between partners, the degree of understanding between partners; The structural Embeddedness sets 4 items to measure: the number of partners in the enterprise innovation, the enterprises are easy to find an ideal partner, most companies are willing to cooperate with us to innovate, enterprises occupy a dominant position in the key resources of innovation.

(4) Controlled variables

In this paper, we choose the size of the enterprise and the age of the enterprise as controlled variables, the larger the size of the enterprise, the higher the coordination costs between members, monitoring costs and management complexity, it will have a direct impact on enterprise innovation. The longer the establishment of the enterprise has, the more conducive it is to the establishment of cooperation and innovation.

# 4 Empirical Analysis and Results of the Study

## 4.1 Reliability and Validity Analysis

First of all, the measured variables in this paper are maturity scales derived from the domestic and foreign, and corrected according to characteristics of technological innovation network, guarantee the reliability and validity of the variables to a certain extent; secondly, we examine the reliability of the variable 'Cronbach' s alpha coefficient using SPSS 18 statistical software, we test validity using KMO value of the sample and the Bartlett sphere test. The results are shown in Table 3.

It can be seen in the results that the 'Cronbach's alpha ranged from 0.724 to 0.831, which are higher than the critical value of 0.700. The KMO values of each variable are greater than 0.665, and the Bartlett test are significant. The load of the measurement factor is greater than 0.615, which indicates that the item can reflect the relevant variables, show a good convergent validity.

The results show that the correlation coefficient of the 5 variables, namely, structural inertia, cognitive inertia, relational embeddedness, incremental innovation and

**Table 3.** The reliability and validity of variables

| Variables | Cronbach' s | KMO | Barlett result |
|---|---|---|---|
| Structural inertia | 0.797 | 0.667 | 0.000 |
| Cognitive | 0.811 | 0.711 | 0.000 |
| Relationship embeddedness | 0.724 | 0.694 | 0.000 |
| Structural embeddedness | 0.754 | 0.686 | 0.000 |
| Incremental innovation | 0.831 | 0.734 | 0.000 |
| Breakthrough innovation | 0.811 | 0.706 | 0.000 |

breakthrough innovation, are less than 0.700, the variance inflation factor (VIF) and tolerance are calculated, and the results showed that the VIF value is less than the upper limit value 10, and the tolerance is greater than the lower limit value 0.100, this means that there is no multicollinearity between the variables.

## 4.2 Data Processing

This paper uses hierarchical regression to test the hypothesis, and establishes 10 sub models, model 1 and 2 analyzes the effect of the control variables on the dual innovation model; 3 and 4 joins the argument structural and cognitive inertia; model 5 and 6 joins the regulation effect of relations embeddedness; model 7 and 8 joins the regulation effect of structural embeddedness; model 9 and 10 joins the interactive items all.

The regression results are shown in Table 4.

**Table 4.** The regression analysis results

| | Gradual innovation | | | | | Breakthrough innovation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1 | Model 3 | Model 5 | Model 7 | Model 9 | Model 2 | Model 4 | Model 6 | Model 8 | Model 10 |
| Control variables | | | | | | | | | | |
| Enterprise scale | 0.242 | 0.223 | 0.202 | 0.244 | 0.240 | 0.251 | 0.312 | 0.276 | 0.221 | 0.251 |
| Enterprise age | 0.102 | 0.114 | 0.117 | 0.122 | 0.127 | 0.131 | 0.105 | 0.136 | 0.121 | 0.111 |
| Independent variable | | | | | | | | | | |
| Structural inertia | | 0.427 | 0.392 | 0.413 | 0.442 | | 0.388 | 0.333 | 0.351 | 0.411 |
| Structural inertia square | | | | | | | −0.112 | −0.127 | −0.142 | −0.288 |
| Cognitive inertia | | 0.393 | 0.377 | 0.208 | 0.221 | | 0.41 | 0.476 | 0.454 | 0.433 |
| Cognitive inertia square | | | | | | | −0.213 | −0.253 | −0.231 | −0.271 |

(*continued*)

**Table 4.** (*continued*)

| | Gradual innovation | | | | Breakthrough innovation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1 | Model 3 | Model 5 | Model 7 | Model 9 | Model 2 | Model 4 | Model 6 | Model 8 | Model 10 |
| Interaction term | | | | | | | | | | |
| Relational embedding* Structural inertia | | | 0.421 | | 0.211 | | | 0.301 | | 0.333 |
| Relational embedding* Structural inertia square | | | | | | | | −0.142 | | −0.162 |
| Relational embedding* Cognitive Inertia | | | 0.333 | | 0.242 | | | 0.419 | | 0.430 |
| Relational embedding* Cognitive inertia square | | | | | | | | −0.256 | | −0.221 |
| Structure embedding* Structural inertia | | | | 0.333 | 0.288 | | | | 0.232 | 0.259 |
| Structure embedding* Structural inertia square | | | | | | | | | −0.141 | −0.132 |
| Structure embedding* Cognitive inertia | | | | 0.270 | 0.229 | | | | 0.209 | 0.241 |
| Structure embedding* Cognitive inertia square | | | | | | | | | −0.333 | −0.266 |
| R square | 0.070 | 0.281 | 0.292 | 0.312 | 0.400 | 0.091 | 0.302 | 0.403 | 0.419 | 0.592 |
| Adjusted R square | 0.065 | 0.271 | 0.283 | 0.292 | 0.366 | 0.085 | 0.292 | 0.391 | 0.401 | 0.544 |
| F | 3.632 | 3.764 | 3.829 | 4.288 | 5.679 | 5.310 | 5.635 | 7.622 | 7.001 | 7.723 |

## 4.3 Empirical Results and Discussion

In model 1 and model 2, the enterprise scale and age has significant positive effect on radical innovation and incremental innovation, the larger the enterprise is and the longer the enterprise establishes, the more abundant the knowledge resources are, the long-term cooperation experience can promote mutual understanding of the network members, improve the performance of the dual innovation.

Model 3 examines the effect of the variables' structural inertia and the cognitive inertia on the incremental innovation. The results of regression analysis in model 3 show that there is a significant positive relationship between the structural inertia, the cognitive inertia and the incremental innovation ($\beta = 0.427$, $\rho < 0.001$; $\beta = 0.393$, $\rho < 0.001$) the regression coefficient of structural inertia is larger than that of cognitive inertia, so the hypothesis H1 has passed the test, which is similar to the results in Men [6] and Sun [2]. Because of the path dependence characteristic of gradual innovation, it is necessary for enterprises to improve their knowledge structure. However, compared with the cognitive inertia, the path dependence of structural inertia is more obvious, and the enterprises are willing to maintain the intrinsic and stable behavior, therefore, the matching of structural inertia and incremental innovation has highly sensibility.

The model 4 examines the variables' influence of the structural inertia and the cognitive inertia and its squared term on the breakthrough innovation. The results of regression analysis in model 4 show that there is a significant inverted U relationship between the structural inertia, the cognitive inertia and its squared term ($\beta = 0.388$, $\rho < 0.001$; $\beta = -0.112$, $\rho < 0.050$; $\beta = 0.418$, $\rho < 0.001$; $\beta = -0.213$, $\rho < 0.010$). The change of direction confirms the threshold effect between the two inertia, and the regression coefficient of the cognitive inertia is larger than that of the structural inertia, so hypothesis H2 has passed the test, this is similar to the conclusions of Shipilov [9] and Lu [3], the breakthrough innovation focuses on the development of the process of organizational innovation. Because the cognitive inertia is based on the organization's successful experience and knowledge information, it will guide the organization carry out new knowledge exploration activities under the cognition and understanding in the past, so the matching of the cognitive inertia and the breakthrough innovation has a strong sensitivity.

The model 5 and model 6 test the moderating effect of the relational embeddedness The model 5 shows that there is a significant positive correlation between relational embeddedness and structural inertia, the cognitive inertia and the incremental innovation, the positive regulation of relational embeddedness is verified. The model 6 regulates significantly the inverted U relationship between the organizational inertia and the breakthrough innovation, and regresses model 9 and model 10 completely, so the hypothesis H3a passes the test, it is similar to the result in Polidoro and Ahuja [1]. On one hand, the stronger the strength of the relationships between enterprises are, and the higher confidence the enterprises have, the higher level of the enterprises' access heterogeneous knowledge resources, this provides a solid foundation for the enterprises to carry out dual innovation; on the other hand, the stable cooperative relationship can promote the efficiency of organizational inertia, reduce the risk of uncertainty in inter organizational cooperation, carry out local search and remote search better, then strengthen the positive role of organizational inertia.

Model 7 and model 8 test the moderating effect of structural embeddedness, the model 7 shows that there is a significant positive relationship between structural embeddedness and organizational inertia and incremental innovation, in model 8, the structural embeddedness adjustment and the structural inertia, the square of the structural inertia and the breakthrough innovation have a inverted U relationship. However there is no significant relationship between the interaction term of the structural embeddedness and the cognitive inertia and its square interaction terms and

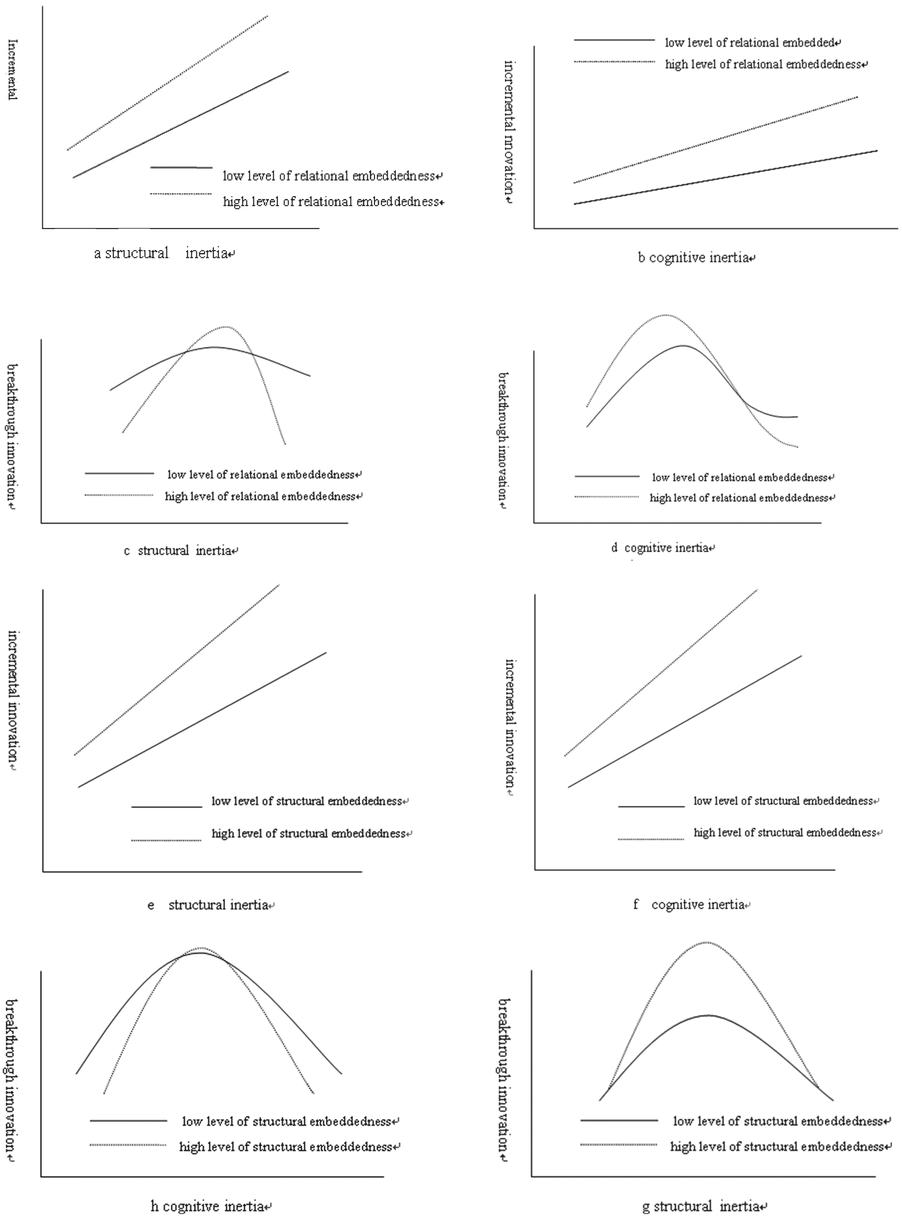

Incremental
low level of relational embeddedness↵
high level of relational embeddedness↵
a structural inertia↵
incremental innovation↵
low level of relational embedded↵
high level of relational embeddedness↵
b cognitive inertia↵
breakthrough innovation↵
low level of relational embeddedness↵
high level of relational embeddedness↵
c structural inertia↵
breakthrough innovation↵
low level of relational embeddedness↵
high level of relational embeddedness↵
d cognitive inertia↵
incremental innovation↵
low level of structural embeddedness↵
high level of structural embeddedness↵
e structural inertia↵
incremental innovation↵
low level of structural embeddedness↵
high level of structural embeddedness↵
f cognitive inertia↵
breakthrough innovation↵
low level of structural embeddedness↵
high level of structural embeddedness↵
h cognitive inertia↵
breakthrough innovation↵
low level of structural embeddedness↵
high level of structural embeddedness↵
g structural inertia↵

**Fig. 1.**

the breakthrough innovation, so the hypothesis H3b passes the test partly. The effect that the structural embeddedness moderates the cognitive inertia and the breakthrough innovation is less remarkable owing to: Firstly, the higher structural embeddedness gives the enterprises a competitive advantage in the network, improves the knowledge transfer and spillover efficiency, gives the enterprises the strong ability to identify opportunities, but it also inhibit the subjective initiative of the organization and the correction of cognitive inertia [10]. Secondly, The higher structural embeddedness will cause over dependence on the organization, which exceeds the carrying capacity of the organization, sinks the organization into a dilemma in the environment of breakthrough innovation, while the cognitive inertia could not play its role, so it limits the organization's exploratory innovation behavior, cannot promote the effect of the cognitive inertia on the breakthrough innovation effectively.

In order to understand regulation of network embeddedness between the organizational inertia and the dual innovation better, this paper selects the variables and the mean of moderating variables add or subtract a standard deviation, then takes them into the regression model, Fig. 1 shows the relationships between the variables interaction diagram. As shown in Fig. 1a–d, the relationship embeddedness plays a significant regulatory role between the organizational inertia and the dual innovation. Figure 1e–h shows that the structural adjustment supports partly, in which the moderating effect of structural embeddedness on the cognitive inertia and the breakthrough innovation is not significant.

## 5 The Conclusions

This paper studies the organizational inertia, distinguishes original difference between the structural inertia and the cognitive inertia in the technological innovation networks, explores the difference influence of organizational inertia on the dual innovation in the technological innovation network, and regulates the relationship embeddedness and structural embeddedness. Basing on the literature review and theoretical analysis, this paper puts forward the research hypothesis and makes the following conclusions on the base of the survey data: (1) The organizational inertia is conducive to the improvement of incremental innovation performance, and the structural inertia is stronger than that of the cognitive inertia. The organizational inertia can not only improve the efficiency of knowledge utilization, but also restrain the uncertainty of the innovation process, and then promote the gradual innovation, where the structural inertia can promote the cooperation between the homogeneous organizations, and enhance learning ability, then influence the progressive innovation stronger.

(2) Moderate organizational inertia can achieve the higher breakthrough innovation performance, and the cognitive inertia is more effective than the structural inertia; but, too much organizational inertia will cause the curing behavior, and inhibit the organization's cooperative innovation ability. Compared with the structural inertia, the cognitive inertia tends to further understand and process the past successful experience and knowledge, guide the implementation of the network innovation strategy, then influence the breakthrough innovation strongly.

(3) The relational embeddedness has a significant effect on the relationship between the organizational inertia and the dual innovation. The higher relational embeddedness can promote the organization obtain a variety of technical and knowledge resources, maintain a stable cooperative relationship, reduce the cost and risk of the organizational cooperation, it can provide the heterogeneous information for the efficiency of organization inertia as the result of the elastic organization behavior adaptability.

(4) The structural embeddedness has a significant effect on the structural inertia and the dual innovation, and partly supports the regulation of the cognitive inertia and the dual innovation. The higher the structural embeddedness is, the stronger the structural inertia is, the more stable the cooperative preference is, the stronger the organization's ability to explore and use, and thus to improve the performance of the dual innovation. Although the structure embeddedness has higher ability to use in the process of incremental innovation, but in the breakthrough innovation environment, it will lead the organization into dependency dilemma, limits the exploratory organizational innovation behavior.

The study, in this paper could help better understand the organizational behavior and environmental adaptability of the innovation network, the theoretical contributions include: (1) On the basis of existing research, we extend the organizational inertia to the technological innovation network, we study the organizational inertia and the technological innovation level of network innovation members, then enrich the research perspective of organizational inertia. (2) We overcome the stereotype of the original logic organization inertia and inert curing description, annotate the innovation influence of the flexible form and adaptive behavior on organizational inertia, enrich the theory of cooperative innovation network, and then give the extension and supplement for the enterprise innovation behavior. (3) This paper studies the organizational inertia, the dual innovation and the network embeddedness in the same framework, analyzes the influence of different types of organizational inertia on the incremental and the breakthrough innovation under different network embeddedness, which deepen the understanding of the technical innovation network's regulation, and enhance the credibility of the theory with the application of China's high-tech enterprises to verify the data, it has important theoretical and practical significance for enhancing the organization's innovation and network stability.

From the perspective of concrete practice, we obtain the following management enlightenment: (1) The organizational inertia as the internal strength for maintaining the network organization, is the key point in coordinating the network movement. From the perspective of the global network, the organizational inertia helps to improve the robustness of the innovation network and reduce the network vulnerability caused by the innovation risk; From the perspective of the relationships, it is helpful to regulate the cooperative behavior and preference of the organization; From the perspective of organization, it is an important path for the development of organizational innovation to improve the embeddedness of organization and reduce the cost of cooperation. (2) The innovation network organization needs to choose a different innovation mode to fit its own inertia characteristic. The structural inertia and the cognitive inertia need the organizational inertia which matches their strong situation dependence to maximize the performance of the organization, then, to improve the organization flexibility and promote the cooperation innovation performance. (3) The innovative network organizations should play its subjective initiative actively in using the network resources, for one thing, it could improve the interaction frequency, form a moderate interactions, further obtain heterogeneous resources; for the other thing, it could help the organization occupy the control position in the network, maintain competitive advantage, and improve the diversification of the organization.

There are some limitations in the study. First of all, we study the effect of organizational inertia on the dual innovation, select the network embeddedness as a moderating variable without considering the microscopic process of topology and network cooperation inertia effect; Secondly, the samples are obtained from the industries which have a strong practical technology, there are limitations in the sample differences. Therefore, in the future research, it is necessary to select the patent database as a data source, and build a technological innovation network, then characterize the corresponding network topology. Moreover, to increase the number of samples and their differences and to analyze the role of organizational inertia in network governance is an important issue to study in the future also needs.

**Acknowledgements.** This work was supported by the Creative Talent Project Foundation of Heilongjiang Province Education Department (UNPYSCT-2015102).

**Declare** The authors declare that there is no conflict of interests regarding the publication of this article.

## References

1. Sun, Y., Chen, J., Song, J.: The study of the dual inertia in the cultural differences. Sci. Sci. **33**(9), 1 (2015)
2. Majchrzak, A., Jarvenpaa, S.L., Bagherzadeh, M.: A review of inter organizational collaboration dynamics. J. Manag. **41**(5), 1338–1360 (2015)
3. Schilling, M.A.: Technology shocks technological collaboration and innovation outcomes. Organ. Sci. **26**(3), 668–686 (2015)
4. Lu, Y., Cheng, L., Su, J.: The impact of organizational inertia on the evolution of cluster network-simulation analysis based on multi agent modeling. J. Manag. Sci. **18**(6), 30–40 (2015)
5. Mckinley, W., Latham, S., Braun, M.: Organizational decline and innovation: turnarounds and downward spirals. Acad. Manag. Rev. **39**(1), 88–110 (2014)
6. Le Mens, G., Hannan, M.T., Pólos, L.: Age-related structural inertia: a distance-based approach. Organ. Sci. **26**(3), 756–773 (2015)
7. Bakker, R.M., Knoben, J.: Built to last or meant to end: inter temporal choice in strategic alliance portfolios. Organ. Sci. **26**(1), 256–276 (2014)
8. Meyer, K.E., Mudambi, R., Narula, R.: Multinational enterprises and local contexts: the opportunities and challenges of multiple embeddedness. J. Manag. Stud. **48**(2), 235–252 (2011)
9. Shipilov, A., Gulati, R., Kilduff, M.: Relational pluralism within and between organizations. Acad. Manag. J. **57**(2), 449–459 (2014)
10. Argote, L., Miron-Spektor, E.: Organizational learning: from experience to knowledge. Organ. Sci. **22**(5), 1123–1137 (2011)

11. Hou, J., Lu, Q., Shi, Y.: The evolution of firm growth based on organizational ecology: a case study of variation and survival. World Manag. **12**, 116–130 (2012)
12. Xue, J.: The impact of regional innovation environment on the innovation of micro enterprises-based on the mediating role of dual learning. Sci. Res. **33**(5), 782–791 (2015)
13. Lavie, D., Kang, J., Rosenkopf, L.: Balance within and across domains: the performance implications of exploration and exploitation in alliances. Organ. Sci. **22**(6), 1517–1538 (2011)
14. O'reilly, C.A., Tushman, M.L.: Organizational ambidexterity: past, present, and future. Acad. Manag. Perspect. **27**(4), 324–338 (2013)
15. Ghosh, A., Rosenkopf, L.: Shrouded in structure: challenges and opportunities for a friction-based view of network research. Organ. Sci. **26**(2), 622–631 (2014)
16. Sa Vinhas, A., Heide, J.B., Jap, S.D.: Consistency judgments, embeddedness, and relationship outcomes in inter organizational networks. Manag. Sci. **58**(5), 996–1011 (2012)
17. Karim, S., Kaul, A.: Structural recombination and innovation: unlocking intra organizational knowledge synergy through structural change. Organ. Sci. **26**(2), 439–455 (2014)

# Concept of Econometric Intelligence System: OLAP Applications in the Ambient Intelligence Environment

Jan Tyrychtr[1](✉), Martin Pelikán[2], Hana Štiková[2], and Ivan Vrana[2]

[1] Department of Information Technologies, Faculty of Economics and Management,
Czech University of Life Sciences in Prague, Prague, Czech Republic
tyrychtr@pef.czu.cz

[2] Department of Information Engineering, Faculty of Economics and Management,
Czech University of Life Sciences in Prague, Prague, Czech Republic

**Abstract.** Econometric analysis is a non-trivial discipline applied to different areas of an enterprise or economy, in order to express economic reality and anticipate economic phenomena. This requires a great deal of econometric knowledge using a number of sophisticated methods and their good capabilities for correct and high quality interpretation of results. Currently, the intelligent system is a solution that is capable of performing highly complex tasks in the same way as people approach these tasks. In the context of ambient intelligence, it is possible to use personalized, contextual awareness and adaptive attributes for the design of an intelligent econometric system. In our work, we focused on the concept of an intelligent econometric system together with the application of OLAP technology for the creation of interactive analytical outputs. This new concept of the system is presented in an example of a data analysis to derive the forecast from the econometric model by designing a multidimensional view of the data.

**Keywords:** Econometric system · Intelligence system · OLAP · Ambient intelligence · Multidimensional data model · Decision support

## 1 Introduction

Over the past period a number of publications dealing with econometric modelling [1] were published but they were mainly focused on the design and the use of models and only marginally on the issues of the development of econometric information systems. The current robust approaches in econometrics require extensive knowledge of econometrists in both economic and mathematical disciplines but also in statistical methods. This knowledgeable and expert personnel is then faced with a number of problems in data processing and data cleaning, the creation of econometric models and their interpretation. The econometric interpretation is a crucial issue in implementing these sophisticated solutions. Currently, there are only few approaches to the design of the econometric systems. The first study [2] was about the creation of an active decision support system for econometric analysis, including the design of the PERM (Progressive Econometric Modelling). In another study [3] the power of expert systems was presented and an examination of how the problem of an expert model could be used in the

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_25

formulation of econometric models. In the recent years several other studies have emerged such as a web-based decision support system for macro-econometric models [4] or a design of automated predictive analytics as a service [5].

Expected benefits in the form of automation or the intelligent behaviour of such systems have not yet been fully implemented. With the development of Ambient Intelligence (AmI), personalised, contextual awareness and adaptive attributes can be used for such expert econometric systems which can be derived from environmental observations (learning and recognition) including acquisition of sensor data and processing of large quantities of econometric and contextual data.

### 1.1 The Ambient Intelligence

At present, the AmI approaches, as a multidisciplinary technological paradigm, are the fastest growing area of intelligent systems [6–10]. AmI is a set of processes, applications, and technologies designed to efficiently and effectively support human needs in many areas (home, business, hospitals, automotive, etc.). The goal of AmI is to create an adaptive, friendly, personalised digital environment that understands, learns, and interacts with the user's needs [11].

As part of such efforts to develop intelligent systems, it is necessary to identify the components of such an econometric system and to design its general architecture. In our understanding of the econometric intelligent system, we consider as crucial both the AmI approach [12] as well as the approaches enabling the online analytical processing of data – the so-called OLAP technology.

### 1.2 The OLAP

OLAP describes a decision support approach that aims to obtain information from a data warehouse or data marketplace [13]. OLAP lets you aggregate and summarise data according to different points of view (dimensions). OLAP acquires aggregated indicators by grouping different relational data from a multidimensional database. Multidimensional databases are suitable for storing large amounts of analytical data on which analyses and surveys are used to support decision-making. Conceptual and logical design of these databases is usually done through the Star or Snowflake scheme. The physical method of data organisation in a multidimensional database is performed through data cubes (MOLAP) or relational databases (ROLAP).

## 2 Methods

The actual derivation of the prognosis from the econometric model precedes the verification of the prognostic properties of the individual equations which can be assessed based on the analysis of the economical interpretability of the calculated parameters, the multi-collinearity between the explanatory variables, the tightness of dependence, the statistical significance of the parameters, the autocorrelation of the residues and the standardized deviations. In the design of the data analysis application to derive a prognosis from an econometric model, we propose OLAP to analyse standardized deviation data and use a multidimensional data view. We identify dimensions, related attributes and measures, which are based on the calculation of standardized deviations.

### 2.1 The Multidimensional Paradigm

In this paper, we adopt the formal apparatus of the data cube according to [14]. Let us have a 6-tuple $\langle D, M, A, f, V, g\rangle$, where four components indicate the properties of the data cube. These properties are:

1. The set of $n$ dimensions $D = \{d_1, d_2, \ldots, d_n\}$, where each $d_i$ is the name of dimension from the particular domain $dom_{\mathrm{dim}(i)}$.
2. The set of $k$ measures $M = \{m_1, m_2, \ldots, m_k\}$, where each $m_i$ is the name of measure from the particular domain $dom_{\mathrm{measure}(l)}$.
3. The set of dimension names and measures is disjoint; i.e. $D \cap M = 0$.
4. The set of $t$ attributes $A = \{a_1, a_2, \ldots, a_t\}$, where each $a_i$ is the name of attribute from the particular domain $dom_{\mathrm{attr}(r)}$.
5. The one-to-many mapping $f{:}D \to A$ exists for every dimension and set of attributes. The mapping is such that attribute sets corresponding are pair wise disjoint, i.e. $\forall i, j, i \neq j, f(d_i) \cap f(d_j) = 0$.
6. The set $V$ represents a set of values used to materialize data cube. Therefore, every element $v_i \in V$ is $k$-tuple $\langle \mu_1, \mu_2, \ldots, \mu_k\rangle$, where $\mu_i$ is instance of $i$-th measure $m_i$.
7. The $g$ represents a mapping $g{:}dom_{\mathrm{dim}(1)} \times dom_{\mathrm{dim}(2)} \times \ldots \times dom_{\mathrm{dim}(n)} \to V$. Thus, $g$ apping intuitively indicates which values are associated with a particular ‘cell’. *Cells* are measures or values based on a set of dimensions.

### 2.2 Calculation of Standardized Deviations

The standardized deviation is the ratio between the offset of the actual value and its standard deviation:

$$N_{it} = \frac{\hat{y}_{it} - y_{it}}{S_{y_i}},\ where\ i = (1 \ldots g)\ and\ t = (1 \ldots n) \tag{1}$$

$\hat{y}_{it}$ is a balanced value of the *i-th* endogenous variable at time $t$.
$y_{it}$ is the real/actual value of the *i-th* endogenous variable at time $t$.

$S_{y_i}$ is the standard deviation of the $i$-th endogenous variable computed as the square root of the total variance.

The standardized deviation of the $i$-th endogenous variable of the model is computed according to the formula:

$$N_i = \sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{n} N_{it}^2}, where\ i = (1 \ldots g) \tag{2}$$

The standardized deviation for individual years of the time series is computed according to the formula:

$$N_t = \sqrt{\frac{1}{g}\sum\nolimits_{i=1}^{g} N_{it}^2}, where\ t = (1 \ldots n) \tag{3}$$

The standardized deviation for the whole model is in the form:

$$N = \sqrt{\frac{1}{g}\frac{1}{n}\sum\nolimits_{i=1}^{g}\sum\nolimits_{t=1}^{n} N_{it}^2} \tag{4}$$

To compute the standardized deviations of the model is, according to the Formula (1), possible to derive the standardized deviation matrix $N_{it}$. The matrix's size is g*n. Computing according to the Formulas (2) and (3) is the quadratic mean of the elements per each row and column. This is important for calculation of a data cube matrix in a multidimensional data model.

# 3 Results

In this section, we propose new econometric intelligence system concept that uses OLAP. Consequently, we show the proposal of a multidimensional approach for the analysis of standardized deviations for forecasting in econometric models.

## 3.1 The Econometric Intelligence System

We define the econometric intelligent system (EIS) as a system capable of implementing econometric tasks in a way in which people approach these tasks on the basis of their intelligence. EIS must include a knowledge base, ambient intelligence and OLAP principles to analyse and process very complex econometric tasks.

Intelligent systems based on the principles of ambient intelligence are new opportunities for the development of intelligent environments. In connection with econometric problems, a problem arises especially in the area of design of rapid and intelligent analyses of econometric data. For these purposes, we suggest using OLAP as an analytical tool which is primarily used in Business Intelligence. OLAP allows you to dynamically analyse a large volume of data coming from econometric variables. We suggest that the OLAP in EIS can be used in the following areas:

1. OLAP as the final presentation tool of EIS - analyses. OLAP outputs will be automated to meet the specific needs in econometric tasks.
2. OLAP as the core component of EIS:

– OLAP for analysis of econometric data.
– OLAP for analysis of knowledge (context-awareness, user preferences etc.).
– OLAP for the support of the construction of econometric models in the following areas:
  - estimation of parameters of the econometric model;
  - testing the significance of structural parameters and tightness of dependence of the chosen function;
  - derivation of the production or demand function and its economic interpretation;
  - creation and computation of nonlinear consumer functions;
  - construction of supply and production functions;
  - analysis of relationships between production factors;
  - verification of prognostic properties of the model.

### 3.2 The Concept of Econometric Intelligence System

The proposed EIS concept consists of two core components: Ambient Intelligence and Business Intelligence. Ambient Intelligence allows to create a smart user environment through multi-agent systems that decide in a cooperative way which action they should prefer to adequately support econometric activities while solving econometric problems (Fig. 1).

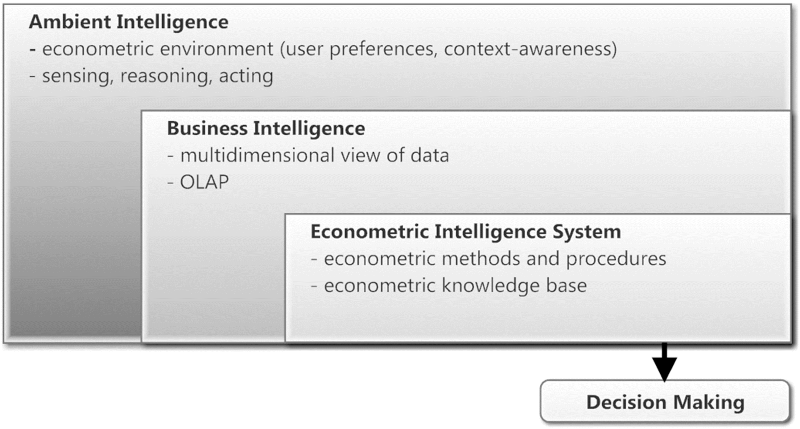

Ambient Intelligence
- econometric environment (user preferences, context-awareness)
- sensing, reasoning, acting
Business Intelligence
- multidimensional view of data
- OLAP
Econometric Intelligence System
- econometric methods and procedures
- econometric knowledge base
Decision Making

**Fig. 1.** Conceptual schema of Econometric Intelligence System

Analytical processes must be mostly considered in the framework of econometric calculations. We suggest using Business Intelligence concepts to support these processes. A large amount of economic and operational data can be transformed through ETL (Extract-Transform-Load) systems into a form suitable for analytical processing. Such analytical data can be stored in data warehouses and subsequently used to create a multidimensional database. From the multidimensional database, it is possible to perform online analytical processing of econometric data through the OLAP technology. Analytical outputs provided by OLAP also enable to perform econometric analyses and create new knowledge that is stored in a knowledge base and used for decision-making processes of econometric issues.

### 3.3 Decision Support of Analyses of Standardized Deviation

As an example, we introduce utilization of the EIS to support the verification of prognostic properties through a multidimensional view of the standardized deviations.

Let *N* be the set of all *elements of standardized deviations* in a particular domain, i.e.

$$N = \{n_1, n_2, \dots n_n\} \tag{5}$$

where $n \in E(\textit{endogenous variable}) \vee n \in T(\textit{time})$. The set of points of view on standardized deviations (*NOV*) is defined as the set of matrices in which each matrix represents the values of standardized deviations of specific econometric models. A *NOV* is defined as:

$$NOV = \{[nv_1],[nv_2]\dots[nv_m]\} \tag{6}$$

where each [*nv*] is a 2-dimensional matrix, as follows:

$$[nv] = \begin{array}{c} \\ e_1 \\ e_2 \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \\ \begin{bmatrix} nv_{11} & nv_{12} & \cdots & nv_{1m} \\ nv_{21} & nv_{22} & \cdots & nv_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ nv_{n1} & nv_{n2} & \cdots & nv_{nm} \end{bmatrix} \end{array} \tag{7}$$

where each *e* represents an *endogenous variable* from *E* to dimension $d_n$ from *D*, each *t* represents *time* from *T* to dimension $d_n$ from *D* and each *nv* is a particular calculated value of the *standardized deviation* to values $v_i$ from *V*. Formula (1) represents the measure $m_k$ from *M* of the data cube.

Now, we will show an example of the output from the multidimensional point of view of the calculated standardized deviation for *i*-th endogenous variable in the econometric model:

Contingency tables represent the most common output of OLAP. This solution allows us to display indicators from two points of view. It is from the point of view of endogenous variables and time in the Table 1. These points of view are possible to switch, filter and in the case of hierarchy of dimensions also carry out the drill-down and roll-up operations. This gives a very effective view on econometric data. It would be possible to monitor more indicators, hundreds of variables and econometric models with millions of records in multidimensional databases in such an EIS.

It is clear that the EIS could determine the quality of construction of econometric models and be important for the work of econometrists.

**Table 1.** Contingency table of the calculated standardized deviations

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | $\sum N_{it}^2$ |
|---|---|---|---|---|---|---|---|---|---|
| $y_1$ | −0.2027 | 1.4217 | 0.5218 | 0.5666 | 0.8632 | −0.9582 | −0.7623 | 0.7174 | 2.1675 |
| $y_2$ | −0.0890 | 0.9759 | 1.4398 | 0.2990 | 1.2320 | −0.2437 | −0.4924 | 0.3787 | 3.5003 |
| $y_3$ | 1.1133 | 1.0187 | −0.7413 | −0.0993 | −0.8575 | 1.0438 | −0.4522 | 1.4406 | 2.4661 |
| $y_4$ | −0.4281 | −0.0525 | 0.7563 | 1.2772 | −0.3475 | 0.9808 | 0.9738 | 1.4077 | 4.5677 |
| $y_5$ | −1.3381 | 3.4564 | 0.0779 | 1.9186 | 2.2474 | −0.3999 | 1.4849 | 0.1908 | 7.6379 |
| $\sum N_{it}^2$ | −0.9447 | 6.8201 | 2.0544 | 3.9622 | 3.1376 | 0.4228 | 0.7519 | 4.1352 | |

## 4 Conclusion

The paper presented a concept of econometric intelligence system based on Ambient Intelligence and Business Intelligence. Such an intelligent system could be a very effective tool to support decision making in the areas of econometric analysis. This research has shown the benefits of using OLAP in econometric methods as a precursor to the current design of an econometric intelligent system. Our research mainly brings the following benefits:

- We have identified the core components of the EIS among which belong:
  - Ambient Intelligence – Intelligent system adaptation based on user preferences and context-awareness;
  - Business Intelligence – OLAP, Multidimensional databases, Data Warehouse, etc.;
  - Econometric Intelligence System – the core of an intelligent system using econometric methods and procedures including recording and utilising knowledge gained from them.
- We have shown the use of the OLAP approach on analyses of the estimation of standardized deviation in prognosis in econometric models.

Further research in this area of intelligent systems is promising. It can allow designing a more adaptive system that will automatically select the most relevant content according to preferences and present it to the user.

Presented results are only the first step of a more complex research. For further research, we plan to address the issue of automatic interpretation of calculated standardized deviations in econometric models.

**Acknowledgements.** This work was conducted within the project DEPIES – Decision Processes in Intelligent Environments funded through the Czech Science Foundation, Czech Republic, Grant No. 15-11724S.

## References

1. Tong, H., Kumar, T.K., Huang, Y.: Developing Econometrics. Wiley, London (2011)

2. Dolk, D.R., Kridel, D.J.: An active modeling system for econometric analysis. Decis. Support Syst. **7**(4), 315–328 (1991)
3. Oxley, L.T.: An expert systems approach to econometric modelling. Math. Comput. Simul. **39**(3–4), 379–383 (1995)
4. Al-Othman, A.N.: Implementing a web-based decision support system for macro-econometric models. Kuwait J. Sci. Eng. **33**(1), 253 (2006)
5. Kridel, D., Dolk, D.: Automated self-service modeling: predictive analytics as a service. IseB **11**(1), 119–140 (2013)
6. Remagnino, P., Foresti, G.L.: Ambient intelligence: a new multidisciplinary paradigm. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **35**(1), 1–6 (2005)
7. Weber, W., Rabaey, J., Aarts, E.H. (eds.): Ambient Intelligence. Springer Science & Business Media, Berlin (2005)
8. Cook, D.J.: Multi-agent smart environments. J. Ambient Intell. Smart Environ. **1**(1), 51–55 (2009)
9. Cook, D.J., Augusto, J.C., Jakkula, V.R.: Ambient intelligence: technologies, applications, and opportunities. Pervasive Mobile Comput. **5**(4), 277–298 (2009)
10. Aarts, E., Wichert, R.: Ambient Intelligence, pp. 244–249. Springer, Berlin (2009)
11. Aly, S., Pelikán, M., Vrana, I.: A generalized model for quantifying the impact of Ambient Intelligence on smart workplaces: applications in manufacturing. J. Ambient Intell. Smart Environ. **6**(6), 651–673 (2014)
12. Tyrychtr, J., Pelikán, M., Štiková, H., Vrana, I.: Multidimensional design of OLAP system for context-aware analysis in the ambient intelligence environment. In: Software Engineering Perspectives and Application in Intelligent Systems, pp. 283–292. Springer (2016)
13. Abelló, A., Romero, O.: On-line analytical processing. In: Encyclopedia of Database Systems, pp. 1949–1954. Springer, US (2009)
14. Datta, A., Thomas, H.: The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decis. Support Syst. **27**(3), 289–301 (1999)

# Earthquake Ground Motion Attenuation Modeling Using Levenberg-Marquardt and Brute-Force Method

Edy Irwansyah[1(✉)], Bayu Kanigoro[1], Priscilia Budiman[1],
and Rokhana D. Bekti[2]

[1] School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
{eirwansyah,bkanigoro}@binus.edu
[2] Department of Statistic, Institut Sains and Teknologi AKPRIND,
Yogyakarta 55222, Indonesia

**Abstract.** In this paper, we discuss the results of research on the optimization modeling of ground motion attenuation in the subduction zone of the model Youngs et al. [1] using two methods: the Levenberg-Marquard and Bruce-Force method. This modeling is particularly important in the case of seismicity. Given that it takes a good model for predicting the strength of earthquakes in order to reduce the risk of the impact of natural disasters. Two major contributions of this study are ground motion attenuation model specific to the subduction zone that has been optimized with the Levenberg-Marquard method and Bruce-Force uses a model Youngs et al. [1] and a proof that the Levenberg-Marquard method for optimization model is better than Bruce-Force method. The Levenberg-Marquardt method has been proven to provide more accurate results on the modeling of ground motion attenuation which is indicated by a very small deviation between the values of PGA predictable results with the PGA actual values.

## 1 Introduction

Indonesia is a country with high-intensity earthquakes. United States Geological Survey (USGS) recorded four episodes of major earthquakes in Indonesia, Banda earthquake (Mw 8.5) 1983 Sumatra earthquake (Mw 9.1) in 2004, Nias earthquake (Mw 8.6) in 2005 and the West Coast Sumatra earthquake (8.6 Mw) in 2012 [2]. High intensity seismic tectonic be the main character of the Indonesian archipelago, located between three major plates, namely the Eurasian plate in the north, the Indo-Australian plate and the Pacific plate to the south in the northeast.

An earthquake with an intensity and a certain magnitude can result in damage to the physical infrastructure and fatalities. Physical infrastructure damage due to the earthquake is the most dominant building damage both supported by the poor quality of construction and due to the environmental conditions in which the building is situated. Significant damage to buildings was recorded in

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_26

the city of Banda Aceh as a result of the Sumatra earthquake in 2004 with the total destruction of the building reaches 35 percent of the existing building [3]. The same condition also occurs in some cities as a result of the earthquake as different as the total damage to the building 140,000 units due to the earthquake in Bantul, Yogyakarta in 2006 [4].

In order to mitigate or reduce the risk of natural disasters, especially earthquakes, we need a model that can predict how big or risk incurred as a result especially in the area of epicenters and surrounding areas. This estimate can be done by determining the value of Peak Ground Acceleration (PGA) using the model as a representation attenuation acceleration that occurs on the surface of the Earth from rest to hit the shock of the earthquake. PGA measure how strong the earth's surface moves in the earthquake which occurred in a region [5].

PGA nonlinear model development has been done, but the values obtained for each model is different and not necessarily suitable for use in different areas. The equation is commonly used to determine the value of PGA uses a model of attenuation is equal Youngs et al. [1] developed from the data of earthquakes in Alaska, Chile, Cascadia, Japan, Mexico, Peru, and the Solomon Islands, with a magnitude ranging from 5.0–8.2 on the Richter scale. Several other researchers conducted a case study to develop a model of ground motion for the PGA is more general such as Atkinson and Boore [6] using the same area by Young et al. [1] but the strength of the earthquake was raised to 5.0–8.3 on the Richter scale. Several other developed equations to calculate the value of PGA among others, as Petersen et al. [7], Gregor et al. [8], Kanno et al. [9], Lin and Lee [10], and others. Youngs et al. [1] and Lin and Lee [10], using Nonlinear Least Square (NLS) for modeling the PGA.

Recent research related to the earthquake has begun using optimization methods such as methods of estimation Levenberg-Marquardt and Brute-Force with goal of improving the calculation results by minimizing the value of the residual sum of squares (RSS). Nascimento [11], in its publication modeling to predict earthquakes like the wave velocity inversion using Levenberg-Marquardt algorithm with ambiguity results in certain parameters for the captured data area central part of Brazil. Underlying the still-optimal existing model, especially to be applied to the area active region of earthquakes in the zone of subduction, this research aims to optimize model ground motion attenuation of [1] to be able to calculate the value of PGA with Levenberg-Marquardt and methods Bruce-Force uses data seismicity in subduction zones in the west coast of Aceh and surrounding areas, Aceh Province, Indonesia.

## 2 Ground Motion Attenuation Model in Subduction Zone

Attenuation relationship is one of the key components of the seismic hazard assessment of an area [12]. Development of various attenuation function to the source of the earthquake due to the subduction (subduction earthquake) has been done as it has been published by [1,6–8,10,12]. Youngs et al. [1] has proposed an attenuation function regression using earthquake data catalogs between the plates with the variation of magnitude 5–8.2 recorded in the subduction area in the area of Alaska, Chile, Cascadia, Japan, Mexico, Peru and the Solomon Islands. Attenuation function is modified by Petersen et al. [7], by comparing the results of observations and predicted using data catalog of earthquakes in 1991 and 2001 in the wider area include New Ireland, New Britain, Kamchatka, Santa Cruz, Peru, Kurile Japan, and Sumatra. Modifications are done attenuation function particularly in the case of the earthquake source within more than 20 Km with the earthquake magnitude 6.8–8.3 Mw.

Gregor et al. [8] have developed an attenuation function to the zone subduction in the Cascadia from the same functions as proposed by Young et al. [1] used the model stochastic finite-fault ground motion from Silva et al. [8] with variation in magnitude higher than 8.0 to 9.0 Mw. The advantage of using this model is that unlike the empirical attenuation relationship, which requires samples to the field and geometry are based on a series of strong-motion data is available, the effect of such finite-fault rupture propagation, direction, and resources to the site geometry, stochastic model with finite- fault can be systematically calculated.

The maximum similarity regression method with a moment magnitude of 5.0–8.3 Mw from various regions of the world such as the Alaskan subduction, Japan, Mexico and Central America are used by Atkinson and Boore [6] to develop attenuation function. Results of analysis of the regional variability of the amplitude of the ground motion using a global database that is available to support the fact that there are significant differences between regional, as shown by the amplitude difference of more than two factors among the Cascadia area and the area of Japan. This model uses only the shortest distance from the earthquake source at a distance of 10–500 Km as used by Youngs et al. [1] and Gregor et al. [8].

Lin and Lee [10], with the reference of the attenuation function previously developed by Crouse [13] and Youngs et al. [6], developed a model of attenuation function regression others use a recording movement of the earthquake in the bedrock in the area subduction of the plates in the Northeast Taiwan and other regions of the value of the low magnitude of about 4.1 to 8.7 on the Richter Scale. The use of a low magnitude value $< 5.0$ Mw is something different from the attenuation function which has been developed by previous researchers. Lin and Lee [10] noted that the use of attenuation function to compute the value of the acceleration in the bedrock or peak ground acceleration (PGA) where the value of ground motion predicted higher than the value produced by using the equation of attenuation that has previously been used in Taiwan and lowers compared to using attenuation equations are used globally, especially in an earthquake zone as a result of subduction.

Development of attenuation function latest local data Sumatra's West Coast done by Megawati and Pan [12] from attenuation function developed by Megawati et al. [12] in the form of regression synthetic movement of the bedrock in an earthquake zone as a result of subduction of the plates using a model of a finite fault kinematic as adopted from Gregor et al. [8]. Validation of the model was done using the data megathrust Sumatra that includes a very large earthquake with a strength of up to 9.0 Mw.

## 3 Levenberg-Marquardt and Brute-Force Method

Nonlinear regression equation has many methods to estimate the parameters. The estimation methods include Ordinary Least Squares (OLS), Nonlinear Least Square (NLS), Generalized Nonlinear Least Square and nonlinear Maximum Likelihood. NLS is a form of least squares analysis used in modeling the nonlinear regression by minimizing the Residual Sum of Squares-RSS [14]. Methods to minimize the value of RSS is the parameter optimization. Some of these methods include the Gauss-Newton, Hartley's Method, Powell's Hybrid Method, Quasi-Newton, Levenberg-Marquardt Method and Brute-Force Method. Levenberg-Marquardt method [15,16] commonly known as the damped least squares method (DLS) which produces a numerical solution to minimize a nonlinear function of the parameters in the function. The Levenberg-Marquardt method is an interpolation between the Gauss-Newton method and the method of Gradient-Descent. The main application of the Levenberg-Marquardt method is the least squares problem that aims to optimize the $\beta$ parameters from $f(x_i, \beta+\delta)$ model, thus RSS became the minimal value.

Levenberg-Marquardt method using the iterative procedure. To start the minimization process, the first step is to estimate the value of the parameter vector, $\beta$. At each stage of iteration, parameter vector, $\beta$, will be replaced with a new estimated value, i.e. $\beta + \delta$. To find the value of $\delta$, the function $f(x_i, \beta + \delta)$ is approached by making linear

$$(x_i, \beta + \delta) \approx f(x_i, \beta) + j_i\delta \tag{1}$$

In where,

$$j_i = \frac{\partial f(x_i, \beta)}{\partial \beta} \tag{2}$$

is the gradient (vector row) on the $f$ of the parameter $\beta$. Approximation of $f(x_i, \beta + \delta)$ will produce,

$$S(\beta + \delta) \approx \sum_{i=1}^{n} [y_i - f(x_i, \beta) - J_i\delta]^2 \tag{3}$$

or in vector notation becomes,

$$S(\beta + \delta) \approx ||y - f(\beta) - J\delta||^2 \tag{4}$$

Levenberg-Marquardt method of modifying step of the Gauss-Newton became,

$$(J^T J + \lambda I)\delta = J^T [y - f(\beta)] \tag{5}$$

where $J$ is the Jacobian matrix that has rows $J_i$ in where $f$ and $y$ is a vector with components $f(x_i, \beta)$ and $y_i$ as much as $i$. $\delta$ value is the value that gives descent direction to the vector parameter $\beta$. $\lambda$ is the damping parameter value that should not be negative and will be adjusted in each iteration.

The brute-force method commonly called Grid-Search. In nonlinear regression, it is a global grid-search method in which the entire space of model parameters is sampled. This method is used to determine the starting value, but can also be used to estimate parameters of this method will eventually choose the estimated value of which generates a smallest RSS value (Residual Sum of Square). Suppose the nonlinear model $f(x_i, \beta)$ with parameter $\beta$, it estimate $\beta$ by selected value from starting value. Some of the research use the selected value by pre-defined grid [17]. Brute-Force method is used if the known range of the estimated value of the parameter. The brute-force method will iterate for each value of the starting value. Iteration stops when all of the starting value has been iterated and subsequently be selected RSS smallest of these iterations [18].

Tsagaan, Nappi, and Yoshida [19] states that brute force optimization has the advantage that it can find a globally optimal solution. The algorithms based on mathematical heuristics could be caught in a local minimum. They use brute force for estimate nonlinear regression for pseudo-enhancement correction in CT colonography. Also, Archontoulis and Miguez [20] use brute force to choose the starting value for estimated parameter in agricultural nonlinear regression. It was done by generating an extensive coverage of possible parameter values and their combinations. Then, it evaluates the model at each one of these parameter combinations.

## 4 Ground Motion Attenuation Model Optimation

Research conducted using secondary data in the form of the variable distance of the location of the epicenters, the depth of the earthquake, the earthquake magnitude, and the value of PGA-year data for 2005 to 2007 are sourced from national meteorological, climatology, and geophysics agency (BMKG). The population and the sample used is the west coast of Sumatra in the area a radius of 500 KM from the center of Banda Aceh city, Aceh province, Indonesia (Fig. 1).

The analysis was conducted on the stage (1) determining descriptive statistics for each variable, such as the calculation of average, variance, median, standard deviation, etc. (2) To test the linearity of the data with the test Ramsey's RESET (3) to determine a nonlinear model for the PGA. The model is a model Youngs et al. [1] as follows:

$$\ln(\mathrm{PGA}) = C_1 + C_2 + C_3 \ln\left[R + e^{C_4 - (\frac{C_2}{C_3})M}\right] \tag{6}$$

(4) Estimating parameters by nonlinear regression method (Brute-Force and Levenberg). Stages in the Brute-Force method is (a) Determine the starting value for each parameter along with the increase, (b) calculate the value of RSS, (c) take any chances starting value on each parameter and (d) comparing RSS smallest of each iteration. Stages in the Levenberg-Marquardt method is (a) determining the starting value for each parameter, (b) Determine the limit

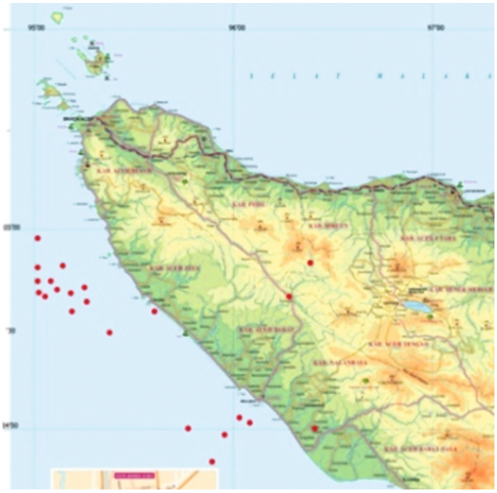
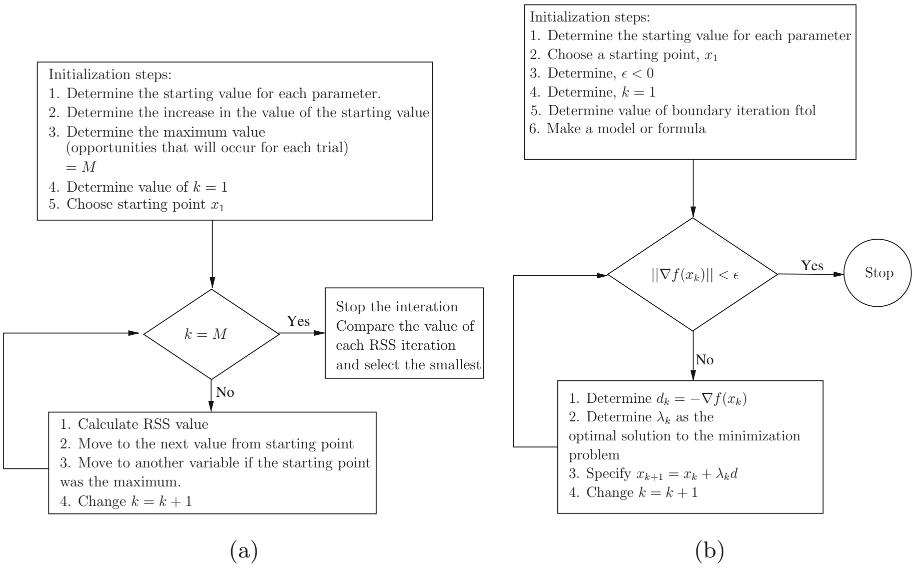


**Fig. 1.** Research area and spatial distribution of the data

iteration with ftol, (c) Calculated Value RSS and (d) Conducting iteration in accordance with the decrease $\lambda_k$, (5) Testing the assumption of residual (identical, independent, and normal distribution) for each model is formed and (6) Comparing the method of nonlinear methods Brute-Force and Levenberg-Marquardt. The comparison is done by comparing the average, variance value, the value of


Initialization steps:
1. Determine the starting value for each parameter.
2. Determine the increase in the value of the starting value
3. Determine the maximum value (opportunities that will occur for each trial) $= M$
4. Determine value of $k = 1$
5. Choose starting point $x_1$
$k = M$
Yes
Stop the interation Compare the value of each RSS iteration and select the smallest
No
1. Calculate RSS value
2. Move to the next value from starting point
3. Move to another variable if the starting point was the maximum.
4. Change $k = k + 1$
Initialization steps:
1. Determine the starting value for each parameter
2. Choose a starting point, $x_1$
3. Determine, $\epsilon < 0$
4. Determine, $k = 1$
5. Determine value of boundary iteration ftol
6. Make a model or formula
$||\nabla f(x_k)|| < \epsilon$
Yes
Stop
No
1. Determine $d_k = -\nabla f(x_k)$
2. Determine $\lambda_k$ as the optimal solution to the minimization problem
3. Specify $x_{k+1} = x_k + \lambda_k d$
4. Change $k = k + 1$

(a)

(b)

**Fig. 2.** Modeling stages using Brute-Forces (a) and Levenberg-Marquardt (b) Methods

the residual standard error, residual assumption test, and plot between the actual value of PGA and the predictive value. Stages of non-linear regression modeling using Brute-Force method (Fig. 2a) and the Levenberg-Marquardt method (Fig. 2b) are as follows:

# 5 Result and Discussion

## 5.1 Nonlinearity Test

Nonlinearity test in this study was conducted to determine whether the data used to follow the pattern of nonlinear models or not. Nonlinearity test conducted by plotting the data between seismicity variable such as distance from earthquake center, depth of the earthquake and earthquake magnitude with the PGA and the nonlinearity pattern. Nonlinearity test can also be done by the method of Ramsey's RESET. The nonlinearity test generate result of $F_{value} = 10.5256$ with $df_1 = 3$ and $df_2 = 16$ with P-value = 0.0004572. Testing concluded that the data follows the nonlinear pattern because the value of $F = 10.5256$ greater than the value of $F(0.05, 3, 16) = 3.24$. The value (P-value = 0.000452) also indicates that the data follows the nonlinear pattern because the P-value less than the value of $\alpha = 0.05$.

## 5.2 Modeling and Model Comparison

Modeling in this study based on the model of Young et al. [1] which is optimized in order to generate value Residual Sum of Square (RSS) is small with Levenberg-Marquardt and Brute-Force methods. In Table 1 we can see the details of each parameter estimation.

**Table 1.** Parameter estimation of Young et al. [1] with Levenberg-Marquardt and Brute-Force Method

| Levenberg-Marquardt method | | Brute-Force method | |
|---|---|---|---|
| Parameter | Estimation | Parameter | Estimation |
| $C_1$ | $-1.101$ | $C_1$ | $-1.1$ |
| $C_2$ | $-0.0008244$ | $C_2$ | 0.002 |
| $C_3$ | 0.005139 | $C_3$ | 0.005 |
| $C_4$ | 11.83 | $C_4$ | 3 |
| $C_9$ | 0.0000283 | $C_9$ | 0.0006 |

After modeling the nonlinear regression have conducted, then performed the assumption that residuals must meet with $\epsilon \sim$ IIND, namely residual must meet identical test using Glejser test, independent test with Durbin-Watson and lag1.plot, and the normal distribution test using Kolmogorov-Smirnov.

Residual assumption test conducted showed that both models tested had fulfilled all residual assumptions, as can be seen in the test results using computer applications and summarized in Table 2.

**Table 2.** Residual assumption of Young et al. [1] model with Levenberg-Marquardt and Brute-Force Method

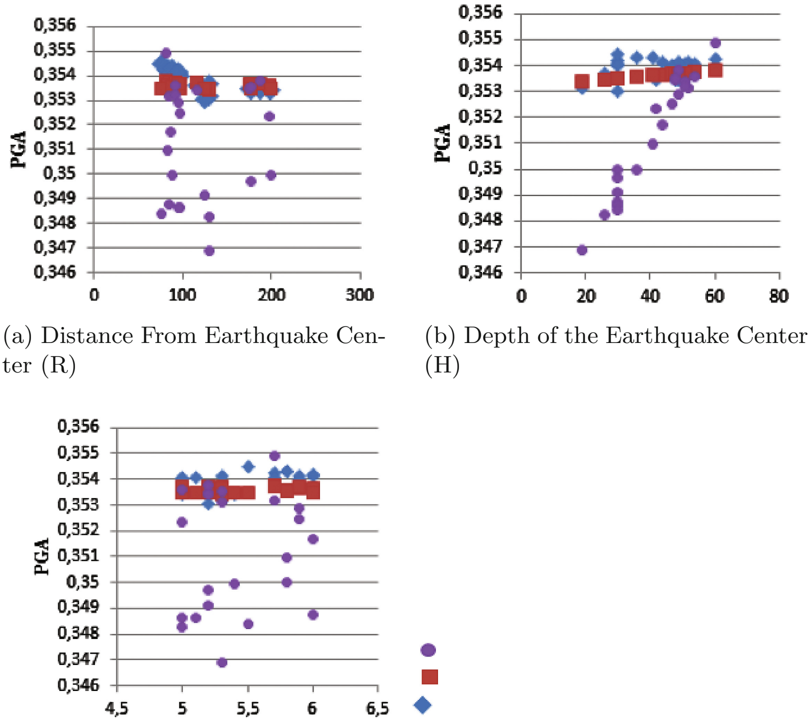
| Residual assumption | Youngs et al. [1] model | |
|---|---|---|
| | Brute-Force | Levenberg-Marquardt |
| Identical | Yes | Yes |
| Independent | Yes | Yes |
| Normal distribution | Yes | Yes |

Once the model was obtained and a residual test showed good results, the next is to compare the model by comparing the actual PGA value with predicted PGA results. This comparison can be seen based on descriptive statistics of data such as average value, variance, and the value of the residual standard error of the model. In Table 3, it can be seen that the model developed has been able to predict the value of PGA well. When viewed from the average of the closest to the average generated from the value of PGA actual model Youngs, et al. 1997, in the estimation by the method of Levenberg-Marquardt is the best because the average value prediction results differ only 0.00004 of an average actual PGA.

**Table 3.** Descriptive statistics comparison between Levenberg-Marquardt and Brute-Force Method

| Descriptive statistics | Actual PGA | Youngs et al. [1] model | |
|---|---|---|---|
| | | Brute-Force | Levenberg-Marquardt |
| Averages | 0.35383 | 0.35304 | 0.35379 |
| Variance | 0.0000002 | 0.00001 | 0.00000001 |
| Residual standard error | | 0.007775 | 0.001251 |

In Fig. 3, shows that the model Youngs et al. [1] in the estimation of the Brute-Force method, different with the numbers large enough rather than PGA actual value. Results scatterplot between the predicted value and the actual value shows that the two major values are further apart. This is also supported by the average value of descriptive statistics in Table 3 that shows the similarity and the difference PGA predictive values of a great range of 0.00079 PGA actual value. Youngs et al. model [1] in the estimation of the Levenberg-Marquardt method can produce PGA predictive value that is almost equal to the value of actual PGA. This is indicated by the closeness between PGA scatterplot point predictions with actual PGA.



(a) Distance From Earthquake Center (R)

(b) Depth of the Earthquake Center (H)

(c) Earthquake Magnitude (M)

**Fig. 3.** Comparison between PGA Actual and PGA Prediction for each Levenberg-Marquardt and Brute-Force Method

## 6 Conclusion

Modeling of ground motion attenuation generated for Youngs et al. [1] and the model estimated by the Levenberg-Marquardt method and Bruce-Force is as follows:

1. Youngs et al. [1] model were estimated by Levenberg-Marquardt method,

$$\ln(\text{PGA}) = -1.101 - 0.00082M + 0.00514\ln\left[R + e^{11.83-\left(\frac{-0.00082}{0.00514}\right)M}\right]$$

2. Youngs et al. [1] model were estimated by Brute-Force method,

$$\ln(\text{PGA}) = -1.1 - 0.002M + 0.005\ln\left[R + e^{3-\left(\frac{-0.001}{0.005}\right)M}\right] + 0.0006H$$

Youngs et al. model [1] in with estimation by Levenberg-Marquardt method proven to provide better results on the modeling of ground motion attenuation shown by the divergence was very small both on the value of the average, variance and residual standard error are generated between the value predicted by actual value. Besides all the residual assumptions are met and the results showed similarities scatterplot predicted results with the actual value.

**Acknowledgments.** We gratefully acknowledge the support of institutions in the implementation of this research. Many thanks to the national meteorological, climatology and geophysical agency (BMKG), Jakarta, Indonesia, which supported the availability of data for testing and data validation. We also thank the research and technology transfer office (RTTO) of Bina Nusantara University, which provided internal funding for the research until the publication of this article.

## References

1. Youngs, R., Chiou, S.J., Silva, W., Humphrey, J.: Strong ground motion attenuation relationships for subduction zone earthquakes. Seismol. Res. Lett. **68**(1), 58–73 (1997)
2. USGS: United States Geological Survey. http://earthquake.usgs.gov
3. Irwansyah, E., Winarko, E., Rasjid, Z., Bekti, R.: Earthquake hazard zonation using peak ground acceleration (PGA) approach. J. Phys.: Conf. Ser. **423**, 012067 (2013). IOP Publishing
4. Miura, H., Wijeyewickrema, A.C., Inoue, S.: Evaluation of tsunami damage in the eastern part of sri lanka due to the 2004 sumatra earthquake using high-resolution satellite images. In: Proceedings of the 3rd International Workshop on Remote Sensing for Post-Disaster Response, Chiba, Japan (2005)
5. Santoso, E., Widiyantoro, S., Sukanta, I.N.: Studi hazard seismik dan hubungannya dengan intensitas seismik di pulau sumatera dan sekitarnya. Jurnal Meteorologi dan Geofisika **12**(2) (2011)
6. Atkinson, G.M., Boore, D.M.: Empirical ground-motion relations for subduction-zone earthquakes and their application to cascadia and other regions. Bull. Seismol. Soc. Am. **93**(4), 1703–1729 (2003)
7. Petersen, M.D., Dewey, J., Hartzell, S., Mueller, C., Harmsen, S., Frankel, A., Rukstales, K.: Probabilistic seismic hazard analysis for sumatra, indonesia and across the southern malaysian peninsula. Tectonophysics **390**(1), 141–158 (2004)
8. Gregor, N.J., Silva, W.J., Wong, I.G., Youngs, R.R.: Ground-motion attenuation relationships for cascadia subduction zone megathrust earthquakes based on a stochastic finite-fault model. Bull. Seismol. Soc. Am. **92**(5), 1923–1932 (2002)
9. Kanno, T., Narita, A., Morikawa, N., Fujiwara, H., Fukushima, Y.: A new attenuation relation for strong ground motion in japan based on recorded data. Bull. Seismol. Soc. Am. **96**(3), 879–897 (2006)
10. Lin, P.S., Lee, C.T.: Ground-motion attenuation relationships for subduction-zone earthquakes in northeastern taiwan. Bull. Seismol. Soc. Am. **98**(1), 220–240 (2008)
11. do Nascimento, P.F., França, G.S., Moreira, L.P., Von Huelsen, M.G.: Application of gauss-marquardt-levenberg method in the inversion of receiver function in central brazil. Revista Brasileira de Geofísica **30**(3) (2012)
12. Megawati, K., Pan, T.C., Koketsu, K.: Response spectral attenuation relationships for sumatran-subduction earthquakes and the seismic hazard implications to singapore and kuala lumpur. Soil Dyn. Earthquake Eng. **25**(1), 11–25 (2005)

13. Crouse, C.: Ground-motion attenuation equations for earthquakes on the cascadia subduction zone. Earthquake Spectra **7**(2), 201–236 (1991)
14. Ritz, C., Streibig, J.C.: Nonlinear Regression with R. Springer Science & Business Media (2008)
15. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. **2**(2), 164–168 (1944)
16. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11**(2), 431–441 (1963)
17. Mudelsee, M.: Ramp function regression: a tool for quantifying climate transitions. Comput. Geosci. **26**(3), 293–307 (2000)
18. Grothendieck, G.: nls2: Non-linear regression with brute force. R package version 0.2 (2013)
19. Tsagaan, B., Näppi, J., Yoshida, H.: Nonlinear regression-based method for pseudoenhancement correction in CT colonography. Med. Phys. **36**(8), 3596–3606 (2009)
20. Archontoulis, S.V., Miguez, F.E.: Nonlinear regression models and applications in agricultural research. Agron. J. **107**(2), 786–798 (2015)

# Advanced Approach for Observability of Distributed Systems Using Internal Pointwise Sensor

Amine Bouaine[1,2](✉) and Mostafa Rachik[2]

[1] Department of Industrial Engineering,
École Nationale Supérieure de Mines de Rabat, Rabat, Morocco
amine.bouaine@gmail.com
[2] Department of Mathematics and Computer Science,
Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco

**Abstract.** This paper will serve as a basic introduction to the observability of diffusion process as an example of distributed parameter systems. The aim of this research is to reconstruct initial state not well known $x_0$, which is known in certain subregions and unknown in others, and to give important results related to internal pointwise sensor in different geometrical situations. Many applications are investigated whether in one-dimensional case or two-dimensional one.

**Keywords:** Distributed systems · Observability · Diffusion process · Strongly continuous semigroups · Sensors

## 1 Introduction

Many systems from science, engineering and different industrial processes belong to distributed parameter systems (DPS). They are modeled by sets of partial differential equations, boundary conditions and initial conditions, which describe the evolution of the state variables in several independent coordinates, e.g. space and time. Most distributed parameter models are derived from first-principles such as conservation of mass, energy and momentum. Advanced technological needs such as semiconductor manufacturing, nanotechnology, biotechnology, material engineering and chemical engineering, have motivated control and observability of material microstructure, fluid flows, spatial profiles and product size distributions [1]. Time and space is the most frequent combination of independent variables, as is the case of some typical examples include thermal process [2,3], fluid process [4–6], convection-diffusion-reaction process [7] and flexible beam [8–10]. But, other combinations of independent variables are possible as well, for instance, time and individual size occur in population models used in ecology, or to describe some important industrial processes like polymerization, crystallization or material grinding.

Observability and control of linear and nonlinear distributed systems is a cross-disciplinary and rapidly growing research area that brings together

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_27

fundamental modeling, numerical simulation, nonlinear dynamics and control theory [1]. In this paper, we consider a class of distributed parameter systems and give various results connected with internal pointwise measurement applied to diffusion process in different geometrical situations. Obviously, we are concerned with the state observation not in the whole domain $\Omega$ where the system is defined but only in a some subregions. This situation occurs in many practical applications where we may be concerned with the knowledge of the state only in critical subregions.

Let $x$ be the state of a linear system (1) with the state space $X = L^2(\Omega)$, and suppose that the initial state $x_0$ is not well known, in other words, $x_0$ is unknown in some subregions whereas is known in other subregions. Suppose now that measurement are given by means of an output $y \in Z$ (depending on the number and the structure of the sensors, the measurement interval, etc.) [11].

This paper is organized as follows. The problematic of this research is revealed and explained in the following section. The Sect. 3 deals with reconstructing initial state $x_0$, not well known, and introduces improved approach to figure out this problem. The Sect. 4 sheds light on diffusion process in one dimension where three cases are studied: initial state unknown on one subregion, on two subregions, and on several subregions. The Sect. 5 addresses diffusion process in two dimensions where two cases are treated: initial state unknown on several subregions and on one subregion. To end, conclusions are summarized and different perspectives are revealed in the last section.

## 2 Problem Statement

Let $\Omega$ be an open regular bounded subset of $R^n$, with boundary $\Gamma = \partial\Omega$ and time horizon $T$. Consider evolution system

$$\begin{array}{ll} \dot{x}\,(t) = Ax\,(t)\,; & x(t) \in L^2(\Omega) \\ x\,(0) \text{ is not well known} & \end{array} \tag{1}$$

with the corresponding measurements

$$y(t) = Cx(t) \tag{2}$$

where $C$ is the measurement operator, which can be bounded or not (depending on the nature of the sensors).

$$C:\, L^2(\Omega) \longrightarrow R \tag{3}$$

We introduce $A$ infinitesimal generator of a strongly continuous semi-group $(S\,(t))_{t\geq 0}$, then

$$x(t) = S(t)x_0,\;\; x_0 \in L^2(\Omega) \tag{4}$$

The observability question deals with the question of while not knowing the initial state, whether one can determine the state from the input and output.

This is equivalent to the question of whether one can determine the initial state $x(0)$ when given the input $x(t)$ and the output $y(t)$.

The originality of this work is to assume that $x_0$ is not well known, in detail, $x_0$ is unknown on subregions $\omega_1, \omega_2, .., \omega_p$, whereas $x_0$ is known on subregions $\omega_{p+1}, .., \omega_n$., thus the expression of initial condition is given by

$$x_0 = \sum_{i=1}^{p} \alpha_i \Pi_{\omega i} + \sum_{j=p+1}^{n} \beta_j \Pi_{\omega j} \tag{5}$$

with these conditions

$$\begin{array}{l} \Omega = \omega_1 \cup \omega_2 \cup ... \cup \omega_n \\ \omega_i \cap \omega_j = \emptyset \ for \ 0 \leq i \prec j \leq n. \\ \omega_i \neq \emptyset \ for \ i = 1..n. \end{array} \tag{6}$$

The vector $(\alpha_1, \alpha_2, .., \alpha_p)$ is unknown, but $(\beta_{p+1}, .., \beta_n)$ is known. In other words

$$x_0(\theta) = \begin{array}{l} \alpha_i \ if \ \theta \in \omega_i \quad for \ 1 \leq i \leq p \\ \beta_j \ if \ \theta \in \omega_j \ for \ p+1 \leq j \leq n \end{array} \tag{7}$$

The problem consists of knowing if Eq. (1) together with the output (2) is sufficient to observe the initial state in subregions $(\omega_1, \omega_2, .., \omega_p)$, then it will be possible to observe the system at any time t.

In the light of observability of distributed systems, the problematic is the possibility of reconstructing this vector $(\alpha_1, \alpha_2, .., \alpha_p)$ just from measurements $y(t)$.

## 3 Advanced Approach for Observability of Distributed Parameter Systems

$A$ infinitesimal generator of strongly continuous semi-group $(S(t))_{t \geq 0}$ given by

$$S_t : \begin{array}{ll} L^2(]0,1[) & \longrightarrow L^2(]0,1[) \\ & \\ f & \longrightarrow \sum_{n=1}^{\infty} e^{\lambda_n t} \prec f, \Phi_n \succ \Phi_n \end{array} \tag{8}$$

with

$$\begin{array}{rcl} \lambda_n & = & -n^2\pi^2 \\ & & \\ \Phi_n(.) = \sqrt{2} sin(n\pi.) & & \end{array} \tag{9}$$

In addition to this, we have

$$C : \begin{array}{ll} L^2(]0,1[) & \longrightarrow R \\ x(.) & \longrightarrow Cx(.) = x(\delta);\ \delta \in ]0,1[ \end{array} \tag{10}$$

We have $y(t) = Cx(t) = CS_t x_0 = CS_t \left(\sum_{i=1}^{p} \alpha_i \Pi_{\omega i}\right) + CS_t \left(\sum_{j=p+1}^{n} \beta_j \Pi_{\omega j}\right)$
Then, we put

$$\bar{y}(t) = y(t) - CS_t \left( \sum_{j=p+1}^{n} \beta_j \Pi_{\omega j} \right) \tag{11}$$

This quantity is measurable and well known, and the unknown vector $(\alpha_1, \alpha_2, .., \alpha_p)$ obeys to this equation

$$\sum_{i=1}^{p} \alpha_i C S_t \Pi_{\omega i} = \bar{y}(t) \tag{12}$$

Furthermore $K\left(\alpha_1, \alpha_2, .., \alpha_p\right)^T = \bar{y}(.)$ and $K$ is the operator defined as follows

$$K : \begin{array}{ll} R^p & \longrightarrow L^2(0, T, R) \\ \\ (\gamma_1, \gamma_2, .., \gamma_p)^T & \longrightarrow \sum_{i=1}^{p} \gamma_i C S_t \Pi_{\omega i} \end{array} \tag{13}$$

For the purpose of giving a signification to this problem, we should guarantee a single initial state leading to $\bar{y}(t)$.

**Theorem 1.**

$$x_0 = \sum_{i=1}^{p} \alpha_i \Pi_{\omega i} + \sum_{j=p+1}^{n} \beta_j \Pi_{\omega j} \tag{14}$$

*with these conditions*

$$\begin{array}{l} \Omega = \omega_1 \cup \omega_2 \cup ... \cup \omega_n \\ \omega_i \cap \omega_j = \emptyset \ for \ 0 \leq i \prec j \leq n. \\ \omega_i \neq \emptyset \ for \ i = 1..n. \end{array} \tag{15}$$

*$x(0)$ is not well known: the vector $(\alpha_1, \alpha_2, .., \alpha_p)$ is unknown, but $(\beta_{p+1}, .., \beta_n)$ is known.*

*The system*

$$\begin{array}{l} \dot{x}(t) = Ax(t) \,; \ x(t) \in L^2(\Omega) \\ x(0) \ \textit{is not well known} \end{array} \tag{16}$$

*provided with $y(t) = Cx(t)$, is recalled observable on $[0, T]$ if*

$$K : \begin{array}{ll} R^p & \longrightarrow L^2(0, T, R) \\ (\gamma_1, \gamma_2, .., \gamma_p)^T & \longrightarrow \sum_{i=1}^{p} \gamma_i C S_t \Pi_{\omega i} \end{array} \tag{17}$$

*injective. Furthermore, $(\gamma_1, \gamma_2, .., \gamma_p)^T = (K^* K)^{-1} K^* \bar{y}(.)$*
*$K^*$ is the adjoint operator of $K$.*

*Proof.* We have $K(\gamma_1, \gamma_2, .., \gamma_p)^T = \bar{y}(.)$
$K^* K(\gamma_1, \gamma_2, .., \gamma_p)^T = K^* \bar{y}(.)$.
We mention that $K^* K : R^p \longrightarrow R^p$.
$\prec K^* K x, x \succ = \|Kx\|^2 \geq 0$ and if $\prec K^* K x, x \succ = \|Kx\|^2 = 0$ then $Kx = 0$.
$K$ injective then $x = 0$.
As result, $K^* K$ definite positive.
Now, we conclude $(\gamma_1, \gamma_2, .., \gamma_p)^T = (K^* K)^{-1} K^* \bar{y}(.)$.

# 4 Diffusion Process in one Dimension with Internal Pointwise Sensor

The diffusion equation is a partial differential equation which describes density fluctuations in a material undergoing diffusion. The equation can be written as

$$\frac{\partial x(r,t)}{\partial t} = \nabla.\left(D\left(x(r,t),r\right)\nabla x(r,t)\right) \tag{18}$$

where $x(r,t)$ is the density of the diffusing material at location $r = (x_1, x_2, x_3)$ and time $t$. $D(x(r,t),r)$ denotes the collective diffusion coefficient for density $x$ at location $r$. If the diffusion coefficient doesn't depend on the density, i.e., $D$ is constant, then Eq. (18) reduces to the following linear equation

$$\frac{\partial x(r,t)}{\partial t} = D\nabla^2 x(r,t) \tag{19}$$

Equation (19) is also called the heat equation.

Consider the diffusion equation in one dimension, where the initial state $x_0$ is not well known.

$$\tfrac{\partial x(x_1,t)}{/}\partial t = D\tfrac{\partial^2 x(x_1,t)}{\partial x_1^2} \quad ]0,1[\times]0,T[ \tag{20}$$

The boundary and initial conditions are given as well

$$\text{x(0,t)=x(1,t)=0 } ]0,1[\times]0,T[ \tag{21}$$

$$x(x_1,0) = x_0(x_1) \text{ supposed not well known} \tag{22}$$

The system (20)–(22) can be written as a state-space system

$$\dot{x}(t) = Ax(t);\, x(0) = x_0 \tag{23}$$

where $A = D\partial^2/\partial x_1^2$.

## 4.1 One-Dimensional System: $x_0$ Unknown on one Subregion

The interval $]0,1[$ is divided into two parts ]0, 1/2] and ]1/2, 1[.

We assume that the density of the diffusing material evolves very slowly in each subregion as we can consider it constant. The value of the diffusing material in subregion ]1/2, 1[ is known and we look for its value in the second one.

**Theorem 2.** *$x_0$ is given by*

$$x_0(\theta) = \begin{array}{l} \alpha\; if\;\; \theta \in ]0,1/2] \\ \beta\; if\;\; \theta \in ]1/2,1[ \end{array} \tag{24}$$

*$x\,(0)$ is not well known : $\alpha$ is unknown whereas $\beta$ is known.*
*The system*

$$\begin{array}{l} \dot{x}\,(t) = Ax\,(t)\,;\;\; x(t) \in L^2(]0,1[) \\ x\,(0)\;\; is\ not\ well\ known \end{array} \tag{25}$$

*provided with* $y(t) = Cx(t)$; *is recalled observable on* $[0, T]$.
*Furthermore,* $\alpha = (K_1^* K_1)^{-1} K_1^* \bar{y}(.)$
*where*

$$K_1 : \begin{array}{l} R \longrightarrow L^2(0, T, R) \\ \gamma \;\longrightarrow \gamma C S_t(.) \varPi_{]0,1/2]} \end{array} \tag{26}$$

$K_1^*$ *is the adjoint operator of* $K_1$.

*Proof.* In this case, we put

$$K_1 : \begin{array}{l} R \longrightarrow L^2(0, T, R) \\ \gamma \;\longrightarrow \gamma C S_t(.) \varPi_{]0,1/2]} \end{array} \tag{27}$$

Consider internal pointwise sensor defined by

$$C : \begin{array}{ll} L^2(]0,1[) & \longrightarrow R \\ x(.) & \longrightarrow Cx(.) = x(\delta);\ \delta \in ]0, 1[ \end{array} \tag{28}$$

Now, we study the injectivity of $K_1$
$K_1\gamma = 0 \ \ \forall t \in ]0, T[$, so $\gamma C S_t(.) \varPi_{]0,1/2]} = 0 \ \ \forall t \in ]0, T[$
$\gamma C \left(\sum_{n=1}^{\infty} e^{\lambda_n t} \prec \varPi_{]0,1/2]}, \varPhi_n \succ \varPhi_n\right) = 0 \ \ \forall t \in ]0, T[$
$\gamma \left(\sum_{n=1}^{\infty} e^{\lambda_n t} \prec \varPi_{]0,1/2]}, \varPhi_n \succ \varPhi_n(\delta)\right) = 0 \ \ \forall t \in ]0, T[$
$\sum_{n=1}^{\infty} \left(\gamma e^{\lambda_n t} \left(\int_0^{1/2} \sqrt{2} sin(n\pi x)dx\right) \sqrt{2} sin(n\pi\delta)\right) = 0.$
Based on analyticity $\gamma e^{\lambda_n t} \left(\int_0^{1/2} \sqrt{2} sin(n\pi x)dx\right) \sqrt{2} sin(n\pi\delta) = 0 \ \ \forall t \in ]0, T[$
$\gamma \left(\int_0^{1/2} \sqrt{2} sin(n\pi x)dx\right) \sqrt{2} sin(n\pi\delta) = 0$; $\gamma \frac{\sqrt{2}}{n\pi} \left[1 - cos(\frac{n\pi}{2})\right] \left(\sqrt{2} sin(n\pi\delta)\right) = 0$
$n\pi\delta \neq k\pi$ then $\sqrt{2} sin(n\pi\delta) \neq 0$, we deduce $\gamma \frac{\sqrt{2}}{n\pi} \left(1 - cos(\frac{n\pi}{2})\right) = 0$.
For $n = 1$, we have $\gamma = 0$, so, $K_1$ is injective.
Further, $K_1\alpha = \bar{y}(.)$, so $K_1^* K_1 \alpha = K_1^* \bar{y}(.)$
$K_1$ is injective then $K_1^* K_1$ is definite positive.
Finally, we conclude $\alpha = (K_1^* K_1)^{-1} K_1^* \bar{y}(.)$

### 4.2 One-Dimensional System: $x_0$ Unknown on two Subregions

In this subsection, the interval $]0, 1[$ is divided into three parts ]0, 1/4], ]1/4, 1/2] and ]1/2, 1[. We assume that the density of the diffusing material evolves very slowly in each subregion as we can consider it constant. The values of the density of the diffusing material in subregion ]1/2, 1[ is known and we look for its value in subregions ]0, 1/4] and ]1/4, 1/2].

**Theorem 3.** $x_0$ *is given by*

$$x_0(\theta) = \begin{array}{ll} \alpha_1 & if\ \theta \in ]0, 1/4] \\ \alpha_2 & if\ \theta \in ]1/4, 1/2] \\ \beta & if\ \theta \in ]1/2, 1[ \end{array} \tag{29}$$

$x\,(0)$ *is not well known:* $\alpha_1 \geq 0$ *and* $\alpha_2 \geq 0$ *are unknown but* $\beta$ *is known.*
*The system*

$$\begin{array}{l} \dot{x}\,(t) = Ax\,(t)\,;\ x(t) \in L^2(]0, 1[) \\ x\,(0)\ \ is\ not\ well\ known \end{array} \tag{30}$$

*provided with $y(t) = Cx(t)$; is recalled observable on $[0,T]$.*
*Furthermore,* $(\alpha_1, \alpha_2)^T = (K_2^* K_2)^{-1} K_2^* \bar{y}(.)$
*where*

$$K_2 : \begin{array}{l} (R^+, R^+) \longrightarrow L^2(0,T,R) \\ (\gamma_1, \gamma_2)^T \ \longrightarrow \gamma_1 C S_t(.) \Pi_{]0,1/4[} + \gamma_2 C S_t(.) \Pi_{]1/4,1/2[} \end{array} \qquad (31)$$

*$K_2^*$ is the adjoint operator of $K_2$.*

*Proof.* In this case, we put

$$K_2 : \begin{array}{l} (R^+, R^+) \longrightarrow L^2(0,T,R) \\ (\gamma_1, \gamma_2)^T \ \longrightarrow \gamma_1 C S_t(.) \Pi_{]0,1/4[} + \gamma_2 C S_t(.) \Pi_{]1/4,1/2]} \end{array} \qquad (32)$$

Consider internal pointwise sensor defined by

$$C : \begin{array}{ll} L^2(]0,1[) & \longrightarrow R \\ x(.) & \longrightarrow Cx(.) = x(\delta);\ \delta \in ]0,1[ \end{array} \qquad (33)$$

Now, we study the injectivity of $K_2$
$K_2(\gamma_1, \gamma_2)^T = 0\ \forall t \in ]0,T[$, so $\gamma_1 C S_t(.) \Pi_{]0,1/4]} + \gamma_2 C S_t(.) \Pi_{]1/4,1/2]} = 0$
$\gamma_1 C \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]0,1/4]}, \Phi_n \succ \Phi_n \right) +$
$\gamma_2 C \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]1/4,1/2]}, \Phi_n \succ \Phi_n \right) = 0$
$\gamma_1 \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]0,1/4]}, \Phi_n \succ \Phi_n(\delta) \right) +$
$\gamma_2 \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]1/4,1/2]}, \Phi_n \succ \Phi_n(\delta) \right) = 0 \ \ \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} \left( \gamma_1 e^{\lambda_n t} \left( \int_0^{1/4} \sqrt{2} sin(n\pi x) dx \right) \sqrt{2} sin(n\pi\delta) \right) +$
$\sum_{n=1}^{\infty} \left( \gamma_2 e^{\lambda_n t} \left( \int_{1/4}^{1/2} \sqrt{2} sin(n\pi x) dx \right) \sqrt{2} sin(n\pi\delta) \right) = 0 \ \ \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} \left( \gamma_1 \frac{\sqrt{2}}{n\pi} e^{\lambda_n t} \left( (1 - cos(\frac{n\pi}{4}) \right) \sqrt{2} sin(n\pi\delta) \right) +$
$\sum_{n=1}^{\infty} \left( \gamma_2 \frac{\sqrt{2}}{n\pi} e^{\lambda_n t} \left( cos(\frac{n\pi}{4}) - cos(\frac{n\pi}{2}) \right) \sqrt{2} sin(n\pi\delta) \right) = 0 \ \ \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} (\gamma_1 \frac{\sqrt{2}}{n\pi} e^{\lambda_n t} ((1 - cos(\frac{n\pi}{4})) \sqrt{2} sin(n\pi\delta) +$
$\gamma_2 \frac{\sqrt{2}}{n\pi} e^{\lambda_n t} (cos(\frac{n\pi}{4}) - cos(\frac{n\pi}{2})) \sqrt{2} sin(n\pi\delta)) = 0 \ \ \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} \left( e^{\lambda_n t} \frac{2}{n\pi} sin(n\pi\delta) \left( \gamma_1 - \gamma_1 cos(\frac{n\pi}{4}) + \gamma_2 cos(\frac{n\pi}{4}) - \gamma_2 cos(\frac{n\pi}{2}) \right) \right) = 0.$
Based on analyticity
$e^{\lambda_n t} \frac{2}{n\pi} sin(n\pi\delta) \left( \gamma_1 - \gamma_1 cos(\frac{n\pi}{4}) + \gamma_2 cos(\frac{n\pi}{4}) - \gamma_2 cos(\frac{n\pi}{2}) \right) = 0\ \forall t \in ]0,T[$
$sin(n\pi\delta) \left( \gamma_1 - \gamma_1 cos(\frac{n\pi}{4}) + \gamma_2 cos(\frac{n\pi}{4}) - \gamma_2 cos(\frac{n\pi}{2}) \right) = 0$
$n\pi\delta \neq k\pi$ then $sin(n\pi\delta) \neq 0$
we deduce $\left( \gamma_1 - \gamma_1 cos(\frac{n\pi}{4}) + \gamma_2 cos(\frac{n\pi}{4}) - \gamma_2 cos(\frac{n\pi}{2}) \right) = 0.$
For $n = 2$, we obtain $\gamma_1 + \gamma_2 = 0$
$\gamma_1, \gamma_2 \geq 0$ then $\gamma_1 = \gamma_2 = 0$, so, $K_2$ is injective.
We have $K_2(\alpha_1, \alpha_2)^T = \bar{y}(.)$
$K_2^* K_2(\alpha_1, \alpha_2)^T = K_2^* \bar{y}(.)$
$K_2$ is injective then $K_2^* K_2$ is definite positive.
Finally, we conclude $(\alpha_1, \alpha_2)^T = (K_2^* K_2)^{-1} K_2^* \bar{y}(.)$

### 4.3 One-Dimensional System: $x_0$ Unknown on Several Subregions

Subdivision of the interval $]0,1[$ in just two or three parts may not be sufficient to assume that the density of the diffusing is constant in these subregions. Therefore, we divide the interval $]0,1[$ into $m$ equal parts where the density is supposed constant in each ones.

The values of the density of the diffusing material are unknown in $p$ areas, whereas they are known in $m-p$ areas.

**Theorem 4.** *$x_0$ is given by*

$$x_0(\theta) = \begin{array}{l} \alpha_i \quad if\ \theta \in \omega_i = \left]\frac{i-1}{m}, \frac{i}{m}\right] \quad for\ 1 \leq i \leq p \\ \beta_j \quad if\ \theta \in \omega_j = \left]\frac{j-1}{m}, \frac{j}{m}\right]\ for\ p+1 \leq j \prec m \\ \beta_m\ if\ \theta \in \omega_m = \left]\frac{m-1}{m}, 1\right[ \end{array} \tag{34}$$

*with $\alpha_i \alpha_{i+1} \leq 0$ for $1 \leq i \leq p-1$.*
*The system*

$$\begin{array}{l} \dot{x}\,(t) = Ax\,(t)\,;\ x(t) \in L^2(]0,1[) \\ x\,(0)\ \ is\ not\ well\ known \end{array} \tag{35}$$

*provided with $y(t) = Cx(t)$; is recalled observable on $[0,T]$.*
*Furthermore, $(\alpha_1, \alpha_2, ..., \alpha_p)^T = (K_p^* K_p)^{-1} K_p^* \bar{y}(.)$*
*where*

$$K_p : \begin{array}{ll} (R_1^s, R_2^s, .., R_p^s) & \longrightarrow L^2(0,T,R) \\ (\gamma_1, \gamma_2, .., \gamma_p)^T & \longrightarrow \sum_{i=1}^{p} \gamma_i C S_t \Pi_{\omega i} \end{array} \tag{36}$$

*$K_p^*$ is the adjoint operator of $K_p$.*
*$R_i^s = R^+ or R^-$, in addition to this $R_i^s$ and $R_{i+1}^s$ have different signs for $1 \leq i \leq p-1$.*

*Proof.* In this case, we put

$$K_p : \begin{array}{ll} (R_1^s, R_2^s, .., R_p^s) & \longrightarrow L^2(0,T,R) \\ (\gamma_1, \gamma_2, .., \gamma_p)^T & \longrightarrow \sum_{i=1}^{p} \gamma_i C S_t \Pi_{\omega i} \end{array} \tag{37}$$

with $R_i^s = R^+ or R^-$, further, $R_i^s$ and $R_{i+1}^s$ have different signs for $1 \leq i \leq p-1$. Consider internal pointwise sensor defined by

$$C : \begin{array}{ll} L^2(]0,1[) & \longrightarrow R \\ x(.) & \longrightarrow Cx(.) = x(\delta);\ \delta \in ]0,1[ \end{array} \tag{38}$$

Now, we study the injectivity of $K_p$
$K_p(\gamma_1, \gamma_2, ..., \gamma_p)^T = 0 \ \ \forall t \in ]0,T[$
$\sum_{i=1}^{p} \gamma_i C S_t(.) \Pi_{\omega_i} = 0 \ \ \forall t \in ]0,T[$
$\sum_{i=1}^{p} \gamma_i C \left(\sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{\omega_i}, \Phi_n \succ \Phi_n\right) = 0 \ \ \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} \left(\sum_{i=1}^{p} \gamma_i e^{\lambda_n t} \prec \Pi_{\omega_i}, \Phi_n \succ \Phi_n(\delta)\right) = 0.$
Based on analyticity $e^{\lambda_n t} \Phi_n(\delta) \left(\sum_{i=1}^{p} \gamma_i \prec \Pi_{\omega_i}, \Phi_n \succ\right) = 0 \ \ \forall t \in ]0,T[$

$n\pi\delta \neq k\pi$ then $\Phi_n(\delta) \neq 0$
we deduce $\sum_{i=1}^{p} \gamma_i \int_{\frac{i-1}{m}}^{\frac{i}{m}} sin(n\pi x)dx = 0$, $\sum_{i=1}^{p} \gamma_i \left[cos(\frac{n\pi(i-1)}{m}) - cos(\frac{n\pi i}{m})\right] = 0.$
For $n = m$, we obtain $\sum_{i=1}^{p}(-1)^{i+1}\gamma_i = 0$
knowing that $\gamma_i\gamma_{i+1} \leq 0$ for $1 \leq i \leq p-1$, we get $\gamma_i = \gamma_{i+1} = 0$ for $1 \leq i \leq p-1$.
Then, $K_p$ is injective.
We have $K_p(\alpha_1, \alpha_2, ..., \alpha_p)^T = \bar{y}(.)$, so $K_p^* K_p(\alpha_1, \alpha_2, ..., \alpha_p)^T = K_p^* \bar{y}(.)$
$K_p$ is injective then $K_p^* K_p$ is definite positive.
Finally, we conclude $(\alpha_1, \alpha_2, ..., \alpha_p)^T = (K_p^* K_p)^{-1} K_p^* \bar{y}(.)$

## 5 Diffusion Process in Two Dimensions with Internal Pointwise Sensor

Consider the diffusion equation in two dimensions, where the initial state $x_0$ is not well known.

$$\begin{gathered}\partial x(x_1, x_2, t)/\partial t = D\left(\frac{\partial^2 x(x_1,x_2,t)}{\partial x_1^2} + \frac{\partial^2 x(x_1,x_2,t)}{\partial x_2^2}\right) \\ ]0,1[^2 \times ]0,T[\end{gathered} \tag{39}$$

The boundary and initial conditions are given as well

$$x(x_1, 0, t) = x(x_1, 1, t) = x(0, x_2, t) = x(1, x_2, t) = 0 \;]0,1[^2 \times ]0,T[ \tag{40}$$

$$x(x_1, x_2, 0) = x_0(x_1, x_2) \text{ supposed not well known} \tag{41}$$

The system (39)–(41) can be written as a state-space system

$$\dot{x}(t) = Ax(t);\, x(0) = x_0 \tag{42}$$

where $A = D\left(\partial^2/\partial x_1^2 + \partial^2/\partial x_2^2\right)$.
$A$ infinitesimal generator of strongly continuous semi-group $(S\,(t))_{t\geq 0}$ given by

$$S_t : \begin{array}{ll} L^2(]0,1[^2) \longrightarrow L^2(]0,1[^2) \\ f \qquad\qquad \longrightarrow \sum_{n=1}^{\infty} e^{\lambda_n t} \prec f, \Phi_n \succ \Phi_n \end{array} \tag{43}$$

with

$$\begin{array}{rcl} \lambda_n & = & -(n^2+1)\pi^2 \\ \Phi_n(x, y) & = & 2sin(n\pi x)sin(\pi y) \end{array} \tag{44}$$

In addition to this, we consider internal pointwise sensor given by

$$C : \begin{array}{ll} L^2(]0,1[^2) \longrightarrow R \\ x(.) \qquad\quad \longrightarrow Cx(.) = x(\delta_1, \delta_2);\, (\delta_1, \delta_2) \in ]0,1[^2 \end{array} \tag{45}$$

### 5.1 Two-Dimensional System: $x_0$ Unknown on Several Subregions

In this subsection, the area $]0,1[\times]0,1[$ will be divided into $mm'$ rectangular subregions. We assume that the density of the diffusing material evolves very slowly in each subregion as we can consider it constant. The values of the density of the diffusing material are known in $mm'-pq$ subregions $\omega_{ij}$ where $p+1\leq i\leq m$ and $q+1\leq j\leq m'$, but we look for its value in $pq$ subregions $\omega_{ij}$ where $1\leq i\leq p$ and $1\leq j\leq q$.

**Theorem 5.** *$x_0$ is given by*

$$x_0(\theta)=\begin{cases}\alpha_{ij} & \textit{if } \theta\in\omega_{ij}=\left]\frac{i-1}{m},\frac{i}{m}\right]\times\left]\frac{j-1}{m'},\frac{j}{m'}\right] \quad \textit{for } 1\leq i\leq p \textit{ and } 1\leq j\leq q\\ \beta_{ij} & \textit{if } \theta\in\omega_{ij}=\left]\frac{i-1}{m},\frac{i}{m}\right]\times\left]\frac{j-1}{m'},\frac{j}{m'}\right] \quad \textit{for } p+1\leq i\prec m \textit{ and } q+1\leq j\prec m'\\ \beta_{mm'} & \textit{if } \theta\in\omega_{mm'}=\left]\frac{m-1}{m},1\right[\times\left]\frac{m'-1}{m'},1\right[\end{cases} \tag{46}$$

*$x(0)$ is not well known: the matrix $B=(\alpha_{i,j})_{\substack{1\leq i\leq p\\1\leq j\leq q}}$ is unknown, but $\beta_{i,j}$ with $p+1\leq i\leq m$; $q+1\leq j\leq m'$, are known.*
*The system*

$$\begin{array}{l}\dot{x}(t)=Ax(t)\,;\ x(t)\in L^2(]0;1[^2)\\ x(0)\ \textit{is not well known}\end{array} \tag{47}$$

*provided with $y(t)=Cx(t)$, is recalled observable on $[0,T]$ if*

$$K_{pq}:\begin{array}{ll} M_{p,q} & \longrightarrow L^2(0,T,R)\\ B=(\gamma_{ij})_{\substack{1\leq i\leq p\\1\leq j\leq q}} & \longrightarrow \sum_{j=1}^{q}\sum_{i=1}^{p}\gamma_{ij}CS_t(.)\Pi_{\omega ij}\end{array} \tag{48}$$

*injective. Furthermore, $B=(K_{pq}^{*}K_{pq})^{-1}K_{pq}^{*}\bar{y}(.)$*
*$K_{pq}^{*}$ is the adjoint operator of $K_{pq}$.*

*Proof.* We have

$$\begin{aligned}y(t)&=Cx(t)\\&=CS_tx_0\\&=CS_t\left(\textstyle\sum_{j=1}^{q}\sum_{i=1}^{p}\alpha_{ij}\Pi_{\omega ij}\right)+CS_t\left(\textstyle\sum_{j=q+1}^{m'}\sum_{i=p+1}^{m}\beta_{ij}\Pi_{\omega ij}\right)\end{aligned}$$

in this case, we put

$$\bar{y}(t)=y(t)-CS_t\left(\sum_{j=q+1}^{m'}\sum_{i=p+1}^{m}\beta_{ij}\Pi_{\omega ij}\right) \tag{49}$$

and

$$K_{pq}:\begin{array}{ll} M_{p,q} & \longrightarrow L^2(0,T,R)\\ B=(\gamma_{ij})_{\substack{1\leq i\leq p\\1\leq j\leq q}} & \longrightarrow \sum_{j=1}^{q}\sum_{i=1}^{p}\gamma_{ij}CS_t(.)\Pi_{\omega ij}\end{array} \tag{50}$$

we obtain $K_{pq}B=\bar{y}(.)$.
If $K_{pq}$ injective then $K_{pq}^{*}K_{pq}$ definite positive, we can conclude $B=(K_{pq}^{*}K_{pq})^{-1}K_{pq}^{*}\bar{y}(.)$.

### 5.2 Two-Dimensional System: $x_0$ Unknown on one Subregion

The area $]0,1[\times]0,1[$ is divided into two rectangular subregions $]0,1/2]\times]0,1[$ and $]1/2,1[\times]0,1[$. The value of the diffusing material in subregion $]1/2,1[\times]0,1[$ is known and we look for its value in the second one.

**Theorem 6.** *$x_0$ is given by*

$$x_0(\theta) = \begin{matrix} \alpha \; if \;\; \theta \in ]0,1/2]\times]0,1[ \\ \beta \; if \;\; \theta \in ]1/2,1[\times]0,1[ \end{matrix} \tag{51}$$

*$x\,(0)$ is not well known : $\alpha$ is unknown whereas $\beta$ is known.*
*The system*

$$\begin{matrix} \dot{x}\,(t) = Ax\,(t)\,; \;\; x(t) \in L^2(]0,1[^2) \\ x\,(0) \;\; is \; not \; well \; known \end{matrix} \tag{52}$$

*provided with $y(t) = Cx(t)$; is recalled observable on $[0,T]$.*
*Furthermore, $\alpha = (K_d^* K_d)^{-1} K_d^* \bar{y}(.)$*
*where*

$$K_d : \begin{matrix} R \longrightarrow L^2(0,T,R) \\ \gamma \;\; \longrightarrow \gamma C S_t(.) \Pi_{]0,1/2]\times]0,1[} \end{matrix} \tag{53}$$

*$K_d^*$ is the adjoint operator of $K_d$.*

*Proof.* In this case, we put

$$K_d : \begin{matrix} R \longrightarrow L^2(0,T,R) \\ \gamma \;\; \longrightarrow \gamma C S_t(.) \Pi_{]0,1/2]\times]0,1[} \end{matrix} \tag{54}$$

Now, we study the injectivity of $K_d$
$K_d \gamma = 0 \;\; \forall t \in ]0,T[$
$\gamma C S_t(.) \Pi_{]0,1/2]\times]0,1[} = 0 \;\; \forall t \in ]0,T[$
$\gamma C \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]0,1/2]\times]0,1[}, \Phi_n \succ \Phi_n \right) = 0 \;\; \forall t \in ]0,T[$
$\gamma \left( \sum_{n=1}^{\infty} e^{\lambda_n t} \prec \Pi_{]0,1/2]\times]0,1[}, \Phi_n \succ \Phi_n(\delta_1,\delta_2) \right) = 0 \;\; \forall t \in ]0,T[$
$\sum_{n=1}^{\infty} \left( \gamma e^{\lambda_n t} \left( \int_0^{1/2} 2sin(n\pi x)dx \right) \left( \int_0^1 sin(\pi y)dy \right) 2sin(n\pi\delta_1) sin(\pi\delta_2) \right) = 0.$
Based on analyticity
$\gamma e^{\lambda_n t} \left( \int_0^{1/2} 2sin(n\pi x)dx \right) \left( \int_0^1 sin(\pi y)dy \right) 2sin(n\pi\delta_1) sin(\pi\delta_2) = 0 \;\; \forall t \in ]0,T[$
$n\pi\delta_1 \neq k\pi$ and $\pi\delta_2 \neq k\pi$, then $sin(n\pi\delta_1)sin(\pi\delta_2) \neq 0$
we deduce $\gamma \left( cos(\frac{n\pi}{2}) - 1) \right) = 0$.
For $n = 1$, we have $\gamma = 0$, so $K_d$ is injective.
We have $K_d \alpha = \bar{y}(.)$, so $K_d^* K_d \alpha = K_d^* \bar{y}(.)$
$K_d$ is injective, then $K_d^* K_d$ is definite positive.
Finally, we conclude $\alpha = (K_d^* K_d)^{-1} K_d^* \bar{y}(.)$.

## 6 Conclusions and Perspectives

This paper establishes various results related to observability of thermal process in two different geometrical situations: one-dimensional and two-dimensional system. Various interesting results concerning reconstruction of initial state $x_0$, not well known, are given and illustrated in different specific situations.

It's very important to clarify that this new approach is not limited to diffusion process, but can be extended to many other distributed parameter systems such as: fluid process, convection-diffusion-reaction process, flexible beam, polymerization process, etc. On top of that, we can also make use of different kind of measurements depending on the number and structure of the sensors: internal zone sensors, zone boundary sensors, pointwise boundary sensors, etc.

## References

1. Christofides, P.D.: Control of nonlinear distributed process systems: recent developments and challenges. AIChE J. **47**(3), 514–518 (2001)
2. Banerjee, S., Cole, J.V., Jensen, K.F.: Nonlinear model reduction strategies for rapid thermal processing systems. IEEE Trans. Semicond. Manuf. **11**(2), 266–275 (1998)
3. Li, H.X., Guan, S.P.: Hybrid intelligent control strategy-Supervising a DCS - controlled batch process. IEEE Control Syst. Mag. **21**(3), 36–48 (2001)
4. Holmes, P., Lumley, J.L., Berkooz, G.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry, 1st edn. Cambridge University Press, New York (1996)
5. Li, H.X., Liu, J., Chen, C.P., Deng, H.: A simple model-based approach for fluid dispensing analysis and control. IEEE/ASME Trans. Mechatron. **12**(4), 491–503 (2007)
6. Hong, Y.P., Li, H.X.: Comparative study of fluid dispensing modeling. IEEE Trans. Electron. Packag. Manuf. **26**(4), 273–280 (2003)
7. Christofides, P.D.: Nonlinear and Robust Control of Partial Differential Equation Systems: Methods and Applications to Transport-Reaction Processes. Birkhauser, Boston (2001)
8. Fleming, A.J., Moheimani, S.O.R.: Spatial system identification of a simply supported beam and a trapezoidal cantilever plate. IEEE Trans. Control Syst. Technol. **11**(5), 726–736 (2003)
9. Halim, D., Moheimani, S.O.R.: Spatial resonant control of flexible structures - application to a piezoelectric laminate beam. IEEE Trans. Control Syst. Technol. **9**(1), 37–53 (2001)
10. Demetriou, M.A.: Integrated actuator-sensor placement and hybrid controller design of flexible structures under worst case spatiotemporal disturbance variations. J. Intell. Mater. Syst. Struct. **15**, 901–921 (2004)
11. Amourox, M., Eljai, A., Zerrik, E.: Regional observability of distributed systems. Int. J. Syst. Sci. **2**, 301–313 (1994)

# Using the Method of System Dynamics to Forecast Additional Manpower Needs in Murmansk Region

Vitaliy Bystrov[1], Svetlana Malygina[1,2], and Darya Khaliullina[1(✉)]

[1] Institute for Informatics and Mathematical Modelling of Technological Processes,
Kola Science Center Russian Academy of Sciences, Apatity, Murmansk Region, Russia
{bystrov,malygina,khaliullina}@iimm.ru

[2] Apatity Branch of Murmansk Arctic State University, Apatity, Murmansk Region, Russia

**Abstract.** The research has application-oriented character and is aimed at solving practical task of forecasting additional manpower needs in the regional economy of the Murmansk region. The authors propose to integrate the existing methods for determining manpower needs with simulation modeling. Models of system dynamics (simulation models) are used to design tools to determine a number of employed people and job vacancies for each type of economy activity. The peculiarity of the developed tools is the possibility to consider large regional investment projects.

**Keywords:** System dynamics · Forecasting · Manpower needs · Multimodel complex

## 1 Introduction

At present time, management of the regional workforce capacity remains an urgent task. This is due to the fact that such management is the most important factor influencing the development of social and economic spheres of a region. There are many different researches in the field of formation of regional personnel policy but some issues still are disputable, in particular, the problem of forecasting manpower needs.

When using the notion of manpower needs (MN) it is necessary to understand accurately what this notion means. So the notion of general MN means total number of workers, specialists and clerks that are required for release of the planned volume of production (goods and/or services) in the region. This notion includes current (basic) MN and additional MN. The notion of additional manpower needs (AMN) reflects the number of workers, specialists and clerks that are required in planning period in addition to the existing number of workers, specialists and clerks at the beginning of the period. AMN arise from workers retirement from enterprises for various reasons, and from creation of new jobs caused by launch of new production facilities or expansion of existing enterprises, including as a result of implementation of investment projects.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_28

## 2 Background

Additional manpower needs are the most interesting for the regional authorities to plan personnel because AMN are the reason of fluctuations in the annual balance of manpower resources. It should be noted that a change of this type of needs occurs in conditions of partial uncertainty which is caused by the following factors: presence of migration processes in the region, changing of social and economic attractiveness of the region, climatic conditions and ecological situation, etc.

Currently several Russian researches techniques allowing to estimate and/or forecast the correspondence between employment market and personnel training of secondary professional and higher education are offered. These techniques are based in different approaches and can be divided into 3 groups [1]:

1. Techniques are based on expert assessments (Matushkina N.N., Stolbova I.D., Shcheglov, P.E., Nikitina N.Sh., etc.).
2. Analytical techniques are based on data from statistical offices and take into account the development programs of the region and the country (Korovkin A.G., Gurtov V.A., etc.).
3. Mixed techniques combine statistical reports and expert knowledge (Sangadiev Z.G., Skotnikov S.N., Zhirnov A.Yu., etc.).

Foreign works also represent researches in the field of determining of balance of manpower resources and macroeconomic modeling for medium-term forecasting of manpower component that is necessary for stable development of national economies. Researches in the following countries are most brightly presented: United States (Bureau of labor statistics of U.S. Department of Labour), Germany (INFORGE and Ifo models), Australia (MONASH model), UK (model MDM) and others.

The considered models of forecasting of manpower demand have a number of general characteristics. Most of these techniques are applied at the regional level and use econometric approach based on the notion of “required manpower resources”. Also they use the results of the macroeconomic forecast of production of goods and services according to economy branches as input parameters for models [2].

Shortcoming of the existing techniques is the lack of consideration of the following parameters: processes of natural migration of population, the level of demand for specialists, the existing educational structure and others. In particular, techniques based only on statistical data are not accurate enough to forecast future indicators because they do not take into account the development strategy of enterprises and the region as a whole. Expert methods are more accurate but more labor-consuming and more time-consuming; they are also very subjective [1]. These methods (expert methods and techniques based only on statistical data) cooperatively used allow assessing the correspondence between employment market and system of personnel training.

## 3 Technique to Determine Additional Manpower Needs

This paper represents computation of forecast values of manpower needs of socio-economic system in Murmansk region. The computation was carried out with combined use of several calculation methods:

- the official technique for development of the forecast balance of manpower resources;
- the technique of Budget monitoring center (BMC) of Petrozavodsk state university (PetrSU);
- methods of data extrapolation;
- simulation modeling;
- expert assessments;
- compliance matrixes.

The process of determining of additional manpower needs of economy branches in Murmansk region can be dividing into seven stages [3]:

1. Processing of data (the number released jobs, of vacancies, employed people, of appeals to Employment Center of Murmansk region) in the context of OKPDTR (Russian Classification of Workers' and Employees' Occupations and Wage Grades) that were obtained in the period from 2007 to 2015.
2. Creating of time series of workers professions and employees occupations by means of methods of mathematical statistics.
3. Updating time series by data obtained from questionnaires filled by employers. These data contain information about planning and realization of investment projects in Murmansk region.
4. Obtaining forecasts of manpower needs on enterprises using system dynamics models and agent-based models that are constructed in the conditions of incomplete information about the planned investment projects.
5. Combining data obtained at stages 3 and 4, and carrying out an expert assessment of the generalized results of the computation of AMN.
6. Creating of compliance matrixes. The matrixes are used to transformation of obtained data represented in the context of OKPDTR to data in the context of types of economic activity (OKVED - Russian Classification of Economic Activities). This data are the information about manpower needs of enterprises and organizations in Murmansk region.
7. Computation of additional manpower needs in the context of types of economic activity and education levels using created compliance matrixes.

The main stages of the proposed methodology are presented in Fig. 1.

The proposed technique allowed us to obtain the table of distribution of additional manpower needs on professions of workers and employees occupations in the context of types of economic activity and education levels. The size of the matrix made up 2168 on 160.

In the offered technique simulation modeling is used as means that compliments information about possible scenarios of regional economy development. In particular, the technique compliments information about changes of manpower needs in separate

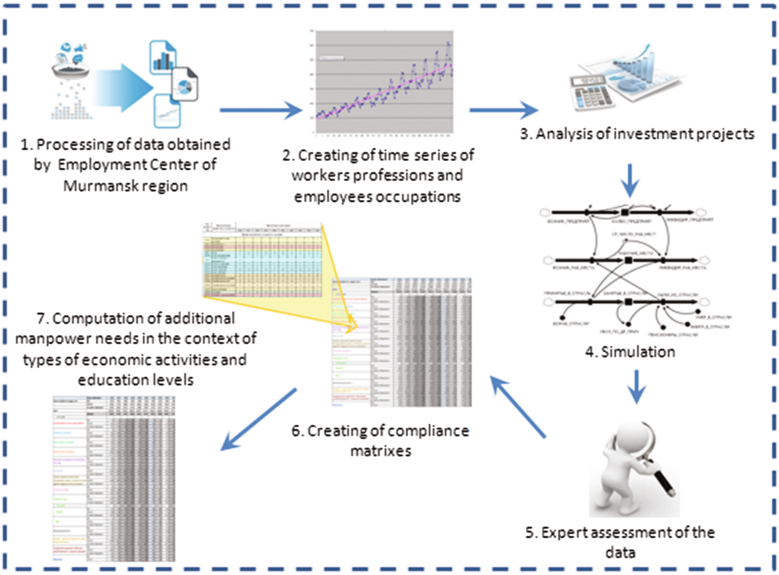

1. Processing of data obtained by Employment Center of Murmansk region
2. Creating of time series of workers professions and employees occupations
3. Analysis of investment projects
4. Simulation
5. Expert assessment of the data
6. Creating of compliance matrixes
7. Computation of additional manpower needs in the context of types of economic activities and education levels

**Fig. 1.** The main stages of the methodology of determining of additional manpower needs of economy branches

enterprises and organizations in Murmansk region. To solve such task, complex of simulation models was developed. The complex can structurally be presented as the following generalized components (units):

- «Regional employment market (official and illegal)»;
- «System of education (training/retraining of personnel)»;
- «Demography»;
- «Regional economy»;
- «Investment project».

A conceptual scheme of the model complex is presented in Fig. 2.

Simulation model «Regional employment market» allows determining demand and supply of manpower resources in Murmansk region and comparing them. Demand is formed by vacancies at enterprises and organizations of the region. Supply is a set of manpower resources that were received from the unit «System of education (training/ retraining of personnel)» and the unemployed. Manpower resources can be changed by the processes occurring in the unit «Demography».

The component «System of education (training/retraining of personnel)» takes into account the regional processes of qualified personnel training (graduates). Such processes include training with secondary professional and higher education, and vocational training in the enterprise. Input parameters of this unit are the number of matriculates and also the number of the people directed to retraining. The output of the unit is the distribution of qualified manpower resources by level of education and integrated groups of directions of education.

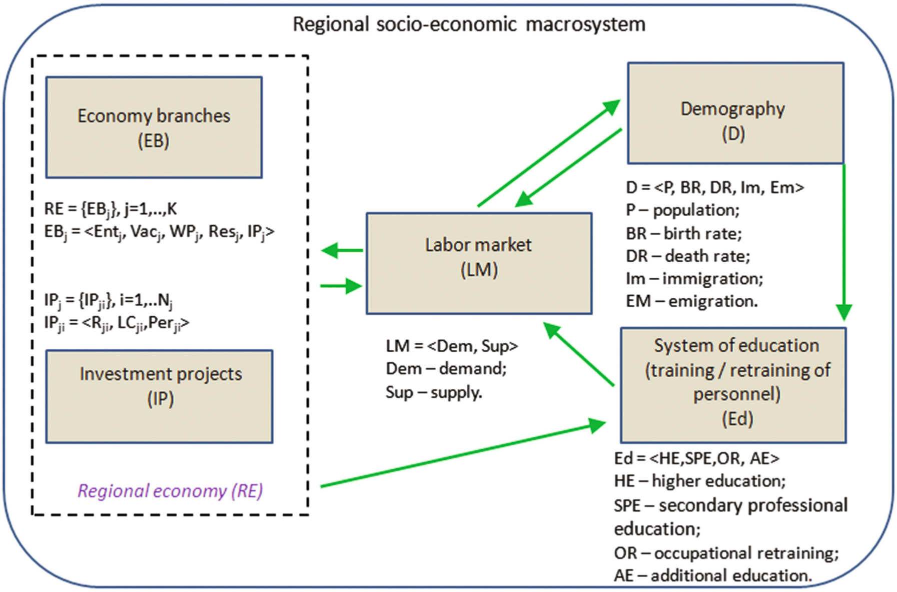

Regional socio-economic macrosystem
Economy branches (EB)
RE = {EB$_j$}, j=1,..,K
EB$_j$ = <Ent$_j$, Vac$_j$, WP$_j$, Res$_j$, IP$_j$>
IP$_j$ = {IP$_{ji}$}, i=1,..N$_j$
IP$_{ji}$ = <R$_{ji}$, LC$_{ji}$,Per$_{ji}$>
Investment projects (IP)
Regional economy (RE)
Labor market (LM)
LM = <Dem, Sup>
Dem – demand;
Sup – supply.
Demography (D)
D = <P, BR, DR, Im, Em>
P – population;
BR – birth rate;
DR – death rate;
Im – immigration;
EM – emigration.
System of education (training / retraining of personnel) (Ed)
Ed = <HE,SPE,OR, AE>
HE – higher education;
SPE – secondary professional education;
OR – occupational retraining;
AE – additional education.

**Fig. 2.** A conceptual scheme of the model complex

The unit «Demography» represents simulation model of population distribution by age groups. This model reflects the demographic processes in the region. Input parameters of the unit can be conditionally divided into such categories as:

- economic attractiveness (average salary in the region, availability of housing, vacancies, etc.),
- social attractiveness (accessibility of medical care, educational services, crime rates, etc.),
- ecological situation.

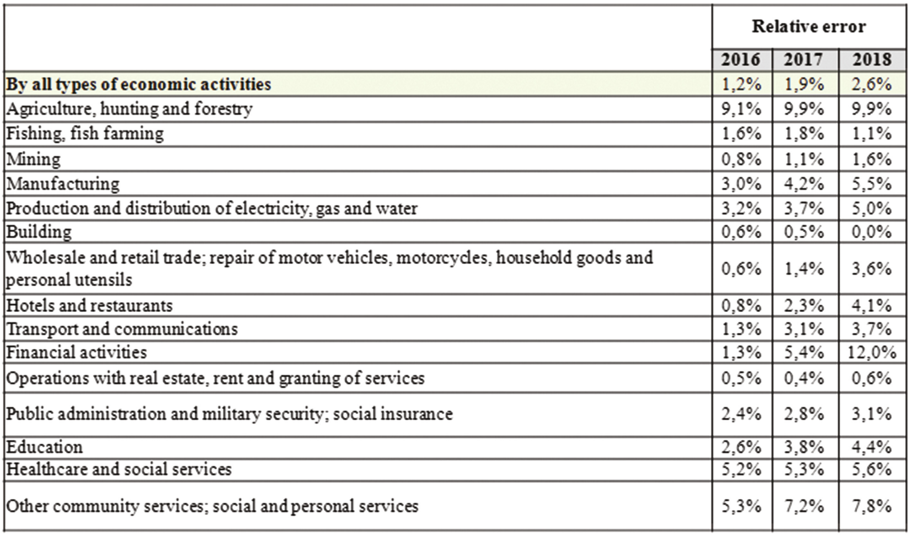
The output of the unit is the number of economically active population and basic demographic indicators (fertility, mortality, migration, etc.).

The component «Regional economy» is a set of agents implemented in the form of simulation models. These models describe the functioning of economic branches in accordance with the main type of economic activity. The main attention is paid to processes of forming demand on manpower resources in enterprises and organizations of the industry. The demand is estimated in the conditions of the normal mode of enterprises work without considering significant changes in their activities. Macroeconomic indicators for each type of economic activity (for example, industry share in GRP, GRP, total energy consumption by the branch of industry, etc.) are used as input parameters. The main output parameter of the unit is vacant workplaces of branch by category of workers.

The unit «Investment project» is intended for modeling of main stages of life cycle of investment project (IP). This unit represents a set of model agents each of which is a system dynamics model. In this research IP is regarded as starting of new enterprises or modernization of the existing enterprises. Such processes are accompanied by the creation of additional workplaces. Input parameters of this unit are the volume of investments, the planned manpower sources, deadlines and others. The main objective of this unit is obtaining the forecast of additional manpower needs necessary to implement planned investment projects.

## 4 Results

The presented units are parts of multimodel complex. Verification of this complex was carried out by comparing the forecasted data distribution of manpower resources with the balance of manpower resources, calculated by the Ministry of Economic Development of the Murmansk region according to the official technique. Manpower resources were categorized by types of economic activity. The forecast was made for the period from 2016 to 2018. On average, the data values of the real research differ from the forecast of the Ministry of Economic Development from 0.5 to 9.6% depending on the type of economic activity (Fig. 3).



| | Relative error | | |
|---|---|---|---|
| | 2016 | 2017 | 2018 |
| **By all types of economic activities** | 1,2% | 1,9% | 2,6% |
| Agriculture, hunting and forestry | 9,1% | 9,9% | 9,9% |
| Fishing, fish farming | 1,6% | 1,8% | 1,1% |
| Mining | 0,8% | 1,1% | 1,6% |
| Manufacturing | 3,0% | 4,2% | 5,5% |
| Production and distribution of electricity, gas and water | 3,2% | 3,7% | 5,0% |
| Building | 0,6% | 0,5% | 0,0% |
| Wholesale and retail trade; repair of motor vehicles, motorcycles, household goods and personal utensils | 0,6% | 1,4% | 3,6% |
| Hotels and restaurants | 0,8% | 2,3% | 4,1% |
| Transport and communications | 1,3% | 3,1% | 3,7% |
| Financial activities | 1,3% | 5,4% | 12,0% |
| Operations with real estate, rent and granting of services | 0,5% | 0,4% | 0,6% |
| Public administration and military security; social insurance | 2,4% | 2,8% | 3,1% |
| Education | 2,6% | 3,8% | 4,4% |
| Healthcare and social services | 5,2% | 5,3% | 5,6% |
| Other community services; social and personal services | 5,3% | 7,2% | 7,8% |

**Fig. 3.** The average error of the forecast of additional manpower needs in the context of types of economic activity

The total error for all types of economic activity amounted to 3.5%. In general, nature of changes in the number of employed population of the Murmansk region in both cases is the same. Exceptions are such types of economic activity as wholesale and retail trade, financial activities, public administration and military security, providing other community services, social and personal services. In such cases there is insignificant divergence in forecasts.

Applying of the developed technique allowed us to obtain the forecast of additional manpower needs in enterprises and organizations in Murmansk region. The forecast was obtained in the different contexts (education level, types of economy activity). In particular, Fig. 4 shows the forecast of the number of workers with different levels of education that are required in the region. The forecast interval is 10 years.

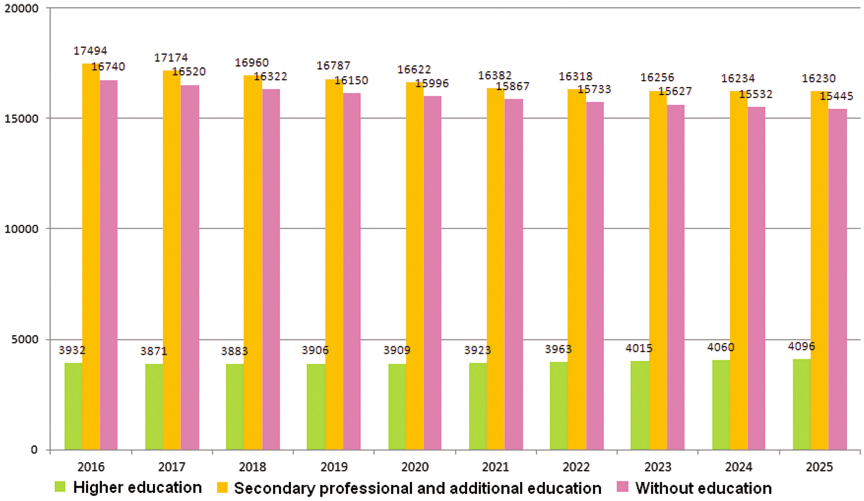


**Fig. 4.** Forecast of additional manpower needs in the context of education level

Analyzing the obtained results, we can conclude that the number of employees with higher education increases slightly (by approximately 100 people to the year 2025). The number of vacancies for other categories of employees at the enterprises and organizations of the Murmansk region will decrease.

## 5 Conclusion

Received long-term forecast of the additional manpower needs shows that the number of required employees of organizations and enterprises of the Murmansk region will gradual decline. Such dynamics of this indicator is observed in the estimates of experts in the field of personnel policy of the region.

However, the reduction of AMN will not cause of deterioration of the social tensions. This is due to the fact that migration processes and processes of natural population decline result to a reduction in the total number of economically active population by 2025 [3].

The nature of the processes occurring in the labour market of the Murmansk region necessitates all stakeholders (employers, municipal authorities and regional authorities, educational institutions) to solve tasks of effective management of available human resources. The solution of this task requires use science-based methods and technologies for the estimation of possible scenarios of regional economy development.

## References

1. Il'ina, L.A., Prosvirina, D.A.: Evolution to develop of techniques of forecasting of regional manpower needs. Vestnik Samara State Univ. Econ. **12**(134), 57–62 (2015)
2. Moroz, D.M., Pitukhin, E.A., Sigova, S.V.: The methodology for forecasting the needs of workers in the context of economy branches. In: Collection of Papers on Materials of the Twelfth all-Russian Scientific-Practical Internet-Conference «Supply and Demand in the Labor Market and Educational Services Market in Russian Regions», vol. 1, pp. 124–143 (2015)
3. Khaliullina, D.N., Bystrov, V.V., Malygina, S.N.: Forecasting additional staffing needs of the Murmansk region economy branches. In: Proceedings of the Kola Science Centre RAS. Information Technology, vol. 7, pp. 94–107 (2016)

# Ordinary Kriging and Spatial Autocorrelation Identification to Predict Peak Ground Acceleration in Banda Aceh City, Indonesia

Rokhana D. Bekti[1], Edy Irwansyah[2(✉)], Bayu Kanigoro[2], and Theodorick[3]

[1] Department of Statistic, Institut Sains and Teknologi AKPRIND, Yogyakarta 55222, Indonesia

[2] School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
{eirwansyah,bkanigoro}@binus.edu

[3] Department of Statistic, Bina Nusantara University, Jakarta 11480, Indonesia

**Abstract.** Peak ground acceleration (PGA) is a measure of earthquake acceleration in the ground. The prediction information about PGA is important to minimize the effect of earthquake. The method for prediction is Ordinary Kriging. It is geostatistic method used to predict data in certain locations which have autocorrelation. The sample data used in this research are PGA in Meuraxa, Banda Aceh 2006. The steps of research methodology consist of autocorrelations identify by Moran's $I$ and LISA, build semivariograms, and prediction by Ordinary Kriging. The results is Ordinary Kriging can be applied to predict PGA. It was shown by evaluate of mean and MSE value. According to mean value of three prediction, all models (Gaussian, Spherical, and Exponential) have mean 0,3534; 0,3584; and 0,3555 which approaches the actual PGA mean 0.34. According to MSE value, it can be seen that all models have small MSE or relatively closed to zero.

## 1 Introduction

There are some location in Indonesia which is vulnerable to natural disasters, such as earthquake. Earthquake is an earth vibrant event which is caused by a sudden explosion of energy inside earth. It is indicated by the cracks of rock layers in earth crust [1]. Sumatera Island is one of the islands that experiences earthquake regularly. It is caused by its position that is near to the path where two tectonic plates collide. In the last six years, there are many earthquakes that happen in Indonesia, the most terrifying one is Aceh Earthquake happened in 2004, which is also followed by tsunami. The impact of the earthquake and tsunami was severe loss of life and property, and severe environmental damage. The most casualties as a result of earthquake followed by tsunami in Aceh are 110,229 people died, 12,123 people missing and 703,518 people evacuated [2].

The earthquake hazard assessment can be performed using the acceleration value in the ground or peak ground acceleration (PGA) [2,3]. PGA is a scale used to measure the speed at ground level. According to Seismic Hazard Analysis

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*, Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_29

(PSHA) research, PGA in Aceh region is 0.3–0.4 g. This number is quite high and it explains that Aceh region is potentially going to experience earthquake in the future. Then, the prediction information about PGA is important to minimize the effect of earthquake.

The information about PGA characteristic caused by earthquake can be obtained from the records of earthquake events in the past. There are some methods for calculate PGA value, such as using the attenuation function [3–5]. This method can be used if there are the data about distance to the location of epicenter, depth of the earthquake, and magnitude. The problem was how to predict PGA in certain locations which no information about it. The spatial statistics methods can be used to perform it.

The current spatial statistic and geostatistic have been developed and able to both explain and analyze variance caused by natural and artificial (human-made) phenomenon in air, under sea, and on earth surface. Geostatistic has been applied in many fields, one of them is geology. Geostatistic is used to predict data at locations that are yet to be measured [6]. Spatial statistics perform based on autocorrelation among locations. One of the prediction methods for geostatistic is Kriging which used to utilize spatial interpolation in a certain region to estimate values in other non-sampled regions. There are several Kriging methods that are commonly used, such as Ordinary Kriging, CoKriging, and Robust Kriging. Ordinary Kriging is use for one variable and can be use to predict PGA. The research about kriging is [2,7–9].

Based on the problems, the purpose of this research is to predict PGA in Banda Aceh based on the desired location points. The method which used is Ordinary Kriging. It predicts PGA based on spatial characteristics or autocorrelation spatial among locations.

## 2 Autocorrelation Spatial and Ordinary Kriging

PGA value at one location has relationship with other locations. The location which closed to earthquake epicenters relatively has high PGA. It was show that there an autocorrelation among locations. It corresponds to the first law of geography by Tobler, "Everything is related to everything else, but near things are more related than distant things" [10]. Spatial method can be used to analyze the autocorrelation in PGA. Spatial method is a method to get information of observations influenced by space or location effect. The spatial methods for identify autocorrelation spatial are Moran's $I$ and Local Indicator of Spatial Association (LISA).

Moran's $I$ coefficient is used to test the spatial dependence or autocorrelation between observations or location [11,12]. The formula of Moran's $I$ is Eq. (1).

$$I = \frac{n}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}} \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum\limits_{i=1}^{n}(x_j - \bar{x})^2} \tag{1}$$

The $x_i$ is the variable in location $i$, $\bar{x}$ is mean of $x_i$, $w_{ij}$ is weighted of $i$ and $j$ location, and $n$ is the total locations. The value of Moran's $I$ is between $-1$ and 1. Value $I > -\frac{1}{n-1}$ is shows the positive autocorrelation and $I < I_o$ is shows the negative autocorrelation [11].

LISA is used to identify autocorrelation in each local locations [12]. It use the hypothesis test

$$H_0 : \ I_i = 0 \text{ (no autocorrelation among location)}$$
$$H_1 : \ I_i \neq 0 \text{ (autocorrelation among location)}$$

The conclusion is reject $H_0$ if P-value of LISA test less than significance level $\alpha$.

Ordinary Kriging is geostatistic method which is used to predict data in certain location. Consider that a random variable $Z$ has been measured at sampling points or locations, $x_i$ with $i$ is location $i = 1, 2, 3, \ldots, n$. It uses to predict or estimate the value at a point $x_0$ [7]. The data point $Z(x_i)$ was obtained from coordinate latitude and longitude in $i$ location. The prediction is,

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2}$$

where,

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{3}$$

The value of $\lambda_i$ is calculate from Eq. (4). $C$ is covariance matrix among data point $x_i$ and $D$ is the vector of covariance between data point $(x_i)$ and the estimation target $(x_0)$.

$$\lambda = C^{-1} D \tag{4}$$

The covariance matrix $C$ can be estimated from variogram model, such as Gaussian, Spherical, Pentaspherical, Exponential, or Stable exponential model. In these models, there are parameter nugget, sill, and range. These parameters can be performing by empirical semivariogram in Fig. 1. Semivariogram is a plot of semivariance versus lag. It commonly represented as a graph that shows the variance in measure with distance between all pairs of sample locations. It also indicates spatial correlation in observations measured at sample locations. The empirical semivariance can be calculated by Eq. (5) [6].

$$\hat{\gamma} = \frac{1}{2n(h)} \sum_{\alpha=1}^{n(h)} [z(x_i + h) - z(x_i)]^2 \tag{5}$$

where $z(x_i)$ and $z(x_i + h)$ are the actual values of $Z$ at location $(x_i)$ and $(x_i + h)$, and $n(h)$ is the number of paired comparisons at lag $h$.

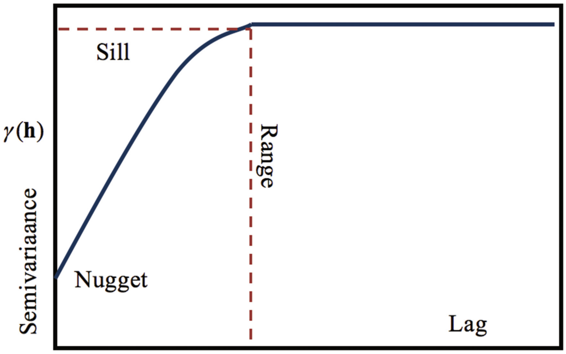


**Fig. 1.** Semivariogram

## 3 Research Methodology

This research was use the secondary data. It was from Sengara research in 2006 about microzonation and hazard mapping of Meuraxa District - Banda Aceh [13]. It was also from Meteorological, Climatological and Geophysics Agency. The methods for predict PGA is Ordinary Kriging. All the calculation was done by *spdep* and *gstat* package in R [14]. The steps are:

1. Descriptive analysis which is done to know the characteristic of PGA through average, minimum value and maximum value of actual data.
2. Calculate Moran's $I$ by Eq. (1) and LISA to identify the autocorrelation spatial.
3. Prediction by Ordinary Kriging.
   (a) Partitions of data by randomly selected data, 17 were actual data and 3 were testing data.
   (b) Determine empirical semivariogram, which its purpose to show the characteristic of spatial correlation among locations and to determine the initial value of sill, range and nugget.
   (c) Determine empiric semivariogram, which consist of Gaussian, Spherical and Exponential.
   (d) Calculate the prediction of PGA.
   (e) Calculate Mean Square Error (MSE) to evaluate the prediction.

## 4 Results and Discussion

Fig. 2 shows PGA explorations, and the location of the actual (training) data and testing data for predict the PGA. In this figure there are 17 actual data and the 3 testing data. The data testing is locations 18,19, and 20. The PGA average of actual data was 0,34. It implies that the average speed of earthquake from earthquake epicenters toward towards the points surrounding area was 0,34. The center point was at the latitude 5,553 and longitude 95,31. The minimum value

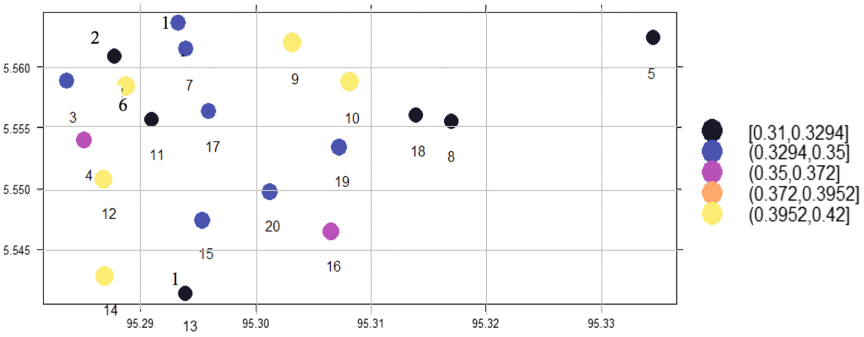


**Fig. 2.** PGA Exploration

**Table 1.** Result of LISA Test

| Location | Latitude | Longitude | PGA | P value | Location | Latitude | Longitude | PGA | P-Value |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.56363 | 95.2933 | 0.33 | 0.296 | 11 | 5.55568 | 95.2909 | 0.32 | 0.425 |
| 2 | 5.56092 | 95.2877 | 0.32 | 0.473 | 12 | 5.55081 | 95.2867 | 0.40 | 0.263 |
| 3 | 5.55884 | 95.2836 | 0.33 | 0.298 | 13 | 5.54139 | 95.2938 | 0.31 | 0.305 |
| 4 | 5.554 | 95.2851 | 0.36 | 0.139** | 14 | 5.54278 | 95.2869 | 0.41 | 0.1998** |
| 5 | 5.56243 | 95.3345 | 0.32 | 0.427 | 15 | 5.54744 | 95.2953 | 0.35 | 0.001* |
| 6 | 5.55841 | 95.2887 | 0.41 | 0.176** | 16 | 5.54643 | 95.3065 | 0.37 | 0.247 |
| 7 | 5.56149 | 95.2939 | 0.33 | 0.306 | 17 | 5.55637 | 95.2959 | 0.34 | 0.203 |
| 8 | 5.55558 | 95.3169 | 0.32 | 0.457 | 18 | 5.55603 | 95.3139 | 0.32 | 0.463 |
| 9 | 5.56201 | 95.3032 | 0.42 | 0.057* | 19 | 5.55344 | 95.3072 | 0.35 | 0.001* |
| 10 | 5.55874 | 95.3082 | 0.41 | 0.204 | 20 | 5.54974 | 95.3012 | 0.35 | 0.001* |

of PGA was 0.31 at latitude 5,541 and longitude 95,294. The maximum value of PGA was 0.41 at latitude 5,562 and longitude 95,303. This figure also shown that there were some locations have PGA which closes together. Example, PGA in location 1, 7, 15, 17, 19, and 20 was between 0,33 and 0,35.

Detection of spatial autocorrelation is very important to show the PGA relationship in every locations. It can perform by Moran's $I$ and LISA. The value of Moran's I by Eq. (1) was -0,038 and expected value of Moran's I was -0.053. Moran's I was greater than expected. It means that there was an autocorrelation between PGA in Meuraxa, Banda Aceh. The autocorrelation shows that PGA has the cluster pattern based on neighbours location. Location which neighborhood have the same characteristics of PGA. It was shown in Fig. 2 and explained by the previous paragraph. Example is PGA in location 1, 7, 15, 17, 19, and 20. They relatively close to location 6, 9, 10, 12, and 14 which have PGA between 0,40 and 0,42. Then location 4 or 16 which have PGA 0,36 and 0,37, was closed to location 9, 10, 12, and 14.

The results of LISA (see Table 1) was performed by P-value. It show the detail autocorrelation characteristics for each location. P-value which less than

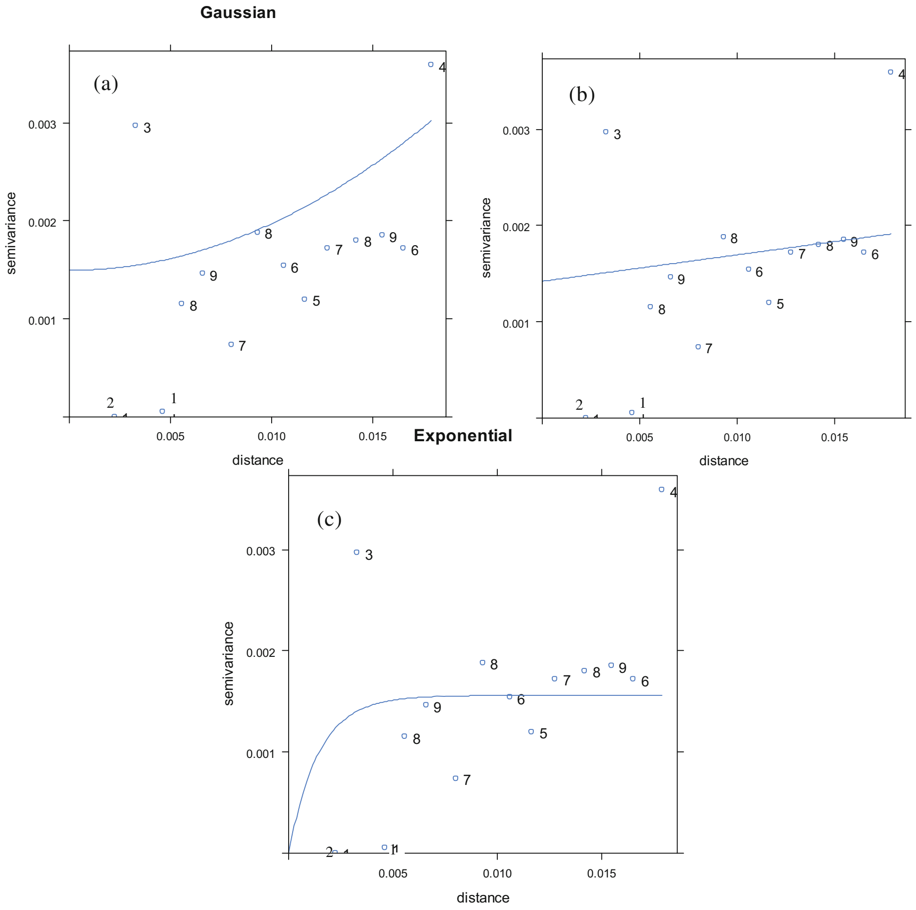


**Fig. 3.** Semivariogram Model: (a) Gaussian, (b) Spherical, and (c) Exponential

significance level ($\alpha$) is show that the location has significant autocorrelation with nearest locations. The LISA test $\alpha = 10\%$ and $\alpha = 20\%$ conclude that there were four and three locations which have significance autocorrelation with nearest locations. That locations tend to have high PGA and close to each other. The results of Moran's I and LISA show that to predict PGA was better perform by kriging which based on autocorrelation spatial.

The first Ordinary kriging process is build empirical semivariogram by Eq. (5). It uses to determine the characteristic of spatial correlation among locations, sill, range, and nugget. After that, semivariogram model will be created. It consists of three models: Gaussian, Spherical, and Exponential. Semivariogram model has different results on each of its plot (see Fig. 3). The blue line shows the semivariance model. The points show that there were 14 groups of data and their spatial autocorrelation. The interpretation of this is very important to identify the characteristic relationship among locations, as same as in Moran's I or LISA. Example in semivariogram, there was a point which has 3 locations. It has the distance among locations about 0,003 and semivariance about 0.003. The points which were at distance more than 0,005 have the constant semivariance and it means that there were no autocorrelation among locations.

After determining semivariogram model, the next step was predicting the PGA value. The prediction result of three testing data was shown in Table 2. Locations with 5,556 latitude and 95,314 longitude has the PGA prediction 0,3499; 0,3554; 0,3526 from Gaussian, Spherical, and Exponential models. The results of prediction was evaluated by mean and MSE. According to mean value, all models have mean 0,3534; 0,3584; and 0,3555 which approaches the actual PGA mean 0.34. According to MSE value, it can be seen that all models have small MSE or relatively closed to zero. This evaluation showed that ordinary kriging give better prediction in PGA or earthquake case. This methods works based on autocorrelation characteristics or spatial methods. The calculation was from semivariance which performs autocorrelation among locations.

**Table 2.** PGA Ordinary Kriging

| No | Locations | | Semivariogram Models | | |
|---|---|---|---|---|---|
| | Latitude | Longitude | Gaussian | Spherical | Exponential |
| 1 | 5.556025 | 95.31391 | 0.3499 | 0.3554 | 0.3526 |
| 2 | 5.553439 | 95.30772 | 0.3533 | 0.3613 | 0.3575 |
| 3 | 5.549742 | 95.30116 | 0.3569 | 0.3584 | 0.3563 |
| Mean | | | 0.3534 | 0.3584 | 0.3555 |
| MSE | | | 0.00095011 | 0.00145588 | 0.00115661 |

## 5 Conclusion

The Ordinary Kriging can be applied to predict Peak Ground Acceleration (PGA). The methods perform prediction based on autocorrelation among locations. The calculation was from semivariance which performs autocorrelation among locations. Based on Moran's I and LISA, it can be conclude that there was an autocorrelation spatial among locations. According to mean value of three prediction, all models (Gaussian, Spherical, and Exponential) have mean 0,3534; 0,3584; and 0,3555 which approaches the actual PGA mean 0.34. According to MSE value, it can be seen that all models have small MSE or relatively closed to zero. This evaluation showed that ordinary kriging give better prediction in PGA.

## References

1. BMKG: Earthquake. http://www.bmkg.go.id/RBMKG_Wilayah_10/Geofisika/gempabumi.bmkg Accessed: 12 Dec 2012
2. Irwansyah, E., Winarko, E., Rasjid, Z., Bekti, R.: Earthquake hazard zonation using peak ground acceleration (pga) approach. In: Journal of Physics: Conference Series, Vol. 423. IOP Publishing (2013) 012067
3. Megawati, K., Pan, T.C.: Ground-motion attenuation relationship for the sumatran megathrust earthquakes. Earthq. Eng. Struct. Dyn. **39**(8), 827–845 (2010)
4. Youngs, R., Chiou, S.J., Silva, W., Humphrey, J.: Strong ground motion attenuation relationships for subduction zone earthquakes. Seismol. Res. Lett. **68**(1), 58–73 (1997)
5. Lin, P.S., Lee, C.T.: Ground-motion attenuation relationships for subduction-zone earthquakes in northeastern taiwan. Bull. Seismol. Soc. Am. **98**(1), 220–240 (2008)
6. Fischer, M.M., Getis, A.: Handbook Of Applied Spatial Analysis: Software Tools, Methods And Applications. Springer, Heidelberg (2009)
7. Davis, E., Ierapetritou, M.: A kriging method for the solution of nonlinear programs with black-box functions. AIChE J. **53**(8), 2001–2012 (2007)
8. Eldeiry, A.A., Garcia, L.A.: Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using landsat images. J. Irrig. Drain. Eng. **136**(6), 355–364 (2010)
9. Kumar, V., et al.: Kriging of groundwater levels-a case study. J. Spat. Hydrol. **6**(1), 81–94 (2006)
10. Anselin, L., Rey, S.J.: Perspectives on spatial data analysis. In: Perspectives on Spatial Data Analysis. Springer (2010) 1–20
11. Bekti, R.: Sutikno. spatial durbin model to identify influential factors of diarrhea. J. Math. Stat **8**, 396–402 (2012)
12. Lee, J., Wong, D.W.: Statistical Analysis With ArcView GIS. Wiley, New York (2001)
13. Sengara, D.I.W.: Microzonation and Hazard Mapping of Meuraxa District-banda Aceh (2009)
14. Bivand, R., Pebesma, E., Gomez-Rubio, V.: Applied Spatial Data Analysis with R. Springer, New York (2008)

# Multicriteria Problem Structuring for the Prioritization of Information Technology Infrastructure Problems

Carolina Ferreira Gomes Silva[1](✉), Plácido Rogério Pinheiro[2], and Odecília Barreira da Silva Benigno[3]

[1] Banco do Nordeste do Brasil, Fortaleza, Brazil
carolfgs@yahoo.com
[2] University of Fortaleza, Fortaleza, Brazil
placido@unifor.br
[3] Centro Universitário Estácio do Ceará, Fortaleza, Brazil
odecilia.benigno@estacio.br

**Abstract.** Technology has become a vital component for organizations. Thus, it is necessary to ensure quality and efficient IT solutions to meet the expectations of the business areas. In this scenario, we realize the need to optimize decision-making in the IT infrastructure problem management process, thus ensuring greater availability of IT solutions that support business. The objective of this work is to propose a model for selection and prioritization of IT infrastructure problems in the Multiple Criteria Decision Analysis. We present the context of the decision of the problem and by applying cognitive mapping we identify the most relevant criteria in order to detect the problems that generate the most impacts in the business.

**Keywords:** IT infrastructure problems · Multicriteria decision analysis · Decision analysis

## 1 Introduction

In the last few decades, technology has become a fundamental part of organizations, so IT areas have become extremely demanding and charged for services with higher quality, performance and availability.

Models focused on management of the operation of IT services have emerged in the last decades with the dissemination of quality, management and IT governance frameworks, among them: ITIL (IT Infrastructure Library) and COBIT (Control Objectives for Information and related Technology). There is a need to combine the practices suggested according to the reality of each company, in order to meet a demand for IT solutions that require increasingly integrated and efficient processes.

The reality of the IT environments of the organizations is that there are countless faults because of the diversity of hardware and software that support the solutions the business demands. The IT areas in the organizations need to act in a timely and assertive manner on IT infrastructure failures to ensure a minimum of degraded or stopped services.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_30

The problem management process proposed by the ITIL methodology aims not only to diagnose the causes of the failures and to correct them, but also to proactively eliminate the recurrences, or even avoid them. However, the decision-making process of on which problems to act and in which order of priority in order to minimize the impact on business, ensuring greater availability of services, is quite complex.

In face of this reality, a study is needed to optimize the decision-making process of problem management.

This work presents the structuring phase of the multicriteria decision-making problem in the problem management process. Through the use of cognitive mapping, we structure the problem in question and identify the criteria and their most relevant instances to be considered in the selection and prioritization of IT infrastructure problems. The purpose of this study is to identify the key-concerns and the Value Tree originated in the structuring process of the multicriteria problem for the selection and prioritization of the IT infrastructure problems idealized by [14].

## 2 The Multicriteria Methodology to Support the Decision

According to [10], the decision is strongly related to the comparison of different points of view, some in favor and others against a certain opinion. This means that the decision is intrinsically related to a plurality of points of view, which can be defined as criteria.

On the other hand, [13] affirms that support for decision is the activity of the person who, through the use of explicit but not necessarily completely formalized models, helps to obtain elements of answers to the questions posed by an interested party in a decision.

At least thirty years ago, researchers and professionals began to look at decision problems in a different way. This new approach explicitly considers the pros and cons of a plurality of viewpoints, called Multiple Criteria Decision Analysis (MCDA). Despite the diversity of approaches, methods and techniques of existing MCDA, three basic elements are common: a finite or infinite set of potential actions (alternatives or solutions), at least two criteria and, of course, at least one decision maker. In this way, the MCDA is an activity that helps to make decisions mainly in terms of choice, classification or ordering of actions [10].

According to [15], the choice of the multicriteria methodology to support the decision-making process depends directly on the subject matter. According to [12, 17], the choice of method should be the result of an evaluation of the parameters chosen, the type and accuracy of the data, the decision maker's way of thinking, and his/her knowledge of the problem.

Thus, [13] states that in a multicriteria decision-analysis study, three concepts play a fundamental role in analyzing and structuring the decision-making process, since they have a close relationship with decision-making itself. They are:

**Potential action** is the object of the decision, or the one to whom the decision support is directed. An action is qualified as potential when it is possible to implement it, or simply when it deserves some interest within the decision support process [13]. Known as an alternative, it can be identified at the beginning of the decision process or in the course of it, and may become a solution to the problem being studied [11].

**Criterion** is a tool built to evaluate and compare potential actions according to a point of view that must be well defined. This assessment shall take into account, for each action, all relevant effects or attributes linked to the point of view considered [13]. Vincke [19] defines a criterion as being a function g, defined in a set A, which assigns order values of set A, and which represents the preferences of the decision maker from a given point of view.

**Decision support problems** – decision support should not be viewed solely in the perspective of solving a choice problem. In some cases, it consists only in elaborating an appropriate *A*-set of potential actions, building a suitable *F*-family of criteria, and determining, for all or some of the $a \in A$, their performances, sometimes supplemented by additional information. To designate this way of conceiving the objective of decision support, without seeking to elaborate any prescription or recommendation, the term Problem of Description ($P.\delta$) is used. Three other reference problems are currently used in practice: Problem of Choice ($P.\alpha$); Problems of Classification ($P.\beta$) and Problems of Ordination ($P.y$) [13]. The problems of Choice, Classification and Ordination lead to a specific result related to the evaluation of alternatives and the Classification problem is based on absolute judgments, where each alternative is associated with a specific group based on a pre-defined rule [8].

According to [2], the decision support process comprises three phases: (i) structuring; (ii) evaluation and (iii) recommendation.

### 2.1 Structuring

For [3], the structuring process should generate a well-defined operational basis to assist the analyst in guiding the decision-maker and other actors to identify key points of view and to operationalize the criteria for evaluating the impacts of the options and to compare the advantages and disadvantages. The authors also suggest that the structuring phase be divided into the following activities: (i) the definition of the problem; (ii) the structuring of the model and (iii) the analysis of the impacts (estimation of the consequences of each of the options considered).

The structuring of the problem is an interactive process of constructing a model of representation that integrates the objective components of the problem and the subjective objectives of the actors, in such a way that the value system of the actors is explicit. Thus, [3] point out that once the problem has been defined, decision support activities should focus on the structuring of the points of view according to which we intend to analyze the impacts and the attractiveness of the options.

A key-concern is an individual isolated point of view, or a set of points of view in which the actors agree upon analyzing their impacts and options separately, independently of the other points of view. The organized list of impacts is the impact table. It provides an overview of the impacts of the options on the main points of view identified in the structuring process. Two conditions are necessary to organize an impact table: (1) it must be possible to describe at least qualitatively the plausible impacts in terms of the key-concerns in the specific context of the problem and (2) it should be possible to estimate, with greater or less certainty, the impact of each option in relation to the key-concern [3].

On the other hand, [3] define “impact descriptor” as being an orderly set consisting of plausible levels of impact, according to a given fundamental point of view that: measures (quantitatively or qualitatively) to what extent the fundamental point of view is satisfied; Describes, as objectively as possible, the impacts of the options in a fundamental point of view; Establishes a plausibility domain for impacts (from a more attractive level to a less attractive level) and verifies the ordinal independence of the corresponding key-concern.

According to [9], the use of maps to describe and explore the cognitive structures of members of organizations that are facing complex issues has become well established in recent years. The author also affirms that in the face of ill-defined and complex situations, involving several actors and with many questions to be addressed, the construction of cognitive or causal maps facilitates the structuring of the points of view and defines cognitive map as a representation of the thought about a problem that follows the mapping process. According to [5], the experiences that people develop in organizational environments are structured into personal patterns of knowledge and can be called cognitive maps, and are used to understand organizational situations and deal with them.

For representational purposes, a cognitive map is usually drawn as small pieces of text attached to one-way arrows to link them together. In the general case, an affirmation at the tip of an arrow is taken to cause or influence the affirmation at the tip of the arrow. In cognitive mapping we try to identify each statement (knot) as having two contrasting poles. The connection between concepts is carried out by means of arrows, each one indicating a different way. Thus, we have: (i) if the arrow enters a concept *A*, then concept *A* can be explained by - or is the cause of - a concept *B*; And (ii) if the arrow comes from a concept *A*, then concept *A* can lead to - or have implications for or have as consequence - a concept *B* [12].

The main result of the structuring process is the family’s definition of the key-concerns. It is common, in complex problems, to organize the areas of concern or interest and the key-concern in a tree structure, often referred to as a tree of values. It is noteworthy that [12] states that based on the family of key-concerns, it is possible to evaluate the attractiveness of the options for each interest.

## 3 Application of the Structuring of the Multicriteral Problem

### 3.1 Decision Context

According to [16], the growth in the use of multicriteria methodologies in decision support is characterized by giving decision-makers the possibility of obtaining the best solution adjusted to the needs of the business.

The multicriteria methodology supports the decision-making of the problem process in the problem prioritization activity by allowing definitions of impact and urgency to be obtained by defining the best order of performance in the existing problems, considering the scope of the failure, impact on the business, available resources and risks that are difficult to identify by the organization.

The multicriteria methodology is inserted in the flow of the problem management process, supporting the decision-making process in the “Problem Prioritization” step,

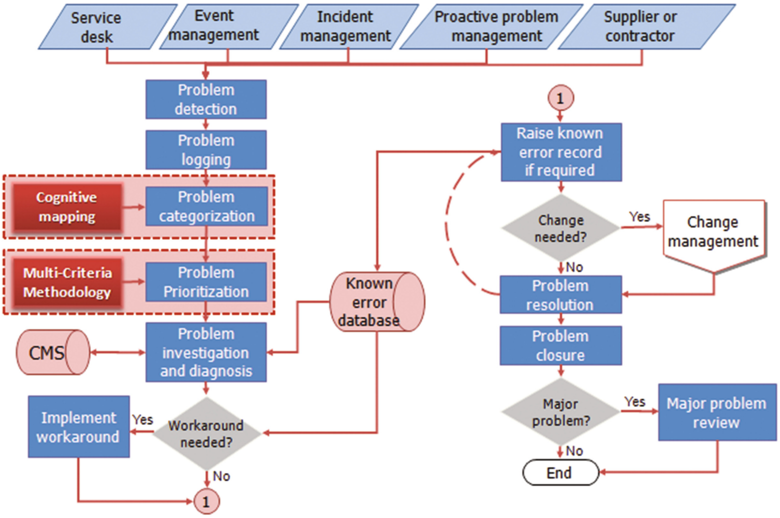


**Fig. 1.** Flow of the ITIL problem management process (*Source*: Adapted from ITIL Service Operation [7])

according to Fig. 1. However, it is of fundamental importance to categorize the problems in the "Categorization" activity, prior to prioritization. In this activity, the importance of cognitive mapping for generating the tree of values is highlighted.

Figure 2 shows the flow of the categorization activity of the problem management process using cognitive mapping to define the relevant criteria for the organization and its descriptors.

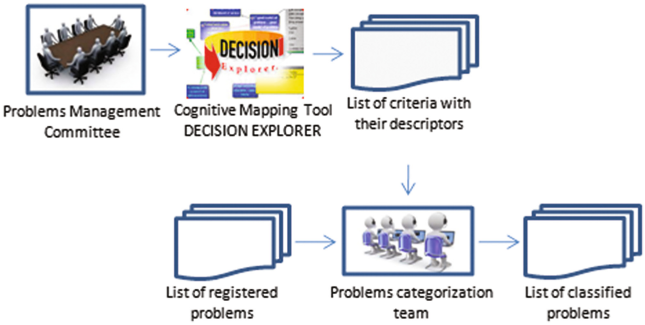
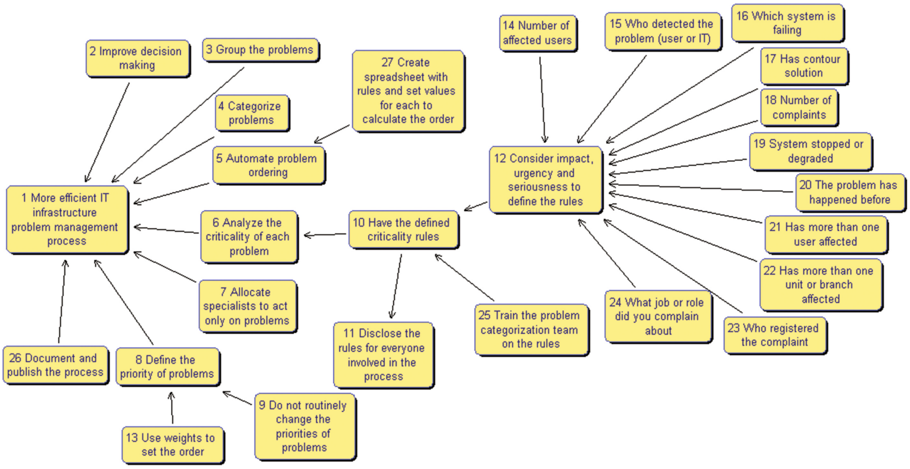

Problems Management Committee
DECISION Explorer
Cognitive Mapping Tool DECISION EXPLORER
List of criteria with their descriptors
List of registered problems
Problems categorization team
List of classified problems

**Fig. 2.** Cognitive mapping in the categorization of problems (*Source*: Formatted by the author)

To structure the problem, initial interviews were conducted with specialists and decision makers of the problem management process of a medium-sized financial organization to map the solutions to the problem in question. Five meetings were held, two with all involved, a third one to refine the ideas with the main decision makers of the process, another with all involved to agree on the final strategic map and a last one to define the fundamental points of view and their descriptors. In the first contact, to contextualize all and to align the problem for which a solution was being sought, the main objectives of the research were clarified and the following key question was posed: **How to optimize the problem management process in order to minimize the impacts on organizations' businesses?**

## 3.2 Cognitive Mapping

In the second meeting, from this question, the participants were encouraged to co-locate ideas to reach the solution, using the cognitive mapping technique that suggests the use of connection between concepts, always seeking cause and effect when structuring the value tree. To support the use of the technique, the Decision Explorer tool [6] was used, in order to record all the information that the actors judged relevant to the problem, according to Fig. 3.



1 More efficient IT infrastructure problem management process
2 Improve decision making
3 Group the problems
4 Categorize problems
5 Automate problem ordering
6 Analyze the criticality of each problem
7 Allocate specialists to act only on problems
8 Define the priority of problems
9 Do not routinely change the priorities of problems
10 Have the defined criticality rules
11 Disclose the rules for everyone involved in the process
12 Consider impact, urgency and seriousness to define the rules
13 Use weights to set the order
14 Number of affected users
15 Who detected the problem (user or IT)
16 Which system is failing
17 Has contour solution
18 Number of complaints
19 System stopped or degraded
20 The problem has happened before
21 Has more than one user affected
22 Has more than one unit or branch affected
23 Who registered the complaint
24 What job or role did you complain about
25 Train the problem categorization team on the rules
26 Document and publish the process
27 Create spreadsheet with rules and set values for each to calculate the order

**Fig. 3.** Initial cognitive mapping on the problem

To organize what had been discussed with the whole group of actors, a third meeting was held only with the key decision makers of the process which resulted in a more organized and non-redundant cognitive map with twenty concepts. Then a new meeting was held with all those involved, where the new improved cognitive map was presented and agreed by all with few adjustments. Figure 4 shows the final cognitive map with the twenty concepts.

Several concepts were analyzed and it was concluded that the key point to solve the problem would be to define the criteria for categorizing the problems, since only then would it be possible to optimize and automate the decision on which problems to act as a priority and thus, minimize the impacts of IT failures on the organizations' businesses.

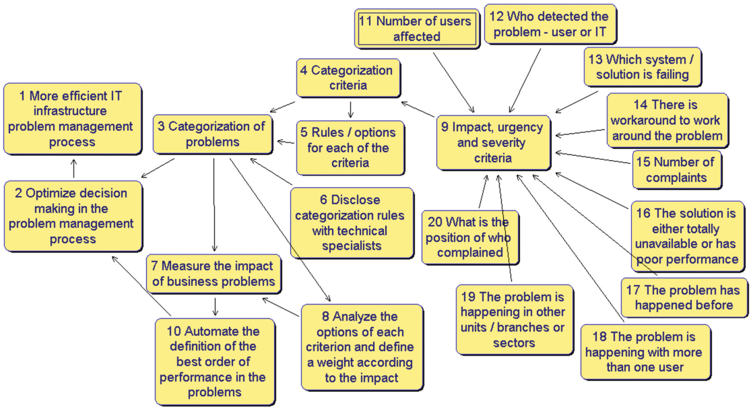

11 Number of users affected
12 Who detected the problem - user or IT
13 Which system / solution is failing
4 Categorization criteria
1 More efficient IT infrastructure problem management process
3 Categorization of problems
5 Rules / options for each of the criteria
9 Impact, urgency and severity criteria
14 There is workaround to work around the problem
15 Number of complaints
2 Optimize decision making in the problem management process
6 Disclose categorization rules with technical specialists
20 What is the position of who complained
16 The solution is either totally unavailable or has poor performance
7 Measure the impact of business problems
17 The problem has happened before
19 The problem is happening in other units / branches or sectors
8 Analyze the options of each criterion and define a weight according to the impact
10 Automate the definition of the best order of performance in the problems
18 The problem is happening with more than one user

**Fig. 4.** Cognitive map of the problem

According to [18], modeling helps decision-making, because it helps to make its objectives explicit, forcing the identification of variables, their forms of measurement and also forces the recognition of limitations.

## 3.3 Key-Concerns

From the strategic map built, it was possible to propose a tree of value, identifying the Points of Fundamental Views. According to [4], the structuring process results in the identification of the family of fundamental points of view, organized and branched into a tree of value. Figure 5 shows the value tree constructed through a mind map tool, highlighting the fundamental points of view grouped into two groups of impact and urgency, as recommended by ITIL for the categorization of IT failures. This tree was structured in the fourth meeting with all the involved in structuring the problem.

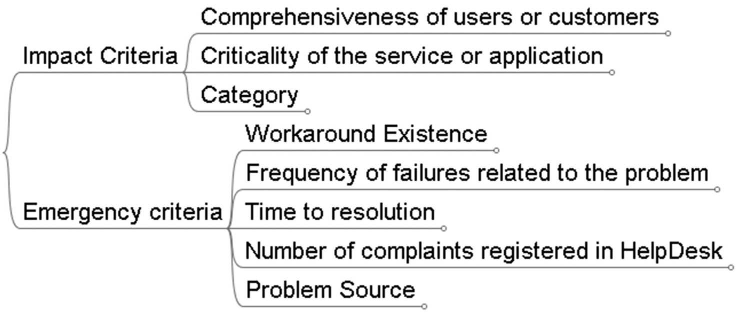

Impact Criteria
Comprehensiveness of users or customers
Criticality of the service or application
Category
Emergency criteria
Workaround Existence
Frequency of failures related to the problem
Time to resolution
Number of complaints registered in HelpDesk
Problem Source

**Fig. 5.** Value tree of key-concerns

## 3.4 Impact Analysis

For [1], the fundamental point of view (FPV) becomes operational if there is a set of levels of impact associated to it, defined by $N_j$, that must be ordered in a decreasing order according to the decision makers. Thus, they constitute a scale of local preference, limited by the upper level $N_{*j}$, which has greater attractiveness, and by the lower level $N_{*j}$, of lower attractiveness, having to meet the following pre-ordering condition: $N_{*j} > \cdots > N_{k+1,j} > N_{k,j} > N_{k-1,j} > \cdots > N_{*j.}$

For each of the eight fundamental points of view of the problem their descriptors were listed, that is, a set of options so that each reflects the weight of the impact and its characteristics of actions for the business of the organization. In this stage of construction of the descriptors, the decisions were taken in a meeting with all the actors involved in the process.

### Comprehensiveness of Users or Customers

This key-concern was operationalized through a qualitative descriptor, fixed and discrete, aiming to evaluate the comprehensiveness of users affected by the problem. For this point of view, three levels with different weights of impact in the business were defined. These levels were defined according to the experience of the actors in the process verifying that the most impactful level is when the problem affects all the users of the organization. Table 1 shows the key-concern descriptors.

**Table 1.** Descriptors for comprehensiveness of users or customers

| NI | Description | Order |
|---|---|---|
| N3 | All customers or users | 1° |
| N2 | Unit/area/part of the customers | 2° |
| N1 | Group of users or customers | 3° |

It is emphasized that the descriptors have a structure so that a higher level is always preferable to a lower level, that is, it is the higher level that most compromises the business of an organization.

### Criticality of the Service or Application

This key-concern was operationalized through a qualitative descriptor, fixed and discrete, aiming to evaluate the characteristic of the problem solution taking into account its impact in the accomplishment of the company's business. An analysis of the types of IT solutions existing in a company was made, grouping them by similarity. In this way, we reached the five levels of groupings, according to Table 2.

**Table 2.** Descriptors for criticality of the service or application

| NI | Description | Order |
|---|---|---|
| N5 | Critical to business (external customers) | 1° |
| N4 | Business support | 2° |
| N3 | Production planning and control | 3° |
| N2 | Administrative systems | 4° |
| N1 | Not critical | 5° |

### Category

This key-concern was operationalized through a qualitative and fixed descriptor, aiming to identify the technical specialty of the failure to evaluate its impact on the company's business. Thus, ten categories were defined according to their criticality to the organization, according to Table 3.

**Table 3.** Descriptors for category

| NI | Description | Order |
|---|---|---|
| N10 | Database | 1° |
| N9 | Connectivity | 2° |
| N8 | Operating environment | 3° |
| N7 | Storage | 4° |
| N6 | Safety | 5° |
| N5 | Application distribution | 6° |
| N4 | Planning and production control | 7° |
| N3 | Backup and restore | 8° |
| N2 | Antivirus | 9° |
| N1 | Email | 10° |

### Workaround Existence

This key-concern was operationalized through a qualitative descriptor, fixed and discrete, aiming to identify the existence of some alternative to overcome the problem. Often, in the technological field, some repetitive action temporarily solves a problem that subsequently occurs again, but it is a way of minimizing the impacts of the problems. In this way, three options were defined, according to Table 4.

**Table 4.** Descriptors for workaround existence

| NI | Description | Order |
|---|---|---|
| N3 | No | 1° |
| N2 | Do not know | 2° |
| N1 | Yes | 3° |

### Frequency of Failures Related to the Problem

This key-concern was operationalized through a qualitative and fixed descriptor, aiming at identifying the frequency with which an IT component is failing and impacting users. This criterion aims to identify the number of times the solution failed. In this way, by discussing with the actors of the process, three groups were defined, according to Table 5.

**Table 5.** Descriptors for Frequency of failures related to the problem

| NI | Description | Order |
|---|---|---|
| N3 | Three (or more) times a week | 1° |
| N2 | Three (or more) times a month | 2° |
| N1 | Less than three times a month | 3° |

### Time to Resolution

This key-concern was operationalized through a qualitative and discrete descriptor, aiming at identifying the impact of the failure from the estimated time for diagnosis and resolution of the problem. Considering the experiences of the actors involved in problem solving, four time slots for definitive resolution of IT failures were reached in order to measure the impact on the affected businesses. Table 6 shows the defined ranges.

**Table 6.** Descriptors for time to resolution

| NI | Description | Order |
|---|---|---|
| N4 | Until 1 h | 1° |
| N3 | Up to 24 h | 2° |
| N2 | Up to 1 week | 3° |
| N1 | Up to 1 month | 4° |

### Number of Complaints Registered in the HelpDesk

This key-concern was operationalized through a quantitative and discrete descriptor, aiming to identify the impact of the failure based on the amount of complaints from users. We attempted to identify ranges of quantities according to the impact generated, in conformation with Table 7.

**Table 7.** Descriptors for number of complaints registered in HelpDesk

| NI | Description | Order |
|---|---|---|
| N4 | $\geq 100$ | 1° |
| N3 | $\geq 50$ e <100 | 2° |
| N2 | >10 e <50 | 3° |
| N1 | $\leq 10$ | 4° |

### Problem Source

This key-concern was operationalized through a qualitative and discrete descriptor, aiming to identify where the complaint of the failure originated, since this origin can affect the priority of action or not. For example, if the complaint came from the user (N5 - Service Desk 1$^{a}$ Order) it is a sign that the failure is already affecting the business front line, but if came from the proactive problem management (N3-3rd order) it can be a failure that will still be felt by the final user. Table 8 shows the options.

**Table 8.** Descriptors for problem source

| NI | Description | Order |
|---|---|---|
| N5 | Service desk | 1° |
| N4 | Incident management | 2° |
| N3 | Proactive problem management | 3° |
| N2 | Event management | 4° |
| N1 | Supplier or contractor | 5° |

## 4 Final Considerations

The problem management process is a key element in ensuring the availability of IT solutions for an organization's business. This process being well structured, implemented following the best market practices and supported by multicriteria methodologies in the decision stage can generate productive and profitable results for the organization. To act in the main problems that are causing major impacts in the generation of business is of fundamental importance in face of the reality of the competitive market that the organizations live in the present day.

The structuring of the multicriteria problem for selection and prioritization of infrastructure problems served the objectives of identifying the key-concerns and creating the value tree with the purpose of supporting those involved in the decision making of which problems to act on as a priority in order to minimize the impact on business.

The structuring of the problem helped process decision making to make it more impartial and less subjective, since criteria and their descriptors were listed for use in the categorization of problems. In this way, the comparison and definition of the most urgent problem will be facilitated.

By optimizing the problem management process and optimizing the decision-making time to prioritize the problems that generate the greatest business impact, we minimize financial losses and save the time and effort of the professionals involved in the process.

One of the main contributions of this work is to allow the decision-making process to be carried out more objectively and efficiently, since there is a standardization of the criteria to be considered in the decision.

As a proposal for improvement and extension of the work, we suggest to apply the fundamental points of view listed in companies as a way of validating the contribution of this work in the decision making process of problem management.

**Acknowledgements.** The first author thanks Banco do Nordeste do Brasil for financially supporting this research and the second author thanks the National Council for Technological and Scientific Development (CNPq) through Grants # 304747/2014-9. The authors thank the Edson Queiroz Foundation/University of Fortaleza and Finep (Funding Agency for Studies and Projects) for all the support provided.

## References

1. Bana e Costa, C.A.: Structuration, Construction et Exploitation d'un Modèle Multicritère d'Aide à la Décision. Tese de doutoramento, Instituto Superior Técnico, IST, Lisboa (1992)
2. Bana e Costa, C.A., Correa, E.C., Corte, J.M.D., Vansnick, J.C.: Facilitating bid evaluation in public call for tenders: a social-technical approach. OMEGA **30**, 227–242 (2002)
3. Bana e Costa, C.A., Beinat, E.: Model-structuring in public decision-aiding. Working Paper LSEOR 05.79. London School of Economics, London (2005)
4. Bana e Costa, C.A., Beinat, E., Vickerman, R.: Introduction and problem definition. CEG-IST Working Paper no 24 (2001)
5. Bastos, A.V.B.: Mapas cognitivos e a pesquisa organizacional: explorando aspectos metodológicos. Estudos de Psicologia **7**(Número especial), 65–77 (2002)
6. Banxia Software: Decision Explorer Software Package. University of Strathclyde, Glasgow (2002)
7. Cabinet Office: ITIL Service Operation. The Sationery Office (TSO), Londres (2011)
8. Doumpos, M., Zopounidis, C.: Multicriteria Decision Aid Classification Methods. Kluwer Academic Publishers, Berlin (2002)
9. Eden, C.: Analyzing cognitive maps to help structure issues or problems. Eur. J. Oper. Res. **159**, 673–686 (2004)
10. Figueira, J., Greco, S., Ehrgott, M.: Multiple Criteria Decision Analysis: State of the Art Surveys. Springer, Boston (2005)
11. Gomes, L.F.A.M., Gomes, C.F.S., Almeida, A.T.: Tomada de Decisão Gerencial: Enfoque Multicritério, 2. Atlas, São Paulo (2006)
12. Pinheiro, P.R., Souza, G.G.C., de e Castro, A.K.A.: Estruturação do problema multicritério para produção de jornal. Pesquisa Operacional **28**(2), 203–216 (2008)
13. Roy, B.: Paradigms and challenges. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys Series: International Series in Operations Research & Management Science, vol. 78, no. XXXVI, pp. 3–24 (2005)
14. Silva, C.F.G., Nery, A., Pinheiro, P.R.: A multi-criteria model in information technology infrastructure problems. Procedia Comput. Sci. **91**, 642–651 (2016)
15. da Silva, O.B.: C3M – Gerenciamento de mudanças estruturado em uma metodologia de otimização multicritério. Dissertação de Mestrado. Universidade de Fortaleza, Fortaleza (2013)
16. da Silva, O.B., Holanda, R., Pinheiro, P.R.: A decision model for change management based on multicriteria methodology. In: Proceedings of the 8th International Workshop on Business-Driven IT Management (BDIM), 2013, pp. 1241–1244 (2013)
17. Pinheiro, P.R., de Souza, G.G.C.: A multicriteria model for production of a newspaper. In: The 17th International Conference on Multiple Criteria Decision Analysis, vol. 17, pp. 315–325. Simon Fraser University, British Columbia (2004)
18. Tamanini, I., Castro, A.K.A., Pinheiro, P.R., Pinheiro, M.C.D.: Verbal decision analysis applied on the optimization of alzheimer's disease diagnosis: a study case based on neuroimaging. Adv. Exp. Med. Biol. **696**(7), 555–564 (2010)
19. Vincke, P.: Multicriteria Decision-Aid. Wiley, Chichester (1992)

# Agent Based Modelling Approach of Migration Dynamics

Samira Boulahbel-Bachari[1,2(✉)], Nadjia El Saadi[1], and Alassane Bah[3]

[1] LAMOPS, Higher National School of Statistics and Applied Economics Koléa (ENSSEA), Higher National School of Statistics and Applied Economics Koléa (ENSSEA), Tipaza, Algeria
boulahbelsamira@gmail.com, enadjia@gmail.com
[2] National Center of Research in Applied Economics for Development (CREAD), Bouzaréah, Algiers, Algeria
[3] UMI 209, UMMISCO (IRD-Paris 6), ESP-UCAD, Dakar, Senegal
alassanebah@gmail.com

**Abstract.** We propose an agent-based model to simulate internal migration of workers. The model is based on mathematical equations describing the socio-economic characteristics of the agents and their behaviors. The main assumption of the model is the objective of individuals to find decent (formal) employment.

The model simulates the migration of agents from a rural-type region of origin to a destination region with more important economic activities. The model consists of two main stages: before and after migration. In the first stage, potential migrants and migrants are determined. The second stage involves job search and demographic and socio-economic updates. The model is divided into several modules: (i) Initiation and migration, (ii) Job search module, and (iii) Module of economic and demographic transitions.

**Keywords:** Agent-based model · Internal migration · Mathematical equations labor market

## 1 Introduction

Thanks to the evolution of computers in the late eighties, computational and simulation methods have invaded all fields of research from robotics to ecology and from medicine to geography. Among these methods, Agent Based Models (ABM) represent a (relatively) new branch of artificial intelligence with a continuously growing interest. ABMs use metaphors inspired from biology and sociology to implement intelligent artificial systems to model real processes.

ABMs consider the basic elements constituting a process and allow analysing the emerging global dynamics of this system. This ability to reconcile the analysis' levels microscopic and macroscopic, is one of the most important advantages of their use in studying heterogeneous systems. In addition, ABMs allow the integration of new agents and the test of different hypotheses without attempting real experiments.

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*, Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_31

The facilities that ABMs offer, make them a powerful tool used in various disciplines, notably in economics.

For economic analysis, ABMs have gained notoriety and interest in recent years. ABMs allow for more flexibility in economic modelling and more realistic analysis than conventional methods. ABMs were used in economic modelling at the end of the seventies with the model of [1] on spatial segregation. The model was realized with a chessboard and pieces were moved by hand. Nether less, this model highlights the individual preference of agents to have a majority of neighbors belonging to the same ethnic group.

Generally, ABEMs use a bottom-up approach to explain the emerging global dynamics of an economic system from the individual dynamics of autonomous and heterogeneous agents that act and interact according to certain rules in a particular environment [2].

The economic models based on agents overcome several limits of standard economic modelling. If conventional economic models adhere to an infinite maximization of individual utility, a complete information and a perfect rationality of individuals, the ABEMs allow us to overcome these simplistic and restrictive assumptions by taking into account agents' heterogeneity in terms of characteristics, behavior and objectives. So agents in ABEMs live their own experience, have a limited perception of their environment and learn from their previous actions.

ABEMs constitute an extension of conventional economic approaches and contribute to their renewal. They are also closer to experimental and behavioral sciences than deductive logic methods.

ABEMs have a wide range of applications ranging from labor market modelling and employment policies [3, 4] to simulation of financial market dynamics [5].

Recently, migration movements have been modeled by ABMs. Generally, Migration ABMs are based on behavioral theories (random utility and planned behavior) and their decision to migrate depends mainly on maximization of expected utility.

Some models integrate the individual's life cycle by taking into account some life events (marriage, divorce, graduation or retirement) that change the status of the individual and thus his objectives. Given the impact that migrations can have on individuals' trajectories, studies taking simultaneously into account migration and the life cycle have known a growing interest.

Even if the use of ABEMs in migration analysis still in its early stages, it allows migration model to be more and more realistic. Few authors have been interested in modeling the decision to migrate by minimalist models or more sophisticated models. For example, the work of [6] who takes into account the expected benefits, wealth, age and migrants' network in the migration decision. Naqvi and Rehm [7] considers migrations resulting after natural disasters and leading to a decrease of incomes and an increase of food prices. Hassani-Mahmooei and Parris [8] made a model in which the decision to migrate takes two steps: selection of potential migrants then determination of migrants. Espındola et al. [9] and El Saadi et al. [10] translate an economic model [11] into ABM.

Unlike the majority of ABM of migration that are based on its demographic aspect, we propose an ABEM of internal migrations of workers. The model differs from conventional economic migration models by: (i) taking into account factors other than economic factors in the decision to migrate such as the socio-demographic characteristics of Agents; (ii) considering individuals' heterogeneity; (iii) and by introducing the selective character of migration (little exploited in the literature). Compared to ABEM of migration, this work introduces the evolution of individuals throughout their life cycle. In fact, the proposed model simulates the main events they will face throughout their lives (studies, marriage, divorce, etc.), allowing perpetuation of individuals' heterogeneity.

The structure of the paper is as follows. Section 2 introduces the proposed model. Section 3 presents the results of the simulations. A conclusion summarizing the essentials of the work done and model's prospects finishes the article.

## 2 The Agent Based Migration Model

The aim of this work is to model the process of labor migration using agent based model. The main assumption of the model is the objective of individuals to find decent (formal) job and we opt for the maximization of income in the job search process. Decisional processes and agents' characteristics are described by mathematical equations.

The model involves two main stages: before and after migration. In the first stage, potential migrants and migrants are determined. The second step involves job search process and updating demographic and socio-economic characteristics.

For each step, a system of relations presenting different behaviors of the agents is defined. The model is decomposed into several modules: (A) Initiation and migration module, (B) Job search module, and (C) Module of economic and demographic transitions.

### 2.1 Initiation and Migration Module

**Initiation**

The model includes two types of agents: Agent Individu and Agent Region.

***Agent Individu***

We consider the person and not the household as the main actor of the model even if the new migration theory rejects the individualistic character of the neoclassical models (like the model of [11]) and places the decision to migrate within a wider societal context taking into account the impact of household support [12]. From this point of view, the model developed here is a step back from this theoretical advance, but it is a deliberate simplification. Modelling household support remains difficult in practice. Also, the main objective of the model is to describe individual dynamics of labour migration and therefore mainly concerns the job seeker. Agent Individu may be either an employee affiliated to the social security system, an employee who is not affiliated to the social security system, an unemployed person or an inactive person. He is described by a set of socio-economic characteristics such as age, marital status, employment status and educational attainment.

### *Agent Region*

Agent Region represents the firms through the creation and destruction of employment. Two types of regions are considered. An origin region with a low rate of decent job creation, a large dispersion in incomes, an undeveloped formal sector, and a destination region with more potential (more important economic activities). These characteristics generally describe the two environments rural and urban.

Each Agent Region is described by the unemployment rate, the tension of the labour market, the probability of finding a job, wages and a set of demographics transitions rates. Firms are not explicitly simulated, but jobs are created according to a job creation rate specific to each region. As the first objective of the model is to analyze the migration of workers and not the recruitment process within firms, it seems preferable to simplify this dynamic by creating and destroying employment within the regions.

## Migration Process

Decision to migrate takes place in two stages. Initially, potential migrants are determined on the basis of certain socio-economic characteristics. Next, potential migrants estimate the expected profit of migration by taking into account wages and the probability of finding a job.

### *Determination of Potential Migrants*

The decision to migrate depends strongly on the individual characteristics of migrants. This is called self-selection that determines the ability to migrate for different categories of people.

In this model, the individual characteristics chosen are age, gender, employment status, educational level, marital status and the existence or absence of migratory networks. All these variables have a great selective influence in the propensity to migrate.

The determination of potential migrants is based on the calculation of a discriminant function relatively to these variables for unemployed or informally employed individuals aged between 16 and 59 and living in the rural. This function is presented as a linear combination of these variables ($F_j(t)$). A worker is considered as a potential migrant if he satisfies the following condition:

$$F_j(t) \times u \times \theta > \mathit{Threshold} \tag{1}$$

With Threshold a constant determining by user.

Once potential migrants are determined, they will have to choose between migrating or staying.

### *Decision to Migrate*

Potential migrant individuals now have to calculate the expected gains with or without migration. These gains are weighted by the probability of finding formal job, which depends on agent's qualification. Skilled workers may apply in contrast to unskilled workers for skilled and unskilled job positions, increasing their chances of having a formal job:

$$G_i^j = \pi_i^j \times W_{i,j}^e \tag{2}$$

Where

- $G_i^j$ is the expected gain for individual i in region j;
- $\pi_i^j$ is the probability of finding a decent job in region j;
- $W_{i,j}^e$ is the expected wage for individual i in region j. The expected urban and rural wages are calculated as follows:

$$W = SNMG \times Q + \sigma W \times R(0,1) \tag{3}$$

With

- $SNMG$ the National Minimum Wage;
- $Q$ Individu's qualification;
- $\sigma W$ Wage Dispersion;
- $R(0,1)$ Realization of the uniform law (0, 1).

Wage differentials between the region of origin and destination regions are calculated. If wage differential is greater than zero, then the potential migrant will decide to migrate.

After determining the migrants, job prospecting is the next step in the model for migrants and non-migrants.

## 2.2 Job Searching

All unemployed or informal workers aged between 16 and 59, are concerned by job search. Those who are qualified can apply for skilled and unskilled job positions and those who are not qualified will have to settle for unskilled job positions.

In each region and every year, a number of jobs are released following retirements, others are created and some can be destroyed.

Individuals with the highest probability of finding formal employment will be the first to be hired.

However, individuals will not be able to remain indefinitely in a situation of unemployment, they will be discouraged after a certain number of attempts.

We suppose that the probability to be required increase with the qualification and the time spent on the search. This probability is defined by [13]:

$$P_i^j(\mathrm{t}) = P_i^j(t-1) + (1 - P_i^j(t-1)) \times \pi_i^j(t) \tag{4}$$

Where

- $P_i^j(t)$ is the probability to be required for agent $i$ in region $j$ at time $t$,
- $\pi_i^j(t)$ is the probability to find a job for agent $i$ in region $j$ at time $t$.

Once the job search process is completed, an update of the demographic and socio-economic variables begins. The module starts a new cycle.

### 2.3 Demographic and Economic Updates

The individuals' state evolves in time and changes throughout their life cycle. Since our model depends largely on individual characteristics of the population and on the current and expected demographic changes, a particular attention is made to the sociodemographic updates. The main simulated demographic events are births, deaths, marriages, divorces, promotions and school dropouts.

First, death individuals are removed from the database and the age of the remaining individuals is increased by one year. Then new agents are created, others divorce while some get marry. Children who have reached the age of six are enrolled in school. Some students (primary, secondary or high school) and students (university or vocational) drop out of school and become inactive. The rest of the students move on to the next higher level. At the end of each model cycle, students complete their studies. Some of them start looking for work and become unemployed and others become inactive.

Economic transitions of individuals are also simulated. Unemployed persons who have spent a long period seeking employment, will abandon and become inactive while a part of the inactive persons will undertake a job search and become unemployed. Some more ambitious unemployed will embark on the entrepreneurial adventure, while some of the entrepreneurs will be forced to give up their business and will be among the new unemployed.

## 3 Simulation and Results

### 3.1 Simulator

The migration model proposed is realized on GAMA platform.[1] GAMA is a free modelling and simulation development environment dedicated to agent-based modelling. GAMA allows the modelling and simulation of several processes of different fields ranging from epidemiology, land use to urban mobility or reconstruction of geo-historical events. The development environment offered by GAMA includes a GAML code editor that facilitates the writing of templates (auto-compilation of models, auto detection and correction of errors, etc.). GAML is an agent-oriented programming language that is simple to use and assimilate and even allows novice modellers to create templates. The GAMA platform has been enriched since its creation in 2007 and incorporates many powerful tools for space management, 3D visualization, or geographic data integration.

The simulator realized is an experimental tool for visualizing evolution of rural and urban labour market under labour migration. Two databases are necessary to model such evolutions. The first database is relative to individuals and regions. The second base regroups the set of probabilities of transitions between states as well as the probabilities of occurrence of the events. The simulator performs the following tasks for each agent:

[1] gama-platform.org.

```
Migration process
For each unemployed (or informal) worker in the rural area
     Calculate the score;
     Is it a potential migrant?
       Yes
         Add to the collection of potential migrants.
For each agent in the collection of potential migrants
     Calculate the expected salaries (rural and urban);
     Calculate the wage differential.
       Is wage differential > 0
         Yes
          Define as a migrant;
          Initialize job search time;
          Migrate
          Delete agent from rural population;
          Add the agent to the list of unemployed agents in the urban area.
Job Search Process
For each unemployed (or informal) worker in the two regions
     Calculate the probability to find a job;
     Classify the unemployed by the probability of recruitment
     Assigning these unemployed persons to vacant posts as follows:
     Assign qualified individuals to qualified positions first;
     Assign the rest of qualified individuals to unskilled positions;
     Assign unskilled individuals to the remaining unskilled positions;
     Increment the search time
       Search time > maximum search time
        Set Employment Status = inactive
Demographic Update
For each region
    Delete death Agents;
    Increase ages of remaining agents;
    Add new agents;
    Get divorced a part of married agents;
    Get married a part of single agents
    Let a proportion of students leave school;
    Promote the remaining students
    Do the Economic transitions between different employment statuses.
    Update global variables
```

### 3.2 The Scenarios Tested

The proposed agent-based migration model allows simulations for a large set of parameters (economic or demographic parameters). We chose to focus on those related to job creation in presence and absence of workers' migration. The realized simulator considers two regions:

- Origin Region representing Rural. The agricultural employment predominates the labor market. Employment in this region is therefore mostly seasonal and precarious;
- Destination region with urban character and preponderance of services and industrial activities.

The number of individuals considered in each region is 9900. The evolution of unemployment and employment are studied according to three rates of job creation: 3%, 5%, and 10%. For demographic characteristics, the table below summarize the most important (Table 1).

**Table 1.** Demographic settings

| Variable | Rural | Urban |
|---|---|---|
| Birth rate | 0.02478 | 0.02478 |
| Death rate | 0.00441 | 0.00441 |
| Marriage rate | 0.01 | 0.01 |
| Divorce rate | 0.001 | 0.0015 |
| Success rate baccalaureate | 0.455 | 0.554 |

### 3.3 Simulations Results

By looking at the evolution of the unemployment rate in the urban area, we can see that developments with or without migration are similar in terms of trend. In fact, for a job creation rate (UJC) of 3%, the unemployment rate remains very high with or without migration (69.1% with migration and 66.7% without migration). For UJC values of 5 and 10%, the rate of unemployment decreases and stagnates around 3% (in the presence or absence of migration). However, it converges faster towards this value for an UJC of 10% in the absence of migrations (Fig. 1).

Concerning urban employment, we find that the higher is the UJCs, the higher is the occupancy rate. When considering migration, the proportions are more important for UJCs above 3%. For an UJC of 3%, the proportion of the occupied is greater without migration. This is certainly due to the fact that an UJC of 3% does not allow the absorption of urban unemployment (native + migrants) and therefore decreases the proportion of the population occupied (Fig. 2).

For a Rural Job Creation rate (RJC) of 3%, rural employment decreases. This rate of job creation does not allow for the expansion of the rural sector either in the presence or in the absence of labour migration. For a RJC of 5%, there is a stabilization of rural employment in the absence of mobility of workers and a decline with migration. The RJC of 5% lets the population's occupancy rate around 20%.

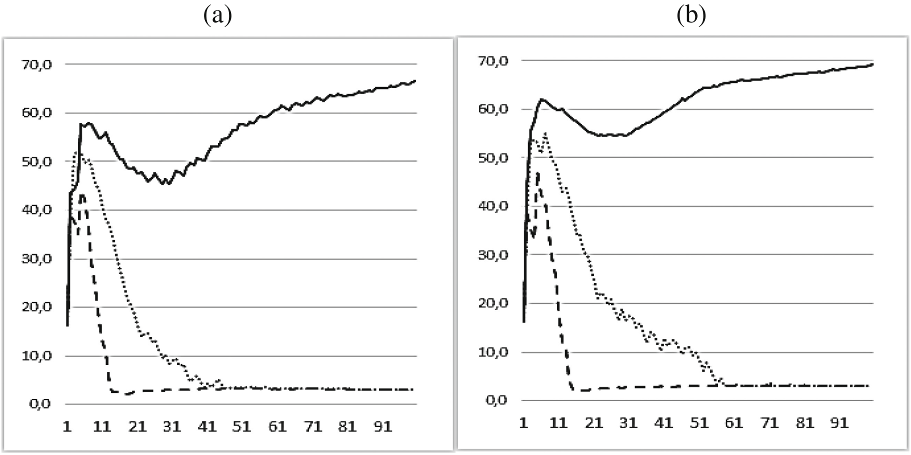


**Fig. 1.** Variations of unemployment rate in urban area for different values of UJC: UJC = 0.03 (solid line); UJC = 0.05 (dotted line) and UJC = 0.1 (dashed line). (a) In presence of massive labour migration, (b) in absence of labour migration

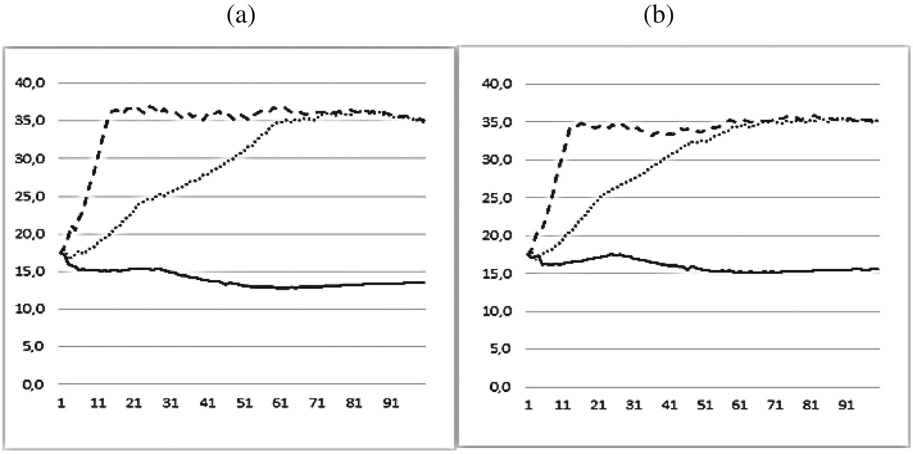


**Fig. 2.** Variations of employment in urban area for different values of UJC: UJC = 0.03 (solid line); UJC = 0.05 (dotted line) and UJC = 0.1 (dashed line). (a) In presence of massive labour migration, (b) in absence of labor migration

When considering migration, this RJC decreases the occupancy rate less rapidly than for a RJC of 3%. The chances of finding a formal rural job with a 3% RJC are lower than in urban area. This incites the rural unemployed to migrate and deprives the rural sector of a potential workers. For a RJC of 10%, the rural sector develops more

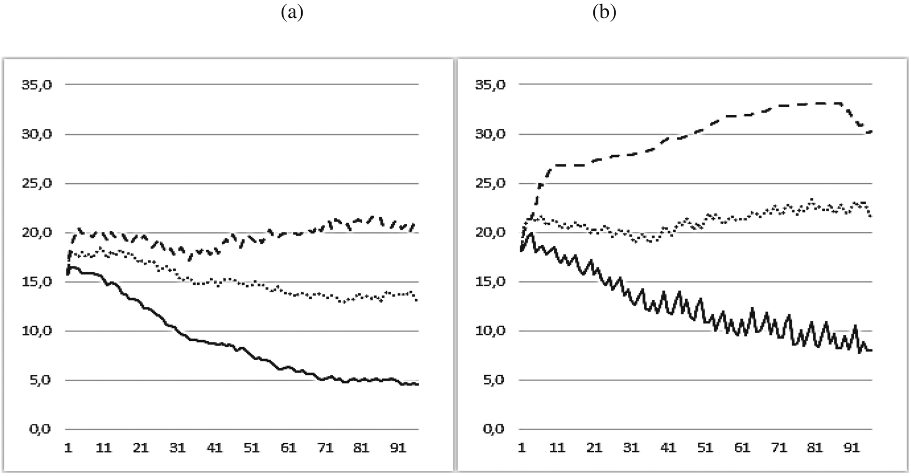


**Fig. 3.** Variations of employment in rural area for different values of RJC: RJC = 0.03 (solid line); RJC = 0.05 (dotted line) and RJC = 0.1 (dashed line). (a) In presence of massive labor migration, (b) in absence of labour migration

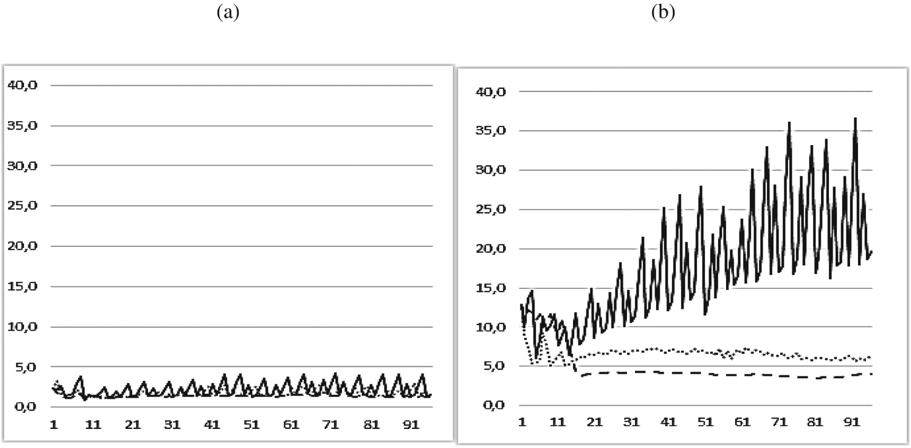


**Fig. 4.** Variations of unemployment in rural area for different values of RJC: RJC = 0.03 (solid line); RJC = 0.05 (dotted line) and RJC = 0.1 (dashed line). (a) In presence of massive labour migration, (b) In absence of labour migration

markedly in the absence of migration (occupancy rate in the presence of migration is 21%, and in their absence 32%) (Fig. 3).

For rural unemployment, the higher is the RJC, the lower is the unemployment rate. In the absence of migration and for a RJC of 3%, the unemployment rate exceeds 35%, and for RJCs of 5 and 10%, unemployment rate decreases and stabilizes around 7 and 4% respectively. However, in presence of migrations, the unemployment rate remains at a very low level and does not reach 5% (Fig. 4).

These results show the importance of migration flows and their effects on the two regions of departure and arrival in terms of employment and unemployment in particular for rural area which is deprived of a potential labour force by migration.

For Urban area, the incidence of migrations is more important in terms of volume than in terms of trends.

## 4 Conclusion and Perspectives

The proposed model represents a first prototype of an agent based labour migration model with heterogeneous agents. We attempt in this first version to introduce the agents' heterogeneity in the decision to migrate to reflect the selective nature of migration and in the job search through hiring probabilities and searching times. The model allows us to identify the incidence of migration on labour markets of origin and destination areas and their evolution through several scenarios in contrast to the most migration models (conventional or agent based models).

The simulation results show an impoverishment of the origin area in favour of destination's region. We have also remarked the existence of thresholds in terms of Job creation rates (rural or urban) above which rural workers do not migrate.

We notice also that despite very high job creation rates, residual unemployment persists. This is certainly due to a bad matching between jobs and agents qualification.

Some results of the model are not exploited because of the lack of pertinence of data (such education and creation of firms).

A second version of the model is in progress improving the process of job searching. Workers will have the choice between integrating an informal job and remaining unemployed. We will introduce also the effect of social network in the job search.

## References

1. Schelling, T.C.: Dynamic models of segregation. J. Math. Sociol. **1**(2), 143–186 (1971)
2. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press, Washington (1996)
3. Neugart, M.: Labor market policy evaluation with ACE. J. Econ. Behav. Organ. **67**(2), 418–430 (2008)
4. Ballot, G., Kant, J.D., Goudet, O.: An agent-based model of the french labor market, applied to the evaluation of employment policies: the example of the "generation contract". Rev. écon. **67**(4), 733–771 (2016)

5. Janssen, M.A., Jager, W.: Simulating market dynamics: Interactions between consumer psychology and social networks. Artif. Life **9**(4), 343–356 (2003)
6. Klabunde, A., Willekens, F.: Decision-making in agent-based models of migration: state of the art and challenges. Eur. J. Popul. **32**(1), 73–97 (2016)
7. Naqvi, A.A., Rehm, M.: A multi-agent model of a low income economy: simulating the distributional effects of natural disasters. J. Econ. Interac. Coord. **9**(2), 275–309 (2012)
8. Hassani-Mahmooei, B., Parris, B.W.: Climate change and internal migration patterns in Bangladesh: an agent-based model. Environ. Dev. Econ. **17**(06), 763–780 (2012)
9. Espındola, A.L., Silveira, J.J., Penna, T.J.P.: A Harris–Todaro agent-based model to rural-urban migration. Braz. J. Phys. **36**(3A), 603 (2006)
10. El Saadi, N., Bah, A., Belarbi, Y.: An agent-based implementation of the Todaro model. In: Advances in Practical Multi-Agent Systems, pp. 251–265. Springer, Berlin (2010)
11. Todaro, M.P.: A model of labor migration and urban unemployment in less developed countries. Am. Econ. Rev. **59**(1), 138–148 (1969)
12. Lauby, J., Stark, O.: Individual migration as a family strategy: young women in the Philippines. Popul. Stud. **42**(3), 473–486 (1988)
13. Harris, J.R., Todaro, M.P.: Migration, unemployment and development: a two-sector analysis. Am. Econ. Rev. **60**(1), 126–142 (1970)

# Applying of the Classifications Trees Method in Forecasting of Risk Groups of Intolerant Behavior

I.V. Vicentiy[1], S.M. Eliseev[2], and A.V. Vicentiy[1,3](✉)

[1] Branch of the Murmansk Arctic State University in the Apatity City,
Lesnaya Street, 29, Apatity 184205, Russia
felysite@yandex.ru, alx_2003@mail.ru

[2] St. Petersburg State University, Smolny Street, 1/3, St. Petersburg 191124, Russia
1956eliseev@inbox.ru

[3] Institute for Informatics and Mathematical Modelling of Technological Processes of the Kola Science Center RAS, Fersman Street, 24A, Apatity 184209, Russia

**Abstract.** In this paper the application capabilities of the heuristic method of mathematical statistics - classifications trees - are considered. Based on the analysis of sociological research results of the political tolerance of students in the Murmansk region, an algorithm for identifying and forecasting risk groups for intolerant behavior is described.

**Keywords:** Math statistics · Classifications trees · Sociological research · Factors of intolerance · Risk groups of intolerance

## 1 Introduction

Modern social practice demonstrates examples of extremism. Extremism refers to the extreme forms of manifestation of intolerance in society. Nowadays, this problem has a global character and is actualized at the international level. The British Institute of Economics and Peace published the “Global Terrorism Rating of 2015” [1]. Researchers found that in 2014, the terrorist threat in the world has increased significantly: the number of attacks and the number of their victims has increased. Since 2000, the number of deaths in terrorist attacks has increased nine-fold: from 3329 to 32658 people [2]. Russia is no exception. According to the “Portal of Legal Statistics” website, the total number of registered crimes in the territory of the Russian Federation in 2016 declined. Nevertheless, the number of crimes of a terrorist nature increased by 44.8% and the number of extremist crimes increased by 9.1% (Fig. 1) [3]. For the period from 2010 to December 2016, the number of extremist crimes in the Russian Federation increased from 656 to 1410, that is more than 2 times (Fig. 2) [3].

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_32

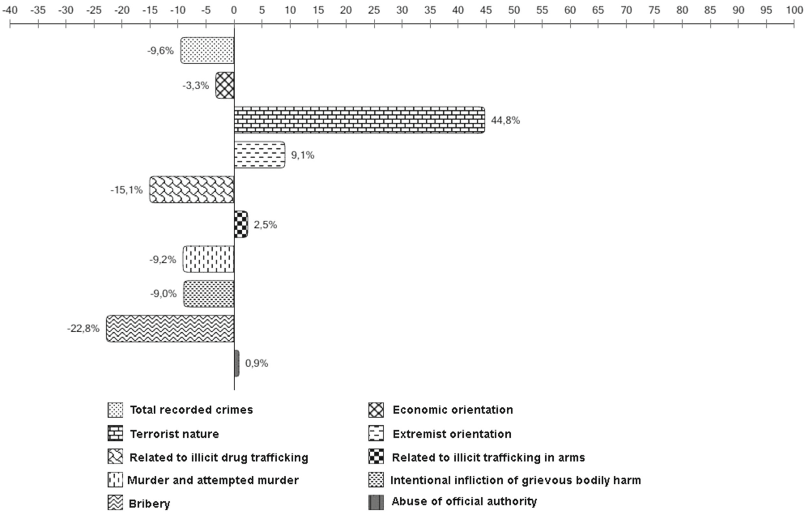


**Fig. 1.** Dynamics of certain types of crime in Russia in 2016. Source: The state of crime in Russia in January–December 2016. - M., 2016. - 52 p. Access mode: Portal of legal statistics http://crimestat.ru/

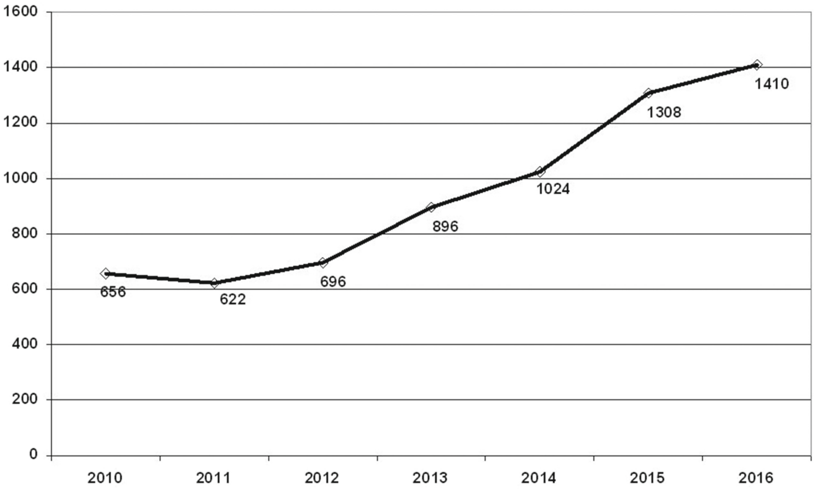


**Fig. 2.** The dynamics of registered crimes of extremist orientation in the Russian Federation, in absolute terms. Source: According to the Portal of Legal Statistics. URL: http://crimestat.ru/

In the context of statistics of extremism, the methods of preventing intolerance in all its forms become important. Tolerance in this context is seen as a means of conflict-free coexistence of representatives of diverse social groups. The classic definition of tolerance was given by Peter Nicholson, who defines tolerance as: “the virtue of refraining from exercising one’s power to interfere with others’ opinion or action although that deviates from one's own over something important and although one morally disapproves of it." [4] The scientific basis for preventive measures for intolerance is its sociological support for identifying centers of intolerance in society - the so-called risk groups of intolerance. These risk groups of intolerance are target groups for specialists of different profiles: social psychologists, social educators, representatives of law enforcement officials and others.

Mass polls of the population using adequate empirical indicators of tolerance and methods of mathematical statistics in the process of statistical analysis of a large volume of primary empirical data make it possible to capture moods in the society, analyze it in dynamics, measure the level of social tension and social distance, conduct comparative analysis on various sociological criteria in specific social groups. Such a statistical analysis makes it possible to diagnose and predict intolerant behavior in specific socio-demographic groups. These data provide a basis for timely and prompt preventive work to prevent negative socio-political processes, extremism, xenophobia, violence and hatred in society and specific socio-demographic groups [5].

## 2 Methods

At the same time, the problem of choosing adequate methods of mathematical statistics in the process of analyzing a large array of primary sociological data for identifying possible social factors of intolerant behavior is being actualized. Wrong choice of methods of mathematical statistics for the processing of sociological research data can lead to a non-valid result. At present, there is no universal method that allows us to study the statistical relationships between variables. Therefore, different researchers choose the most suitable methods for mathematical statistics. Among the methods of studying social factors, traditionally popular among sociologists are:

- statistical analysis of the conjugacy of variables attributes and the calculation of the coefficients of mutual conjugation of Chuprov, Pierson, Cramer, Goodman, and Kruskal, Fechner, and others;
- correlation analysis and calculation of the Spearman and Kendall coefficients;
- regression analysis;
- statistical methods of comparing the average;
- cluster analysis;
- discriminant analysis;
- factor analysis;
- analysis of the main components (component analysis) and others [6–11].

The peculiarity of studying the factors of intolerant behavior is that it is necessary to analyze large volumes of social information. This is a large set of hypothetical variables of influence, describing the socio-demographic status of respondents. In the process of sociological research, large amounts of data are collected on various parameters, for example, sex, age, financial position, characteristics of socialization conditions, professional status, educational level and other characteristics of respondents. Using traditional methods to identify the relationships between variables is an extremely routine job that requires significant time, material and other resources. Among the various statistical methods for studying the relationship between variables, the classification tree method is beneficially different. The purpose of this paper is to describe the possibilities of using the method of classification trees in predicting the risk groups of intolerant behavior.

The method of classification trees can be considered as a method of exploratory analysis or as a "last resort" when all traditional methods are rejected. Classification trees, according to many researchers, is one of the best for researching a large amount of data. Classification trees are a classification analysis method that allows you to predict the belonging of objects to a particular class, depending on the corresponding values of the characteristics that characterize the objects. Classification trees are able to perform one-dimensional branching on variables of various types (categorical, ordinal, interval). There are no restrictions on the law of distribution of quantitative variables. The method makes it possible to analyze the contributions of individual variables to the classification procedure. Classification trees sometimes have a very complex structure, which depends on the data. However, the use of special graphical procedures makes it possible to simplify the interpretation of results even for very complex trees. The ability to graphically represent the results and simplify the interpretation largely explains the great popularity of classification trees in applied fields. But the most important distinguishing features of classification trees are its hierarchical structure and wide applicability. The method allows the user to construct trees of arbitrary complexity, using controlled parameters and minimizing classification errors. A wide range of applicability of classification trees makes them a very attractive tool for data analysis.

To date, there are several different algorithms for constructing classification trees (or decision trees). The most famous of them are implemented in the SPSS (Statistical Package for the Social Sciences) Classifications Trees: CHAID, ECHAID, CRT, QUEST. It is generally accepted that CHAID is the most flexible algorithm, allowing to build both classification trees and regression trees with any number of branches in nodes. In this paper we used a method of constructing decision trees, called CHAID (Chi square automatic interaction detection).

In classification studies, in particular, limited to studying demographic and behavioral information or based on data extracted from customer databases, often only categorical (nominal) information can be available. In addition, the number of variables of this type can be extremely large. To study the relationship between a single predictor and an interesting response, you can use the contingency table. However, if there are many potential predictors, then it is necessary to build a set of tables. When considering two-factor, three-factor interactions, and also interactions of higher order, the number of required tables increases with the rate of geometric progression. CHAID offers a method of effectively searching for the relationship between predictor variables and categorical response. The resulting output is a tree graph that visualizes combinations of predictor categories (which provide the most significant difference in percentages of observations) that fall into the variable response category. Therefore, CHAID is an effective means of identifying segments that maximize the interest of the response.

CHAID is a heuristic statistical method based on a decision tree that examines the relationship between the set of categorical predictor variables and the categorical target variable. It forms a tree diagram that reveals the categories of predictors that most significantly predict the interest categories of the target variable. More detailed information on the method of classification trees can be obtained in the special literature [12–14].

## 3 Results

In this article, we would like to share our experience of applying the method of classification trees in the process of sociological research of students' political tolerance. This is a long-term sociological study covering the period from 2007 to 2016. This study makes it possible to demonstrate the possibilities and applied value of the method of classification trees in solving the specific applied sociological problem. This is the task of diagnosing and predicting the risk groups of intolerant behavior as a way of early warning of extremism in the youth environment.

The research included three stages of collecting primary sociological data. Each stage implemented a specific research task:

1. "Representations of tolerance in the subculture of students of the Murmansk Region". This stage included: (a) the collection of field information (2007); (b) questionnaire survey of full-time students of The Kola branch of Petrozavodsk State University (Apatity, Murmansk region). The survey was conducted on a stratified sample (proportional selection), the volume of which was 319 people.
2. "Attitudes of political tolerance in the political subculture of students in the Murmansk region". This stage included: (a) the collection of field information (2007); (b) questionnaire survey of full-time students of The Kola branch of Petrozavodsk State University (Apatity, Murmansk region). The survey was conducted on a stratified sample (proportional selection), the volume of which was 449 people.
3. "Values and attitudes of political tolerance in the subculture of students in the Murmansk region". This stage included: (a) the collection of field information (2011); (b) questionnaire survey of full-time students of The Kola branch of Petrozavodsk State University (Apatity, Murmansk region). The survey was conducted on a random sample (repetition-free selection), the volume of which was 421 people.

The main scientific task of the research is to study the political tolerance in the subculture of student youth in conditions of the Russian border region (Murmansk region) and the identification of risk groups for intolerant behavior [15].

The general direction of empirical sociological research was determined by the strategy of quantitative sociology. The students of the full-time department of the Kola branch of Petrozavodsk State University (Apatity, Murmansk Region) were questioned.

Based on the idea of a verbal representation of values and attitudes, the method of "pencil and paper" was used in the process of studying the students' political tolerance. For the purpose of research, special scales were designed. These scales allow measuring the political tolerance of the respondents. Respondents were offered judgments, each of which was to be evaluated on a numerical scale. The technique represents a set of judgments and five-member scales. Using scales, respondents need to express their degree of agreement with the judgments. The methodology includes 26 scales formulated as tolerant and intolerant value judgments.

We designed these scales specifically for our research and tested its validity. These scales allow measuring the level of political tolerance among respondents. The basis for the development of our scales, which fix values and attitudes of political tolerance, was the scale of R. Likert (the scale of summary estimates). Our methodology "The level of political tolerance", designed according to the rules of Likert's scaling and allows to reveal the existence of differences in the indicators of political tolerance in different socio-demographic groups.

For these purposes, in the questions of the questionnaire, we have included empirical indicators that measure the socio-demographic status of the respondent, and hypothetically are factors that determine the intolerant behavior of individuals:

- Demographic characteristics (gender and age);
- Educational capital (course of study, faculty of training, form of training, etc.);
- Economic capital;
- The configuration and characteristics of the family in which the respondent was brought up (type of family, parents' education, professional status of parents, political activity of parents, etc.);
- Religious identity (religion, religious practices);
- Socio-political practice (or integration into the system of social relations, which are represented by labor, political, leisure associations of individuals);
- Social environment (which is represented by agents of primary and secondary socialization, such as the institution of the family, the institution of education, the institution of religion, socio-political associations and the mass media).

The target (dependent) variable in our research is the variable "political tolerance". A predictor (independent) variables are variables that describe the socio-demographic status of the respondents.

In such researches, the problem of processing and analyzing large amounts of data often arises. These data are necessary to obtain predictive information about the most likely groups of intolerant behavior and the formation of an effective strategy for the prevention of intolerance and extremism in the social environment. As a result of our empirical research, we obtained data on whether students demonstrate tolerance or intolerant attitudes on each of the 26 scales, as well as socio-demographic information about them (age, sex, income category, etc.). It was necessary to find out which subgroups are more likely to demonstrate intolerance attitudes in order to focus on them within the framework of social prevention programs for intolerance.

The traditional approach to solving this problem is the approach related to the construction of a set of conjugacy tables and the identification of variables that are statistically related to the variables of the installations for political intolerance. The disadvantage of this approach is that all variables should be analyzed separately. However, since the intolerance attitudes can depend on combinations of socio-demographic characteristics, it is necessary to study three-dimensional, four-dimensional tables and, possibly, tables of even greater dimension. Studying such multidimensional tables requires a lot of effort, time, and can lead to confusion. As a result, this way of data analysis is often ineffective. Also, for constructing a model of the relationship of variables in a multi-input contingency table, we can use log-linear models, but they usually require large samples. CHAID allows us to solve this problem without ever considering the complete tables of large dimension. Instead of processing high-dimensional tables, CHAID examines all two-input tables and looking for the social characteristic most closely related to the target variable. In the next step, CHAID examines the splitting of the categories of this predictor using the next most important predictor.

In this way, the CHAID method can provide useful information about which demographic characteristics are associated with an interesting response in a situation where the use of formal models encounters difficulties (such as a large number of categorical predictor variables or an insufficient sample size). Questions that can be answered by the CHAID procedure can be stated as follows. At first, Which of the predictor variables are interrelated with the target variable? In other words, which of the predictors can help predict the target variable. Secondly, which combinations of categories of these predictors give the highest percentage of hits in the desired category of response variable? These predictors are the target groups most interesting for further study.

CHAID is the most popular of the methods of predictive modeling and data mining, from implemented in AnswerTree. It is not binary, that is, it can form more than two categories at any level of the tree. Therefore, it seeks to build a tree with more branches than binary methods of constructing a tree. It works with all types of variables and allows both observation weights and frequency variables. The decision tree methods allows the researcher to perform sequential separation operations automatically until the specified criterion is met. Thus, the decision tree method provides a means of “searching” (“ransacking”) a data set to discover key relationships.

Formalization, processing and statistical analysis of the data obtained during the questionnaire survey was carried out in the SPSS Base 17.0 software environment, as well as with the help of an additional module called “Classifications Trees”. Carrying out a statistical analysis to identify the relationship between attitudes to a certain type of demonstrated behavior and the socio-demographic characteristics of students, we checked the connection on the scales that measured the attitudes of tolerance to political parties.

As a method of statistical analysis, which allows us to identify relationships between variables, their tightness and direction, we used dependency analysis based on the decision tree model by the CHAID method. At the stage of selecting the statistics used for our target nominal variable, we used the likelihood ratio test. This is a statistical test that is used to check the limitations on the parameters of statistical models evaluated on the basis of sample data. This method is considered more stable than Pierson’s chi-square, but requires more time-consuming calculations (but this does not really matter much when very powerful processors have appeared in modern computers). In addition, for small samples this method is preferable, in contrast to Pearson’s Chi-square method.

Further, as an example of the implementation of the method of classification trees in the process of revealing the factors of intolerance, the analysis of factors of intolerance is presented according to one of the 26 appraisal judgments (scales) of the methodology.

Figure 3 shows the final diagram of the tree, which shows the variables that were found to be statistically significant in the study, as well as a system of hierarchical relationships between them. The tree represented in Fig. 3 is implemented by a special set of commands. A fragment of the batch file is shown in Fig. 4.

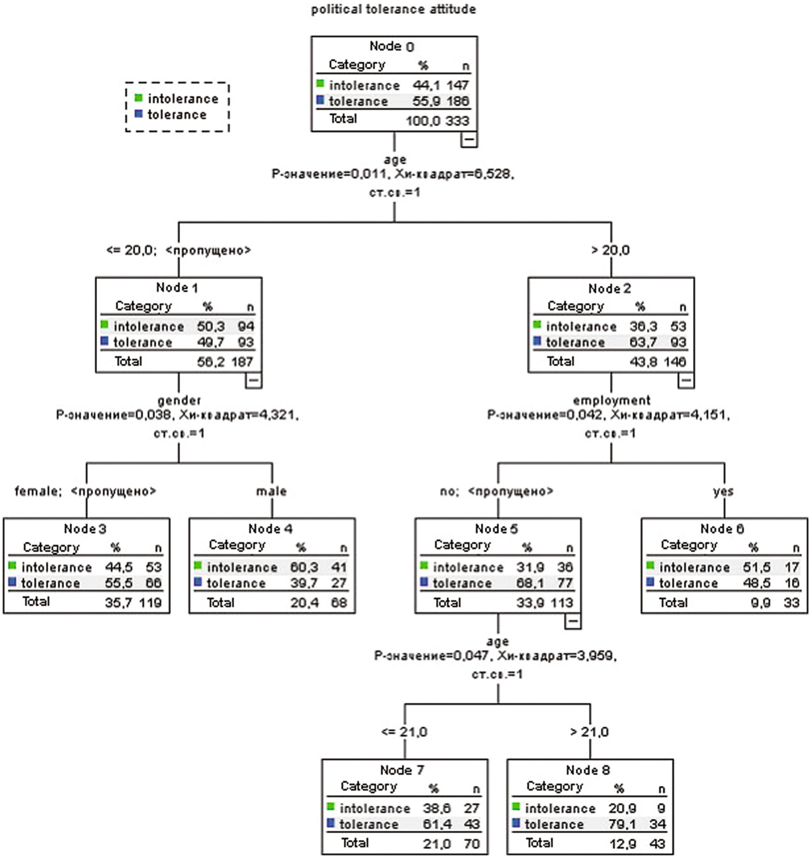
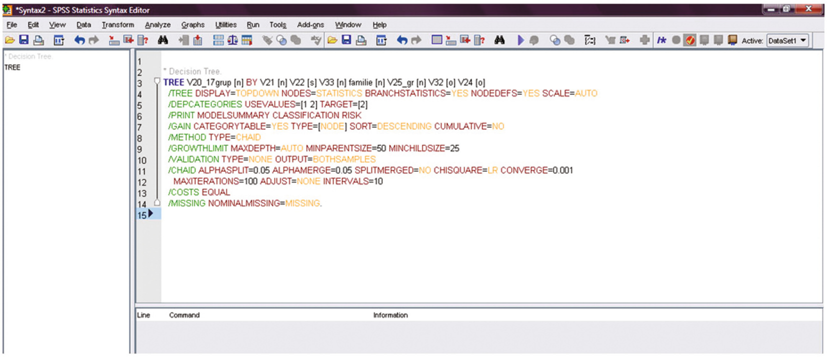


**Fig. 3.** The tree diagram "Factors of political intolerance".

Statistical analysis of empirical data using the CHAID decision tree model has made it possible to identify several social factors that influence the actualization of the attitudes of students' political intolerance. They are associated with the socio-demographic and socio-cultural characteristics of the individual. With regard to stratification within the student subculture, the variables that determine the intolerance attitudes were male sex, younger respondents, labor employment (involvement in the labor relations system).

Table 1 presents the node ranking and shows that node 8 is the best. The "response" cell shows that 79.1% of observations from node 8 belong to the category of intolerance attitude, which is described by such characteristics of respondents as unemployed respondents over twenty-one years old.


*Syntax2 - SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Tools Add-ons Window Help
Active: DataSet1
* Decision Tree.
TREE
Line Command Information

```
* Decision Tree.
TREE V20_17grup [n] BY V21 [n] V22 [s] V33 [n] familie [n] V25_gr [n] V32 [o] V24 [o]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[1 2] TARGET=[2]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /GAIN CATEGORYTABLE=YES TYPE=[NODE] SORT=DESCENDING CUMULATIVE=NO
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=50 MINCHILDSIZE=25
  /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO CHISQUARE=LR CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=NONE INTERVALS=10
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.
```

**Fig. 4.** Fragment of the batch file of the classification tree in the process of identifying the risk groups of intolerance.

**Table 1.** Winnings for nodes, target category intolerance.

| Winnings for nodes | | | | | | |
|---|---|---|---|---|---|---|
| Node | Node | | Winning | | Response | Index |
| | Number of observations | Percent | Number of observations | Percent | | |
| 8 | 43 | 12,9% | 34 | 18,3% | 79,1% | 141,6% |
| 7 | 70 | 21,0% | 43 | 23,1% | 61,4% | 110,0% |
| 3 | 119 | 35,7% | 66 | 35,5% | 55,5% | 99,3% |
| 6 | 33 | 9,9% | 16 | 8,6% | 48,5% | 86,8% |
| 4 | 68 | 20,4% | 27 | 14,5% | 39,7% | 71,1% |

Method of construction: CHAID.
Dependent variable: political intolerance attitude.

The results of the statistical analysis made it possible to conclude that the risk group for intolerant behavior includes, first of all, male students 20 years or younger (that is, the demographic factor of intolerance was actualized) and also working students older than twenty years.

## 4 Discussion

The method of classification trees is an effective tool for processing sociological data, which was confirmed by the application of this method in predicting the risk groups of intolerant behavior. Sociological studies based on methods of mathematical statistics are an operative means of predicting and early warning of intolerant behavior and extremism in the youth environment. The results of the research have practical importance, since they serve the purposes of identifying the risk groups of intolerant behavior using statistical methods.

The described algorithm for processing sociological data can be used as the basis for creating an information and analytical system for predicting intolerant behavior in society. Replenishment of the bank of sociological data (including an increase in the sample size) based on the proposed list of variables (or even extended ones) will allow further classification with accuracy and identify variables that are more conducive to intolerance in society.

## References

1. Global index of terrorism. Humanitarian Encyclopedia [Electronic resource]. Center for Humanitarian Technologies, 2006–2016 (2016). http://gtmarket.ru/ratings/global-terrorism-index/info. Last revised 30 Oct 2016, (in Russian)
2. Eliseev, S., Vicentiy, I., Gluchich, V.: Monitoring of political tolerance as a tool for early warning on youth extremism. Indian J. Sci. Technol. **10**(2), 115–123 (2017)
3. Portal of legal statistics. [Electronic resource]. The state of crime in Russia in January–December 2016, 52 p. (2016). http://crimestat.ru. (in Russian)
4. Nicholson, P.: Toleration as a moral ideal. In: Horton, J., Mendus, S. (eds.) Aspects of Toleration: Philosophical Studies, p. 162. Methuen, London (1985)
5. Vicentiy, I.V.: The student body political culture (according to a study of the Murmansk region). Int. J. Cult. Res. **1**(14), 76–81 (2014)
6. Ermolaev, O.J.: Mathematical Statistics for Psychologists: The Textbook. Flint, Moscow (2006). (in Russian)
7. Glass, J., Stanly, J.: Statistical Methods in Pedagogy and Psychology. Progress, Moscow (1976). (in Russian)
8. Gmurman, V.E.: The Theory of Probability and Mathematical Statistics: A Manual for Schools. Higher School, Moscow (2003). (in Russian)
9. Grabar, M., Krasnyanskaya, K.A.: Application of Mathematical Statistics in Educational Research Non-parametric Methods. Pedagogika, Moscow (1977). (in Russian)
10. Granichina, O.: Mathematical and Statistical Methods of Psychological and Educational Research: Study Guide. Publishing House of VVM, St. Petersburg (2012). (in Russian)
11. Hollender, M., Wolfe, D.: Nonparametric Statistical Methods. Finance and Statistics, Moscow (1983)
12. Arabie, P., Hubert, L.J., De Soete, G.: Clustering and Classification, 500 p. World Scientific, Singapore (1996)
13. Yohannes, Y., Hoddinott, J.: Classification and Regression Trees: An Introduction, 29 p. (1999)
14. Buntine, W.: Learning classification trees. Stat. Comput. **2**, 63–73 (1992)
15. Eliseev, S.M., Ustinova, I.V.: The characteristics of college students' political tolerance. Russ. Educ. Soc. **53**(9), 71–82 (2011)

# Prediction of Attacks Against Honeynet Based on Time Series Modeling

Pavol Sokol[1(✉)] and Andrej Gajdoš[2]

[1] Faculty of Science, Institute of Computer Science,
Pavol Jozef Safarik University in Kosice, Jesenna 5, 040 01 Kosice, Slovakia
pavol.sokol@upjs.sk

[2] Faculty of Science, Institute of Mathematics,
Pavol Jozef Safarik University in Kosice, Jesenna 5, 040 01 Kosice, Slovakia
andrej.gajdos@student.upjs.sk

**Abstract.** Honeypots are unconventional tools to study methods, tools, and goals of attackers. In addition to IP addresses, these tools collect also timestamps. Therefore, time series analysis of data collected by honeypots can bring different view for prediction of attacks. In the paper, we focus on the model AR(1) and bootstrap based on AR(1) model to predict attacks against honeynet. For this purpose, we used data collected in CZ.NIC honeynet consists of Kippo honeypots in medium-interaction mode. The prediction of attacks is based on 75 weeks data and it has been verified by five weeks data. In the paper, we have shown that prediction model AR(1) and bootstrap based on AR(1) model are suitable for prediction of attacks.

**Keywords:** Honeypot · Attack · Prediction · Time series analysis · Bootstrap

## 1 Introduction

There are increasing security threats in current information society. Therefore, an important part of information security is the protection of information. Common security tools, methods and techniques used before are ineffective against these new threats and it is necessary to use other tools and techniques. It seems that the network forensic tools, especially honeypots and honeynets, are very useful.

A **honeypot** is "a computing resource, whose value is in being attacked" [1]. Lance Spitzner defines honeypots as "an information system resource whose value lies in unauthorized or illicit use of that resource" [2]. Honeypots are very useful for learning about attackers and their objectives, methods and tools. The honeypots are classified based on the **level of interaction**. We know low-interaction, medium-interaction honeypots, and high-interaction honeypots. The difference between these types is the extent, to which the attacker is allowed to interact with the system. For purpose of this paper, the **medium-interaction honeypots** are important. These types of honeypots emulate network services

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_33

and allow attackers to access to emulated environment. Example of this type of honeypot is SSH honeypot Kippo [3]. In our research we used **honeynet**, which extends "the concept of a single honeypot to a highly controlled network of honeypots" [4].

Collection of data from honeypots and honeynets and subsequent analysis of these data is the main purpose of using these security tools. Learning new unconventional information about the attackers and their methods helps with protection of the organizations. Honeypots and honeynets capture a number of different data. Each record collected by these security tools contain at least a timestamp, type of service, IP addresses of honeypot and attacker and specific data per type of honeypots. Important collected data are timestamps. **Timestamp** can be defined as an unambiguous representation of some instant in time [5]. Honeypots add the timestamp to each record. Therefore, data collected by honeypots can be considered as **time-oriented data**. In the time-oriented data collected by honeypots, there is a relationship between attacks and time period [6]. In other words, attacks in one period affect attacks in the others time periods. Based on this, data collected by honeypots will be used for **time series analysis and modeling**. Time series models "attempt to make use of the time dependent structure present in a set of observations" [7].

In this paper, we apply very well-known and useful approach, **Box-Jenkins methodology** [8,9]. It is based on ARMA/ARIMA modeling. Since we deal with stationary data, we do not discuss ARIMA models here. **ARMA(p,q) models** are time series models, which can be expressed as:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-p} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-q} \tag{1}$$

where $X_t$ is the sum of a weighted average of the previous $p$ values in time with weights $\phi_i$ and a weighted average of the previous $q$ random error terms with weights $\theta_i$. $\varepsilon_t$ is an additional random error term. The error term is generally assumed to be sampled from a white noise process of a zero mean and constant variance.

Special cases of general ARMA(p,q) model are Auto-regressive (AR) model and Moving average (MA) model. **AR(p) model** can be expressed as:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-p} \tag{2}$$

On the other hand, **MA(q) model** can be expressed as:

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-q} \tag{3}$$

In the paper, we also apply **bootstrap** [10–13] as an alternative approach for point predictions and as a tool to gain prediction intervals which cannot be obtained directly from model when residuals are non − Gaussian.

Bootstrap methods are computational methods with statistical background and employ the computational power of computers while they are based on Monte Carlo simulations. They can be helpful in cases when theory says nothing or is hard to develope.

The main aim of this paper is to **predict number of attacks** using the time series analysis. This prediction also helps with protection of the organizations, since the administrators are better informed and they can be better prepared for security incidents in their organizations. To formalize the scope of our work, we state following research questions:

- Is it possible to use Box-Jenkins methodology for attack prediction?
- Is it possible to use model-based (residual) bootstrap to obtain accurate predictions of attacks?
- Which of the above approaches is better for attack predictions?

This paper is organized into five sections. Section 2 focuses on the review of published research related to time series analysis and prediction of attacks and security incidents. Section 3 outlines the dataset and methods used for experiment. Section 4 focuses on prediction of attacks in honeypots and discusses the results. The last section contains conclusions and our suggestions for the future research.

## 2 Related Works

As it was mentioned before, the main aim of safety tools like honeypots and honeynets is to capture an enormous number of time-related data order to detection of attacks or prediction of attacks. The most part of the papers focuses on detection of attacks rather than prediction of attacks [14]. This section provides an overview of papers that focus on the prediction of attacks or security incidents using time series analysis.

Papers [7,15] focus on prediction of security incident based on **ARIMA time series model**. In [7] authors discuss time series models with a large set of security incident data. Authors also compare the forecasts from time series models with forecasts from Non-Homogeneous Poisson Process (NHPP) software reliability growth (SRG) models. Another paper focused on prediction of attack based on ARIMA time series is [15]. Paper aims to exploit temporal correlations between the number of attacks per day in order to predict the future intensity of cyber incidents. They were interested in 4 types of attack - denial of service (DOS), malicious email, malicious URL, and attack on internet facing service (AOIFS). For this purpose, they proposed system used ARIMA time series forecasting on all previously collected incidents.

On the other hand, paper [16] focuses on prediction of attack based on **ARMA time series model**. It proposes an intrusion detection system for wireless networks for industrial automation-process automation (WIA-PA). In the paper, authors focused on modeling and analyzing traffic flow data by time-sequence techniques. Also, they proposed a data traffic prediction model based on autoregressive moving average (ARMA) using the time series data.

Another research groups conduct research in field of prediction of attack based on **generalized ARCH (GARCH) models**. In [17] authors propose a framework for statistically analyzing long-term vulnerability time series between January 1999 and January 2016. For this purpose, generalized ARCH (GARCH) models and SARIMA model are used for National Vulnerability Database. Another example is paper [18]. Authors use gray-box FARIMA+GARCH models and discuss the integration of Extreme Value Theory (EVT) and the Time Series Theory (TST). In this paper, they show that EVT can offer long-term predictions (e.g., 24-hour ahead-of-time), while graybox TST models can predict attack rates 1-hour ahead-of-time at an accuracy that can be deemed practical.

In the field of prediction of attacks it is necessary to mention also paper [19]. Authors focus on the problem of forecasting attack sources based on past attack logs from several contributors. They evaluate and combine several factors - attacker-victim history using time-series, attackers and/or victims interactions using neighborhood models and global patterns using singular value decomposition. In terms of time series analysis, they use an **Exponential Weighted Moving Average (EWMA) model**.

## 3 Data Collection and Analysis Methodology

The time series data were collected from the honeynet located in the **CZ-NIC network**. This honeynet consists of Kippo [3] honeypots, which run on 22 port in medium-interaction mode. The honeypots allow attackers to log into shell in this mode and capture data from attackers. These data, which are used for Turris greylist [20], are transformed to the following format:

- **IP** IP address of device which connected to honeypot
- **MD5** MD5 hash of malware captured by honeypot
- **KIPPO_COMMANDS** - number of commands entered to honeypot
- **SHA256** - SHA256 hash of malware captured by honeypot

The honeypots have collected authentication attempts from 2nd November 2014 to 8th May 2016. During this period **179.540 records** were collected. Dataset contains 80 weeks. Data from **75 weeks** (167.862 records) are used **for training** and data from **5 last weeks** (11.678 records) are used to **compare with predicted data**. The first step of analysis (in finding the right model) is a visualisation of data (Fig. 1). Data seems to be stationary due to following aspects: no apparent trend – just constant one, stable variance and no apparent seasonality.

Since we would like to use a class of models ARMA, it is needed to test data for stationarity. For this purpose, we used tool R [21] and two tests implemented in R packages mentioned in parentheses:

- Augmented Dickey-Fuller (ADF) Test (tseries) [22,23] and
- KPSS (Kwiatkowski–Phillips–Schmidt–Shin) Test (tseries) [24].

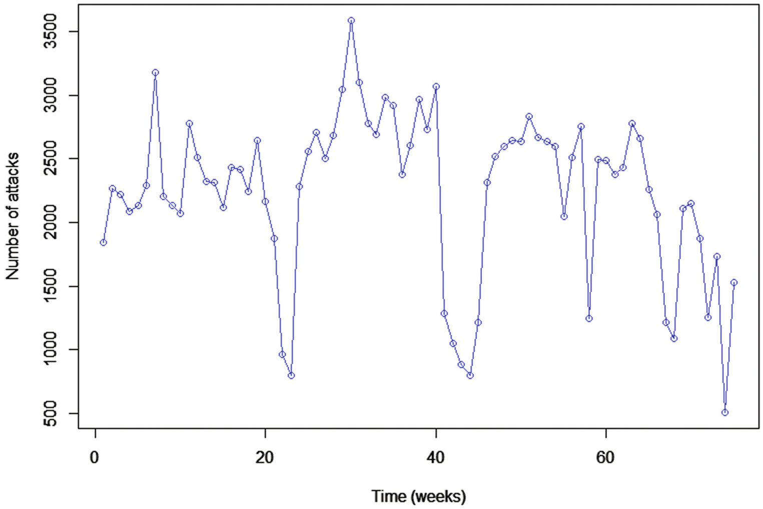


**Fig. 1.** Number of attacks collected by honeypots

Table 1 summarises **p − values** of previous tests for our data. Both tests admit the stationarity of data or refuse the hypotheses of non-stationarity.

The next step of analysis is to **eliminate the trend from data**. In our case, it is sufficient to subtract the constant trend (mean). However, one can try to estimate some quadratic trend, but no better results can be achieved. We also follow **Occam's razor** (law of parsimony) [25]. For this reason, we estimate as little parameters as possible. The goal is to find the simplest adequate model for data.

**Table 1.** P-values of stationarity tests

| Type of test | p-value |
|---|---|
| ADF | 0.010 |
| KPSS | 0.100 |

Figure 2 shows **autocorrelation function** (acf) and **partial autocorrelation function** (pacf). According to results from these plots it is obvious there is some correlation among time series components in different time lags. These functions are also helpful in choosing the appropriate auto-regressive (AR) model or moving average (MA) model. More precisely according to graph of partial autocorrelation function we chose **AR model of the 1st order - AR(1)**. AR(1) is determined by equation (2) for $p = 1$. The same model is chosen by automatic model selection or by considering Akaike Information Criterion – AIC [26] of different models.

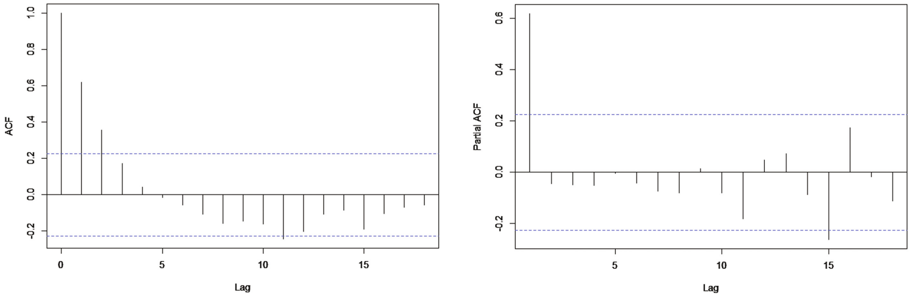


**Fig. 2.** Data without trend with autocorrelation function (left) partial autocorrelation function (right)

Next step of analysis consists of **fitting model** – estimating model parameters from data. We chose AR(1) model. It has just one coefficient as a parameter and its estimate is equal to 0.629. Figure 3 shows that estimated model fits our data very well.

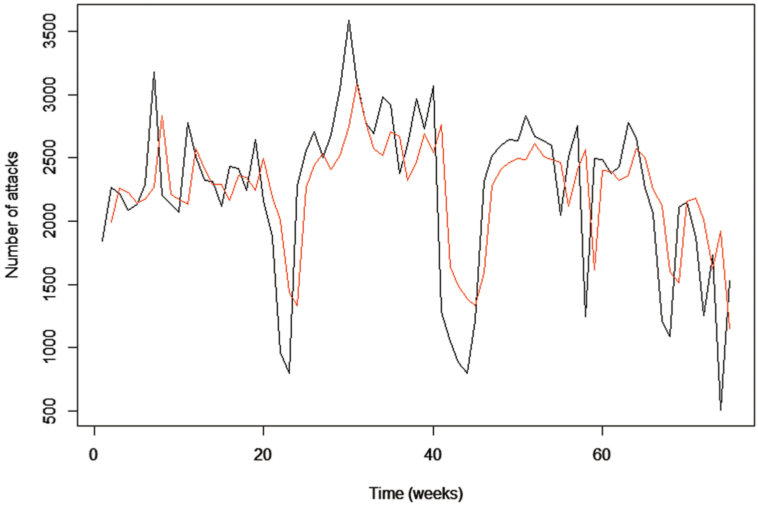


**Fig. 3.** Model fitting data

It is just the first verification that fitted model is appropriate for the data. However, we would like to be more certain of the model so that we can obtain good predictions of future time series values. Therefore, we need other diagnostics of residuals. At first, it is needed to focus on residuals themselves. The residuals of our data after application of AR(1) are shown in Fig. 4. Subsequently, we check the existence of significant correlation, but this time among residuals. The autocorrelation function of residuals are shown in Fig. 5(left) and partial

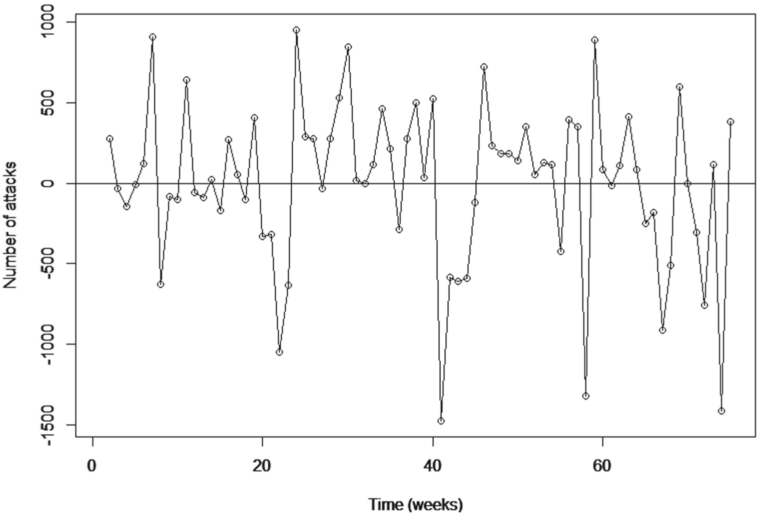


**Fig. 4.** Residuals

autocorrelation function of residuals are shown in Fig. 5(right). In Fig. 5 we can see that there is no significant correlation anymore. Another possible diagnostic plot for residuals is **cumulative periodogram**.

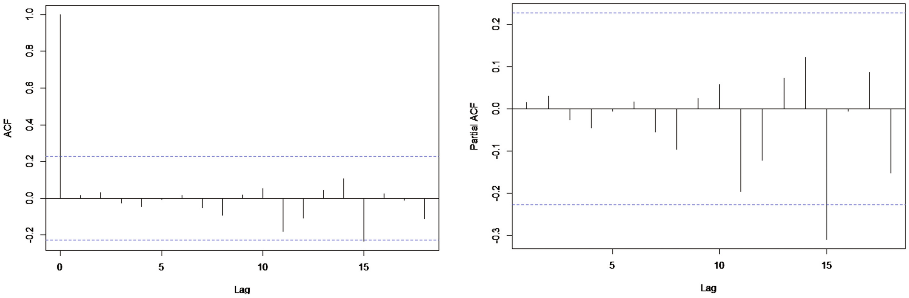


**Fig. 5.** Residuals - autocorrelation function (left) partial autocorrelation function (right)

It is also necessary to do numerical tests of residuals independence. We used **Box-Pierce test** [27] and **Box-Ljung test** [28]. Both tests do not reject null hypothesis about residuals independence, because their p-values are 0.8948 and 0.8927 respectively. All introduced diagnostic tools for residuals confirm that AR(1) model is appropriate for our data.

At this place, we can approximate a prediction. We have found good model for data. Due to this fact it is possible to do predictions for future values of time series.

In our case it will be prediction for number of attacks during the next weeks. We also would like to construct 95% prediction intervals for the numbers of attacks to have some certainty/confidence where should true future values belong (interval) if we assume that the model found is valid. For this purpose, it is needed to test for residuals normality. The first insight can be obtained by histogram of residuals.

We also used **Shapiro-Wilk normality test** [29]. The **p-value** of this test is equal to **0.0041**. It means that **null hypothesis** about normality is **rejected**. In this case, we cannot use traditional derived prediction/forecast intervals [26]. We decided to implement our **own bootstrap procedure** in R language [30] inspired by Chernick and LaBudde bootstrap prediction intervals [30]. This bootstrap procedure uses boot R package [31]. When a theory is not available or fails, bootstrap can be useful. We can gain some information from it, but one should be careful when dealing with bootstrap methods [12,13].

## 4 Prediction of Attacks and Discussion

In this paper, we focus on two point prediction approaches such as prediction based directly on AR(1) model and bootstrap prediction. Moreover, we include interval predictions based on bootstrap. At first, we predict attacks directly from model. Figure 6 shows the green coloured real numbers of attacks and red coloured predicted values in five future time points (five weeks). Numerical results are summarised in the Table 2. It is obvious that real and predicted values are close enough to each other. So, the point predictions are quite good, but we can do somewhat (at some time points of time) better using bootstrap approach.

Figure 7 shows the fact that **bootstrap point predictions** (connected by red line) are also very good. They are sometimes closer to true values than in the

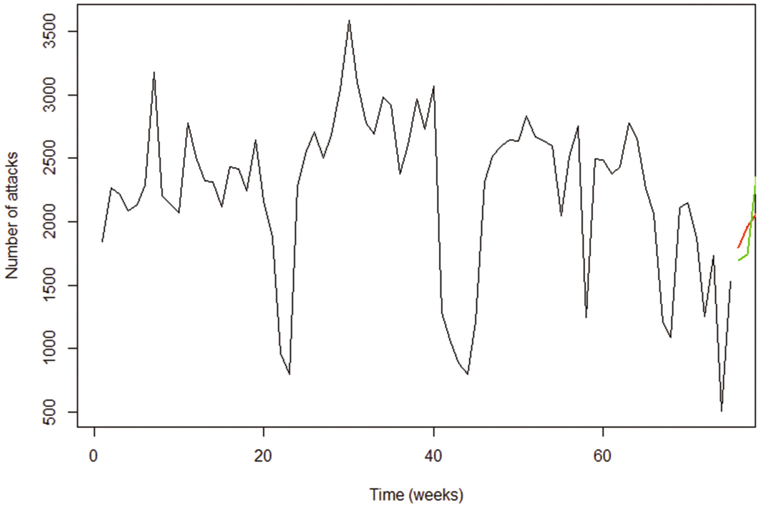


**Fig. 6.** Predictions directly from model

**Table 2.** Results from model

| Week | Predicted number of attacks | Real number of attacks |
|---|---|---|
| 76 | 1795 | 1693 |
| 77 | 1959 | 1743 |
| 78 | 2063 | 2355 |
| 79 | 2123 | 2401 |
| 80 | 2169 | 2486 |

previous approach (prediction based directly on AR(1) model). More important finding is fact that true numbers of attacks (green line) lie in our 95% confidence interval. It gives us a 95% confidence, that this prediction interval - PI (blue lines) will contain real future numbers of attacks if we assume that model we found is valid for our data. Tests confirmed that model is valid for our data. Numerical results are summarised in Table 3.

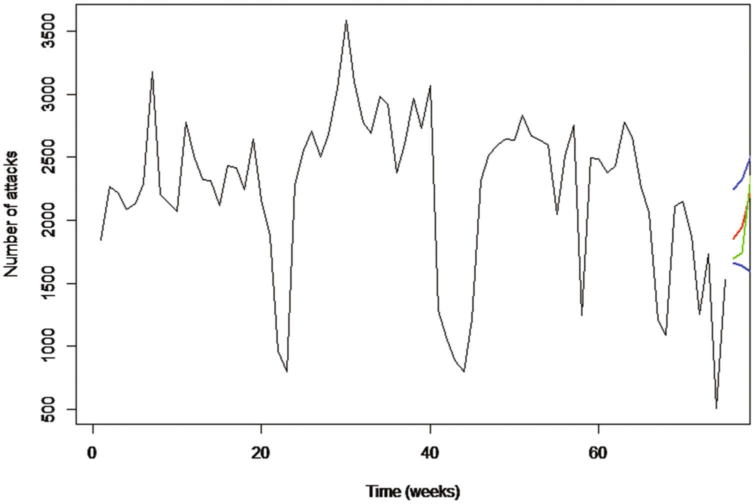


**Fig. 7.** Bootstrap predictions

For comparison of prediction approaches, we computed four forecast accuracy measures [26] in R. We specifically calculated MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error) and MASE (Mean Absolute Scaled Error). Table 4 shows the results of forecast accuracy measure for both approaches. These measures just confirm our previous conclusion that bootstrap point predictions are a little bit more accurate than predictions made directly from model AR(1).

**Table 3.** Results of bootstrap predictions

| Week | Predicted number of attacks | Real number of attacks | Prediction interval (PI) |
|---|---|---|---|
| 76 | 1848 | 1693 | (1659,2246) |
| 77 | 1947 | 1743 | (1642,2324) |
| 78 | 2361 | 2355 | (1590,2509) |
| 79 | 2376 | 2401 | (1604,2462) |
| 80 | 2385 | 2486 | (1578,2541) |

Above mentioned results show prediction of attacks against honeynet. Since honeypots within honeynet collect activities of no-legitimate users, it can be stated that they only collect attacks. If we use data from production network (e.g. data flow), we will have to make pre-processing data (determine attacks). For this reason, prediction of attacks against honeynet can be used like baseline for network intrusion detection system (NIDS) based on anomaly detection.

**Table 4.** Results of forecast accuracy measure

| Approach/Forecast accuracy measure | MAE | RMSE | MAPE | MASE |
|---|---|---|---|---|
| AR(1) model | 241 | 253.02 | 11.03 | 1.22 |
| Bootstrap | 98.20 | 123.70 | 5.24 | 0.50 |

## 5 Conclusion and Future Works

Data collected by honeypots and honeynets are interesting source for further analysis. In this paper we focus on prediction of attacks. Using the model AR(1), bootstrap based on AR(1) model we predicted number of attacks against honeynet. For this purpose, we used data from honeypot Kippo. The prediction of attacks is based on 75 weeks data and it has been verified by five weeks data.

As we have shown above, prediction model AR(1) and bootstrap based on AR(1) model are suitable for prediction of attacks. We also obtained prediction intervals. Predicted attacks based on these approaches are near to the real attacks. For prediction of attacks, it is better to use the bootstrap based on AR(1) model. We can also conclude that prediction intervals (Table 3 and Fig. 7) will contain the real number of attacks with approximately 95 % certainty.

In the future, we are primarily planning to discuss other time series models (integer/count models) applicable in this field of cyber security (e.g. honeypots). Also, we would like to focus on their comparison and multivariate time series analysis to predict attacks. We would like to include some other variables (e.g. publicly known information security vulnerabilities), which could influence the number of attacks.

**Acknowledgments.** We would like to thank colleagues from the Czech chapter of The Honeynet Project for their comments and valuable input. This paper is funded by the Slovak Grant Agency for Science (VEGA) grants under contract No. 1/0142/15 and No. 1/0344/14, VVGS projects under contract No. VVGS-PF-2016-72610 and No. VVGS-PF-2016-72616 and Slovak APVV project under contract No. APVV-14-0598.

## References

1. Spitzner, L.: The honeynet project: trapping the hackers. IEEE Secur. Priv. **1**(2), 15–23 (2003)
2. Spitzner, L.: Honeypots: Tracking Hackers. Addison-Wesley Reading, Boston (2003)
3. Honeynet.org: Kippo project. Accessed 20 May 2017
4. Abbasi, F.H., Harris, R.: Experiences with a Generation III virtual Honeynet. In: 2009 Australasian Telecommunication Networks and Applications Conference (ATNAC), pp. 1–6. IEEE (2009)
5. Klyne, G., Newman, C.: Date and time on the internet: timestamps. Technical report (2002)
6. Sokol, P., Kleinová, L., Husák, M.: Study of attack using honeypots and honeynets lessons learned from time-oriented visualization. In: EUROCON 2015-International Conference on Computer as a Tool (EUROCON), pp. 1–6. IEEE (2015)
7. Condon, E., He, A., Cukier, M.: Analysis of computer security incident data using time series models. In: 2008 19th International Symposium on Software Reliability Engineering, ISSRE 2008, pp. 77–86. IEEE (2008)
8. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. Springer Science & Business Media, New York (2013)
9. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken (2015)
10. Efron, B.: Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics, pp. 569–593. Springer (1992)
11. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC Press, Boca Raton (1994)
12. Kreiss, J.P., Lahiri, S.: Bootstrap methods for time series. Handbook Stat.: Time Ser. Anal.: Methods Appl. **30**(1) (2012)
13. Chernick, M.R., González-Manteiga, W., Crujeiras, R.M., Barrios, E.B.: Bootstrap Methods. Springer, Heidelberg (2011)
14. Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., Liu, M.: Cloudy with a chance of breach: forecasting cyber security incidents. In: USENIX Security, pp. 1009–1024 (2015)
15. Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: Proceedings of the 12th Annual Conference on Cyber and Information Security Research, p. 18. ACM (2017)
16. Wei, M., Kim, K.: Intrusion detection scheme using traffic prediction for wireless industrial networks. J. Commun. Netw. **14**(3), 310–318 (2012)
17. Tang, M., Alazab, M., Luo, Y.: Exploiting vulnerability disclosures: statistical framework and case study. In: 2016 Cybersecurity and Cyberforensics Conference (CCC), pp. 117–122. IEEE (2016)
18. Zhan, Z., Xu, M., Xu, S.: Predicting cyber attack rates with extreme values. IEEE Trans. Inf. Forens. Secur. **10**(8), 1666–1677 (2015)

19. Soldo, F., Le, A., Markopoulou, A.: Blacklisting recommendation system: using spatio-temporal patterns to predict future attacks. IEEE J. Sel. Areas Commun. **29**(7), 1423–1437 (2011)
20. CZ.NIC: Turris greylist project. Accessed 20 May 2017
21. Team, R.C.: R: A language and environment for statistical computing. vienna: R foundation for statistical computing; 2014 (2016)
22. Banerjee, A., Dolado, J.J., Galbraith, J.W., Hendry, D., et al.: Co-integration, error correction, and the econometric analysis of non-stationary data. OUP Catalogue (1993)
23. Said, S.E., Dickey, D.A.: Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika **71**(3), 599–607 (1984)
24. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. **54**(1–3), 159–178 (1992)
25. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Occam's razor. Inf. Process. Lett. **24**(6), 377–380 (1987)
26. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2014)
27. Horowitz, J.L., Lobato, I., Nankervis, J.C., Savin, N.: Bootstrapping the box-pierce q test: a robust test of uncorrelatedness. J. Econom. **133**(2), 841–862 (2006)
28. Harvey, A.C.: Trends and cycles in macroeconomic time series. J. Bus. Econ. Stat. **3**(3), 216–227 (1985)
29. Shapiro, S.S., Francia, R.: An approximate analysis of variance test for normality. J. Am. Stat. Assoc. **67**(337), 215–216 (1972)
30. Chernick, M.R., LaBudde, R.A.: An Introduction to Bootstrap Methods With Applications to R. Wiley, Hoboken (2014)
31. CRAN.R: Package boot. Accessed 20 May 2017

# The Cognitive Approach to the Coverage-Directed Test Generation

Anna Klimenko[1](✉), Galina Gorelova[2], Vladimir Korobkin[3], and Petr Bibilo[4]

[1] SFedU Acad. Kalyaev Scientific Research Institute of Multiprocessor Computer Systems, Taganrog, Russia
anna_klimenko@mail.ru
[2] SFedU, Rostov-on-Don, Russia
[3] Scientific-Research Institute of Multiprocessor Computing and Control Systems, Taganrog, Russia
[4] The State Scientific Institution "The United Institute of Informatics Problems of the National Academy of Sciences of Belarus", Minsk, Belarus

**Abstract.** The important contemporary issue of VLSI design verification is its time-consuming. The hardware model, written, for instance, with VHDL, is verified by formal and dynamic verification approaches. Dynamic verification (simulation) is widely-used due to the possibility of full automation of the process, but takes too much time due to its redundancy.

The concept of the coverage-directed test generation is to redirect the test generator such as to reach uncovered metric points. There are several approaches for this, including genetic algorithms using, Bayesian network, Data mining, etc.

The new cognitive approach to the coverage-directed test generation (CA CDG) is proposed within this paper. It is based on a cognitive map usage. The CA CDG is described, some simulation results are given. Also the future work areas are outlined.

**Keywords:** VHDL · Verification · Simulation · Coverage-directed test generation · Cognitive map

## 1 Introduction

Functional VLSI (very large scale integration) verification is one of the most challenging issues in the last decade [1]. Functional verification is the task of verifying that the logic design conforms to specification. There are two fundamental approaches in functional verification: formal [2] and dynamic verification [3], or, as it is usually called, "simulation".

Each approach has its own pros and cons. Ideally, formal verification is able to locate all design errors, but is too complicated and requires the huge effort of qualified engineers.

Simulation, and random testing in particular, is considered as a mean to discover errors automatically. In practice, around 80% of the target coverage can often be achieved within 20% of the overall verification time. It means that the last 20% of

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_34

target coverage consumes 80% of the overall time [4]. Random testing can be fully automated, so the large number of tests can be easily generated. This allows to exercise repeatedly the RTL(register transfer language) code with varying conditions. It is known that for most coverage metrics, constrained random verification rapidly achieves 80% coverage and then asymptotically approaches 100% [4]. The slow convergence of random conditions to complete coverage is well known and referred to as the "coupon collector problem" [5].

Nowadays no one approach prevails, and, as a consequence, in practice the verification process involves both of them. We concentrate on the simulation approach within this paper.

Random testing is based on several coverage metrics. Coverage metric allows to evaluate the completeness of the testing process. There is a vast number of coverage metrics [6], which has been developed. Some of them are used widely, and some of them are just proposed for particular problem solving. But, independently of metric used, main issue of the random testing – the redundancy and time-consuming and the next effort to normalize the situation was the concept of "coverage-directed test generation" (CDG).

There are two approaches towards CDG: one is by construction using formal methods and the other is based on feedback [7–9]. The main idea of CDG based on feedback is that there is a closed loop between the metric value and the test generator with an element, which automatically directs the test vectors generation to reach the uncovered metric points.

The machine learning techniques are used for such feedback implementation. They are classified according to the learning methods (supervised learning, unsupervised learning and reinforcement learning).

There are several technique examples, which are used in CDG:

- evolutionary algorithms;
- probabilistic methods;
- data mining;
- inductive logic programming approach.

Evolutionary algorithms, and, genetic algorithms in particular, are used in CDG in works [10–12].

Probabilistic methods relate to the Bayesian network and Markov Models usage. The short list of related works includes [13, 14].

Also, Data mining techniques and inductive logic programming are used [15, 16], but much less often then evolutionary algorithms and methods based on the Bayesian network.

The new approach to the coverage-directed test generation is presented within this paper. It is based on cognitive maps usage. Cognitive maps are the convenient modeling tool, usually categorized as a neuro-fuzzy method, for modeling and simulation of dynamic systems. In the scope of this paper the cognitive map is used to implement the feedback between the coverage metric value and the test generator.

The formal model of the approach is given, some simulation results are represented and discussed.

## 2 The Cognitive Maps

A cognitive map is an image of cognitive processes and an attempt to utilize expert opinion and cognition about ill-structured social relationship [17, 18].

Also cognitive map is determined as cause-effect network, with nodes representing concepts articulated by individuals, and directional linkages capturing casual dependencies [19].

The simplest cognitive model is represented by a signed directed graph of basic elements, which are:

- Concept variables (graph vertexes);
- Relationships between concepts;
- The signs of the relations.

Formally the simple cognitive map is described as follows:

$$G = \langle V, E \rangle, \tag{1}$$

where $V$ is the set of concepts, $E$ is the set of relations.

Nowadays much more complex cognitive models are developed: functional vector graph, parametric functional vector graph (2), a modified functional graph and their modifications.

$$\Phi = \{G, X, F, \Theta\}, \tag{2}$$

where $X = \left\{ x_g^{(i)} \right\}$, $g = 1{,}2, \ldots k$, $F = F(X, E) = F(x_i, x_j, e_{ij})$ is the edge transformation functional that assignes a sign (“+”or «-») to each edge, or a weight coefficient $w_{ij}$, or a function $f(x_i, x_j, e_{ij}) = f_{ij}$.

Within this paper the simplest cognitive map (1) is used for the cognitive approach to the coverage-directed test generation (CA CDG). The CA CDG is described in the next section.

## 3 The Cognitive Approach to the Coverage-Directed Test Generation

The CA CDG represented within this paper is based on a cognitive map using, as it was mentioned above. As a metric for the simulation the assertion coverage was chosen. It is proved, that simulation time with assertion usage is up to 10% of the overall project time, while the general time and effort benefit is up to 80% [20].

The preliminary steps of the CA CDG are the forming of the test vector space and the set of assertions.

Test vector is described by a set of values, which are the input variables for the RTL module (or the integrated modules): $v = (v_1, v_2, \ldots v_k)$, where $k$ – the number of input variables, so the test vector space is $k$-dimensional.

We make an assumption, that test vectors are generated according the appropriate predefined constraints. The problem of test vector generation relates to the constraint satisfaction problem and out of the scope of this paper.

Then, we assume that there is a method of test vector space $V$ fragmentation to the set of disjoint areas: $V = V^1 \cup V^2 \ldots V^{\xi}$, $V^1 \cap V^2 \ldots V^{\xi} = \emptyset$.

Every generated test vector $v^i$ belongs to the region $j$ if:

$$v^i = (v_1^i, v_2^i, \ldots, v_k^i) \in V^j \, ec\pi u \, v_1^i \in (V_1^j, V_1^j + \Delta V_1),$$

$$v_2^i \in (V_2^j, V_2^j + \Delta V_2),$$

$$\ldots$$

$$v_k^i \in (V_k^j, V_k^j + \Delta V_k),$$

$$where \ \Delta V_k \ - \ the \, preliminary \, defined \, incrementation \, of \, the \, dimension \, k.$$

Also we define the set of assertions, which are located in the RTL code: $A$ = *{A₁, A₂, ... Aₘ}*.

If $A_j$ is a precessor of $A_i$ independently of input test vector, $A_i$ and $A_j$ are linked with the positive relation $A_j \rightarrow A_i$. Such links can be formed on the stage of assertion writing. At the next stage of the CA CDG the cognitive map is formed. This map describes the relations between test vector space regions and assertions. The cognitive map forming takes place through the testing process.

So, the cognitive map for the coverage-driven testing includes concepts, which describe the test vector space regions, concepts, which describe the assertions, and relations between linked assertions. Every concept is weighted by 0-value (Fig. 1).

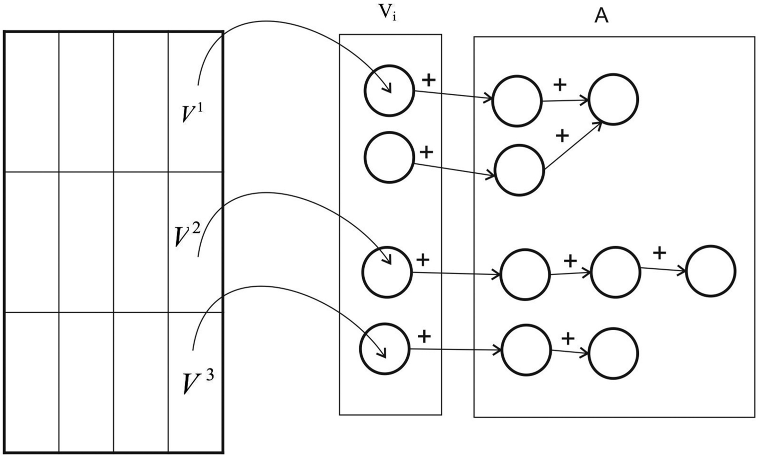


**Fig. 1.** The general structure of CA CDG cofnitive map.

Then, the test generating process begins:

1. The threshold value D is chosen, where D is the sufficient and goal coverage metric value (for instance, D = 60%). The current coverage metric value is $D_{fact}$.
2. While $D_{fact} < D$
   a. Generate test vector $v^i \in V^i$ from the test vector space $V$;
   b. If $A_i$ is reached, unite the concepts $v^i$ and $A_i$ by relation with sign "+";
   c. The value of concept $v^i$ is incremented;
   d. The stimulus from the $v^i$ propagates to the approachable vertexes;
   e. Modify $D_{fact}$.

While coverage metric points reach the chosen coverage threshold, some concepts $V^i$ and $A_k$ have the positive values.

The next important variable of the CA CDG is the critical weight C of the concept $A_i$, which determines that this assertion was reached C times. For example, C = 3 means that from the test vector space $V^i$ the assertion $A_i$ were reached for 3 times. The elimination of $V^i$ from the test vector space allows to direct the test generator to the unexplored test vector space areas. The description of the next stage of the CA CDG is given below:

1. While there are concepts $A_i$ without any relations with $V^i$
2. {Eliminate from the test vector space $V^k$ where $C > num$;
3. Generate new test vector according the new constraints;
4. If $A_j$ is covered with the test vector, increment the concept values $V^j$, unite concepts with positive relation and propagate the stimulus from the $V^j$.}

Following the steps above, the test vector space is shrunk through the simulation process. It adds the new constraints and makes the test generator – if there are any uncovered assertions - explore new test vector space regions.

The proposed approach has its own pros and cons.

+ there is no separate learning stage. The cognitive map is formed through the simulation process till the coverage metric reaches the threshold value D. Hence, the time consuming is reduced.

+ cognitive map proposed is implemented simply.

− There is a possibility to loose some useful test vectors due to the test vector space regions eliminating. This issue can be solved by the increasing of parameter C. It is important to note, that such strategy also increases the time-consuming of the overall testing process.

## 4 Simulation Results

The approach represented above is the effort to use the cognitive maps for the coverage-directed simulation. The cognitive map is used in an unusual manner – to implement the feedback between the coverage metric value and test vector generator. The usage of a cognitive map is simple and doesn't need the supervised learning stage.

Also the main distinguish between the cognitive map and Bayesian Network usage must be emphasized:

- Bayesian network helps to provide the stimulus needed to cover the particular metric point (branch, assertion, etc.)
- The cognitive map eliminates the explored regions of the test vector space to reduce the redundancy of the test generation and to redirect the test generator to the unexplored regions.

Some simulations were made as an experiment: the first one contains the comparison of the simple random test and CA CDG.

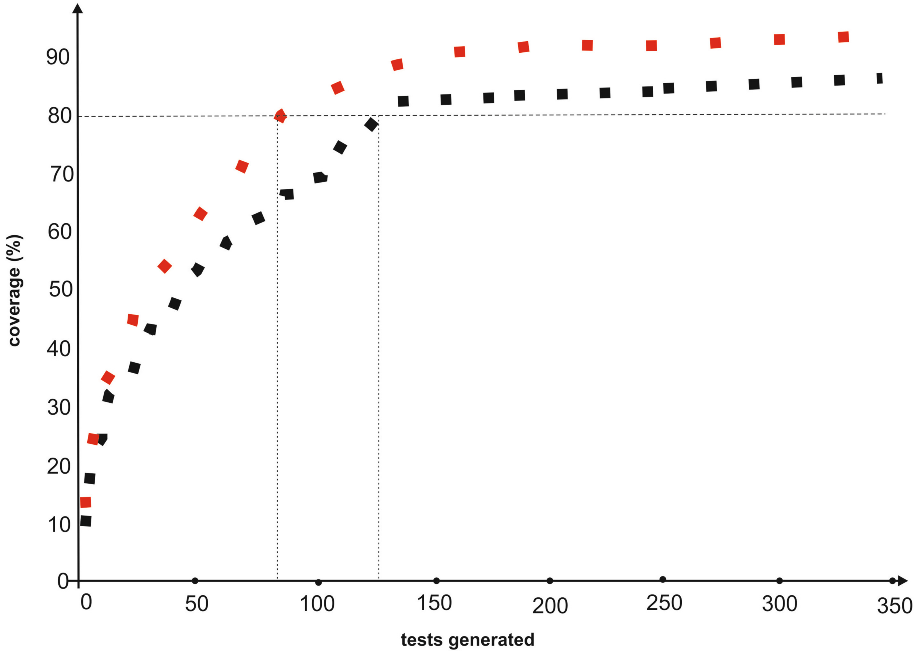


**Fig. 2.** Results of the pseudorandom testing (black) and CA CDG (red).

It is seen (Fig. 2), that with the CA CDG 80% assertion coverage is reached faster then with simple pseudorandom testing. At the beginning of the testing the coverage rates are almost equal, but while explored regions are eliminated from the search space, the coverage rate grows. The next simulation shows the reducing rate of the test vector regions.

It is shown on Fig. 3, that as the threshold C is low, the rate of the vector space region eliminating is high. It leads to the search space reducing and, hence, increases the possibility to generate the required test vectors.

On the other hand, if we exclude test vector regions too fast, it is possible to loose some "good" test vectors from these regions. The graphics on Fig. 4 illustrates the situation. It is seen, that with the fast test vector space reducing the coverage rate is high enough at the beginning, but then stagnates, some assertions are unreachable till the end of the testing. When the test vector space reduces slower, the coverage metric

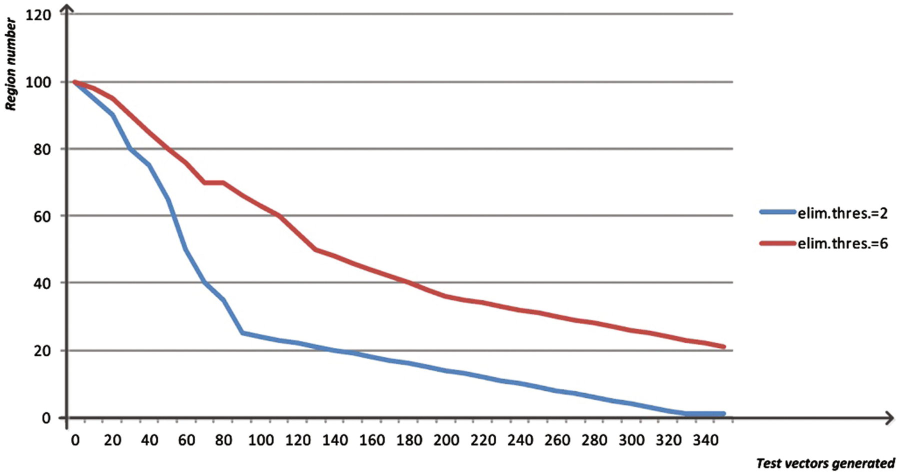


**Fig. 3.** The vector space region eliminating rate. When C = 2, the number of vector space regions reduces faster than with the eliminating C = 6.

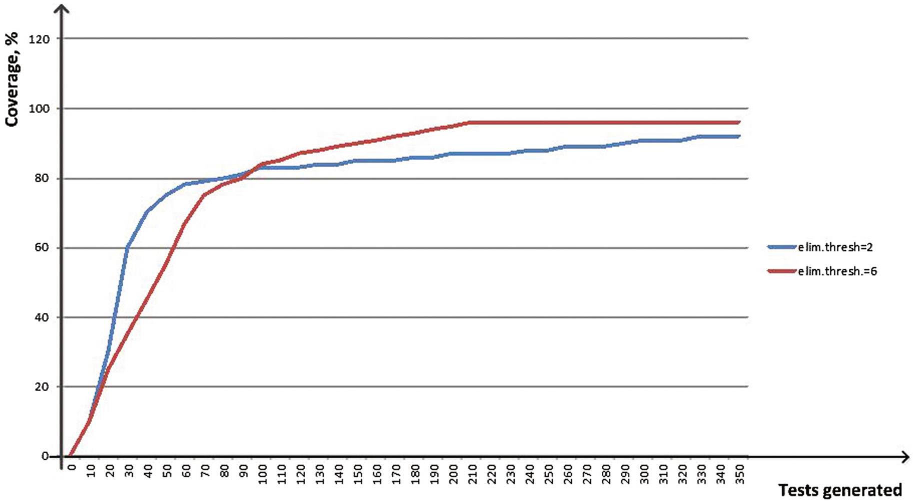


**Fig. 4.** The coverage metric value depends on the region eliminating threshold. When the threshold is small, the test vectors can be lost.

value does not grows as fast as in previous case, but the overall metric coverage becomes better. Such particularities of CA CDG raises the question of the appropriate CA CDG parameters choice.

## 5 Conclusions and Future Work

A new approach to the coverage-directed test generation based on cognitive map usage is represented and described in the current paper.

The cognitive maps are simple enough to implement and does not require the special learning stage, in our particular case. The forming of the cognitive map is spread through the part of testing process and does not consume additional time. The main concept of the CA CDG is to eliminate test vector space regions through the testing procedure and so redirect the test generator to the unexplored ones. The search space reducing and the possible loss of the "good" test vectors are the additional effect of CA CDG, but this is the question of the appropriate parameter value choice.

The paper contains the CA CDG description and some simulation results. It is seen that the coverage rate of the CA CDG is higher than the coverage rate of simple random testing. So, CA CDG seems to be rather prospective.

The proposed future work contains the research of the following areas:

- CA CDG comparison with other based on the machine learning CDG techniques;
- CA CDG optimization through the appropriate parameter C and D values choice;
- Development and research of the stochastic CA CDG.

**Acknowledgement.** The reported study is funded by the 16-58-00191 Bel_a RFBR project.

## References

1. ITRS. International Technology Roadmap for Semiconductors, Design Chapter, 2008 Edition (2008)
2. Wile, B., Goss, J., Roesner, W.: Comprehensive Functional Verification: The Complete Industry Cycle (Systems on Silicon). Morgan Kaufmann Publishers Inc., Burlington (2005)
3. Bergeron, J.: Writing Testbenches: Functional Verification of HDL Models, 2nd edn. Springer, Cham (2003)
4. Butka, B.: Advanced verification methods for safety-critical airborne electronic hardware. Office of Aviation Research and Development. Washington (2013). (https://www.faa.gov/aircraft/air_cert/design_approvals/air_software/media/AdvancedVerifAEHFinalDraft-101813.pdf
5. Blom, G., Holst, L., Sandell, D.: 7.5 Coupon Collecting I, 7.6 Coupon Collecting II, and 15.4 Coupon collecting III. Problems and Snapshots from the World of Probability, pp. 85–87. Springer, New York (2004)
6. Jerinic V., Langer J., Heinkel U., Miller, D.: New Methods and Coverage Metrics for Functional Verification. In: Proceedings of Design, Automation and Test in Europe, pp. 1–6 (2006)

7. Ur, S., Yadin, Y.: Micro architecture coverage directed generation of test programs. In: Proceedings of the 36th Annual ACM/IEEE Design Automation Conference, pp. 175–180. ACM, New Orleans (1999)
8. Bose, M., Shin, J., Rudnick, E., Dukes, T., Abadir, M.: A genetic approach to automatic bias generation for biased random instruction generation, pp. 442–448 (2001)
9. Tasiran, S., Fallah, F., Chinnery, D.G., Weber, S.J., Keutzer, K.: A Functional Validation Technique: Biased Random Simulation Guided by Observability-Based Coverage, pp. 82–88. Institute of Electrical and Electronics Engineers Inc., Piscataway (2001)
10. Habibi, A.: Efficient assertion based verification using TLM. In: Proceedings of Design, Automation and Test in Europe Conference (IEEE Cat. No. 06EX1285C) (2006)
11. Yu, X., et al.: A genetic testing framework for digital integrated circuits. In: Proceedings of 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2002). Washington, DC, USA, pp. 521–526 (2002)
12. Samarah, A., et al.: Automated coverage directed test generation using a cell-based genetic algorithm. In: 11th IEEE International High Level Design Validation and Test Workshop (IEEE Cat. No.06EX1547), pp. 19–26. IEEE, Piscataway
13. Baras D., Ziv. A.: Automating boosting of cross-product coverage using Bayesian networks. In: Proceedings of Conference: Hardware and Software: Verification and Testing, 4th International Haifa Verification Conference, HVC 2008, Haifa, Israel, 27–30 October (2008)
14. Fine, S., Ziv, A.: Coverage directed test generation for functional verification using Bayesian networks. In: Proceedings of IEEE Xplore Conference: Design Automation Conference (2003)
15. Braun M., Rosenstiel W., Schubert, K.-D.: Comparison of Bayesian Networks and data mining for coverage directed verification category simulation-based verification. In: IEEE Xplore Conference: High-Level Design Validation and Test Workshop, Eighth IEEE International (2003)
16. Eder, K., Flach, P., Hsueh, H.-W.: Towards Automating Simulation-Based Design Verification Using ILP. In: Proceedings of DBLP Conference: Inductive Logic Programming. 16th International Conference, ILP 2006, Santiago de Compostela, Spain, 24–27 August, Revised Selected Papers (2006)
17. Axelrod, R.: Structure of Decision: The Cognitive Maps of Political Elites. Princeton University Press, Princeton (1976)
18. Lee, S., Courtney Jr., J.F., O'Keefe, R.M.: A system for organizational learning using cognitive maps. OMEGA **20**(1), 23–36 (1992)
19. Srinivas, V., Shekar, B.: Applications of uncertainty-based mental models in organizational learning: a case study in the Indian automobile industry. Account. Manag. Inf. Technol. **7**(2), 87–112 (1997)
20. Functional Verification of a Multiple Issue, Out of Order, Superscalar Alpha Processor. DAC (1998)

# An Agent Based Model to Study the Impact of Intra-annual Season's Variability on the Dynamics of Aedes Vexans and Culex Poicilipes Mosquito Populations in North Senegal (Ferlo)

Python Ndekou Tandong Paul[1,2,3](✉), Alassane Bah[1,2,3],
Papa Ibrahima Ndiaye[1,2,3], and Jacques André Ndione[1,2,3]

[1] Ecole Superieure Polytechnique-UCAD, UMMISCO, Dakar, Senegal
pppython@yahoo.fr, alassane.bah@gmail.com, papaibra@yahoo.com
[2] Université Alioune Diop, Bambey, Senegal
[3] Centre de Suivi Ecologique, Dakar, Senegal
jacques-andre.ndione@cse.sn
http://www.springer.com/lncs

**Abstract.** We present an agent-based model for studying the dynamics of Aedes vexans and Culex poicilipes mosquito populations taking into account interactions with animal herds, water ponds and climate factors in the Ferlo region (Senegal). The main objective of this work is to show the impact of intra-annual season's variability on the Aedes vexans and Culex poicilipes mosquito populations dynamics. We have designed a UML class diagram representing interactions between animal hosts, mosquitoes, and water ponds and used this diagram to create an agent based model which helps us to carry out the sensitivities analysis on dry spells. The obtained results show that there is a growth of Aedes vexans mosquito populations at the end of each dry spell following by rainfall and the appearance of Culex poicilipes mosquito populations at the end of August coinciding with the disappearance of Culex poicilipes mosquito populations. The developed model provides a tool for understanding and predicting the dynamics of Aedes vexans and Culex poicilipes mosquito populations in the short and long term. It can also be used to study the sensitivities analysis of daily rainfalls of other types of vectors.

**Keywords:** Rift Valley fever · Mosquito populations modeling · Agent based model · Multi-agent · Dry spell · Rainfall variability

## 1 Introduction

RVF is an infectious disease transmitted by mosquito species called Aedes vexans and Culex poicilipes. It is transmitted through interactions between mosquitoes, humans and animals. The eradication of this disease is almost impossible, what can be done is to control the dynamics of their populations by finding a way to

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_35

shorten their life cycle duration. The Modeling of the life cycle of Aedes vexans and Culex poicilipes mosquitoes taking into account climatic factors could help to understand the different mechanisms of RVF outbreak and to know the periods of excessive growth of the mosquito populations. Several studies have been carried out on the modeling of the dynamics of mosquito populations [6,12,16]. The control of the life cycle of mosquito populations studied by Becker N. et al. (2003) [1] has allowed researchers to develop different strategies to reduce the number of mosquitoes that can leave the aquatic phase to the adult phase. Climatic changes plays a major role in the emergence and re-emergence of vector-borne diseases [3,12,13]. Several researchers have developed different types of models to describe the dynamics of mosquito populations [15,16]. Intra-seasonal climatic variability influences the egg's hatching rate of mosquitoes that transmit RVF [13]. To predict and map areas at risk of Rift Valley fever virus in Ferlo (Senegal), it is essential to understand the epidemiology of the disease. This includes the understanding of interactions between hosts, vectors and the environment. The Knowledge about spatial and temporal behaviors of mosquitoes could help to plan and control this disease. Mosquitoes are the main vectors of RVF transmission [9]. Several types of mosquitoes (Aedes vexans, Culex poicilipes) are naturally infected [7,12]. An entomological campaign carries out in Mauritania and all along the Senegal river valley highlighted the role of mosquito Culex poicilipes in the spread of RVF virus. The infected Aedes vexans female can transmit the virus to the eggs [4]. The population of Aedes vexans, which is mostly present at the beginning of the rainy season, declines in the middle of the rainy season [10,13]. The eggs of Mosquitoes are able to survive on the dry soil until the next rainy season, which will produce a favorable condition for the hatching of mosquito eggs [5]. The impact of environmental factors on the epidemiology of RVF is well known, the impact of climate factors (daily precipitation, daily temperature, humidity, wind speed) influences the life cycle of mosquitoes [8]. The multi-agent systems have been used to model complex systems [2]. They have also been used in epidemiology in the field of vector-borne diseases [14]. The agent-based models based allow to model and simulate many complex phenomena [2]. These models are increasingly used in environmental science [17]. Nowadays, very few studies have used agent-based systems for studying the epidemiology of vector-borne diseases [11] and the dynamics of mosquito populations. Previous studies have suggested a correlation between precipitations and the abundance of mosquitoes [15,16], but no study in West Africa region have used agent-based models to study the impact of a dry spell after rainfalls on the mosquito population dynamics. The aim of our work is to design a static model based on an unified modeling language (UML) and build an agent-based model on the CORMAS multi-agent platform that will allow us to study the impact of dry spell after rainfall and their sensitivities on the dynamics of the populations of the mosquitoes Aedes vexans and Culex poicilipes.

## 2 Presentation of the Ferlo Area

The region of Ferlo is a Sahelian zone in northern Senegal, covered with vegetation and water ponds during the winter. This zone is characterized by a Sahelian climate. Our study area is about 25 km$^2$ and includes different types of water ponds and vegetation during the period of winter. Each water pond is a habitat for Aedes vexans and Culex poicilipes mosquitoes. The climate of the Ferlo region is characterized by two main seasons: a dry season and a rainy season. The mean annual rainfall is ranges from 300 mm to 500 mm. From the beginning of the first rains in June, the first proliferation of mosquitoes Aedes vexans appears, but this proliferation will increase with the dry spell which will allow the desiccation of the mosquito eggs. The first Culex poicilipes mosquitoes occur in August. The eggs of Culex poicilipes mosquito populations do not need dry spells before hatching. The number of animal herd arriving in the Ferlo region during the winter is very important, they are exposed to the biting of the Aedes vexans and Culex pocilipes mosquitoes.

## 3 Description of the Model

The environment of study is composed of several interacting entities that can be modeled by a UML class diagram. We have identified the following classes: the Mosquito class, it is the superclass of the MosquitoAedes and MosquitoCulex classes. A mosquito is characterized by its physiological state (egg, larva, pupa), the duration of the desiccation of eggs, its health status, the minimum amount of water within the ponds required for the hatching of eggs. Each Mosquito class has the following methods: searchHost( ), bitingHost( ), beInfected( ), layEgg( ), beDrying( ), dead( ), detectWaterLevel( ). DetectDayNumber( ). The Host class is a class that allows to model animals and human species, a Host class is characterized by its: degree of mobility, the list of the ponds where it can move, the list of resources, the name of its settlement, Its physiological state, its period of Incubation. Each Host class must perform the following methods: searchWater( ), searchPasture( ), consummeWater( ), consummePasture( ), MoveToWaterPond( ), returnToSettlement( ), beInfected( ). The Pond class is a class which allows managing the dynamics of the water pond characterized by: the list of Mosquito objects it contains, its water level, the status of the water pond, the list of the daily climatic variables. Each water pond class must perform the following methods:waterPonddynamic( ), waterPondState( ), evaporation( ), infiltration( ). The climate class is a class that allows managing the climatic data essential for the survival of mosquitoes and animals. This class is characterized by: daily temperature, humidity, daily precipitation. All those methods are implemented in the CORMAS [2] platform.

## 4 Experimental Description and Results

In the model, a simulation was carried out in 215 steps corresponding to 215 days of the wintering period when studying the dynamics of the Culex poicilipes mosquito populations. Another 90 steps of simulation (90 days) was also carried out for studying the population dynamics of Aedes vexans mosquitoes. The first part consists of initializing the virtual environment of study, the parameters allowing to create the different agents that will have to interact, the number of water ponds, the number of settlements, the climatic parameters such as the intensity of the daily rainfall, the daily temperature, and the daily humidity. The second part consists of integrating the climatic data of year 2010 in the multi-agent platform. The last part consists of carrying out several sensitivities analysis on the dry spells. The corresponding rainfall data was collected during the wintering period of the year 2010. This period starts from June to October or December (about 215 days). All mosquitoes at the beginning of each simulation are In the form of eggs. The hatching of Aedes vexans eggs begins at the beginning of the first rains of June while the hatching of the eggs of the Culex poicilipes mosquitoes will only take place towards the end of August to amplify the infections of the rift valley fever. Each mosquito has its life duration between 21 and 28 days. The different agents are randomly placed in the environment at the beginning of the simulation. Only the hosts move towards the water ponds,

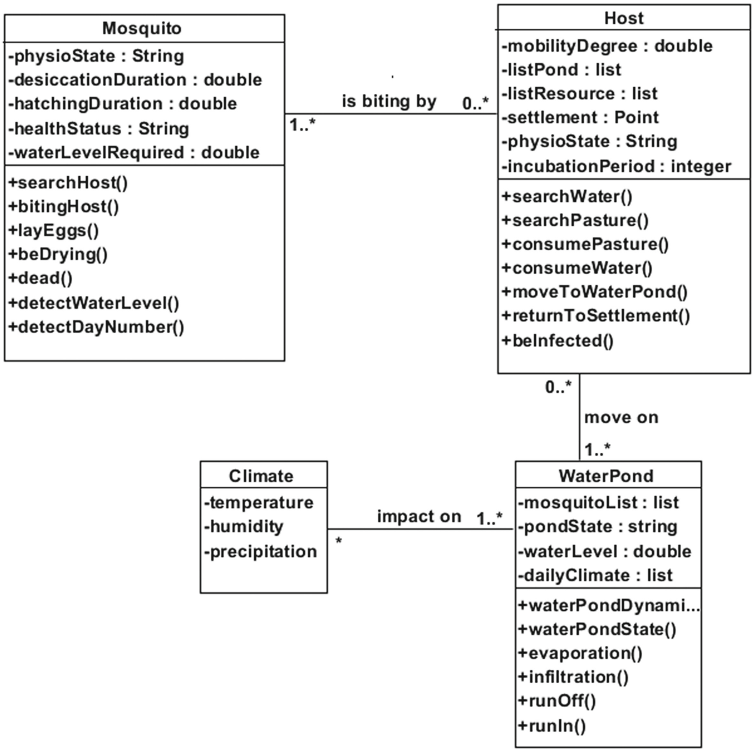


**Fig. 1.** A Simplified UML diagram showing the interactions between mosquitoes, animals, water ponds and climate.

the mosquitoes in this model do not leave the water ponds. The results are structured as follows:

Figure 1 shows The UML class diagram of the interactions between Hosts, mosquitoes, water ponds and climate.

Figure 2 shows the dynamics of the water pond in Ferlo (Senegal) during the winter season 2010 on a period of 215 days (Fig. 3).

Figure 4 shows the dynamics of the water pond and the dynamics of the population of Aedes vexans mosquitoes when occurring two dry spells during the winter season.

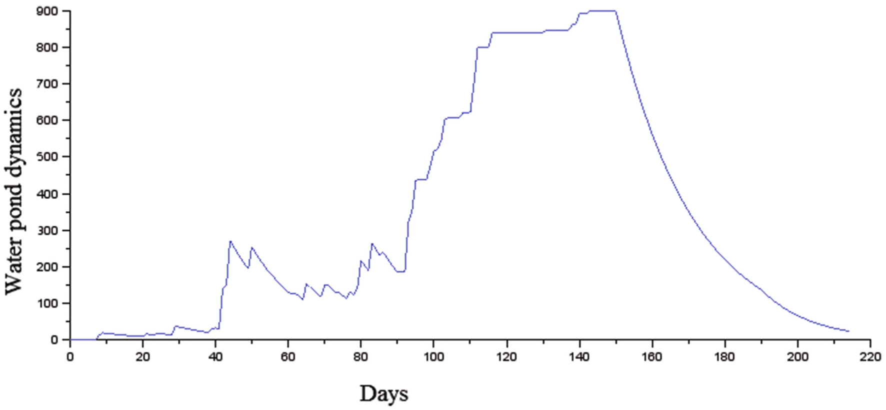


**Fig. 2.** Dynamics of each water pond using the climatic data of year 2010.

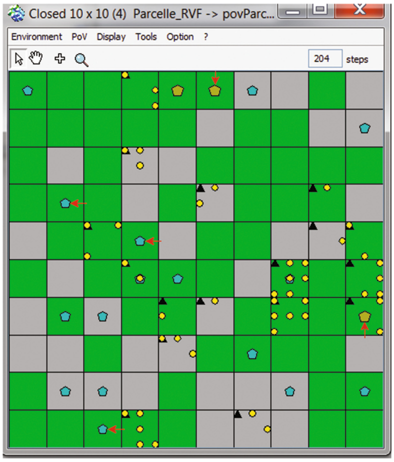

Closed 10 x 10 (4) Parcelle_RVF -> povParc...
Environment PoV Display Tools Option ?
204 steps

**Fig. 3.** Presentation of the screenshot of a simulation showing hosts in yellow, mosquitoes in red, settlements in black triangle, water ponds in blue and gray yellow.

Figure 5 shows the dynamics of the water pond and the dynamics of the population of Culex poicilipes mosquito populations during the winter season 2010.

Figure 6 shows the dynamics of the water pond and the dynamics of the population of Aedes vexans mosquito populations for a single dry spell at the beginning of the winter season during the year 2010.

Figure 7 shows the dynamics of water pond and the dynamics of the population of Aedes vexans mosquito populations for a single dry spell in early July during the winter season of year 2010.

Figure 8 shows the dynamics of water pond and the dynamics of the Aedes vexans mosquito populations without dry spell during the winter season in year 2010.

## 5 Discussion

The life cycle of Aedes vexans and Culex poicilipes mosquito populations depends on the dynamics of water ponds (Fig. 1). The wintering period of 2010 in the Ferlo region has 23 days of dry spell (23 consecutive days without rain) during the month of June, followed by a second (Fig. 4) days of a dry spell between the beginning of July. Each dry spell leads to a decrease of the level of the water within ponds. The eggs of Aedes vexans mosquitoes which are laid from the first day of each dry spell will undergo a desiccation for at least 7 days and will hatch as soon as they are completely immersed. (Fig. 4) shows that the rate of the evolution of Aedes vexans mosquito populations between the start of the first dry spell in June and the end of this first dry spell is 200%. The rate of change between the start of the second dry spell and the end of this second dry spell is 250% (Fig. 4). We can say that the more the dry spell succeed each other, the

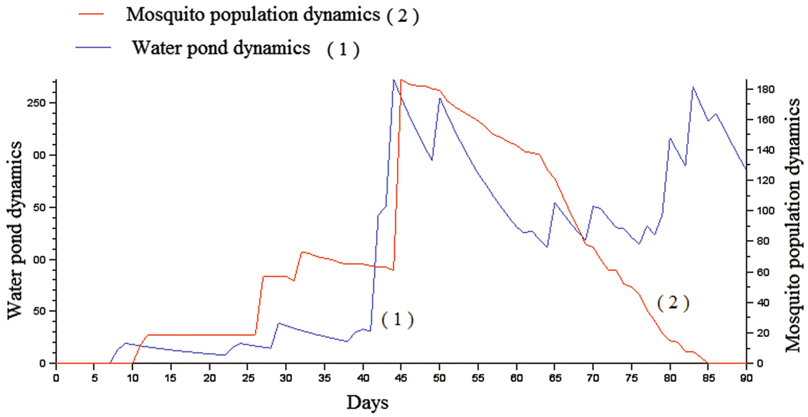


**Fig. 4.** Dynamics of the water pond (in blue color) and dynamics of the Aedes vexans mosquito populations(red color) for two dry spell during the year 2010.

more the growth of the Aedes vexans mosquito populations tends to be exponential. It is also observed that the presence of the two dry spell leads an evolution rate of the Aedes mosquito populations of 850% compared to the numbers of Aedes vexans mosquitoes obtained from the date of the beginning of the first dry spell. It should also be noted that these rates of evolution take into account the death of adult mosquitoes. During the winter of 2010, after the complete disappearance of the mosquitoes Aedes vexans at the end of August, we noted the appearance of Culex poicilipes mosquito populations (Fig. 5) because at this

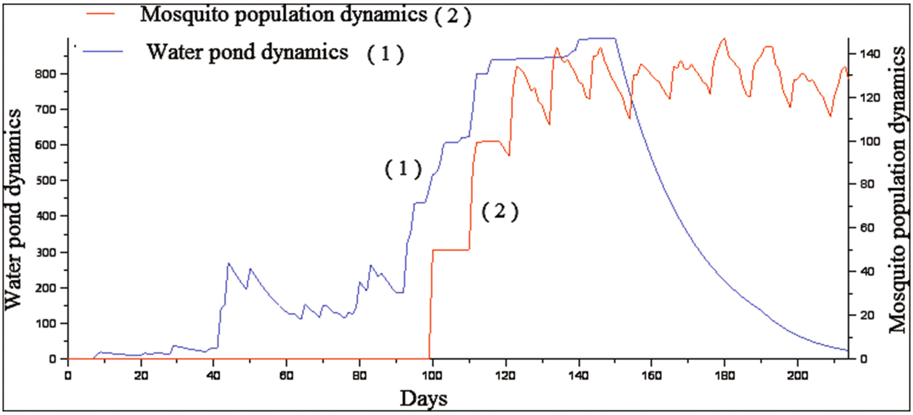


**Fig. 5.** Dynamics of the water pond and dynamics of culex mosquito populations during the winter period 2010. This curve shows the appearance of Culex mosquitoes 10 days after the beginning of August.

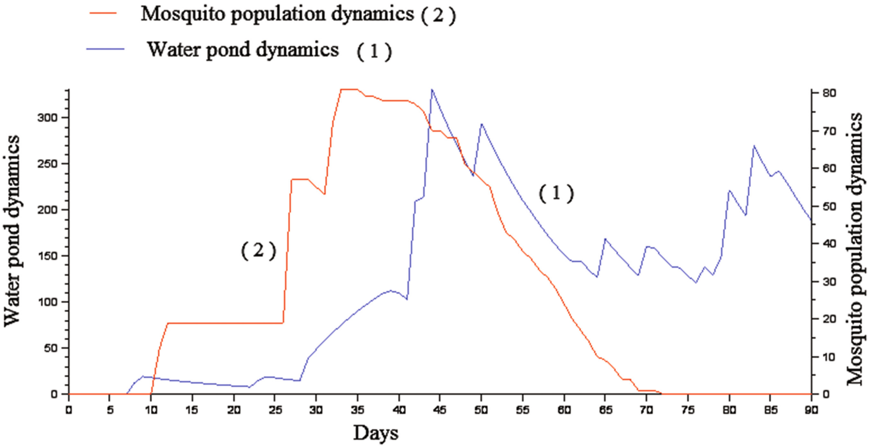


**Fig. 6.** Dynamics of the water pond and the dynamics of the Aedes vexans mosquito populations when there exist a single dry spell at the beginning of winter season during the year 2010.

period all the ponds are full. The results confirm the work of several researchers who state that the growth of mosquito populations plays an important role in the transmission of valley fever Rift in Senegal [18]. We have found that the year where there are several infected animals coincides with the fact that there are too many mosquitoes. The abundance of rains does not necessarily lead to an excessive multiplication of mosquitoes. The variability of daily rainfall is more important than the abundance of successive rains over several days in the process of the hatching of eggs. Our results confirm what other researchers have proved in Senegal that the alternation between rainfall (rainfall variability) causes dry spells that which lead to the multiplication of Aedes vexans mosquitoes. For 90 steps of simulation corresponding to 3 months from the beginning of the wintering period of the year 2010, a large population of Aedes vexans mosquitoes was observed after the second dry spell (Fig. 4). This population of Aedes vexans was relayed by the Culex poicilipes population at the end of August (Fig. 5). Until the appearance of the Culex poicilipes mosquitoes at the end of August, we noted a complete disappearance of Aedes vexans mosquito populations in the absence of dry spells. We carried out sensitivities analysis on dry spells duration during the 2010 wintering period, first, we asked ourselves how many adult mosquitoes can be obtained during the first three months of winter if the dry spell Between the 30th day and the 38th day were canceled by the intermittent rainfall? If this were the case, it would remain the only dry spell that goes from the 10th day to the 22nd day. A simulation performed in this case for a single dry spell at the beginning of the rainy season gives a population of 81 mosquitoes Aedes vexans (Fig. 6) between the beginning of the dry spell and the end of the dry spell. Assuming that a dry spell from 10 June to 22 June was canceled by intermittent rainfall, the only dry spell between 30 June and 8 July 2010 allows mosquitoes

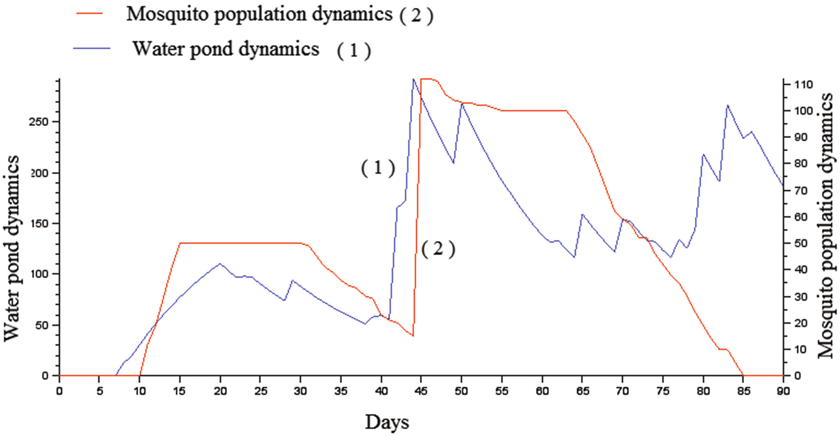


**Fig. 7.** Dynamics of the water pond and the dynamics of the population of Aedes vexans mosquito population when there exist a single dry spell in early July during the year 2010.

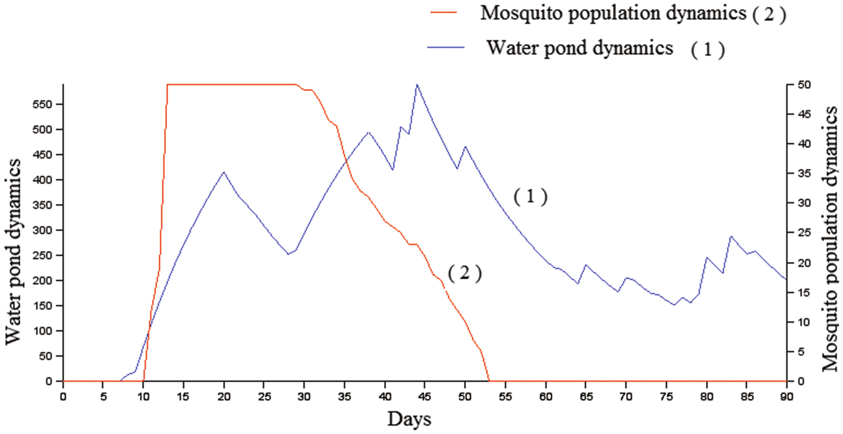


**Fig. 8.** Dynamics of the water pond and the dynamics of the Aedes vexans mosquito populations without dry spell during the winter season of the year 2010.

to reach a population of 103 (Fig. 7). It can, therefore, be said that, as soon as the first mosquitoes were born, the late arrival of the first dry spell caused a considerable increase in the Aedes vexans mosquito populations. In the situation where there are two dry spells (Fig. 4), 186 Aedes mosquitoes were obtained at the end of the second dry spell, so the sequence of dry spells followed by intermittent rains amplifies the growth of Aedes vexans mosquito populations. In the absence of dry spells (Fig. 8), the 50 eggs of mosquitoes used in the initialization of the simulation will emerge as adult mosquitoes will bite animals and lay the eggs again, these eggs will never hatch due to a lack of dry spell, the population of the Aedes mosquitoes will disappear before The beginning of August. The Culex poicilipes mosquitoes will appear (Fig. 5) when Aedes vexans mosquitoes have disappeared 2 weeks or 3 weeks in advance which is an advantage for the spread of RVF.

## 6 Conclusion

This work allows us to show that the abundance of mosquitoes is not directly related to the abundance of rainfall in the Ferlo region. This justification was done by modeling the interactions between rainfall, water ponds, mosquitoes Aedes vexans and Culex poicilipes using the multi-agent approach. The results showed that the continuity of the infections of the animals during the wintering period persist until the end of August because of the presence of the Culex mosquito populations which did not need dry spells for the desiccation of their eggs. The developed model can be used to estimate the number of mosquitoes capable of reaching the adult phase as a function of daily rainfall and dry spell period. To model the dynamics of the mosquito populations, we have only considered the water pond dynamics taking into account evaporation and infiltration.

This model could be used to study the impact of rainfall variability on other species of insects living in water ponds in Senegal. We also observed that the eggs of Aedes vexans at the end of the dry spell are closely related to the water level of the ponds responsible for the immersion of these eggs. Agent-based modeling is a powerful tool because it gives the privilege to simulate the behaviors of mosquitoes in the environment. The developed model thus allows us to better understand the notion of a dry spell, as well as its impact on the filling and drying of ponds and their effect on the life cycle of mosquitoes.

## References

1. Becker, N.: Life strategies of mosquitoes as an adaptation to their habitats. Bull. Soc. Vector Ecol. **14**, 6–25 (1989)
2. Bousquet, F., Le Page, C.: Multi-agent simulations and ecosystem management: a review. Ecol. Model. **176**, 313–332 (2004)
3. Confalonieri, U., Menne, B., Akhtar, R., Ebi, K.L., Hauengue, M., Kovats, R.S., Revich, B., Woodward, A.: Climate Change 2007: Impacts, Adaptation and Vulnerability. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (eds.) Human Health. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 391–431. Cambridge University Press, Cambridge, UK (2007)
4. Davies, F.G., Jacobsen, P., Sylla, D.: Laboratory manual on Rift Valley fever: isolation and identification techniques. In: Report of FAO/WHO Group on Emergency Preparedness for Rift Valley Fever Control in West Africa. Report no. WHO-VPH/88.77, pp. 1–134. World Health Organization, Geneva (1998)
5. Davies, F.G., et al.: Rainfall and epizootic Rift Valley fever. World Health Org. Rep. **63**, 941–943 (1985)
6. Focks, D.A., Haile, D.G., Daniels, E., Mount, G.A.: Dynamic life table model for Aedes aegypti (Diptera: Culicidae): simulation results and validation. J. Med. Entomol. **30**, 1018–1028 (1993a)
7. Hoch, A.L., Turell, M.J., Bailey, C.L.: Replication of Rift Valley fever virus in the sand fly Lutzomyia longipalpis. Am. J. Trop. Med. Hyg. **33**(2), 295–299 (1984)
8. House, J.A., Turell, M.J., Mebus, C.A.: Rift Valley fever: present status and risk to the Western Hemisphere. Ann. N. Y. Acad. Sci. **653**, 233–242 (1992)
9. Meegan, J.M., Bailey, C.L.: Rift Valley fever. In: Arboviruses Epidemiology and Ecology, vol. IV, pp. 51–76 (1988)
10. Mondet, B., et al.: Relations entre la pluviométrie et le risque de transmission virale par les moustiques: cas du virus de la Rift Valley fever (RVF) dans le Ferlo (Sénégal). Environnement Risques et Santé 4, 125–129 (2005)
11. Muller, G., Grébaut, P., Gouteux, J.: An agent-based model of sleeping sickness: simulation trials of a forest focus in Southern Cameroon. C. R. Biol. **327**, 1–11 (2004)
12. Ndiaye, P.I.: Modélisation de la dynamique de population des moustiques Aedes en zonne sahélienne. Exemple des Aedes vexans arabiensis (Diptera: cilicidae) vecterus de la fièvre de la du Rift en Afrique de l'ouest. Mathematics, Metz University and Gaston Berger de Saint Louis University (2006)

13. Ndione, J.-A., et al.: Variabilite intra-saisonniere de la pluviometrie et emergence de la fievre de la vallee du rift dans la vallee du fleuve Senegal: nouvelles considerations. Climatologie 5. 8397. Environnement et épidémiologie de la fièvre de la vallée du Rift (FVR) dans le bassin inférieur du fleuve Sénégal. Environnnement, Riques et Santé 2, 176–182 (2008)
14. Roche, B., et al.: Multi-agent systems in epidemiology: a first step for computational biology in the study of vector-borne disease transmission. BMC BioInformatics (2008)
15. Shaman, J., Day, J.F.: Reproductive phase locking of mosquito populations in response to rainfall frequency. PLoS ONE **2**, e331 (2007)
16. Shone, S.M., Curriero, F.C., Lesser, C.R., Glass, G.E.: Characterizing population dynamics of Aedes sollicitans (Diptera: Culicidae) using meteorological data. J. Med. Entomol. **43**, 393–402 (2006)
17. Hare, M., Deadman, P.: Further towards a taxonomy of agent-based simulation models in environmental management. Math. Comput. Simulat. **64**, 25–40 (2004)
18. Pin, R.: Spatialisation du risque de transmission de la fièvre de la Vallee du Rift en milieu agropastoral sahélien du Sénégal septentrional. Thèse d'Université, Géographie, environnement, aménagement, Orléans (2006)

# Author Index

A
Abdelhafid, Zitouni, 50
Abdul Aziz, Izzatdin, 173
Adewumi, Aderemi O., 204
Almonacid, Boris, 102
Alves, Tiago, 37
Arki, Oussama, 62
Astorga, Gino, 126
Aziz, Izzatdin Abdul, 195

B
Badr, Amr, 86
Bah, Alassane, 338, 381
Bekti, Rokhana D., 287, 318
Berrazega, Ines, 238
Bibilo, Petr, 372
Bouaine, Amine, 298
Boulahbel-Bachari, Samira, 338
Budiman, Priscilia, 287
Budniak, Zbigniew, 250
Bystrov, Vitaliy, 310

C
Caballero, Hugo, 156
Chaudhuri, Arindam, 114
Chern-Tong, Han, 195
Čížek, Ondřej, 13
Coelho, Jorge, 37
Cordeiro, Raul, 185
Crawford, Broderick, 102, 126, 156

D
da Silva Benigno, Odecília Barreira, 326
Djoudi, Mahieddine, 139

E
El Saadi, Nadjia, 338
Elfida, Maria, 220
Eliseev, S.M., 350
Estrada, Sandra, 74
Ezugwu, Absalom E., 204

F
Faiz, Rim, 238
Formánek, Tomáš, 1
Forte, Sergio Alexandre Barreira, 227
Forte, Sergio Henrique Arruda Cavalcante, 227

G
Gajdoš, Andrej, 360
García, José, 126
Ghosh, Soumya K., 114
Gorelova, Galina, 372

H
Hasan, Mohd Hilmi, 173
Hernandez, Leonel, 74
Hu, Longying, 262

I
Irwansyah, Edy, 287, 318

J
Jaafar, Jafreezal, 173

K
Kacalak, Wojciech, 250
Kanigoro, Bayu, 287, 318
Khaliullina, Darya, 310
Klimenko, Anna, 372
Korobkin, Vladimir, 372
Kularbphettong, Kunyanuth, 150, 167

M
Maghawry, Ahmed Mohamed, 86
Majewski, Maciej, 250

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0

Malygina, Svetlana, 310
Manuel, José, 185
Mohammad Zubir,
Wan Mohammad Aflah, 173

**N**
Nasution, Mahyuddin K.M., 220
Ndiaye, Papa Ibrahima, 381
Ndione, Jacques André, 381
Nogueira, Luís, 37

**O**
Olivares, Rodrigo, 102
Olusanya, Micheal O., 204
Omar, Yasser, 86
Ortega, Héctor, 102

**P**
Paul, Python Ndekou Tandong, 381
Pelikán, Martin, 279
Pinheiro, Plácido Rogério, 227, 326

**Q**
Qasem, Mais Haj, 24
Qatawneh, Mohammad, 24

**R**
Rachik, Mostafa, 298
Rahab, Hichem, 139
Ramdane, Maamri, 50
Riad, Bouaita, 50

**S**
Silva, Carla, 185
Silva, Carolina Ferreira Gomes, 326
Sokol, Pavol, 360
Song, Cheng, 262
Soto, Ricardo, 102, 126, 156
Štiková, Hana, 279
Syah, Rahmad, 220

**T**
Theodorick, 318
Tyrychtr, Jan, 279

**V**
Vicentiy, A.V., 350
Vicentiy, I.V., 350
Villanueva, Humberto, 74
Vrana, Ivan, 279

**Y**
Yuan, Haiyan, 262

**Z**
Zitouni, Abdelhafid, 62, 139