

Liew-Kee Kor · Abd-Razak Ahmad ·
Zanariah Idrus · Kamarul Ariffin Mansor
Editors

Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)

Transcending Boundaries, Embracing
Multidisciplinary Diversities

 Springer

Proceedings of the Third International Conference
on Computing, Mathematics and Statistics
(iCMS2017)

Liew-Kee Kor · Abd-Razak Ahmad ·
Zanariah Idrus · Kamarul Ariffin Mansor
Editors

Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)

Transcending Boundaries, Embracing
Multidisciplinary Diversities

 Springer

Editors

Liew-Kee Kor
Universiti Teknologi MARA (UiTM) Kedah
Sungai Petani, Kedah, Malaysia

Abd-Razak Ahmad
Universiti Teknologi MARA (UiTM) Kedah
Sungai Petani, Kedah, Malaysia

Zanariah Idrus
Universiti Teknologi MARA (UiTM) Kedah
Sungai Petani, Kedah, Malaysia

Kamarul Ariffin Mansor
Universiti Teknologi MARA (UiTM) Kedah
Sungai Petani, Kedah, Malaysia

ISBN 978-981-13-7278-0 ISBN 978-981-13-7279-7 (eBook)
<https://doi.org/10.1007/978-981-13-7279-7>

Library of Congress Control Number: 2019935829

© Springer Nature Singapore Pte Ltd. 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Organizing Committee

Local Organizing Committee

Abd Razak Ahmad, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Asmahani Nayan, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Fazillah Bosli, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Illiasaak Ahmad, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Kamarul Ariffin Mansor, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Kor Liew Kee, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Noor Hafizah Zainal Aznam, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Rosidah Ahmad, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Shahida Farhan Zakaria, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Siti Fairus Mokhtar, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Siti Nurbaya Ismail, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Suhardi Hamid, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Wan Siti Esah Che Hussain, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Wan Zulkipli Wan Salleh, Universiti Teknologi MARA (UiTM) Kedah, Malaysia
Zanariah Idrus, Universiti Teknologi MARA (UiTM) Kedah, Malaysia

International Scientific Committee

Tudorel Andrei, National Institute of Statistics and Bucharest, University of Economic Studies, Bucharest, Romania
Matthias Templ, Statistics Austria & Vienna University of Technology
Ralf Munnich, Statistics Austria & Vienna University of Technology
Risto Lehtonen, University of Helsinki, Finland
Gergely Daroczi, Easystats Ltd., Budapest, Hungary

Lucian Liviu Albu, Romanian Academy, Institute for Economic Forecasting, Bucharest, Romania
 Gheorghe Zaman, Romanian Academy, Institute of National Economy, Bucharest, Romania
 Valentina Vasile, Romanian Academy, Institute of National Economy, Bucharest, Romania
 Dorin Jula, Romanian Academy, Institute for Economic Forecasting and Ecological University of Bucharest, Romania
 Bogdan Oancea, The University of Bucharest and National Institute of Statistics, Bucharest, Romania
 Monica Roman, Bucharest University of Economic Studies, Romania
 Nicoleta Caragea, National Institute of Statistics and Ecological University of Bucharest, Romania
 Antoniadă-Ciprian Alexandru, Ecological University of Bucharest, Romania
 Adrian Dusa, University of Bucharest, Romania
 Elena Druica, University of Bucharest, Romania
 Nicolae-Marius Jula, Nicolae Titulescu University, Bucharest, Romania
 Ana Maria Dobre, National Institute of Statistics
 Gerald Cheang, University of South Australia, Australia
 Suhaidi Hassan, Universiti Utara Malaysia, Malaysia
 Martin Everett, University of Manchester, UK
 Nicolaas Jan Dirk Nagelkerke, University of Amsterdam, Netherland
 Poo Kuan Hoong, AC Nielson, Malaysia
 Nurhuda Ismail, Universiti Teknologi MARA, Malaysia

Reviewers

Abd Razak Ahmad, Universiti Teknologi MARA (UiTM), Malaysia
 Ab Razak Mansor, Universiti Teknologi MARA (UiTM), Malaysia
 Afdallyna Fathiyah Harun, Universiti Teknologi MARA (UiTM), Malaysia
 Ahmad Farid bin Osman, Universiti Malaysia (UM), Malaysia
 Amar Faiz Zainal Abidin, Universti Teknikal Malaysia (UTEM), Malaysia
 Azrul Hazri Jantan, Universiti Putra Malaysia (UPM), Malaysia
 Chew Cheng Meng, Universiti Sains Malaysia (USM), Malaysia
 Ciprian Antoniadă Alexandru, Ecological University of Bucharest, Romania
 Dahlia Ibrahim, Universiti Teknologi MARA (UiTM), Malaysia
 Daud Bin Mohamad, Universiti Teknologi MARA (UiTM), Malaysia
 Erny Arniza Ahmad, Universiti Teknologi MARA (UiTM), Malaysia
 Goh Kim Leng, Universiti Malaysia (UM), Malaysia
 Ida Normaya Mohd Nasir, Universiti Teknologi MARA (UiTM), Malaysia
 Idham Arif Hj Alias, Universiti Putra Malaysia (UPM), Malaysia
 Illiasaak Ahmad, Universiti Teknologi MARA (UiTM), Malaysia
 Jasmin Ilyani Ahmad, Universiti Teknologi MARA (UiTM), Malaysia

Jehan Zeb Shah, Optics Laboratories, Islamabad, Pakistan
Kamarul Ariffin Mansor, Universiti Teknologi MARA (UiTM), Malaysia
Kartini Kasim, Universiti Teknologi MARA (UiTM), Malaysia
Khairul Adilah Ahmad, Universiti Teknologi MARA (UiTM), Malaysia
Kor Liew Kee, Universiti Teknologi MARA (UiTM), Malaysia
Kumaresan Nallasamy, Universiti Malaysia (UM), Malaysia
Leony Tham Yew Seng, Universiti Malaysia Kelantan (UMK), Malaysia
Maznah Mat Kasim, Universiti Utara Malaysia (UUM), Malaysia
Michael Khoo Boon Chong, Universiti Sains Malaysia (USM), Malaysia
Mohd Bakri Adam, Universiti Putra Malaysia (UPM), Malaysia
Mohd Hafiz bin Mohd, Universiti Sains Malaysia (USM), Malaysia
Mohd Rijal Ilias, Universiti Teknologi MARA (UiTM), Malaysia
Mohd Tahir Ismail, Universiti Sains Malaysia (USM), Malaysia
Muhammad Rozi Malim, Universiti Teknologi MARA (UiTM), Malaysia
Nicoleta Caragea, Ecological University of Bucharest, Romania
Noor Hasnita Abdul Talib, Universiti Teknologi MARA (UiTM), Malaysia
Noor Rasidah Ali, Universiti Teknologi MARA (UiTM), Malaysia
Noor Wahida Yunus, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Nor Hafizah Abdul Razak, Universiti Teknologi MARA (UiTM), Malaysia
Nor Idayu Mahat, Universiti Utara Malaysia (UUM), Malaysia
Norashikin Nasaruddin, Universiti Teknologi MARA (UiTM), Malaysia
Norhashidah Awang, Universiti Sains Malaysia (USM), Malaysia
Norin Rahayu Shamsuddin, Universiti Teknologi MARA (UiTM), Malaysia
Nur Syibrah Muhamad Naim, Universiti Sains Malaysia (USM), Malaysia
Nurazlina Abdul Rashid, Universiti Teknologi MARA (UiTM), Malaysia
Nurul Akmal Mohamed, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Nurul Hila Zainuddin, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Phoong Seuk Yen, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Raja Noor Farah Azura Binti
Raja Ma'amor Shah, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Ramayah Thurasamy, Universiti Sains Malaysia (USM), Malaysia
Rogayah Abdul Majid, Universiti Teknologi MARA (UiTM), Malaysia
Rosma Mohd Dom, Universiti Teknologi MARA (UiTM), Malaysia
Rozaini Roslan, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia
Sayang Mohd Deni, Universiti Teknologi MARA (UiTM), Malaysia
Shahida Farhan Binti Zakaria, Universiti Teknologi MARA (UiTM), Malaysia
Shaifizat Mansor, Universiti Teknologi MARA (UiTM), Malaysia
Shamsul Jamel Elias, Universiti Teknologi MARA (UiTM), Malaysia
Shazlyn Milleana Shaharudin, Universiti Pendidikan Sultan Idris (UPSI), Malaysia
Shukri Shamsuddin, Universiti Teknologi MARA (UiTM), Malaysia
Siti Ainor Binti Mohd Yatim, Universiti Sains Malaysia (USM), Malaysia,
Siti Amirah Abd Rahman, Universiti Sains Malaysia (USM), Malaysia
Siti Fairus Binti Mokhtar, Universiti Teknologi MARA (UiTM), Malaysia
Siti Nurbaya Ismail, Universiti Teknologi MARA (UiTM), Malaysia
Siti Rafidah Muhamat Dawam, Universiti Teknologi MARA (UiTM), Malaysia

Siti Suhana Jamaian, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia
Suhardi Hamid, Universiti Teknologi MARA (UiTM), Malaysia
Sumarni Abu Bakar, Universiti Teknologi MARA (UiTM), Malaysia
Suraya Masrom, Universiti Teknologi MARA (UiTM), Malaysia
Taniza Tajuddin, Universiti Teknologi MARA (UiTM), Malaysia
Teh Yuan Ying, Universiti Utara Malaysia (UUM), Malaysia
Waaail Mahmood Lafta Al-Waely, Al-Mustafa University College, Iraq
Wan MuhamadAmir W Ahmad, Universiti Sains Malaysia (USM), Malaysia
Wan Siti Esah Che Hussain, Universiti Teknologi MARA (UiTM), Malaysia
Wan Zawiah Wan Zin, Universiti Kebangsaan Malaysia (UKM), Malaysia
Yanti Aspha Ameira Mustapha, Universiti Teknologi MARA (UiTM), Malaysia
Yap Bee Wah, Universiti Teknologi MARA (UiTM), Malaysia
Zahayu Md Yusof, Universiti Utara Malaysia (UUM), Malaysia
Zabidin Salleh, Univerisit Malaysia Terengganu (UMT), Malaysia
Zainura Idrus, Universiti Teknologi MARA (UiTM), Malaysia
Zanariah Idrus, Universiti Teknologi MARA (UiTM), Malaysia

Preface

This book of Conference Proceedings consists of papers presented at The 3rd International Conference on Computing, Mathematics and Statistics (iCMS2017) held in conjunction with the 5th Joint International Conference on New Challenges for Statistical Software: The Use of R in Official Statistics (uRos2017)—Asia Pacific on 7–8 November 2017 in Langkawi Island, Malaysia. The Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kedah, Malaysia hosted the iCMS2017 along with the Ecological University of Bucharest, Romania co-organizing the uRos2017 Asia-Pacific.

Fuelled by the momentum created at the previous conference series iCMS2015 which successfully attracted paper contributors from around the world, the iCMS2017 together with uRos2017 have adopted the theme *Transcending Boundaries, Embracing Multidisciplinary Diversities* to provide a common ground for educators from diverse settings to discuss common challenges and to seek innovative solutions.

This book is divided into four parts: Computing, Mathematics, Statistics, and Applications. All papers have undergone blind reviews and revised in response to reviewers' comments to ensure that the article is of a high quality. There are 74 papers in this volume.

In putting this book together, we would like to thank all contributors to this Conference Proceedings. We extend our sincere appreciation to the reviewers for the effort and time put into the review of the papers.

Sungai Petani, Malaysia

Liew-Kee Kor
Abd-Razak Ahmad
Kamarul Ariffin Mansor
Zanariah Idrus

About This Book

This Proceedings consists of refereed papers presented at The 3rd International Conference on Computing, Mathematics and Statistics (iCMS2017) held in conjunction with the 5th Joint International Conference on New Challenges for Statistical Software: The Use of R in Official Statistics (uRos2017)—Asia Pacific on 7–8 November 2017 in Langkawi Island, Malaysia. This book is divided into four parts: Computing, Mathematics, Statistics, and Applications. All chapters of the book were written in response to the conference theme *Transcending Boundaries, Embracing Multidisciplinary Diversities*.

The growth of scientific and technical knowledge in recent decades has produced new solutions to solve complex problems. iCMS2017 together with uRos2017 have paved a common ground for researchers from diverse settings to discuss common challenges and to seek innovative solutions. This book is set aside for readers who are enthusiastic about solving problems of global importance with new methods and technologies that go beyond borders of any subject matter or discipline.

Contents

Part I Computing

1	Determining Influential Household Routines for Domestic Water Consumption Estimation via Genetic Algorithm	3
	Nurul Nadia Hani, Khairul Anwar Rasmani, Noor Elaiza Abd Khalid and Ahmad Firdaus Ahmad Fadzil	
2	Sport Suitability Prediction Based on Physical Fitness Components Using k-Nearest Neighbors Algorithm	11
	Muhammad Nabil Fikri Jamaluddin, Mohd Syafiq Miswan, Shukor Sanim Mohd Fauzi, Ray Adderley JM Gining, Noor Fadlyana Raman and Mohd Zaid Mohd Ghazali	
3	Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE	19
	Nurulfitrah Noorhalim, Aida Ali and Siti Mariyam Shamsuddin	
4	Blood Vessels Segmentation of Retinal Fundus Image via Stack-Based Object-Oriented Region Growing	31
	Ahmad Firdaus Ahmad Fadzil, Shafaf Ibrahim and Noor Elaiza Abd Khalid	
5	Similarity Measures of Intuitionistic Fuzzy Sets for Cancer Diagnosis: A Comparative Analysis	39
	Lazim Abdullah and Sook Wern Chan	
6	Comprehensive Performance Assessment on Open Source Intrusion Detection System	45
	Fuad Mat Isa, Shahadan Saad, Ahmad Firdaus Ahmad Fadzil and Raihana Md Saidi	

7	Mobile Ad-Hoc Network (MANET) Routing Protocols: A Performance Assessment	53
	Nur Amirah Mohd Saudi, Mohamad Asrol Arshad, Alya Geogiana Buja, Ahmad Firdaus Ahmad Fadzil and Raihana Md Saidi	
8	Multimedia Data Archive Application in Cloud Environment	61
	Syarilla Iryani A. Saany, M. Nordin A. Rahman, A. Rasid Mamat and Ahmad Fakhri Ab Nasir	
9	Thalas Test Android Application	69
	Iliana Mohd Ali and Nooraida Samsudin	
10	Dimensions of Mobile Information Behavior	77
	Zuraidah Arif, Abd Latif Abdul Rahman and Asmadi Mohamed Ghazali	
Part II Mathematics		
11	A Promising Method to Approximate Fractional Derivatives Under Uncertainty	87
	Ali Ahmadian, Norazak Senu, Fudziah Ismail and Soheil Salahshour	
12	Energy Dissipation of Free Convection Boundary Layer Flow in a Jeffrey Fluid Across a Horizontal Circular Cylinder with Suspended Nanoparticles	93
	Syazwani Mohd Zokri, Nur Syamilah Arifin, Abdul Rahman Mohd Kasim, Nurul Farahain Mohammad and Mohd Zuki Salleh	
13	Generalized Half-Step Hybrid Block for Solving Second Order Ordinary Differential Equations Directly	101
	Kamarun Hizam Mansor, Zurni Omar and Azizah Rohni	
14	Mixed Convection Boundary Layer Flow on a Solid Sphere in a Viscoelastic Micropolar Fluid	111
	Laila Amera Aziz, Abdul Rahman Mohd Kasim, Mohd Zuki Salleh and Sharidan Shafie	
15	Mathematical Modelling of Bank Financial Management in Malaysia with Goal Programming Approach	119
	Chen Jia Wai, Lam Weng Siew and Lam Weng Hoe	
16	Numerical Solutions on Boundary Layer of Casson Micropolar Fluid Over a Stretching Surface	127
	Abdul Rahman Mohd Kasim, Hussein Ali Mohammed Al-Sharifi, Nur Syamilah Arifin, Mohd Zuki Salleh and Sharidan Shafie	

17 One-Step Third-Derivative Block Method with Two-Hybrid Points for Solving Non-linear Dirichlet Second Order Boundary Value Problems 135
 Mohammad Alkasassbeh and Zurni Omar

18 Pricing Asian Option by Solving Black–Scholes PDE Using Gauss–Seidel Method 147
 W. S. Koh, R. R. Ahmad, S. H. Jaaman and J. Sulaiman

19 Reinforcement Learning for Computing Power Grid Network Operating Functions 153
 Shivam Sharma, Pinki Gupta and Laxminarayan Das

20 Similarity Measure for Fuzzy Number Based on Distances and Geometric Shape Characteristics 159
 Nur Amira Mat Saffie, Khairul A. Rasmani and Nor Hashimah Sulaiman

21 Solving Fourth Order Linear Initial and Boundary Value Problems Using an Implicit Block Method 167
 Oluwaseun Adeyeye and Zurni Omar

22 Stability Analysis of Explicit and Semi-implicit Euler Methods for Solving Stochastic Delay Differential Equations 179
 Norhayati Rosli, Noor Amalina Nisa Ariffin, Yeak Su Hoe and Arifah Bahar

23 Stability Analysis of 4-Stage Stochastic Runge-Kutta Method (SRK4) and Specific Stochastic Runge-Kutta Method (SRKS1.5) for Stochastic Differential Equations 187
 Noor Amalina Nisa Ariffin, Norhayati Rosli and Abdul Rahman Mohd Kasim

24 The VIKOR Method with Pythagorean Fuzzy Sets and Their Applications 195
 Wan Rosanisah Wan Mohd and Lazim Abdullah

25 Topological Properties of Flat Electroencephalography 201
 Tan Lit Ken, Tahir Ahmad, Nor Azwadi Che Sidik, Chuan Zun Liang, Lee Kee Quen, Gan Yee Siang, Goh Chien Yong and Tey Wah Yen

26 Two-Phase Mixed Convection Flow of Dusty Williamson Fluid with Aligned Magnetic Field over a Vertical Stretching Sheet 209
 Nur Syamilah Arifin, Syazwani Mohd Zokri, Abdul Rahman Mohd Kasim, Mohd Zuki Salleh and Nurul Farahain Mohammad

27	Unveiling the Asymmetric Adjustments of Policy Reaction Function in Indonesia	217
	Lavaneesvari Manogaran and Siok Kun Sek	
Part III Statistics		
28	A Comparative Study of Outlier Detection Methods in Poisson Regression	227
	Faten Nabila Rustam Affandy and Sanizah Ahmad	
29	A Modified Long Memory Model for Modeling Interminable Long Memory Process	235
	Rosmanjawati Abdul Rahman and Sanusi A. Jibrin	
30	Application of Functional Data Analysis in Streamflow Hydrograph	245
	Jamaludin Suhaila	
31	Bayesian Random Forest for the Classification of High-Dimensional mRNA Cancer Samples	253
	Oyebayo Ridwan Olaniran and Mohd Asrul Affendi Bin Abdullah	
32	Bayesian Statistical Modeling: Comparisons Between Poisson and Its Zero-Inflated Regression Model	261
	Muhammad' Afif Amir Husin and Mohd Fadzli Mohd Fuzi	
33	BayesRandomForest: An R Implementation of Bayesian Random Forest for Regression Analysis of High-Dimensional Data	269
	Oyebayo Ridwan Olaniran and Mohd Asrul Affendi Bin Abdullah	
34	Bivariate Weibull Exponential Model Based on Gaussian Copula	277
	Zakiah Ibrahim Kalantan and Mervet Khalifah Abd Elaal	
35	Claim Assessment of a Rainfall Runoff Model with Bootstrap	287
	Wen Jia Tan, Lloyd Ling, Zulkifli Yusop and Yuk Feng Huang	
36	Comparing the Influences of Monetary Versus Fiscal Policy on the Economy: The Case of Malaysia	295
	Siti Fatimah Ismail and Sek Siok Kun	
37	Comparison Between k-Means and k-Medoids for Mixed Variables Clustering	303
	Norin Rahayu Shamsuddin and Nor Idayu Mahat	
38	Data Analysis Comparison Logit and Probit Regression Using Gibbs-Sampler	309
	Subanar	

39 Dirichlet-Multinomial Estimation of Small Area Proportions of Socio-Economic Classes 317
 Shirlee R. Ocampo, Harley Garcia and Mariel Uy

40 Efficiency of Fishery Production in Malaysia Using Data Envelopment Analysis 325
 Anis Atiqah Abdul Rais, Siti Shaliza Mohd Khairi, Zalina Zahid and Noor Asiah Ramli

41 Examining the PPP Theory for ASEAN-5 with Panel Data Analysis 333
 Niri Martha Choji and Siok Kun Sek

42 Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach 343
 M. Fariz Fadillah Mardianto, Sri Haryatmi Kartiko and Herni Utami

43 Hydrological Trend Analysis in Johor 351
 Norazlina Ismail, Werda Yelling and Nurfarhana Hassan

44 Implementing Correlation Dimension: K-Means Clustering via Correlation Dimension 359
 Zakiah Ibrahim Kalantan

45 Job Satisfaction Among Academic Staff: A Structural Equation Modeling Approach 367
 Haslinda Ab Malek, Farah Farhana Mazli, Hafizah Sharif, Noor Azira Mohammad and Isnewati Ab Malek

46 Machine Learning Using H2O R Package: An Application in Bioinformatics 375
 Azian Azamimi Abdullah and Shigehiko Kanaya

47 Modeling of Risk for Diabetes Mellitus and Hypertension Using Bi-response Probit Regression 383
 Suliyanto and M. Rifada

48 On Bootstrapping Using Smoothed Bootstrap 391
 Sulafah Binhimd

49 On the Markov Chain Monte Carlo Convergence Diagnostic of Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification 397
 Nur Iriawan, Kartika Fithriasari, Brodjol Sutijo Suprih Ulama, Irwan Susanto, Wahyuni Suryaningtyas and Anindya Apriliyanti Pravitarsari

50 Road Fatalities Using Logistic Regression 405
 Isnewati Ab Malek, Nurul Najihah Mohd Salim, Siti Naffsikah Alias,
 Nurul Akilah Mohd Zaki and Haslinda Ab Malek

51 Robust Clustering on Spatial Torrential Rainfall Patterns 413
 Shazlyn Milleana Shaharudin and Norhaiza Ahmad

52 Robust Logistic Regression in Application to Divorce Data 421
 Sanizah Ahmad and Rosa Shafiq Azureen Mohamad Rosni

**53 SMOTE Approach to Imbalanced Dataset in Logistic Regression
 Analysis 429**
 Amirah Hazwani Abdul Rahim, Nurazlina Abdul Rashid,
 Asmahani Nayan and Abd-Razak Ahmad

**54 Statistic for Outlier Detection in Circular Functional Relationship
 Model 435**
 Mohd Syazwan Mohamad Anuar, Abdul Ghapor Hussin
 and Yong Zulina Zubairi

55 Structural Breaks in Malaysian Shariah Compliant Indices 441
 Ida Normaya Mohd Nasir and Mohd Tahir Ismail

**56 Teenage Driving Behavior Modeling Using Deep Learning
 for Driver Behavior Classification 449**
 Muhamad-Husaini Abu-Bakar, Rizal Razuwan and Syafiq Kamal

**57 Temporal Patterns Analysis of Paddy Production in Sri
 Lanka 457**
 N. B. W. I. Udeshika and T. M. J. A. Cooray

**58 The Effects of Awareness Level on Littering Behaviour
 on Campus: Family Income as Moderator 465**
 Mas'udah Asmui, Sharifah Norhuda Syed Wahid,
 Suhanom Mohd Zaki, Noorsuraya Mohd Mokhtar
 and Siti Suhaila Harith

**59 The Trends of Age and Gender Specific Mortality Rates by
 Ethnic Groups 475**
 Saiful Azril Ishak, Syazreen Niza Shair,
 Wan Nor Ayunni Wan Ahmad Shukiman, Nurazliyana Mat Radzi
 and Nur Salbiah Abdul Rahman

Part IV Application

60 A Markov Chain Model for Diabetes Mellitus Patients 483
 Muhammad Rozi Malim, Faridah Abdul Halim,
 Farah Wahidah Md Aris, Nur Musrsyiddah Azizuddin,
 Raihannah Othman, Siti Nur Azyyati Rosli
 and Siti Fairuz Kamaruzaman

**61 A New Method to Forecast TAIEX Based on Fuzzy Time
 Series with Trapezoidal Fuzzy Numbers and Center of Gravity
 Similarity Measure Approach** 489
 Siti Musleha Ab Mutalib, Nazirah Ramli, Daud Mohamad
 and Norhuda Mohammed

62 A Proposed Conceptual of Derivative Games Based Learning 497
 Ainon Syazana Ab Hamid, Izni Syamsina Saari,
 Samsiah Abdul Razak and Aslina Omar

**63 AHP Ranking of CSR Human Resource Theme of Takaful
 Operators** 505
 Shahida-Farhan Zakaria and Abd-Razak Ahmad

**64 Assessing Malaysian Teachers’ Perception on Computational
 Thinking Concepts Using SEM** 513
 Ung L. Ling, Tammie C. Saibin, Jane Labadin
 and Norazila Abdul Aziz

**65 Chaotic Stochastic Lee-Carter Model in Predicting Kijang Emas
 Price Movements: A Machine Learning Approach** 521
 Siti Nurasyikin Shamsuddin, Nur Haidar Hanafi,
 Muhammad Hilmi Samian and Mohd Nazrul Mohd Amin

**66 Efficiency of General Takaful Industry in Malaysia: A Two-Stage
 Data Envelopment Analysis** 529
 Hui Shan Lee, Fan Fah Cheng, Annuar Md Nassir,
 Nazrul Hisyam Ab Razak and Wai Mun Har

67 Enhancement of DNA Gel Images 537
 CT Munnirah Niesha Mohd Shafee, Ahmad Khudzairi Khalid
 and Zarith Sofiah Othman

**68 Factors Affecting Entrepreneurial Intention Among IKN
 Students** 545
 Faridah Abdul Halim, Muhammad Rozi Malim, Siti Iliyana Hamdan,
 Atika Salehan and Farhana Syahirah Kamaruzzaman

**69 Improving the Food Manufacturing System by Using
 Simulation and DEA** 555
 Noor Fatin Kamarudin, Ruzanita Mat Rani and Faridah Abdul Halim

70 Modelling Multi-dimensional Contingency Tables: LASSO and Stepwise Algorithms 563
Nur Huda Nabihan Md Shahri and Susana Conde

71 Microwave Tomography: A Numerical Study of Solving Linear Equations in the Non-linear Inverse Scattering Problem 571
Latifah Mohamed, Nur Adyani Mohd Affendi, Azuwa Ali and Nooraihan Abdullah

72 Multimedia Learning Tools for Autism Children 579
Jasmin Ilyani Ahmad, Suhailah Mohd Yusof and Noor Hasnita Abdul Talib

73 Relapse Cases Among Drug Addicts Using Logistic Regression Modeling 585
Siti Fairus Mokhtar, Fazillah Bosli, Norashikin Nasarudin and Fathiyah Ahmad@Ahmad Jali

74 The Influence of Fixed Rhythm Auditory Icon on Food Intake Mimicry 591
Suzilah Ismail, Norhayati Yusof and Hanif Baharin

Part I

Computing

Chapter 1

Determining Influential Household Routines for Domestic Water Consumption Estimation via Genetic Algorithm



Nurul Nadia Hani, Khairul Anwar Rasmani, Noor Elaiza Abd Khalid and Ahmad Firdaus Ahmad Fadzil

Abstract Domestic water consumption can be affected by many factors. Household routines that involve the use of water appliances such as number of time the occupants of a household took bath, flushing toilet, washing clothes and others ultimately regulate the amount of residential household's monthly water consumption. Accurately estimating the amount of domestic water consumption is a very challenging task as these household routines differs from one another with one routine may be more influential than the others. This paper therefore proposes the employment of Genetic Algorithm (GA) in order to optimize the coefficient of micro-components of water consumption (coMC) to determine which micro-component of water consumption (household routines) is more influential than the others. This is accomplished by encoding the chromosome data in GA to incorporate the coMC values to minimize the domestic water consumption estimation error and subsequently enabling increased accuracy towards estimating the amount

N. N. Hani (✉) · N. E. A. Khalid
FSKM, Universiti Teknologi MARA, Kampus Shah Alam, Shah Alam, Selangor, Malaysia
e-mail: nurulnadiahani@gmail.com

N. E. A. Khalid
e-mail: elaiza@tmsk.uitm.edu.my

K. A. Rasmani
FSKM, Universiti Teknologi MARA, Kampus Seremban, Seremban, Negeri Sembilan, Malaysia
e-mail: khairulanwar@ns.uitm.edu.my

A. F. A. Fadzil
FSKM, Universiti Teknologi MARA, Kampus Jasin, Jasin, Melaka, Malaysia
e-mail: firdausfadzil@melaka.uitm.edu.my

of monthly water consumption. Using household's characteristics data and monthly water consumption from 80 residential households, it is discovered that there exist micro-components that are more influential towards the water consumption than the others.

Keywords GA • Domestic water consumption • Household routines

1 Introduction

Water consumption signifies one of the most important characteristic that indirectly affect a country's growth. Water consumption can be divided into domestic, agricultural and industrial. Domestic water consumption is a significant contributor towards a nation's overall water consumption which consists of indoor consumption, outdoor consumption and leakage [1]. Factors that contributes to domestic water consumption varied in terms of behavior and attitude, climate and seasonal changes, socio-economics characteristics and socio-demographics characteristics [1–4]. Most research focus only on these factors whereas water consumption are influence by household routines involving water-using appliances transpired from residential household. These includes bath, flushing toilet, washing clothes and other common residential household routines. Each routine and household vary quantitatively in water consumption. Consequently, this presents a challenge on deciding which household routines has high influence in estimating water consumption.

Genetic algorithm (GA) is an optimization technique to solve various types of problem [5, 6]. The evolutionary approach of GA allows candidate solutions to continually evolve to achieve the problem's final objective. This research proposes to determine which household routines highly influence the monthly domestic water consumption by employing GA. Selected household routines will be paired with random coefficient of micro-components of water consumption (coMC) values to find influential household routines for monthly domestic water consumption estimation.

2 Methodology

2.1 Data Collection

Data are collected by using questionnaire survey randomly selected residential households in Seremban, Negeri Sembilan, Malaysia. Household characteristic's data such as house type, house age, household size, number of adults and children, frequency of household routines involving water-using appliances and monthly

Table 1 Pre-determined water consumption value for each household routines

Household routines	Identifiers	Pre-determined water consumption (in Liters)
Shower/bath	P ₁	180
Brush teeth/wash hand/wash face	P ₂	10
Flush toilets	P ₃	15
Wash clothes (by hand)	P ₄	25
Wash clothes (by washing machine)	P ₅	120
Cooking	P ₆	12
Water plants	P ₇	80
Wash cars	P ₈	400
Others	P ₉	50

water consumption data (January 2014–December 2014) were obtained from 80 residential households with a total of 367 occupants.

The household routines involving water-using appliances were divided into two groups: daily (P₁ – P₃) and weekly routines (P₄ – P₉) as depicted in Table 1. The pre-determined water consumption value used in this study is collected from various sources¹ across different countries and has been tested against the collected monthly water consumption as shown in Table 1.

2.2 Genetic Algorithm

The algorithm comprises of seven different steps in order to achieve its objective. The steps are deliberated in this whole section below;

- Step 1: Chromosome is encoded based on Fig. 1.
- Step 2: Each chromosome is evaluated based on the fitness function (Eq. 4).
- Step 3: All chromosome is sorted based on its error value.
- Step 4: Fifty percent of the population is selected for crossover (single-point crossover), where a combination of the coMC values are inherited from Parent 1 and Parent 2 by a new child chromosome.
- Step 5: During mutation, new coMC values is randomly generated to produce more variation to the population.
- Step 6: All chromosome is ranked from best to worst. Only fifty percent of the fitter chromosome will be used for next generation.
- Step 7: The processes are repeated until the error value reaches convergence.

In chromosome population, each chromosome is encoded based on house ID, household size, micro-components of water consumption (household routines),

¹<https://water.usgs.gov/edu/qa-home-percapita.html>; <http://www.dwrcymru.com>, etc.

House ID	Household Size (h)	p_1	c_1	...	p_n	c_n	Average Water Consumption ($AVGWC$)	Estimated Water Consumption (EWC)	Fitness (f)
----------	------------------------	-------	-------	-----	-------	-------	---------------------------------------	---------------------------------------	-----------------

Fig. 1 Chromosome encoding

coMC values, average of monthly water consumption, estimated water consumption and fitness (total error). The population size used in this algorithm is 80.

The micro-components of water consumption selected are identified as P_1 to P_n , while the coMC values is denoted as C_1 to C_n . The values of coMC are in the form of randomized floating-point numbers, ranging from 0 to 1, where 0 represent lowest influence whereas 1 as highest influence on the water consumption estimation. The average monthly water consumption in the chromosome denotes the real water consumption value for the residential households. This value will be compared with the estimated water consumption value that is produced by using the coMC values generated and the pre-determined water consumption value (Table 1).

In the chromosome selection stage, each chromosome is evaluated by calculating the monthly water consumption by employing the estimated daily water consumption (EWC_d), estimated weekly water consumption (EWC_w) and estimated total water consumption (EWC_t) which are defined in Eqs. 1–4:

$$EWC_d = (h p_n a_n)(30)(c_n) \quad (1)$$

$$EWC_w = (p_n a_n) \left(\frac{30}{7} \right) (c_n) \quad (2)$$

$$EWC_t = EWC_d + EWC_w \quad (3)$$

where h is household size, p_n is micro-component of water consumption, a_n is pre-determined water consumption value, and c_n is coMC values. The fitness or error value is then established by using:

$$f = EWC_t - AVGWC \quad (4)$$

where EWC_t is estimated monthly water consumption and $AVGWC$ is the average of residential household's monthly water consumption. The value f is considered fitter if the value is lower as the algorithm's objective function is to find low volume of error for each chromosome.

3 Results and Findings

Once the algorithm reaches convergence, coMC values for each micro-component of water consumption is determined by analyzing which coMC values has the highest frequency when executing the algorithm for 100 times. Each coMC values is evaluated to obtain the lowest error values. The finalized coMC values used in this research compared with the R^2 values illustrated as follows:

Table 2 shows the correlation or coefficient of determination (R^2) between each micro-component of water consumption and the residential households monthly water consumption are very low in dependencies. The highest correlation is P_7 : 0.0383; followed by P_3 : 0.0348 and P_4 : 0.0302. It is not sufficient to build statistical model using this dataset [7]. However, coMC show a reliable prediction models.

The coMC values acquired allows the establishment of a set of values that represents how each micro-components of domestic water consumption (household routines) that influence the overall water consumption. Based on Table 2, it has been determined that micro-components of water consumption P_5 : 0.7550; P_6 : 0.6250; P_3 : 0.5600; P_1 : 0.3650 and P_2 : 0.2550 has significant influence for estimating the residential household's monthly water consumption. On the other hand, P_4 : 0.0600; P_7 : 0.0650; P_8 : 0.0600; P_9 : 0.0400 only depicts marginal influence towards estimating the water consumption. The results also suggest that GA has identified the three most influential micro-components of water consumption to be P_5 : washing clothes by washing machine (27.11%); P_6 : cooking (22.44%); and P_3 : flushing toilets (20.11%).

The coMC values established is then utilize in calculation to estimate the monthly water consumption value for 80 different residential households. This estimation value is then compared with two estimation method; pre-determined water consumption value (directly estimates the water consumption based on Table 1 in earlier section), and by employing the R^2 value in Table 2 as c_n in the

Table 2 Coefficient of determination (R^2) and coMC values for each household routines

Household routines	Identifiers	R^2	R^2 (%)	coMC	coMC (%)
Shower/bath	P_1	0.0084	5.83	0.3650	13.11
Brush teeth/wash hand/wash face	P_2	0.0004	0.28	0.2550	9.16
Flush toilets	P_3	0.0348	24.15	0.5600	20.11
Wash clothes (by hand)	P_4	0.0302	20.96	0.0600	2.15
Wash clothes (by washing machine)	P_5	0.0103	7.15	0.7550	27.11
Cooking	P_6	0.0024	1.67	0.6250	22.44
Water plants	P_7	0.0383	26.58	0.0650	2.33
Wash cars	P_8	0.0038	2.64	0.0600	2.15
Others	P_9	0.0155	10.76	0.0400	1.44

Table 3 Error rates of estimating 80 residential households monthly water consumption

	Pre-determined water consumption value	R ²	coMC
Total error (m ³)	3933.6249	2381.0883	785.2679

Eqs. (1) and (2). Table 3 illustrates the total error of estimating 80 residential households monthly water consumption.

It is indicated that by using coMC values shows significantly lower error estimation rates when being compared with the other methods as shown in Table 3. This demonstrates that the algorithm implemented enables more accurate estimation towards domestic water consumption estimation.

4 Conclusion

This research proposes the employment of GA in order to optimize the coMC values that provides valuable insights on which micro-component of water consumption that affect residential household's monthly water consumption. The findings of this study appear that washing clothes by washing machine, cooking and flushing toilets is the most influential micro-component of water consumption. It has been shown that GA is able to produce precise monthly water consumption estimation when being compared to other estimation methods in this study. Nevertheless, the algorithm can be further refined to incorporate more data in order to produce more concrete coMC values.

Acknowledgements The author acknowledges with gratitude to the Ministry of Higher Education (MOHE) under the Fundamental Research Grant Scheme (FRGS) grant Fuzzy Sets Approach to Per Capita Domestic Water Consumption Estimation with reference number 600-RMI/FRGS TD 5/3 (1/2015).

References

1. Makki, A.A., Stewart, R.A., Beal, C.D., Panuwatwanich, K.: Novel bottom-up urban water demand forecasting model: revealing the determinants, drivers and predictors of residential indoor end-use consumption. *Resour. Conserv. Recycl.* **95**, 15–37 (2015)
2. Altunkaynak, A., Nigussie, T.A.: Monthly water consumption prediction using season algorithm and wavelet transform-based models. *J. Water Resour. Plan. Manag.* **143**(6), 04017011 (2017)
3. Chen, X., Yang, S.H., Yang, L., Chen, X.: A benchmarking model for household water consumption based on adaptive logic networks. *Procedia Eng.* **119**, 1391–1398 (2015)
4. Fielding, K.S., Russell, S., Spinks, A., Mankad, A.: Determinants of household water conservation: the role of demographic, infrastructure, behavior, and psychosocial variables. *Water Resour. Res.* **48**(10) (2012)

5. Li, X., Parrott, L.: An improved Genetic Algorithm for spatial optimization of multi-objective and multi-site land use allocation. *Comput. Environ. Urban Syst.* **59**, 184–194 (2016)
6. Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., Wei, Y.: A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Math. Comput. Model.* **58**(3), 458–465 (2013)
7. Mohd Hanif, H., Rasmani, K. A., Mohamed Ramli, N.: Challenges in determining attributes to generate models for estimation of residential water consumption based on consumer data. In: *AIP Conference Proceedings AIP*, vol. 1522, no. 1, pp. 1306–1311 (2013)

Chapter 2

Sport Suitability Prediction Based on Physical Fitness Components Using k-Nearest Neighbors Algorithm



Muhammad Nabil Fikri Jamaluddin, Mohd Syafiq Miswan,
Shukor Sanim Mohd Fauzi, Ray Adderley JM Gining,
Noor Fadlyana Raman and Mohd Zaid Mohd Ghazali

Abstract Various type of sport requires different levels of physical fitness capability to achieve optimal performance. Physical fitness components such as endurance and physical characteristics shown to have great influence in performance of athlete. Data collected from Physical Fitness Test (PFT) by trainers are usually for record keeping and monitoring, it also consists rich of data attributes of athletes and sports they play. However, the relationship between these components and type of sports are poorly understood. Analysis such as cross tabulation for under-standing the relationships have not been explored. In this project, 16 attributes from PFT have been recognized to contribute to the type of sport they play. These data are used to predict suitable sports type for athletes based on their physical fitness score. The development begins with data preparations, digitization and cleaning and three data sets are prepared for this purpose. Data sets divided into male, female and combination male and female athletes, the results shown male athlete data sets outperform others with 81.3% accuracy. Overall, the physical fitness components influence to the type of sport athletes play.

Keywords Sport suitability prediction · Physical fitness · k-Nearest neighbors · Athlete development · Physical fitness components

M. N. F. Jamaluddin (✉) · S. S. M. Fauzi · R. A. J. Gining · N. F. Raman
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Perlis Branch, 02600 Arau, Perlis, Malaysia
e-mail: nabilfikri@perlis.uitm.edu.my

M. S. Miswan
Faculty of Sports Science and Recreation, Universiti Teknologi MARA,
Perlis Branch, 02600 Arau, Perlis, Malaysia

M. Z. M. Ghazali
Professional Education, Research and Education Division, National Sports Institute
of Malaysia, 57000 Bukit Jalil, Kuala Lumpur, Malaysia
e-mail: mzaid@isn.gov.my

1 Introduction

It is undoubtedly various factors could lead to the success of athlete in sports. Various type of sport requires different levels of physical fitness capability to achieve optimal performance. For example, marathon athlete needs endurance while sprinter in athletic requires speed [1]. Sukan Prestasi Tinggi Sekolah (SPTS) is a program implemented in three centers which includes Pusat Latihan Daerah (PLD), Sekolah Sukan Negeri (SSN) and Sekolah Sukan Malaysia (SSM). As a routine, there are trainers who perform data collection from tests on athletes known as Physical Fitness Test (PFT).

The data collected from physical fitness assessments are usually for record keeping and monitoring, it consists rich of data attributes of athletes and sports they play which is beneficial for understanding athletes, future decision making and determining the athletes' future progression. However, relationship between these components and type of sports are poorly understood and have not been explored.

Based on literature, massive papers produced for predicting results of sporting events [2] and framework also developed [3]. Efforts has been put in predicting football matches result which help strategize winning preparation [4]. Prediction of match scores in premier league of American football based on the evaluation of criteria including Kelly betting strategy is done by [5]. Prediction of result from college football games using data mining technique by analyzing past game results and statistics [6]. The prediction in sporting events seems to be benefited to the betting market, but not relatively contribute to the athletes' performance.

This paper presents development of prediction prototype using k-Nearest Neighbors algorithm to determine suitable sport for athlete based on attributes from physical fitness components. Discussions begins with PFT, K-Nearest Neighbors Algorithm, design and implementation, results and discussions and finally conclusion.

2 Physical Fitness Test

Physical Fitness components can be identified as well-being (health) and aptitudes (skills). This can be depicted in Table 1. According to Caspersen et al. [7], in health-related fitness, muscular endurance, fat, bone, and body parts are essentials to determine health and weight. Components from the health-related fitness must be accompanied by skill-related fitness to produce athletes with best ability and performance.

Several studies show positive correlation in multi-dimensional assessment including physical fitness component for predicting future sport talent. Reference [8] have studied about multi-dimensional assessment, which consisted of physical, technical and perceptual-cognitive performance tests between two groups classified as talent identified and non-talent identified. Louzada et al. [9] also proposed

Table 1 Physical fitness components [7]

Health-related fitness	Skill-related fitness
Muscular endurance	Agility
Body composition	Balance
Flexibility	Coordination
Cardiorespiratory endurance	Power
Muscular strength	Speed
	Reaction time

web-based expert system for soccer sport talent identification based on multivariate statistical analysis. Multi-dimensional assessments also been carried out for different type of sports such as rugby [10], American football [8] and soccer [9].

3 K-Nearest Neighbors Algorithm

K-Nearest Neighbors (KNN) is one of the well-known algorithm for prediction which is categorized as supervised learning algorithm. It is one of famous and least difficult calculation and works well in practice [11]. KNN is a non-parametric calculation which means it does not make any presumptions on the distributed data. These algorithms utilizes a theory of distance between training data and new element to perform classification [12]. Thus, the new element that represents the closest among classes are categorized to the class. Hence, the selection of parameters that contributes to specific classes are important [11].

Distance in KNN algorithm represents value of length between two points. Several formula for calculating distance are available which includes Euclidean, Cityblock (Manhattan), Cosine and Correlation [11]. To evaluate the performance of KNN, sample data is separated into two, which are train data set and test data set. The train data set is used to calculate the distance between each point of test data set. The distance is then sorted in increasing order, first K neighbors are selected from the list. Then, class of test data can be determined by counting majority class and assign test data to the class.

4 Design and Implementation

In this section, development of proposed prototype is discussed. The development of prototype begins with data preparations and experimental configurations. The prototype is developed using JAVA programming language and Eclipse as an Integrated Development Environment (IDE).

4.1 Data Preparation

Raw data are obtained from Sukan Prestasi Tinggi Sekolah (SPTS) program which collected by trainers from selected schools of each state in Malaysia. The raw data stored in Microsoft Excel, consists of multiple sheets. They are divided into male and female athletes, then classified into four types of sport they play which are athletics, ball games, racquet sport, and target sport. Each record consists of seven groups of attributes which are personal information, anthropometric, agility, flexibility, speed, strength and endurance which total up to 33 attributes. Sample data are depicted in the Fig. 1.

In data preparation stage, we begin by removing record with missing values which caused by athlete’s injury. Availability of attributes is not consistent as some of sport type did not conduct certain tests. The data preparation also involves determining intersection and removing symmetric difference of attributes (column) among sport types. The agility test and personal information are excluded except ‘age’. Two types of sport are excluded due to small number of learning instances.

Data preprocessing results into two classes of sport type with 16 attributes considered for prediction consisting of components such as age, anthropometric, speed, strength, power and endurance. Each attribute is normalized to ensure proper data scaling for distance calculations. Three data sets are produced. It encompasses records of male athletes, female athletes and combination of male and female athletes which are separated into training data (60%) and testing data (40%) and can be presented in Table 2. The combination data set consists of 17 attributes that add ‘gender’ attribute.

SPORT	POSITION	CODE	ANTHROPOMETRIC					SPEED			STRENGTH & POWER			STRENGTH		ENDURANCE		
			WEIGHT (kg)	HEIGHT (cm)	SITTING HEIGHT (cm)	ARMSPAN (cm)	BMI	40 METER SPRINT (sec)			VERTICAL JUMP (cm)	STANDING BROAD JUMP (cm)	MEDICINE BALL THROW (m)	MAXIMUM PUSH UP	1 MINUTE SIT UP	YO-YO TEST		
								10 meter	20 meter	40 meter					LEVEL	SHUTTLE	PREDICTED VO2 MAX	
OLAHRAGA	LOMPAT TINGGI	3	42.30	163.20	80.90	168.10	15.88	1.88	3.25	5.94	49.00	210.00	6.76	25.00	25.00	11.00	8.00	49.90
OLAHRAGA	LOMPAT TINGGI	3	57.40	178.00	86.50	181.40	18.12	1.77	3.00	5.43	63.00	220.00	9.46	15.00	32.00	9.00	7.00	42.70
OLAHRAGA	PELUT	1	58.50	167.00	83.50	172.50	20.98	1.89	3.23	5.80	63.00	192.00	8.45	43.00	34.00	10.00	1.00	44.20
OLAHRAGA	JALAN KAKI	2	34.75	155.00	75.00	157.50	14.46	2.02	3.45	6.43	44.00	195.00	6.08	8.00	18.00	9.00	5.00	42.00
OLAHRAGA	THROWER	4	76.40	175.30	90.70	180.40	24.86	1.92	3.31	5.95	56.00	190.00	10.60	20.00	12.00	6.00	3.00	31.00
OLAHRAGA	LOMPAT JAUH	3	42.00	156.50	77.20	164.30	17.15	1.86	3.22	5.85	55.00	220.00	7.65	28.00	17.00	8.00	9.00	40.20
OLAHRAGA	BOOM	1	52.40	165.30	85.50	166.50	19.18	1.29	3.10	5.56	58.00	248.00	8.29	39.00	25.00	10.00	11.00	47.40
OLAHRAGA	400M BOOM	1	60.30	168.50	89.00	172.40	21.24	1.71	2.96	5.25	63.00	255.00	11.05	45.00	30.00	12.00	2.00	51.40
OLAHRAGA	400M BOOM	3	62.70	174.20	88.00	185.60	20.66	1.79	3.07	5.50	71.00	280.00	11.07	37.00	9.00	10.00	1.00	44.20
OLAHRAGA	SPRINTER	1	50.80	160.90	77.50	170.00	19.62	1.78	3.11	5.53	49.00	189.00	28.00	29.00	11.00	9.00	49.90	
OLAHRAGA	SPRINTER	1	50.50	172.00	85.00	177.90	17.07	1.85	3.17	5.66	48.00	216.00	26.00	22.00	11.00	2.00	48.20	
OLAHRAGA	SPRINTER	1	54.40	165.40	80.20	173.70	19.89	1.90	3.24	5.79	39.00	195.00	7.75	22.00	31.00	5.00	3.00	27.60
OLAHRAGA	SPRINTER	1	57.30	169.50	82.00	178.50	19.94	1.82	3.15	5.54	46.00	226.00	20.00	20.00	10.00	7.00	46.00	
OLAHRAGA	THROWER	4	59.90	168.60	80.80	180.10	21.07	1.78	3.18	6.23	46.00	201.00	8.38	30.00	10.00	10.00	46.50	
OLAHRAGA	LONG DISTANCE	2	46.80	155.20	81.50	161.20	19.43	1.60	2.95	5.02	56.00	224.00	8.65	28.00	40.00	12.00	6.00	52.60
OLAHRAGA	LONG DISTANCE	2	30.00	141.80	69.80	141.80	14.92	1.73	3.28	6.45	49.00	180.00	5.04	15.00	26.00	11.00	5.00	48.90
OLAHRAGA	SPRINTER	1	46.95	160.30	82.80	157.70	18.27	1.99	2.70	5.31	47.00	171.00	8.24	11.00	18.00			
OLAHRAGA	JUMPER	3	48.85	170.90	85.10	175.50	16.73	1.47	2.77	5.26	47.00	234.00	8.16	18.00	18.00			
OLAHRAGA	SPRINTER	1	52.60	167.00	82.50	165.50	18.86	1.51	2.81	5.37	51.00	172.00	8.02	10.00	24.00			
OLAHRAGA	JUMPER	3	33.95	142.90	71.40	139.00	16.63	1.87	3.45	6.59	44.00	182.00	5.51	21.00	25.00			
OLAHRAGA	SPRINTER	1	45.94	164.60	76.60	163.00	18.09	1.61	2.84	6.14	40.00	191.00	6.92	11.00	24.00			

Fig. 1 Sample raw data

Table 2 Number of record for each data sets prepared

Data set	Male	Female	Combination
Training	376	301	677
Testing	252	203	455
Total	628	504	1132

4.2 Experimental Configurations

Prediction engine is developed based on k-NN algorithm to receive 16 (male and female data sets) and 17 (combination data set) attributes as input and produce one output (type of sport) as a result of classification from two classes which are ‘athletics’ and ‘ball games’. Euclidean distance calculation is used to find distance between instances of test data and for each of train data. Results of distance calculation are stored and sorted to determine the nearest neighbors based on K-th minimum distance from the test instance. The predicted value is compared to determine the accuracy. To find the value of K with maximum prediction accuracy, simulation feature is added to prototype which perform operations as discussed above repeatedly for three different data sets from K = 1 to K = 50.

5 Result and Discussion

Results obtained from the simulations can be depicted in Fig. 2. There are three graphs in the chart. Blue, orange and grey graphs present results from male, female and combination data sets respectively and prediction accuracy obtained from K = 1 to K = 50. The horizontal axis presents value of K and the vertical axis show accuracy percentage. It can clearly be seen, male data set scores highest. The male data set seems in increasing trend until K = 18 then stabilize. Female data set shows

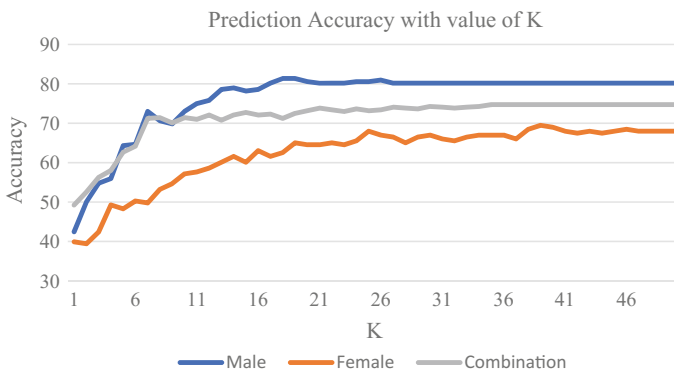


Fig. 2 Prediction results

Table 3 Results for different data sets

Data set	Value of K	Accuracy percentage (%)
Male	18	81.3
Female	39	69.5
Combination	35	74.7

an upward trend with increase and decrease intermittently and achieve its best accuracy when $K = 39$. Combination data set achieve its best accuracy with $K = 35$.

The summary of results can be presented in Table 3. Value of K represent K value used when the first occurrence of best prediction accuracy is achieved. From the table, best accuracy obtained is from male data set. Combination data set outperform female data set with 74.7% and 69.5% accuracy respectively. The best prediction accuracy from male data sets might be due to the amount of data involved and composition of data itself. The combination data set which adds one attributes seems to achieve better accuracy than female data set. This might be results from composition of male and female records in the data set.

6 Conclusion

In this paper, prediction of suitable sport type for athlete using k-NN based on PFT components is presented. Based on three data sets prepared, male data set scores highest with 81.3% accuracy. The physical fitness components do influence the type of sport athletes play. Low accuracy on female data set is due to the volume and the availability of learning instances in the data itself. The combination data sets also did not yield better results as the data is composed of male and female records. The results also suggest for separating records according to gender, because of score in physical fitness components are different between male and female athletes. Missing data for agility component also contributes to the data inconsistency and lead to reduced prediction accuracy.

References

1. Fister, I., Ljubič, K., Suganthan, P.N., Perc, M., Fister, I.: Computational intelligence in sports: Challenges and opportunities within a new research domain. *Appl. Math. Comput.* **262**, 178–186 (2015)
2. Stekler, H.O., Sendor, D., Verlander, R.: Issues in sports forecasting. *Int. J. Forecast.* **26**(3), 606–621 (2010)
3. Bunker, R.P., Thabtah, F.: A machine learning framework for sport result prediction. *Appl. Comput. Inf.* (2017)

4. Prasetio, D., Harlili, D.: Predicting football match results with logistic regression. In: 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, pp. 1–5 (2016)
5. Baker, R.D., McHale, I.G.: Forecasting exact scores in National Football League games. *Int. J. Forecast.* **29**(1), 122–130 (2013)
6. Leung, C.K., Joseph, K.W.: Sports data mining: predicting results for the college football games. *Procedia Comput. Sci.* **35**, 710–719 (2014)
7. Caspersen, C.J., Powell, K.E., Christenson, G.M.: Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Rep.* **100**(2), 126–131 (1985)
8. Woods, C.T., Raynor, A.J., Bruce, L., McDonald, Z., Robertson, S.: The application of a multi-dimensional assessment approach to talent identification in Australian football. *J. Sports Sci.* **34**(14), 1340–1345 (2016)
9. Louzada, F., Maiorano, A.C., Ara, A.: ISports: a web-oriented expert system for talent identification in soccer. *Expert Syst. Appl.* (2016)
10. Till, K., Cobley, S., Morley, D., O’Hara, J., Chapman, C., Cooke, C.: The influence of age, playing position, anthropometry and fitness on career attainment outcomes in rugby league. *J. Sport. Sci.* 1–6 (2015)
11. Medjahed, S.A., Saadi, T.A., Benyettou, A.: Urinary system diseases diagnosis using machine learning techniques. *Int. J. Intell. Syst. Appl. Intell. Syst. Appl.* **7**(5) 1 (2015)
12. Giri, A., Bhagavath, M.V.V., Pruthvi, B., Dubey, N.: A placement prediction system using k-nearest neighbors classifier. In: Second International Conference on Cognitive Computing and Information Processing CCIP 2016, pp. 3–6 (2016)

Chapter 3

Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE



Nurulfitrah Noorhalim, Aida Ali and Siti Mariyam Shamsuddin

Abstract There are many issues regarding datasets classification. One such issue is class imbalance classification, which often occurs with extreme skewness across many real-world domains. The issue presents itself as one of the fundamental difficulties to form robust classifiers. In this paper, a sampling method was used to identify the performance of classification for k-NN classifier and C4.5 classifier with a ten-fold cross validation. Experimental results conducted showed that sampling greatly benefited the performance of classification in class imbalance problem, by improving class boundary region especially with extremely imbalanced datasets (extreme number of imbalanced ratio). This result demonstrates that class imbalance can affect many domains in real-world applications.

Keywords Sampling · Synthetic minority over-sampling · Imbalanced dataset

1 Introduction

The use of today's technology has contributed significantly to the revenue of millions of data within seconds in scientific research, business, and industry. This has created a huge opportunity, especially for researchers to analyse and interpret data into a form that can give implication to an organisation's profitability, especially in consumerism-specific domain, attributed to the growth in science and technology field. This phenomenon has grown drastically and gained substantial popularity, largely owed to heavy interest in artificial intelligence behaviour research, particularly in machine learning, which can be very useful in administrating customer

N. Noorhalim (✉) · A. Ali · S. M. Shamsuddin
Faculty of Computing, Universiti Teknologi Malaysia, UTM, 81310 Skudai, Johor, Malaysia
e-mail: nurulfitrah2@live.utm.my

A. Ali
e-mail: aida@utm.my

S. M. Shamsuddin
e-mail: mariyam@utm.my

relationship management (CRM). For instance, collected business data can be used and classified in numerous ways including, in the form of descriptive analytics, diagnostic analytics, prescriptive analytics, and predictive analytics.

However, there exist several issues surrounding datasets collection classification, such as imbalanced class classification that often faces extreme skewness, which occurs across numerous real-world domains, contributing to be one of the fundamental difficulties to form a robust classifier [1–4]. Class imbalance refers to a situation or condition where the number of minority class instances (positive class) is far lesser or smaller than the number of majority class instances (negative class) or not adequately represented.

According to Chawla et al. [4], class imbalance appeared attributed to the evolution of science development into applied technology in machine learning. From then on, various classifier learning algorithms were developed, which assume that datasets possess a comparatively balanced distribution, neglecting a serious complexity in imbalanced class distribution [5, 6]. Such imbalanced, real-world classification problems can be found throughout various domains such as, diagnostic problems [7, 8], software quality prediction [9], software defect prediction [10], and activity recognition [11]. Against a backdrop of the notion that various methodologies extraction produces a significant knowledge to many general practitioners, as indicated by Bach et al. [12], focusing on class imbalance topic is, thus, crucial.

There exist substantial cases where imbalanced datasets require two-class classification in learning classifiers, when the natural distribution of the datasets is dominated by one class or when a positive majority class exceeds other class (majority or negative class). For instance, a dataset of breast tumour patients exhibits a dual-class classification problem for breast cancer diagnoses, comprising benign and malignant cancer. In medical cases, there are lesser patients diagnosed with malignant cancer (minority class) than patients diagnosed with benign cancer (majority class). Lack of instances (minority class) could lead to difficulties in obtaining a precise classification.

The skewness of data distribution has inspired the formation of more robust classifiers, which is a basic problem in data mining [6, 13]. In addition, standard classifiers can be biased in evaluation measurements, which typically produce bad performance for minority class, and vice versa for majority class. This is because the intention of classifiers is to identify noise in positive samples and ignore them during learning process [14, 15]. Traditional class imbalance classification accuracy also degrades, due to the presence of class imbalance difficulties including, (i) borderline samples, (ii) small disjuncts, (iii) noisy data, (iv) small number of training data, (v) overlapping instances, and (vi) differences of distribution on test and training data (or dataset shift).

There are two methods to handle class imbalance problem including, data-level and algorithm-level methods. Typically, data-level will be run as pre-processing steps to make sure unbalanced datasets can be adjusted and skewness of the distributions can be reduced, by using any types of sampling methods [16–18]. According to Lee, Sheen [16], random sampling imposes some limitations to most data pre-processing. Major differences in pre-processing results could end up with severely biased data distribution. Meanwhile, algorithm-level involves more on

modifying algorithms such as, ensemble approach and cost-sensitive learning that can manage imbalanced class distribution [19, 20]. Ali et al. [1] claimed that, even though both approaches are able to change and balance data distribution to become normal, there remains theoretical gaps in existing approaches that can bear all the aforementioned problems.

1.1 *Imbalanced Ratio*

In class imbalanced classification, there are a few challenges that might be faced during the classification processes. This difficulties often occur when the datasets have imbalanced distribution between classes such as minority class or often known as positive class and majority class or familiar as negative class [21, 22], which also known as imbalance ratio. Imbalance ratio (IR) is a proportion samples in the number of majority class (negative class) to the number of minority class (positive class) [15, 23]. In binary class datasets, the problems appear based on the number of samples i.e. the number of samples for majority class has monopolized the distribution of minority class samples which also important as class of interest in classification. This condition probably assumed minority class has very low predictive accuracy [24]. However, the error that comes from the minority class ought be important as stated in many research works [21]. This scenario obviously shows that most of the datasets in classification has imbalanced problem with its classes. It has widely happened in various studies involving with the issue of lack of training sample set in data classification [13] which is also noted as another issue for class imbalance problem. Class imbalanced could be happened when the datasets has insufficient amount of samples that can lead to another problem in identifying the pattern regularities [1] in which the pattern itself could help to improve the decision boundary of the classes [25]. The best classification performance should have balanced distribution and sufficient number of samples to represent the training samples that can provide more knowledge that can be useful for learning processes. Otherwise, it can degrade the classification performance.

Besides, another issue is, where most of the standard learning algorithms thought that all datasets has same number of data class distributions [6]. This overlapping problem has a higher tendency to create another related issue that can affect the classifier performance greater than the problem of imbalanced distribution [1, 26–31].

2 **Data Collection**

Three datasets were used in this experiment, which are available on KEEL datasets repository [32]. The datasets were chosen because they possess an imbalanced ratio (IR) of positive and negative classes. These datasets have been used and reported in

previous studies such as [22–25] using SMOTE and other sampling techniques with different types of datasets such as, datasets with mildly imbalanced ratio between 1.5 and 9, and datasets that have extremely imbalanced ratio of more than nine. These datasets were divided into two, in order to identify the extent that SMOTE helps C4.5 and k-NN. Apart from that, different types of IR may also help in identifying whether SMOTE only helps in classifying extremely imbalanced datasets, or both mild and extreme datasets. This is due to different values of IR have different distribution pattern, attributed to implemented sampling method. The datasets used in this experiment referred to the imbalance ratio between 1.5 and 9. There are 12 datasets out of 22 datasets available which were randomly selected, as listed in Table 1. All the datasets are already set up into a sufficient number of minority samples using 5-folds stratified cross validation for the test partition.

Referring to Table 1, the datasets are arranged accordingly from lowly imbalanced to highly imbalanced datasets, based on ascending imbalanced ratio (IR) values with additional details including, number of instances (#Ex), number of attributes (#Atts) in real (R), integer (I), and nominal (N) valued, and number of classes (accounting both minority and majority classes).

Table 1 consists of 12 different types of fungi that have different number of IR, in which all are less than nine, and different number of instances. All datasets have two types of classes comprising minority class (positive class) and majority class (negative class). Most of the instances are real attributes. None of nominal attributes were utilized in this dataset.

3 Data Pre-processing

In order to make sure the classification process runs smoothly and efficiently, all data needed to be pre-processed by using Waikato Environment for Knowledge Analysis (WEKA) [33] before the experiment proceeded to classification processes.

Table 1 Summary description datasets with imbalance ratio between 1.5 and 9

Data-set	IR	#Ex.	#Atts. (R/I/N)	#Class
Glass1	1.82	214	9 (9/0/0)	2
Wisconsin	1.87	683	9 (0/1/0)	2
Glass0	2.06	214	9 (9/0/0)	2
Yeast1	2.46	1484	8 (8/0/0)	2
Glass-0-1-2-3_Vs_4-5-6	3.2	214	9 (9/0/0)	2
Vehicle0	3.25	846	18 (0/18/0)	2
New-Thyroid1	5.14	215	5 (4/1/0)	2
Ecoli2	5.46	336	7 (7/0/0)	2
Segment0	6.02	2308	19 (19/0/0)	2
Glass6	6.38	214	9 (9/0/0)	2
Ecoli3	8.6	336	7 (7/0/0)	2
Page-blocks0	8.79	5472	10 (4/6/0)	2

Normalization is used to synchronize the interval value, as first step, in the pre-processing for all three datasets. As stated by Kotsiantis et al. [34], this process can be essential for k-Nearest Neighbor (k-NN) algorithm.

4 Classification

Two classification algorithms were used in this experiment consisting C4.5 and k-NN. Every dataset was run with a ten-fold cross-validation, with a single repetition. Table 2 shows specification of parameters for the classification process.

In this experiment, SMOTE sampling utilized five nearest neighbours with its default percentage, 100%, for instances to be created. C4.5 was used with its default parameter of confidence factor = 0.25. Meanwhile, for k-NN, the authors used $KNN = 10$, instead of one, as its default. To make sure the readers understand this experiment, Table 3 is constructed to summarize the list of algorithms, according to two types of classifiers; sampling and non-sampling. The abbreviation for every method with its short description is included in Table 3.

The analysis of classification task for performance measure was based on confusion matrix of the results for each instance, as shown in Table 4, which was used to predict positive class.

Table 2 Specification of parameters

Parameters	
Experiment type	Cross-validation
Number of folds	10
Number of repetitions	1

Table 3 Algorithms used in experimental design

Non-sampling classifier		
<i>Abbr.</i>	<i>Method</i>	<i>Short description</i>
C45	C4.5	Class generating pruned C4.5 decision tree and previously normalized datasets
k-NN	K-Nearest Neighbour	Selects appropriate value of K based on cross-validation and performs distance weighting, pre-processed with normalization
Sampling classifier		
<i>Abbr.</i>	<i>Method</i>	<i>Short description</i>
CSMT	C4.5 + SMOTE	Applies C4.5 on datasets after pre-processing with normalization and SMOTE
KSMT	K-Nearest Neighbor + SMOTE	Selects appropriate value of K based on cross validation and performs distance weighting after pre-processing, with normalization and SMOTE

4.1 Synthetic Minority Over-Sampling

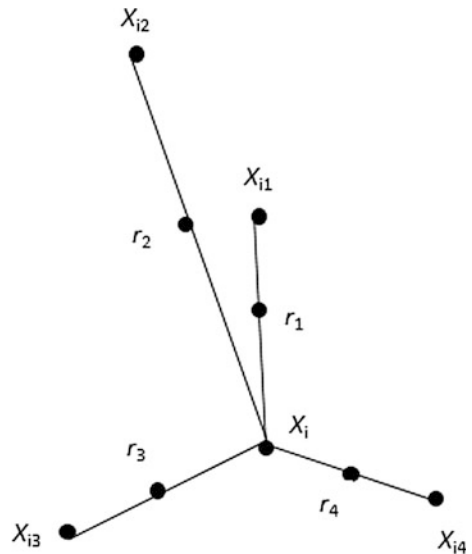
Synthetic minority over-sampling or widely known as SMOTE [35] is a most sophisticated technique in handling problems of over- and under-sampling techniques as mention previously which required minority samples to create its impostor synthetically and typically will be merged with an under-sampling approach from majority instances.

In this approach, the over-sampling process of the minority instances used selection approach and also using iterative search until it reach its amount needed for every observation [17]. The new samples which produced synthetically will used selected k nearest neighbours (k-NN) at random along the line segments that joining any/all of its neighbours [15] as illustrated in Fig. 1.

Figure 2.10 shows point X_{i1} to point X_{i4} as selected nearest neighbours in order to synthetically produced data point r_1 to r_4 at random.

This research using SMOTE sampling as pre-processing technique because of it can produce a better performance in practice which is excellent either for basic or hybrid methods [36].

Fig. 1 Illustration on how the synthetic data points created in SMOTE algorithm [15]



5 Results

The objective of the experimental study is to identify which datasets could give a balance, robust, and good trade-off, or conversely sacrifice performance; with sampling and without sampling by using decision trees with C4.5 [37] and the widely known k-Nearest Neighbor (k-NN) [38], as learning classifiers.

In this experiment, the authors investigated the performance of methods in terms of robustness and ability in handling imbalanced datasets with different IR. The authors also took into consideration the improvement to the results and justified it with respect to two algorithms used, to determine whether it can provide a better trade-off or otherwise, for both C4.5 and k-NN algorithms.

The authors utilized datasets with closely related imbalance ratio (IR) ranging between 1.5 and 9, across 12 different data. The datasets were divided into three categories comprising, mild imbalance with IR less than three, medium imbalance with IR ranging from three to six, and extreme imbalance with IR of more than six. WEKA [24] was used in this experiment to perform k-NN and C4.5 learning classifiers across all datasets. It is used due to its ability to set up large-scale experiments.

Based on the results, with close observation on the scores of performance measures, the authors found that imbalance ratio does not affect the findings for the datasets. The results for experiment using k-NN algorithm did yield an improvement, superior than C4.5 on datasets with IR between 1.5 and 9. Similar results' patterns were observed for both algorithms on sampling and non-sampling approaches as shown in Tables 4 and 5.

From the experiments, it is observed that SMOTE improved both C4.5 and k-NN, ranging from 0.45 to 0.18. It can be deduced that such improvement is achieved due to increasing number of training data points which formed better decision boundary representation between majority class and minority class.

Figures 2 and 3 illustrate the performance of classifiers for both C4.5 and k-NN algorithms; with and without sampling for *segment0* dataset. The classifiers correctly classified all data in sampling with SMOTE, with extremely imbalanced IR. Hence, it is safe to conclude that class imbalance problem does benefit from sampling approach, especially when dealing with extremely skewed datasets, as in this case with SMOTE. The results also indicated trade-off between recall and precision. The classifiers have completely learned the data.

Table 4 Average performance measure on ten-fold CV of C4.5 and K-NN in non-sampling approach for datasets with imbalance ratio between 1.5 and 9

		IR		Sensitivity		Specificity		Precision		Accuracy		G mean		F-measure	
		C4.5	k-NN	C4.5	k-NN	C4.5	k-NN	C4.5	k-NN	C4.5	k-NN	C4.5	k-NN	C4.5	k-NN
Non-sampling															
IR < 3															
<i>Glass1</i>	1.82	0.613	0.611	0.842	0.864	0.795	0.818	0.737	0.727	0.718	0.726	0.692	0.699		
<i>Wisconsin</i>	1.87	0.941	0.967	0.968	0.975	0.968	0.975	0.955	0.971	0.955	0.971	0.954	0.971		
<i>Glass0</i>	2.06	0.757	0.914	0.840	0.785	0.826	0.809	0.850	0.799	0.798	0.847	0.790	0.859		
<i>Yeast1</i>	2.46	0.464	0.490	0.869	0.857	0.780	0.774	0.667	0.673	0.635	0.648	0.582	0.600		
Mean		0.694	0.745	0.880	0.870	0.842	0.844	0.787	0.808	0.776	0.798	0.755	0.782		
Standard deviation		0.204	0.232	0.060	0.079	0.086	0.089	0.124	0.131	0.136	0.141	0.158	0.165		
3 > IR < 6															
<i>Glass-0-1-2-3_vs_4-5-6</i>	3.2	0.823	0.770	0.975	0.932	0.971	0.919	0.899	0.851	0.896	0.847	0.891	0.838		
<i>Vehicle0</i>	3.25	0.873	0.899	0.950	0.941	0.946	0.939	0.912	0.920	0.911	0.920	0.908	0.919		
<i>New-thyroid1</i>	5.14	0.942	0.800	0.989	1.000	0.988	1.000	0.965	0.900	0.965	0.894	0.964	0.889		
<i>Ecoli2</i>	5.46	0.750	0.907	0.972	0.968	0.964	0.966	0.861	0.953	0.854	0.937	0.844	0.936		
Mean		0.847	0.844	0.972	0.961	0.967	0.956	0.909	0.906	0.906	0.900	0.902	0.895		
Standard deviation		0.081	0.069	0.016	0.030	0.017	0.035	0.043	0.042	0.046	0.039	0.050	0.043		
IR > 3															
<i>Segment0</i>	6.02	0.957	0.982	0.997	0.996	0.997	0.996	0.977	0.989	0.977	0.989	0.977	0.989		
<i>Glass6</i>	6.38	0.800	0.767	0.978	0.968	0.974	0.959	0.889	0.867	0.885	0.861	0.878	0.852		
<i>Ecoli3</i>	8.6	0.567	0.717	0.957	0.964	0.930	0.952	0.762	0.840	0.736	0.831	0.704	0.818		
<i>Page-blocks0</i>	8.79	0.859	0.696	0.987	0.987	0.985	0.981	0.923	0.841	0.920	0.829	0.917	0.814		
Mean		0.796	0.790	0.980	0.979	0.971	0.972	0.888	0.884	0.880	0.877	0.869	0.868		
Standard deviation		0.166	0.131	0.017	0.016	0.029	0.020	0.091	0.071	0.103	0.076	0.117	0.082		
Mean		0.779	0.793	0.944	0.936	0.927	0.924	0.861	0.866	0.854	0.858	0.841	0.849		
Standard deviation		0.158	0.150	0.058	0.067	0.079	0.079	0.101	0.092	0.109	0.098	0.125	0.111		

Table 5 Average performance measure on ten-fold CV of C4.5 and K-NN in sampling approach for datasets with imbalance ratio between 1.5 and 9

Sampling	IR	Sensitivity		Specificity		Precision		Accuracy		G mean		F-measure		
		CSMT	KSMT	CSMT	KSMT	CSMT	KSMT	CSMT	KSMT	CSMT	KSMT	CSMT	KSMT	
IR < 3														
	<i>Glass1</i>	1.82	0.795	0.914	0.732	0.685	0.748	0.744	0.763	0.800	0.763	0.791	0.771	0.820
	<i>Wisconsin</i>	1.87	0.983	0.998	0.950	0.968	0.952	0.969	0.967	0.983	0.967	0.983	0.967	0.983
	<i>Glass0</i>	2.06	0.893	0.957	0.814	0.675	0.828	0.747	0.854	0.816	0.853	0.804	0.859	0.839
	<i>Yeast1</i>	2.46	0.711	0.853	0.779	0.680	0.763	0.727	0.745	0.766	0.744	0.761	0.736	0.785
	Mean		0.846	0.931	0.819	0.752	0.823	0.797	0.832	0.841	0.832	0.835	0.833	0.857
	Standard Deviation		0.118	0.062	0.094	0.144	0.093	0.115	0.101	0.097	0.102	0.100	0.103	0.087
3 > IR < 6														
	<i>Glass-0-1-2-3_vs_4-5-6</i>	3.2	0.914	0.952	0.928	0.927	0.927	0.929	0.921	0.939	0.921	0.939	0.920	0.940
	<i>Vehicle0</i>	3.25	0.925	0.985	0.946	0.904	0.945	0.911	0.935	0.945	0.935	0.944	0.935	0.947
	<i>New-thyroid1</i>	5.14	0.914	0.957	0.983	0.983	0.982	0.983	0.949	0.970	0.948	0.970	0.947	0.970
	<i>Ecoli2</i>	5.46	0.855	0.932	0.969	0.958	0.965	0.957	0.912	0.945	0.910	0.945	0.906	0.944
	Mean		0.902	0.956	0.956	0.943	0.954	0.945	0.929	0.950	0.928	0.949	0.927	0.950
	Standard Deviation		0.032	0.022	0.025	0.035	0.024	0.031	0.016	0.014	0.017	0.014	0.018	0.013
IR > 6														
	<i>Segment0</i>	6.02	0.995	0.986	0.998	0.991	0.998	0.991	0.997	0.989	0.997	0.989	0.997	0.989
	<i>Glass6</i>	6.38	0.930	0.863	0.957	0.962	0.956	0.958	0.943	0.913	0.943	0.911	0.943	0.908
	<i>Ecoli3</i>	8.6	0.800	0.871	0.943	0.934	0.934	0.929	0.872	0.902	0.869	0.902	0.862	0.899
	<i>Page-blocks0</i>	8.79	0.905	0.853	0.979	0.974	0.978	0.970	0.942	0.914	0.942	0.912	0.940	0.908
	Mean		0.969	0.965	0.966	0.962	0.939	0.929	0.938	0.928	0.935	0.926	0.969	0.965
	Standard deviation		0.024	0.024	0.028	0.026	0.051	0.040	0.053	0.040	0.056	0.042	0.024	0.024
	Mean		0.885	0.927	0.915	0.887	0.914	0.901	0.900	0.907	0.899	0.904	0.899	0.911
	Standard deviation		0.082	0.054	0.088	0.127		0.100	0.078	0.074	0.078	0.077	0.079	0.066

Fig. 2 ROC curve of sampling for *segment0* datasets on false positive rate (x-axis) and true positive rate (y-axis)

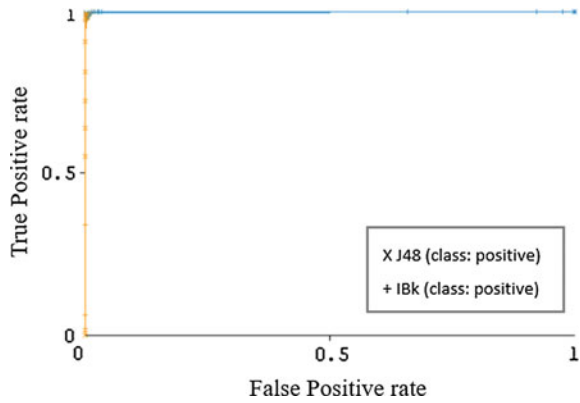
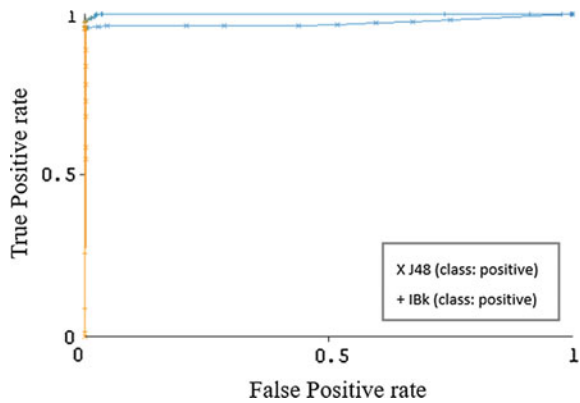


Fig. 3 ROC curve without sampling of *segment0* on false positive rate (x-axis) and true positive rate (y-axis)



6 Conclusion

This study focused on the effect of sampling on C4.5 and k-NN algorithms in handling imbalanced datasets. It was observed that SMOTE sampling could lead to improvements in classification. The findings of the experiment could help inform general practitioners of the main problem surrounding efforts to improve decision boundary of data region with extreme IR values, especially, concerning the skewness of data, which can be resolved through identifying factors that cause imbalance ratio to occur. The study also observed that extreme IR benefited from sampling, while other ranges of IR did not have much improvement on their classification rate.

Acknowledgements The authors would like to express appreciation to the UTM Big Data Centre of Universiti Teknologi Malaysia and Y.M. Said for their support in this study. The authors greatly acknowledge the Research Management Centre, UTM and Ministry of Higher Education for the financial support through Research University Grant (RUG) Vot. No. Q.JI30000.2528.13H30.

References

1. Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem: a review. *Int. J. Adv. Soft Comput. Appl.* **7**(3) (2015)
2. Beyan, C., Fisher, R.: Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recogn.* **48**(5), 1653–1672 (2015)
3. Cleofas-Sánchez, L., Sánchez, J.S., García, V., Valdovinos, R.: Associative learning on imbalanced environments: An empirical study. *Expert Syst. Appl.* **54**, 387–397 (2016)
4. Al-Stouhi, S., Reddy, C.K.: Transfer learning for class imbalance problems with inadequate data. *Knowl. Inf. Syst.* **48**(1), 201–228 (2016)
5. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **6**(1), 1–6 (2004)
6. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: a review. *Int. J. Pattern Recognit Artif Intell.* **23**(04), 687–719 (2009)
7. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009)
8. Bruha, I., Kočková, S.: A support for decision-making: cost-sensitive learning system. *Artif. Intell. Med.* **6**(1), 67–82 (1994). [https://doi.org/10.1016/0933-3657\(94\)90058-2](https://doi.org/10.1016/0933-3657(94)90058-2)
9. Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., Fettich, J.: Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artif. Intell. Med.* **16**(1), 25–50 (1999). [https://doi.org/10.1016/S0933-3657\(98\)00063-3](https://doi.org/10.1016/S0933-3657(98)00063-3)
10. Gao, K., Khoshgoftaar, T.M., Napolitano, A.: An empirical investigation of combining filter-based feature subset selection and data sampling for software defect prediction. *Int. J. Reliab. Qual. Saf. Eng.* **22**(6) (2015). <https://doi.org/10.1142/s0218539315500278>
11. Gao, K., Khoshgoftaar, T.M., Napolitano, A.: Aggregating data sampling with feature subset selection to address skewed software defect data. *Int. J. Soft. Eng. Knowl. Eng.* **25**(09n10), 1531–1550 (2015)
12. Abidine, M.B., Fergani, B., Ordóñez, F.J.: Effect of over-sampling versus under-sampling for SVM and LDA classifiers for activity recognition. *Int. J. Des. Nat. Ecodynamics* **11**(3), 306–316 (2016). <https://doi.org/10.2495/DNE-V11-N3-306-316>
13. Bach, M., Werner, A., Żywiec, J., Pluskiewicz, W.: The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Inf. Sci.* **384**, 174–190 (2017). <https://doi.org/10.1016/j.ins.2016.09.038>
14. Ando, S.: Classifying imbalanced data in distance-based feature space. *Knowl. Inf. Syst.* **46**(3), 707–730 (2016)
15. Lee, W., Jun, C.-H., Lee, J.-S.: Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification. *Inf. Sci.* **381**, 92–103 (2017). <https://doi.org/10.1016/j.ins.2016.11.014>
16. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **250**, 113–141 (2013)
17. Lee, C.S., Sheen, D.: Nonconforming generalized multiscale finite element methods. *J. Comput. Appl. Math.* **311**, 215–229 (2017)
18. Rivera, W.A., Xanthopoulos, P.: A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Syst. Appl.* **66**, 124–135 (2016)
19. Vluymans, S., Triguero, I., Cornelis, C., Saeys, Y.: EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data. *Neurocomputing* **216**, 596–610 (2016). <https://doi.org/10.1016/j.neucom.2016.08.026>
20. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **42**(4), 463–484 (2012). <https://doi.org/10.1109/TSMCC.2011.2161285>
21. Downs, R.: Beware the aliased signal! *Electron. Des.* **59**(4) (2011)

22. Visa, S.: Fuzzy classifiers for imbalanced data sets. University of Cincinnati (2006)
23. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: a review. *GESTS Int. Trans. Comput. Sci. Eng.* **30**(1), 25–36 (2006)
24. García, V., Sánchez, J.S., Mollineda, R.A.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowl.-Based Syst.* **25**(1), 13–21 (2012)
25. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *Int. J. Approx. Reason.* **50**(3), 561–577 (2009)
26. Phung, S.L., Bouzerdoum, A., Nguyen, G.H.: Learning pattern classification tasks with imbalanced data sets (2009)
27. Xiong, H., Wu, J., Liu, L.: Classification with class overlapping: a systematic study. In: *The 2010 International Conference on E-Business Intelligence*, pp. 491–497 (2010)
28. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **42**(4), 463–484 (2012)
29. Longadge, R., Dongre, S.: Class imbalance problem in data mining review (2013). arXiv preprint [arXiv:1305.1707](https://arxiv.org/abs/1305.1707)
30. Japkowicz, N.: Learning from imbalanced data sets: a comparison of various strategies. In: *AAAI Workshop on Learning from Imbalanced Data Sets*, pp. 10–15, Menlo Park, CA (2000)
31. Batista, G.E., Prati, R.C., Monard, M.C.: Balancing strategies and class overlapping. In: *International Symposium on Intelligent Data Analysis*, pp. 24–35. Springer, Heidelberg (2005)
32. Prati, R.C., Batista, G.E., Monard, M.C.: Learning with class skews and small disjuncts. In: *Brazilian Symposium on Artificial Intelligence*, pp. 296–306. Springer, Heidelberg (2004)
33. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Multiple-Valued Logic Soft Comput.* **17** (2011)
34. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009)
35. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Data preprocessing for supervised learning. *Int. J. Comput. Sci.* **1**(2), 111–117 (2006)
36. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
37. Fernández, A., García, S., del Jesus, M.J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets Syst.* **159**(18), 2378–2398 (2008)
38. Salzberg, S.L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **16**(3), 235–240 (1994). <https://doi.org/10.1007/bf00993309>

Chapter 4

Blood Vessels Segmentation of Retinal Fundus Image via Stack-Based Object-Oriented Region Growing



Ahmad Firdaus Ahmad Fadzil, Shafaf Ibrahim
and Noor Elaiza Abd Khalid

Abstract Retinal fundus image is an important medical imaging modality that provide different information on eye diseases. Eye disease such as glaucoma can be diagnosed by evaluating different features of retinal fundus images such as optic disc, macula, and blood vessels. Various image segmentation algorithm has been employed to segment blood vessels features of retinal fundus image as it is the hardest feature to segment compare to optic disc and macula. Region growing segmentation algorithm is one of the segmentation algorithm that is proven to be able to segment various type of medical imaging modalities. However, region growing segmentation algorithm has massive caveats especially in dealing with memory stacks, and restricted flexibility due to its recursive programming nature. In this paper, a segmentation algorithm that utilizes stack-based object-oriented region growing (SORG) algorithm is proposed. This algorithm proposes the combination of stack data structure and object-oriented programming paradigm in order to negate the utilization of recursive programming. The result shows that SORG able to provide decent segmentation when assessed in terms of accuracy when being evaluated with 30 manually annotated images

Keywords Retinal fundus image • Segmentation • Region growing algorithm • . stack-based. object-oriented

A. F. A. Fadzil (✉) · S. Ibrahim
FSKM, Universiti Teknologi MARA Kampus Jasin, Jasin, Melaka, Malaysia
e-mail: firdausfadzil@melaka.uitm.edu.my

S. Ibrahim
e-mail: shafaf2429@melaka.uitm.edu.my

N. E. A. Khalid
FSKM, Universiti Teknologi MARA Kampus Shah Alam, Shah Alam, Selangor, Malaysia
e-mail: elaiza@tmsk.uitm.edu.my

1 Introduction

Retinal fundus medical imaging modality is one of the most widely utilized data for detecting and diagnosing eye diseases. This imaging modality allow medical practitioner to evaluate the condition of a patient's eye by extracting and segmenting different features contains within the image [1, 2]. Retinal fundus image contains crucial internal features of the eyeball namely the optic disc, macula, and blood vessels [3]. Optic disc and macula features are easier to be identified as optic disc feature basically represented by the brightest and macula the darkest area of the fundus image respectively.

Segmenting blood vessels on the other hand, represents greater challenge due to its tendency to blend with the rest of the images. The ambiguity of blood vessels feature therefore necessitates a robust algorithm in order to be automatically detected. Various automated segmentation algorithms are presented and deliberated for segmenting retinal fundus imaging blood vessel features. Most algorithms however, requires a very substantial image pre-processing and enhancement towards the original datasets [4, 5] that may lead to long and exhaustive execution time

Region growing segmentation algorithm is one of the more prevalent algorithm that can be used for segmenting various types of medical modalities. This algorithm has also been shown to be able to produce decent segmentation for retinal fundus blood vessels features. Alas, the algorithm is largely presented as a recursive algorithm is largely inefficient due to the amount of memory that needs to be stored in stack [6]. Therefore, employing stack data structure is shown to be a more elegant solution due to the fact that this technique allow easier access towards data when being compared with the traditional recursive method [7].

Nevertheless, employing such method can only be implemented to the utmost efficiency if the way of accessing the pixel of the images provides more information than normal data structure such as bitmap or graphics. Object-oriented programming paradigm is one of the major programming paradigm that revolutionize programming languages by providing a robust and efficient data readability and maintainability. Consequently, this paper proposes the implementation of object-oriented approach for stack-based region growing algorithm. Stack-based Object-oriented region growing segmentation (SORG) is an algorithm that implement region growing segmentation via object-oriented stack-based approach.

2 Methodology

2.1 Data Collection

Retinal fundus images are collected from DRIVE dataset collection [3]. This data-base provide a collection of retinal fundus image data with the fixed size of 565×584 pixels resolution. This collection came with the ground truth data

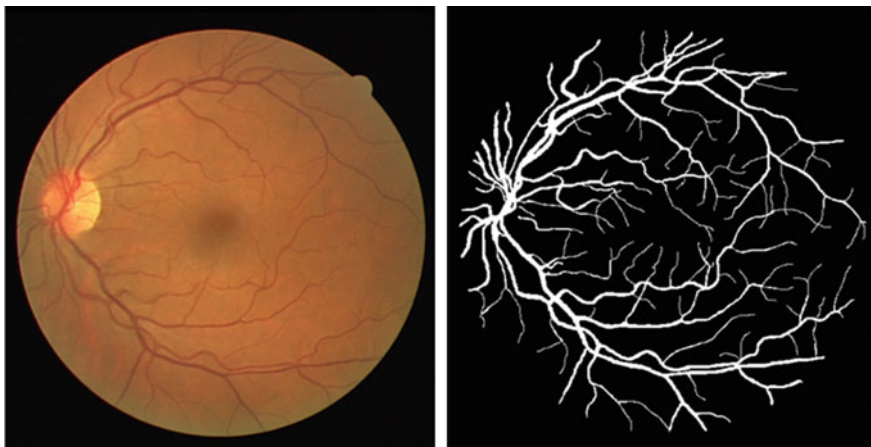


Fig. 1 Sample DRIVE dataset; retinal fundus image (left) and ground truth (right)

annotated by experts for every image provided. Sample data collected are depicted in Fig. 1.

Figure 1 depicts the sample data utilized for this research. The brightest area of the image on the left shows the optic disc and cup of the eye. The disc and cup ratio of the image allow practitioner to diagnose the different type of eye diseases. The dark spot in the middle of the image otherwise represents the macula of the eyes, the place where light is focused by the structures in the front of the eye. Finally, blood vessels are the feature that structurally covers most of the retinal image. Notice how some part of the images vaguely blends with the color of the retinal image.

2.2 *SORG Algorithm Flow*

Most retinal fundus image segmentation on blood vessels rely heavily on the pre-processing stage. For pre-processing stage of the SORG algorithm, only one filter is applied to the original image in order increase the visibility of the blood vessels features. This filter is referred as the photocopy filter, which darkens the dark features within an image and transform the image into black and white photocopy-inspired image. From here, the image is then converted into an object-oriented based data in order to allow the algorithm manipulates different types of data related to the sample image. This data structure contains essential information concerning each pixels with information such as the current point

location of the pixel, where the algorithm currently execute its process on. This data structure also employ the 3×3 kernel of neigh-boring pixels that allow detailed thresholding and boundary checking.

Information regarding the previous point on which the pixel are placed on the stack is also stored in order to allow the algorithm to find the current pixel’s predecessor. This allows comparisons between the predecessor and successor pixels. Finally, the indicator on whether the current pixel selected has already entered the stack is also stored in order to prevent the algorithm from repeatedly grow the pixel that has already selected. Figure 2 illustrates the object-oriented based data structure that has been employed in this research.

The object-oriented structure stored plays a pivotal role in the overall implementation of SORG algorithm. This structure allows more elaborate boundary checking by providing detailed information on each and every pixels.

Figure 3 demonstrates the overall flow of SORG segmentation algorithm. After every pixels are converted into the object-oriented data structures, random seed points are generated to initialize the algorithm. The algorithm starts with a single point and continuously checks the boundary in order to allow the pixel to grow its region. In this algorithm, the boundary is determined by a threshold value that subtracts the total value of red channel kernel of the current pixel with the total value of red channel kernel of its predecessor. Four more pixel point will be pushed into the stack if this criteria is fulfilled, and the algorithm will continue until there are no more point available in the stack.

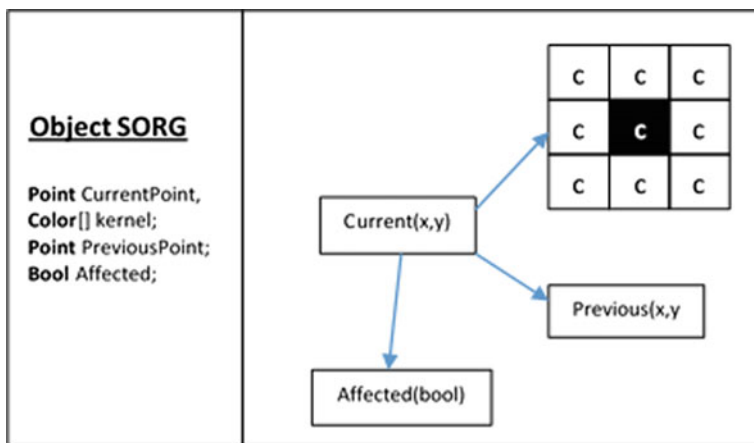


Fig. 2 Object-oriented based data structure for SORG

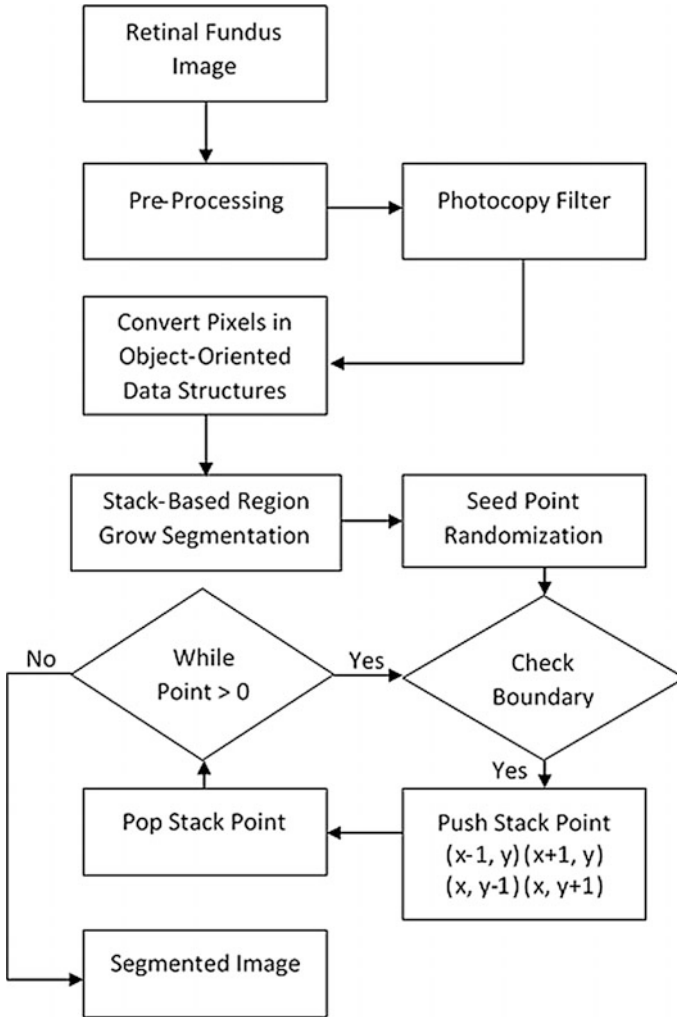


Fig. 3 SORG algorithm flow

3 Results and Discussion

The performance of SORG segmentation algorithm is determined by using receiver operating characteristic (ROC) [8]. The performance of the segmentation produced in terms of accuracy is evaluated by comparing the automated segmentation with the ground truth data manually annotated by experts as cited in earlier section.

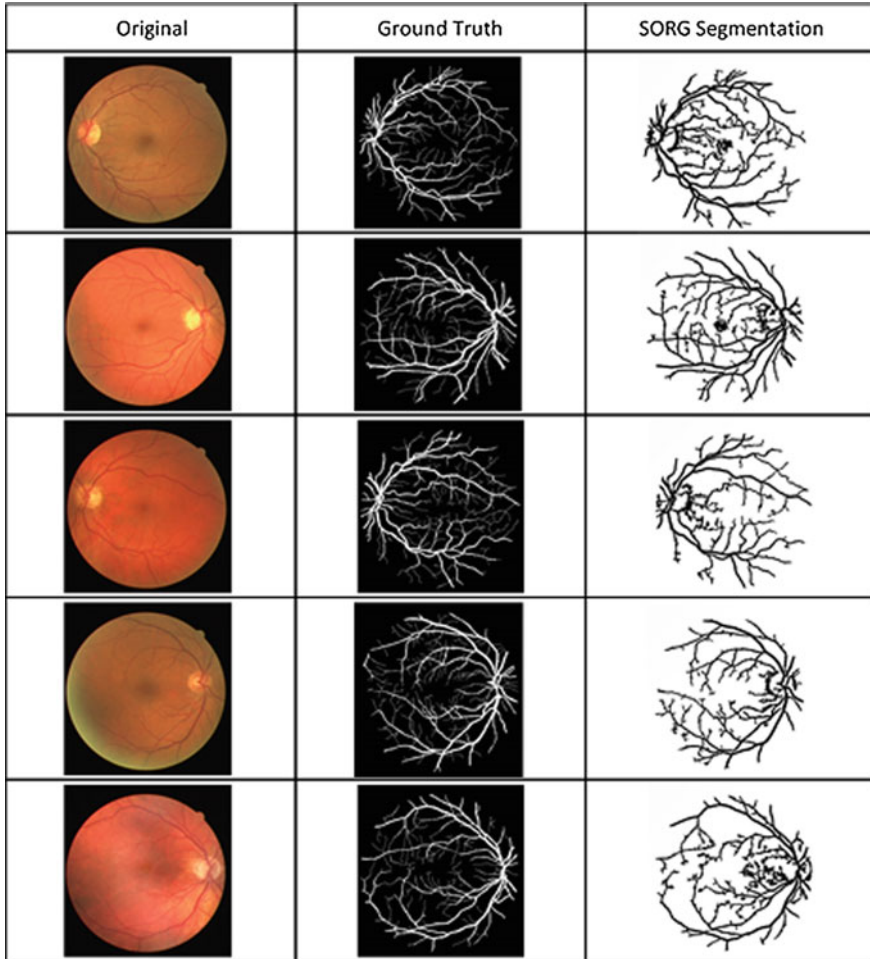


Fig. 4 SORG segmentation result

Figure 4 exemplifies the best 5 results of the total 30 DRIVE dataset that has been executed using SORG segmentation to extract the blood vessels from the fundus image.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Using equation above [8], SORG manages to achieve 75–83% accuracy for the 5 selected images. This figure implies that SORG algorithm able to produce decent segmentation of blood vessels features in retinal fundus images.

4 Conclusion and Recommendation

Nonetheless, this result also suggests that this algorithm has more room to be improved. Further enhancement such as implementing more complex boundary checking by employing larger number of kernel instead of the 3×3 kernel or by employing statistical method to find more accurate threshold value. The seed point randomization can also be replaced with evolutionary computing algorithm such as genetic algorithm in order to optimize the seed point selection for better results.

References

1. Mary, M.C.V.S., Rajsingh, E.B., Naik, G.R.: Retinal fundus image analysis for diagnosis of glaucoma: a comprehensive survey. *IEEE Access* **4**, 4327–4354 (2016)
2. Odstrcilik, J., Kolar, R., Budai, A., Hornegger, J., Jan, J., Gazarek, J., Angelopoulou, E.: Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. *IET Image Process.* **7**(4), 373–383 (2013)
3. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**(4), 501–509 (2004)
4. Kayal, D., Banerjee, S.: A new dynamic thresholding based technique for detection of hard exudates in digital retinal fundus image. In: 2014 International Conference on Signal Processing and Integrated Networks (SPIN), pp. 141–144. IEEE (2014)
5. Usman, A., Khitran, S.A., Akram, M.U., Nadeem, Y.: A robust algorithm for optic disc segmentation from colored fundus images. In: International Conference Image Analysis and Recognition, pp. 303–310. Springer, Cham (2014)
6. Ye, C.F., Li, Y.Z., Zeng, W.Q.: Study of non-recursive transformation algorithms of recursive problems. In: Applied Mechanics and Materials, vol. 644, pp. 1969–1971. Trans Tech Publications (2014)
7. Hore, S., Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Van Chung, L., Le, D.N.: An integrated interactive technique for image segmentation using stack based seeded 3. Region growing and thresholding. *Int. J. Electrical Comput. Eng.* **6**(6), 2773 (2016)
8. Khalid, N.E.A., Ibrahim, S., Manaf, M., Ngah, U.K.: Seed-based region growing study for brain abnormalities segmentation. In: 2010 International Symposium on Information Technology (ITSim), vol. 2, pp. 856–860. IEEE (2010)

Chapter 5

Similarity Measures of Intuitionistic Fuzzy Sets for Cancer Diagnosis: A Comparative Analysis



Lazim Abdullah and Sook Wern Chan

Abstract Membership degree, non-membership degree and hesitancy degree are the three components that characterized the intuitionistic fuzzy sets (IFSs). Based on information inherited from the three membership degrees, this paper proposes Cosine Similarity Measures (CSM) and Jaccard Similarity Measures (JSM) between IFSs and their application to a case of cancer diagnosis. Three experts in medical fraternity were invited to provide linguistic evaluation pertaining to symptoms with respect to types of cancer diseases, and patients with respect to symptoms using linguistic terms that defined in IFSs. The information of symptoms and type of cancer diseases for each patient was collected and then computed using CSM. The similar set of information was iterated using JSM. A comparative analysis of similarity measures between CSM and JSM is presented to illustrate the consistency of the two similarity measures of IFSs. It is shown that the two similarity measures are consistent in suggesting a cancer diagnosis despite differences in mathematical formulations.

Keywords Similarity measures · Intuitionistic fuzzy sets · Cancer diagnosis · Decision making · Fuzzy relation

1 Introduction

One of the most significant current discussions in decision analysis is the concept of similarity measure. Typically, similarity measure is a critical tool used for deciding the degree of closeness between two objects. It is a distance function where any types of objects could be measured and compared. This measure quantifies the strength of the relationship between two objects. Similarity of 0 implies that the two objects are dissimilar to each other and 1 indicates that the objects are identical [1].

L. Abdullah (✉) · S. W. Chan
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
Kuala Terengganu, Malaysia
e-mail: lazim_m@umt.edu.my

There are abundant of literatures about similarity measures between two sets. One of the widely used similarity measures is the similarity between two intuitionistic fuzzy sets (IFSs). Since Atanassov [2] extended fuzzy sets to IFSs, many different similarity measures of IFSs have been proposed. Liang and Shi [3], for example, proposed several similarity measures of IFSs and discussed the relationships between these measures. Not only literatures about methods of similarity measures of IFS, there are also a considerable amount of literature about its applications to pattern recognition and decision making. Li and Cheng [4] proposed similarity measure between IFSs and applied to pattern recognition. Abdullah and Wan Ismail [5] proposed a similarity measure and a weighted similarity measure of IFSs by considering hesitation degree and its application to pattern recognitions.

One of the most commonly discussed similarity measures is CSM. The CSM is the similarity measure that being built using the concept of the inner vector product. The CSM is defined as the inner product of two vectors divided by the product of their lengths. Another widely used similarity measure is JSM. In contrast to CSM, despite the used of inner vector product, the JSM of the two sets is measured through the comparison of the size of the overlap against the size of the two sets. This method is useful as with CSM under small data conditions and it is well fitted for market-basket data [6]. However, it is hypothesized that similarity measures obtained from CSM and JSM are inconsistent due to the fact that the two similarity measures are computed based on different mathematical definitions and concepts. Considering the information carried by three membership degrees of IFSs as vector representations in the vector space, this paper aims to compare similarity measures between CSM and JSM of IFS for a case of cancer diagnosis.

2 Preliminary

The following definitions are presented prior to applying and comparing CSM and JSM of IFS for the case of cancer diagnosis.

Definition 1: Intuitionistic Fuzzy Sets with hesitation margin [7] For an intuitionistic fuzzy set A ,

$$A = \{(x, \mu_A(x), \nu_A(x), \pi_A(x)) | x \in X\},$$

a hesitation margin

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x),$$

which is the intuitionistic fuzzy index of the element $x \in A$, denotes a measure of the degree of non-determinacy (or uncertainty) of the element $x \in X$. It expresses

the lack of knowledge on whether x belongs to A or not. Apparently, $0 \leq \pi_A(x) \leq 1$, for every $x \in X$. For every fuzzy set A' ,

$$\pi_{A'}(x) = 1 - \mu_{A'}(x) - [1 - \mu_{A'}(x)] = 0$$

where $x \in X$.

Definition 2: Cosine Similarity Measure with Hesitation Degree [7] Assume that there are two IFSs A and B in the universe of discourse $X = \{x_1, x_2, \dots, x_n\}$. For the IFS A is characterized by the degree of membership, $\mu_A(x_i)$, degree of non-membership, $\nu_A(x_i)$ and degree of hesitation, $\pi_A(x_i)$ for $i = 1, 2, 3, \dots, n$, which can be considered as vector representations with n elements: $\mu_A = (\mu_A(x_1), \mu_A(x_2), \dots, \mu_A(x_n))$, $\nu_A = (\nu_A(x_1), \nu_A(x_2), \dots, \nu_A(x_n))$ and $\pi_A = (\pi_A(x_1), \pi_A(x_2), \dots, \pi_A(x_n))$.

For the IFS B it is characterized by the degree of membership, $\mu_B(x_i)$, degree of non-membership, $\nu_B(x_i)$ and degree of hesitation, $\pi_B(x_i)$ for $i = 1, 2, 3, \dots, n$, which can be considered as vector representations with n elements: $\mu_B = (\mu_B(x_1), \mu_B(x_2), \dots, \mu_B(x_n))$, $\nu_B = (\nu_B(x_1), \nu_B(x_2), \dots, \nu_B(x_n))$ and $\pi_B = (\pi_B(x_1), \pi_B(x_2), \dots, \pi_B(x_n))$. Therefore, cosine similarity measure between A and B is defined as follows,

$$C_{IFS}(A, B) = \frac{1}{n} \frac{\sum_{i=1}^n \mu_A(x_i)\mu_B(x_i) + \nu_A(x_i)\nu_B(x_i) + \pi_A(x_i)\pi_B(x_i)}{\sqrt{\sum_{i=1}^n \mu_A^2(x_i) + \nu_A^2(x_i) + \pi_A^2(x_i)} \sqrt{\sum_{i=1}^n \mu_B^2(x_i) + \nu_B^2(x_i) + \pi_B^2(x_i)}}$$

3 Application

This section describes the application of CSM to a case of cancer diagnosing. Linguistic variables used by the experts, the proposed algorithm purposely for the case, and the results are explained in this section.

3.1 Experts and Linguistic Variables

In this research, a committee of three experts was invited to provide qualitative evaluation of cancer diagnosis using linguistic variables. The experts comprise three doctor attached to Malaysia Ministry of Health in Ipoh Malaysia. The experts were interviewed to collect their evaluation towards the symptoms that lead to different types of cancer. Responses from experts were collected using the linguistic terms retrieved from Chang and Chen [8].

3.2 Proposed CSM of IFS

The computation steps that apply similarity measure of IFSs in medical diagnosis of cancer are proposed as follows.

Step 1: Construct a set of patients $P = \{P_1, \dots, P_m\}$

Step 2: Construct a set of symptoms $S = \{S_1, \dots, S_n\}$.

Step 3: Construct an intuitionistic fuzzy relation A from the set of patients to the set of symptoms.

For each patient $P_i, i = 1, \dots, m$, a set of symptoms $S_j, j = 1, \dots, n$ is specified. The fuzzy relation A is defined as

$$A = \{ \langle (p, s), \mu_A(p, s), \nu_A(p, s), \pi_A(p, s) \mid (p, s) \in P \times S \rangle \}$$

where $\mu_A(p, s)$ indicates the degree of the symptom s appears in patient p , $\nu_A(p, s)$ indicates the degree of the symptom s does not appear in patient p , and $\pi_A(p, s)$ denotes the degree of uncertainty of the appearance of the symptom s in patient p . It shows the degree of relationship between the patients and the symptoms.

Step 4: Construct a set of diseases $D = \{D_1, \dots, D_q\}$.

Step 5: Construct an intuitionistic fuzzy relation B from the set of symptoms to the set of diseases. Consider a set of diseases $D = \{D_1, \dots, D_q\}$. For each disease $D_k, k = 1, \dots, q$, a set of symptoms $S_j, j = 1, \dots, n$ is specified. The fuzzy relation B is defined as

$$B = \{ \langle (s, d), \mu_B(s, d), \nu_B(s, d), \pi_B(s, d) \mid (s, d) \in S \times D \rangle \}$$

where $\mu_B(s, d)$ indicates the degree of the symptom s confirms the presence of the disease d , $\nu_B(s, d)$ indicates the degree of the symptom s confirms the non-existence of the disease d , and $\pi_B(s, d)$ denotes the degree of uncertainty of the symptom s confirms the presence of the disease d . It shows the degree of relationship between the symptoms and the diseases, which is also known as confirmability degree.

Step 6: Calculate the similarity measure for all symptoms of the i th patient from the k th disease using the cosine similarity measure equation and Jaccard similarity measure.

Step 7: Identify the diagnosis on the basis of composition of intuitionistic fuzzy relation.

The seven-step procedure is implemented in the case of patients that seeking an appropriate diagnosis based on symptoms that they experienced. The results are presented in the next sub-section.

Table 1 Similarity measures of patients using cosine similarity measure

	Lung Cancer	Breast Cancer	Colorectal Cancer	Nasopharyngeal Cancer	Cervical Cancer
Faty	0.7370	0.7696	0.9451	0.8793	0.9851
Lam	0.7042	0.7452	0.9967	0.7932	0.9769
Kuz	0.8854	0.8829	0.7520	0.9764	0.7884
Sheil	0.9536	0.9825	0.7376	0.8817	0.7641
Hali	0.9910	0.9629	0.7157	0.8543	0.7382

3.3 Implementation

Step 1: A set of patients, $P = \{Faty, Lam, Kuz, Sheil, Hali\}$

Step 2: A set of symptoms

$S = \{Unintended\ Weight\ Loss, Unusual\ Bleeding, Swelling\ or\ Lump, Shortness\ of\ Breath, Persistent\ Cough, Fatigue, Diarrhea, Swallowing\ Problem, Loss\ of\ Appetite\}$

Step 3: An intuitionistic fuzzy relation A from the set of patients to the set of symptoms is computed.

Step 4: A set of diseases $D = \{Lung\ Cancer, Breast\ Cancer, Colorectal\ Cancer, Nasopharyngeal\ Cancer, Cervical\ Cancer\}$

Step 5: An intuitionistic fuzzy relation B from the set of symptoms to the set of diseases is computed.

Step 6: Similarity measure of intuitionistic fuzzy sets for all diagnosis of the i -patient from k -th disease are obtained using fuzzy relation A . The CSM of patients are computed and the results are summarised in Table 1.

Step 7: Identify the diagnosis on the basis of composition of intuitionistic fuzzy relation. The highest similarity gives a proper medical diagnosis for each patient. According to the value of CSM, we can conclude that Faty suffers from cervical cancer, Lam suffers from colorectal cancer, Kuz suffers from nasopharyngeal cancer, Sheil suffers from breast cancer and Hali suffers from lung cancer.

4 Comparative Analysis

The similar data sets are computed using the JSM. The procedures in Sect. 3 are iterated accordingly. However, the CSM in Step 6 is now substituted with Eq (2). The results using JSM are presented in Table 2.

The highest value of similarity indicates the disease that most likely suffered by the patient. From Table 2, we can conclude that Faty suffers from cervical cancer, Lam suffers from colorectal cancer, Kuz suffers from nasopharyngeal cancer, Sheil suffers from breast cancer and Hali suffers from lung cancer. These results are consistent with the results obtained using CSM.

Table 2 Similarity measures of patients using jaccard similarity measure

	Lung Cancer	Breast Cancer	Colorectal Cancer	Nasopharyngeal Cancer	Cervical Cancer
Faty	0.5624	0.5760	0.9032	0.7672	0.9472
Lam	0.5046	0.5240	0.9852	0.5919	0.9418
Kuz	0.7720	0.7337	0.5163	0.9430	0.5801
Sheil	0.8805	0.9383	0.5167	0.7624	0.5689
Hali	0.9752	0.9052	0.5117	0.6000	0.5527

5 Conclusion

Similarity measures are basically providing a clue to see the similarity or dissimilarity between two objects or sets. Each similarity measure method has a unique mathematical formulation and thereby provides different measurement and interpretation. Furthermore, type of sets used in similarity measures is also contributing to the measures. In this paper, cosine similarity measures and Jaccard similarity measures of intuitionistic fuzzy sets were investigated and applied to a case of cancer diagnosis. Information about types of cancer and symptoms was translated into the degree of membership, non-membership and degree of hesitation of intuitionistic fuzzy sets and measured using the two similarity measures. The similarity measures obtained from the two methods were analysed and compared. The comparative analysis suggests that the results obtained using the two similarity measures are consistent despite differences in mathematical formulation.

References

1. Vlachos, M.: Similarity measures. In: Sammut, C., Webb, G.I. (eds.) *Encyclopaedia of Machine Learning*. Springer, New York (2011)
2. Atanassov, K.: Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **20**, 87–96 (1986). [https://doi.org/10.1016/s0165-0114\(86\)80034-3](https://doi.org/10.1016/s0165-0114(86)80034-3)
3. Liang, Z. Shi, P.: Similarity measures on intuitionistic fuzzy sets. *Pattern Recognit. Lett.* **24**, 2687–2693 (2003). [https://doi.org/10.1016/s0167-8655\(03\)00111-9](https://doi.org/10.1016/s0167-8655(03)00111-9)
4. Li, D., Cheng, C.: New similarity measures of intuitionistic fuzzy sets and application to pattern recognition. *Pattern Recognit. Lett.* **23**, 221–225 (2002)
5. Abdullah, L., Wan Ismail, W.K.: Hesitation degree of intuitionistic fuzzy sets in a new cosine similarity measure. *J. Uncertain Syst.* **8**, 109–115 (2014)
6. Jangale, S. Hadsul, D.: Fault detection mechanism for wireless sensor networks. *Int. J. Eng. Sci. Innov. Tech.* **2**, 558–563 (2013)
7. Atanassov, T.: *Intuitionistic fuzzy sets: theory and application*. Physica-Verlag, Heidelberg (1999)
8. Ye, J.: Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Math. Comput. Model.* **53**, 91–97 (2011)

Chapter 6

Comprehensive Performance Assessment on Open Source Intrusion Detection System



Fuad Mat Isa, Shahadan Saad, Ahmad Firdaus Ahmad Fadzil and Raihana Md Saidi

Abstract Several studies have been conducted where authors compared the performance of open source Intrusion detection systems, namely Snort and Suricata. However, these studies were mostly limited to either security indicators or performance measurements under the same operating system. The objective of this study is to provide a comprehensive analysis of both products in terms of several security and performance related indicators under two different operating systems. Experiments were conducted to evaluate the effects of open source intrusion detection and prevention systems; Snort and Suricata running on Windows and Linux operating system. Attack types on system such as resource usage, dropped packets rate and ability to detect intrusions serve as experiment benchmarks. From the result, Snort is shown to work better in Linux platform in terms of intrusion detection compared to Suricata. In terms of performance in windows platforms, Snort demonstrated lesser intrusion detection than its Linux-based execution. This is different with Suricata, with its Linux configuration shown to be unable to detect any attack executed. The results also indicated that Linux-based execution consumes more system resources than its windows-based counterpart.

Keyword Attacks · Intrusion detection systems (IDS) · Network · Traffic · Performance evaluation · Snort · Suricata

F. M. Isa (✉) · S. Saad · A. F. A. Fadzil · R. Md. Saidi
FSKM, Universiti Teknologi MARA, Kampus Jasin, Melaka, Malaysia
e-mail: fuadisa19@gmail.com

S. Saad
e-mail: shahadan@melaka.uitm.edu.my

A. F. A. Fadzil
e-mail: firdausfadzil@melaka.uitm.edu.my

R. Md. Saidi
e-mail: raihana@melaka.uitm.edu.my

1 Introduction

Computer network nowadays has developed into a very complex architecture. The complexity of this architecture necessitates high level security to secure the network. Therefore, intrusion detection system (IDS) plays a pivotal role in securing the network. As the name implies, IDS detect intrusion or unauthorized entries into a network. IDS act as the first line of defense if any attack were made on the network [1]. An IDS monitors network traffic for suspicious activity and alerts network administrators or responds by taking predefined action such as blocking the attacker's IP addresses. There are several IDS systems available on the market. In the open source area, there are two major products, Snort with the most dominant market share [2, 3] and one of its closest rival, Suricata [4].

Snort and Suricata can be implemented on both UNIX-based and Windows operating system. However, recent research only emphasizes the performance of both IDS in terms of the hardware employed e.g. number of processors and utilizing general purpose graphic processor unit (GPGPU) [1, 4–8]. Therefore, experiments in this research are conducted essentially to assess both Snort and Suricata performance in different operating system. This research has focuses on signature-based IDS with an emphasis on evaluating their performance in Windows and Linux (Ubuntu) OS. Five different types of attacks namely the port scan, denial of service (DDOS), bad traffic, and brute-force are evaluated in both operating system. Performance metrics such as true positive, false positive, ability to detect intrusion, and CPU usage are chosen to assess both IDS performance.

2 Methodology

2.1 Hardware and Software Specifications

A specific test bed is employed to evaluate both IDS. The test bed design consists of two computer systems, namely the server and the client. The client will generate and send network traffic for analysis in the server. Client is installed with two operating systems, Windows and Ubuntu Linux. Connection between server and client is established via a switch. Attacks will be generated using Kali Linux virtual operating system. Table 1 below depicts the software specifications for this research;

Table 1 Software specification

Machine type	Description
Operating system	Windows 7, Ubuntu, Kali Linux
Intrusion system detection	Snort 2.9 and Suricata
IDS testing framework	Pytbull

Table 2 Category of attack employed

Category of attack	Description
Port scan	Module identify network services running on a host and exploit vulnerabilities
Denial of Service (DOS)	This module implements various forms of DOS attacks by using tools like scapy and hping
Bad traffic	Module intended for sending network packets of data that are not in compliance with RFC standards
Brute force	Implementation of Brute Force attacks to the FTP server by using the hydra tool

In both operating system the same configuration for each IDS (with differences only at specific operating system-related parameters). Snort and Suricata both functions in IDS (Intrusion Detection System) mode. For testing purpose, the official free Sourcefire VRT (Vulnerability Research Team) signatures and complimentary signatures from Emerging Threats (ETOpen Rules) are employed.

2.2 Categories of Attacks

For each parameter, five attacks are evaluated. In Windows operating system, certain services and applications were disabled such as automatic installation and downloading of updates (Windows Update) and firewall (Windows Firewall) to maintain the system's stability (Table 2).

3 Results and Discussion

3.1 Attack Results

In this section, the results of each categories of attack are deliberated. There are four different configurations tested for each different category; Snort on Windows, Snort on Linux, Suricata on Windows, and finally Suricata on Linux.

Figure 1 illustrates the results of all four configurations for port scan and DOS attack. In terms of port scan performance metric, the attacks are divided into three different categories; intense scan, ping scan, regular scan. From the figure above, Snort on both Windows and Linux is shown to be able to detect the port scan attacks, with significantly more attacks are detected in Windows OS. Suricata meanwhile does not able to correctly detect any port scan attack in both OS.

On the other hand, DOS attack is evaluated to measure the inspection capabilities of both IDS in facing to Denial of Service attempts. DOS is typically accomplished by flooding the targeted machine or resource with superfluous requests to overload systems and prevent some or all legitimate requests from being fulfilled.



Fig. 1 Port scan and DOS attack results

DOS attacks are divided into 2 two parts, Hping Syn Flood and DOS against MSSQL service. From the figure above, it is shown that only the configuration of Snort on Linux and Suricata on Windows that can correctly identify DOS intrusion.

Figure 2 illustrates the results of all four configurations for bad traffic and brute force attack. Bad traffic attack is evaluated to test the behavior of each IDS to face crafted packets that are non-RFC compliant. Bad Traffic consists of Nmap Xmas Scan, Malformed Traffic and Land Attack. The result obtained are shows that in a set of 25 attacks, both Suricata and Snort have unsuccessfully detected any bad traffic when running on both Windows and Linux. Finally, both IDS is evaluated to test their ability to detect multiple bad logins against a service (e.g. Brute force against FTP). Brute force (also known as brute force cracking) is a trial and error

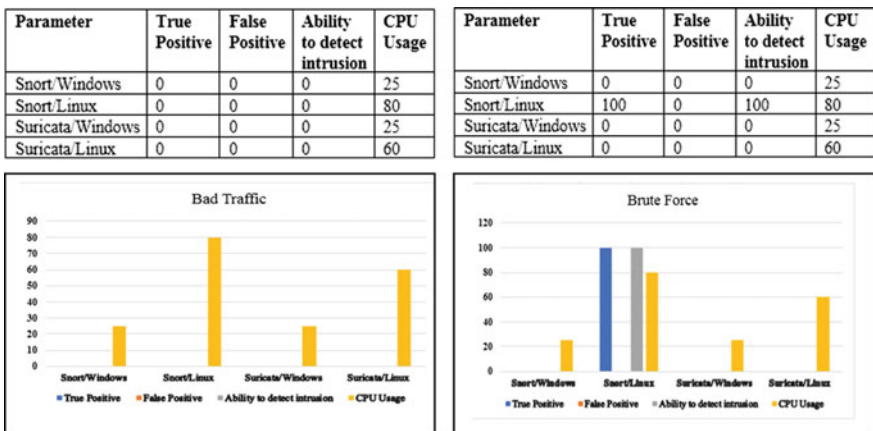


Fig. 2 Bad traffic and brute force attack results

method used by application programs to decode encrypted data such as passwords or Data Encryption Standard (DES) keys, through exhaustive effort (using brute force) rather than employing intellectual strategies. In The result obtained in Fig. 2 shows that only Snort on Linux configuration was able to correctly identify brute force attack.

4 Configuration Analysis

In this section, all four experiment configurations are analyzed to compare both IDS in terms of their respective true positive, false positive, ability to detect intrusion, and CPU usage value acquired from the previous section. As soon as the IDS produces alerts, results were taken in terms of true positive rates, false positive rates, ability to detect intrusion and CPU resources. True positive state when the ids identifies an activity as an attack and the activity is an actual attack. On the hand, false positive or false alarm is a result that indicates a given condition has been fulfilled when it has not. Ability to detect intrusion is the process of identifying and responding to intrusion activities. CPU usage is a term used to describe how much the processor is working while doing tasks.

Based on the experiment conducted, Snort is indicated to work better on Linux platform. It is shown on Fig. 3 that Snort on Linux platform was able to produce True Positive alert on most attack conducted. However lesser amount of detection based on those attacks can be found on both IDS when running on Windows

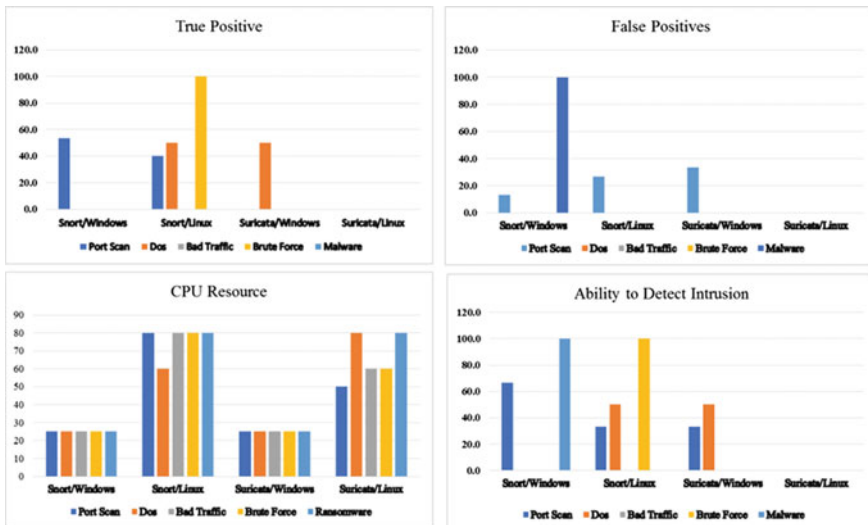


Fig. 3 Snort versus Suricata configuration results

platform. The performance of Suricata on Linux is indicated to be the worst as the configuration does not able to detect any of the attack executed. False positive or false alarm is a result that indicates a given condition has been fulfilled, when it has not. Based on Fig. 3, Snort on Windows platform shows higher amount of false positives alert among others. Suricata on Linux presented no alert this time because it didn't detect any of those attacks.

In terms of ability to detect intrusion, Snort performed well on both platforms. On both Windows and Linux, Snort was able to detect most of the intrusion. On the other hand, Suricata is shown to perform weaker on both platforms when being compared to Snort. Comparison in terms of CPU usage on both platforms indicated that both Snort and Suricata use more resources on the Linux operating system. The CPU usage of each solution depends not only on the simulated attack, but also on the operating system on which the IDS is implemented.

5 Conclusion

The focus of the research was to establish that employing/selecting an IDS requires a thorough evaluation of network characteristics and requirements. The host hardware selected should correspond to the network traffic and the processing needs of the system. Snort indicated that it performs better when it is configured on Linux platform. This is different to Suricata as its performance is better when executed on Windows operating system. Employing Suricata in Linux in this paper shown to return the worst performance recorded with its inability to identify any attack. In terms of CPU resource, it is indicated that Windows operating system returns a lower CPU resource consumption than its Linux counterpart. In conclusion, Snort demonstrates its stability, ease of configuration and excellent documentation. Nevertheless, Suricata can revolutionize IDS/IPS with features such as multi-threading support.

References

1. Albin, E.: A comparative analysis of the Snort and Suricata intrusion-detection systems. Naval Postgraduate School, Monterey CA (2011)
2. Chakrabarti, S., Chakraborty, M., Mukhopadhyay, I.: Study of Snort-based IDS. In: Proceedings of the International Conference and Workshop on Emerging Trends in Technology, pp. 43–47. ACM (2010)
3. Saboor, A., Akhlaq, M., Aslam, B.: Experimental evaluation of Snort against DDoS attacks under different hardware configurations. In: 2013 2nd National Conference on Information Assurance (NCIA), pp. 31–37. IEEE (2013)
4. Day, D., Burns, B.: A performance analysis of Snort and Suricata network intrusion detection and prevention engines. In Fifth International Conference on Digital Society, Gosier, Guadeloupe, pp. 187–192 (2011)

5. Alhomoud, A., Munir, R., Disso, J.P., Awan, I., Al-Dhelaan, A.: Performance evaluation study of intrusion detection systems. *Procedia Comput. Sci.* **5**, 173–180 (2011)
6. Kacha, C., Shevade, K.A.: Comparison of different intrusion detection and prevention systems. *Int. J. Emerg. Technol. Adv. Eng.* **2**(12), 243–245 (2012)
7. Albin, E., Rowe, N.C.: A realistic experimental comparison of the Suricata and Snort intrusion-detection systems. In: 2012 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 122–127. IEEE (2012)
8. White, J.S., Fitzsimmons, T., Matthews, J.N.: Quantitative analysis of intrusion detection systems: Snort and Suricata. *SPIE Defense, Security, and Sensing*, pp. 875704–875704 (2013)

Chapter 7

Mobile Ad-Hoc Network (MANET) Routing Protocols: A Performance Assessment



Nur Amirah Mohd Saudi, Mohamad Asrol Arshad,
Alya Geogiana Buja, Ahmad Firdaus Ahmad Fadzil
and Raihana Md Saidi

Abstract Mobile Ad Hoc network (MANET) is a collection of mobile devices which form a communication network. There are multiple type of routing protocols that de-signed for MANETs. This paper presents an investigation of four MANET protocols' performance, namely the Ad Hoc On-Demand Distance Vector (AODV), Destination-Sequenced Distance-Vector (DSDV), Dynamic Source Routing (DSR) and Ad Hoc On-Demand Multipath Distance (AOMDV). These protocols are evaluated using three difference performance metrics; average end-to-end, throughput and packet delivery ratio. Simulations of MANET is conducted to analyze the behavior of these protocols with different node mobility and node speed. From the results, it is indicated that different protocols performs better than the other on different performance metrics. For Average end-to-end metric, AODV is shown as the best performer even with the increment of speed. All four protocols meanwhile shows similar performance when node speed are increased for the throughput performance metric. For the final metric, it is shown that AOMDV returns the highest packet delivery ratio.

Keywords MANET · AODV · AOMDV · DSDV · Average end-to-end · Throughput · Packet delivery ratio

N. A. M. Saudi (✉) · M. A. Arshad · A. G. Buja · A. F. A. Fadzil · R. Md. Saidi
FSKM, Universiti Teknologi MARA Kampus Jasin, Jasin, Melaka, Malaysia
e-mail: nuramirahmohdsaudi@gmail.com

M. A. Arshad
e-mail: mohamad_asrol@melaka.uitm.edu.my

A. G. Buja
e-mail: geogiana@melaka.uitm.edu.my

A. F. A. Fadzil
e-mail: firdausfadzil@melaka.uitm.edu.my

R. Md. Saidi
e-mail: raihana@melaka.uitm.edu.my

1 Introduction

Mobile Ad Hoc network (MANET) is an infrastructure-less system which has no central server, or specialized hardware and fixed routers [1]. Each device plays the role of an independent router and generates independent data as it functions in dispersed peer to peer style. MANET can be operated as a stand-alone network or connected to the internet via cellular network [2]. This brings different possibilities towards employing MANET technology. In addition, MANET is a Multi-Hop Routing as it does not need a router to operate. There are two types of ad hoc routing which are single-hop and multi-hop routing [3]. Single-hop is simpler than multi-hop as it is low cost, simple structure and easy to implement. Every node plays the role of router and forward packets for information distribution among portable hosts [4].

Previous research shown that MANET are independent from the central network administration as its network is self-configured [5]. Furthermore, MANET does not depend on other networking devices such as servers, switches or even router [6]. This therefore minimizes the number of device used as the nodes themselves can act as routers.

MANET is a technology that allows easier communication due to its dynamic topologies. Nodes in MANET are free to move randomly anywhere in the network [7]. Due to the arbitrary movement of nodes, the network topography (which is classically multi-hop) changes regularly and randomly. Nodes are also free to move with different momentum and speed on unpredictable time [8]. Therefore, it is imperative to know which routing protocols can be employed to provide the best all round performance when connecting multiple nodes in MANET. In this paper, the performance of four different MANET routing protocols namely the Destination-Sequenced Distance-Vector (DSDV), Dynamic Source Routing (DSR), Ad Hoc On-Demand Distance Vector (AODV), and Ad Hoc On-Demand Multipath Distance (AOMDV) are assessed for comparison.

2 Literature Review

2.1 *Destination-Sequenced Distance-Vector (DSDV)*

The DSDV protocol is a table-driven routing based on an update of the classic Bell-man-Ford routing algorithms in MANET [8]. Each node has a routing table that indicates for each destination, which is the next hop and number of hops to the destination. DSDV also based on the distance vector routing that uses a bidirectional link [10]. In this context, DSDV requires each node periodically to update routing. Each DSDV node maintain routing tables to list the “next hop” so that every destination can be reached. The main advantage of DSDV is loop freedom is guaranteed through distance vector routing.

2.2 *Dynamic Source Routing (DSR)*

The Dynamic Source Routing protocol composes of two main mechanisms to allow the discovery and maintenance of source routes in the ad hoc networks [11]. Source routing does not need to maintain a middle node to update the routing information to route packets as all routing decisions are continuously updated inside the mobile nodes. DSR contains two mechanisms, namely the Route Discovery and Route Maintenance. In route discovery, DSR floods Route Request Packet to the network [12]. Route Discovery is a mechanism whereby node S sends packets to the destination D and have access to the source D. For Route maintenance, DSR provides three successive steps [13]. Route Maintenance is the mechanism whereby the packet forwarding S detect if the network topology has change, the route to the destination D cannot be used because the two nodes that are listed in the route have been out of range of each other. Hence, when Route Maintenance indicates source routing damaged, S notified the route error packet. Sender S can try to use any other route to D for requesting Route Maintenance to seek new password again.

2.3 *Ad Hoc On-Demand Distance Vector (AODV)*

AODV is considered as a combination of both DSR and DSDV [8]. This is because AODV borrow the basic mechanism for requesting Route Discovery and Route Maintenance of DSR. In addition, this protocol performs Route Discovery using control messages Route Request (RREQ) and Route Reply (RREP) [9]. When the source node S wants to send data packets to the destination node D but could not find a route in the routing table, the node spreads the message of Route Request (RREQ) to neighboring nodes, including the last known sequence number for the destination. Neighbors and spread the message RREQ to its neighbors if they do not have a good route to the destination node. This process continues until the message RREQ reaches the destination node or an intermediate node that has a good track.

2.4 *Ad Hoc On-Demand Multipath Distance (AOMDV)*

AOMDV is a multipath extension to the AODV protocol [14]. In AOMDV protocols, multiple routes are founded between the source and destination. In this context, When AOMDV builds multiple paths, it will select the main path for data transmission which is based on the time of routing establishment. The earliest one will be regarded the best, and only when the main path is down other paths can be effective. In fact, numerous studies indicate that this scheme does not necessarily produce the best path.

3 Methodology

AODV, DSDV, DSR and AOMDV protocols in different node numbers in MANET are examined in terms of average end to end delay, throughput, and packet delivery ratio performance metrics. The end-to-end delay of a path is the sum of all the above delays incurred at each link along the path. Packet Delivery Ratio is the ratio of the successful data packets to the destination generated by the CBR source. Meanwhile, throughput is the number of packets that pass through the network in one unit of time in kbps size.

This research proposes a development of a network simulations that involves three steps; identifying the required research data, identifying the required software and identifying the parameters that affect the network simulation. For verification purpose, the network simulation is analyzed and evaluated. To develop a network simulation, Network Simulator 2 (NS2) is employed. NS2 software is employed to reduce the range of deviations and data errors while improving the accuracy of the research results.

4 Results and Discussion

The performance in terms of average delay end-to-end, throughput and packet delivery ratio are evaluated and compared for AODV, DSDV, DSR and AOMDV. Simulations are conducted on the Network Simulator 2 (NS-2) with network comprising of 3 nodes, 5 nodes, 10 nodes, 15 nodes, 30 nodes and 40 nodes moving over an area of $800\text{ m} \times 800\text{ m}$ for 150 s of simulated time. Constant bit Rate (CBR) traffic is presumed. A 512-byte data packet with 2 packets/second sending rate is assumed for all the experiments. The simulations are set with three different node speed which are 5, 25 and 55 m/s.

From Fig. 1, it is indicated that the protocols that has the lowest reading of average end-to-end delay is DSR. DSR has the lowest reading of average end-to-end delay when the node number are 30 and 40.

Based on Fig. 2, an average reading of throughput is recorded for all four protocols. AODV and DSDV shows the highest reading of throughput. Therefore, routing protocol that has the highest reading of throughput indicates a good performance as the delay of a packet while transmitting data is low. All the protocol shows the improvement when the speed of the node increases.

Figure 3 shows the packet delivery ratio for each protocol for mobile speed of 5 m/s. AOMDV have the highest reading of packet delivery ratio for all the nodes.

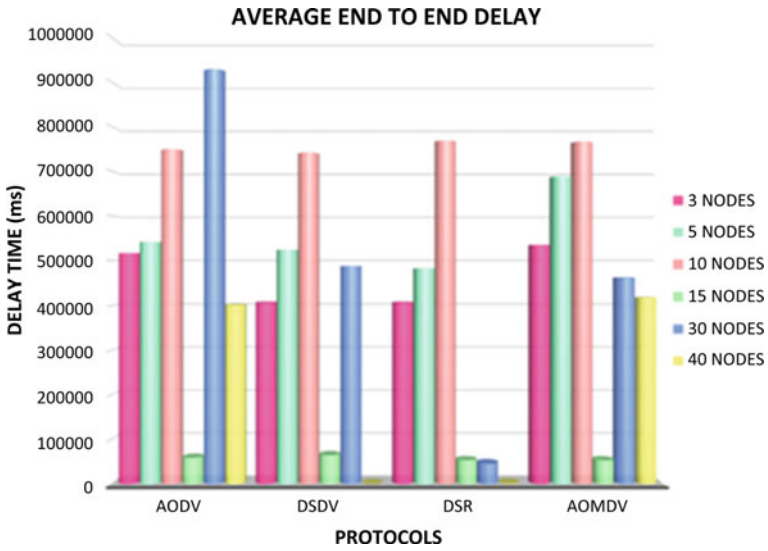


Fig. 1 Average end-to-end delay performance using different protocols in multiple nodes

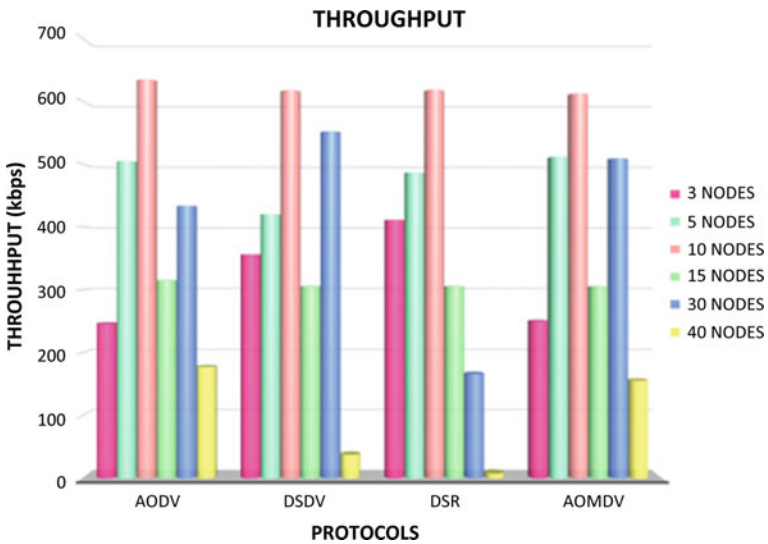


Fig. 2 Throughput performance using different protocols in multiple nodes

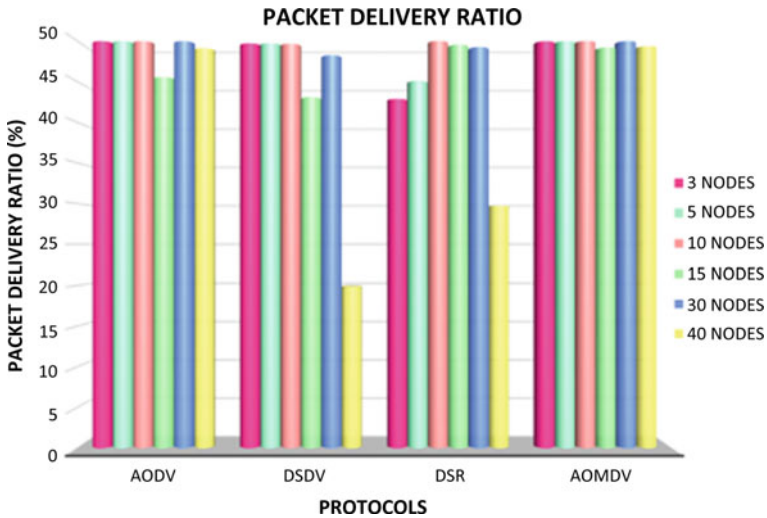


Fig. 3 Packet delivery ratio performance with different protocols in multiple nodes

5 Conclusion

In general, AODV can be considered more efficient routing protocols as the packet delay rate is lower than other protocols. For throughput and packet delivery ratio, we can conclude that AOMDV and DSDV protocols are the two protocols that show the best performance since these two protocols have the highest value. For 15 nodes, better performance of average end to end delay is indicated in general. This is largely due to the movements of the node or the energy of the node that in the simulations.

References

1. Mishra, S., Singh, A.: A novel approach for video transmission. In: Clerk Maxwell, J. (ed.) *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2, pp. 68–73 Clarendon 1892, Oxford (2014) **5**(6), 7270–7275
2. Aarti, Tyagi, D.S.S.: Study of MANET: characteristics, challenges, application and security attacks. *Int. J. Adv. Res.* **3**(5), 252–257 (2013)
3. Rana, A., Gupta, S.: Review on MANETs characteristics, challenges, application and security attacks, **4**(2), 2203–2208 (2015)
4. Yadav, M.: Survey on MANET: routing protocols, advantages, problems and security, **1**(2), (2014). Accessed from <http://ijicse.in/wp-content/uploads/2014/12/12-17.pdf>
5. Bhalia, M.: Analysis of MANET characteristics, applications and its routing challenges, **3**(4), 139–143 (2015)

6. Chavan, A.A., Kurule, D.S., & Dere, P.U.: Performance analysis of AODV and DSDV routing protocol in MANET and modifications in AODV against black hole attack. *Procedia Comput. Sci.* **79**, 835–844, (2016). <https://doi.org/10.1016/j.procs.2016.03.108>
7. Gupta, S.K., Saket, R.K.: Performance metric comparison of AODV and DSDV routing protocol in MANETs using NS-2. *IJRRAS* **7**(3), 339–350 (2011) (7, June)
8. Goto, H., Hasegawa, Y., Tanaka, M.: Efficient scheduling focusing on the duality of MPL representatives. In: *Proceedings of the IEEE Symposium Computational Intelligence in Scheduling (SCIS 07)*, pp. 57–64. IEEE Press (Dec 2007). <https://doi.org/10.1109/scis.2007.35>
9. Khiavi, M.V., Jamali, S., Gudakahriz, S.J.: Performance comparison of AODV, DSDV, DSR and TORA routing protocols in MANETs. *Int. Res. J. Appl. Basic Sci.* **3**(7), 1429–1436 (2012)
10. Keshtgary, M., Babaiyan, V.: Performance evaluation of reactive, proactive and hybrid routing protocols in. *Int. J.* **4**(2), 248–254 (2012)
11. Manohari, P.K., Ray, N.: Multipath routing protocols in MANETs: A Study. In: (ICICCS), pp. 91–96 (2016)
12. Mello, S.D., Patil, P.B., Shaikh, T.P.S., Shaikh, H.S., Malik, S.M.: A survey on wireless routing protocols (AODV, DSR, DSDV), **2**(1), 3–8 (2015)
13. Saad, P., Hasson, T.: Designing a new MANET's environment using computer simulation **1** (3), 370–375 (2013)
14. Yang, B., Chen, Y., Jiang, X.: Multicast delay of mobile ad hoc networks. In: *2014 Second International Symposium on Computing and Networking*, pp. 272–277 (2014). <https://doi.org/10.1109/CANDAR.2014>

Chapter 8

Multimedia Data Archive Application in Cloud Environment



Syarilla Iryani A. Saany, M. Nordin A. Rahman, A. Rasid Mamat
and Ahmad Fakhri Ab Nasir

Abstract The volume and importance of multimedia data archive increase rapidly over the years. This situation creates a requirement for refined multimedia data archive application that enable for interactive analysis, dynamic and large-scale transaction management. The application needs to be modelled with dynamic data design, open architecture that is easy to be accessed by data stakeholders at any time and less dependent on the location-wise. This article introduces the concept of temporal data element in multimedia data archive management model. Two elements of time, transaction time and valid time are embedded into data design for multimedia data archive. These temporal elements associates with real-world facts, events and can be exploited for query reasoning. To evaluate the model, a web service-based application is developed and deployed under public cloud environment. This application prototype can be benefitted by the multimedia production house or researcher to manage a massive size of multimedia data effectively, systematically and more openly.

Keywords Multimedia data archive · Temporal data · Application as a service · Cloud environment

S. I. A. Saany (✉) · M. N. A. Rahman · A. R. Mamat
Faculty of Informatics and Computing, University of Sultan Zainal Abidin, Kuala
Terengganu, Malaysia
e-mail: syarilla@unisza.edu.my

M. N. A. Rahman
e-mail: mohdnabd@unisza.edu.my

A. R. Mamat
e-mail: rasid@unisza.edu.my

A. F. A. Nasir
Faculty of Manufacturing Engineering, Universiti Malaysia Pahang, Gambang, Malaysia
e-mail: afakhri@ump.edu.my

1 Introduction

Recently, an increasing number of scientific fields, such as business intelligence, forensics, medical sciences and environmental studies utilize scientific imagery to advance the growth of multimedia data size [1, 2]. In communication and media social applications, the advent of devices with good multimedia recording capabilities, resulting in the contributing large amounts of social multimedia data shared on cloud platform. The main area of research in multimedia data management system are data modelling, storage management, data retrieval, media integration and archiving process [3]. Multimedia data archiving application is very important in multimedia system industry. Huge size of multimedia data that distributed in multi locations makes multimedia data archiving more complicated. This situation requires for efficient data archiving technique for providing effective multimedia data manipulation process.

In digital era, the modern multimedia production house has been looking for a new approach and flexible IT platform that can manage multimedia data archive system [2]. Cloud computing technology has been explored to address the aforementioned issue. Cloud technology is a computing trend that provide elasticity, scalability, on-demand capabilities, effective cost and efficient delivery performance of services [4, 5]. Web services can be considered as an important emerging technology that is trending used and evolving in many information application developments [6, 7]. Web services technology provides the dynamically of the composition less changing and stability forms of programming. This supports the development of an application that can be acted as a service. A complex business process (e.g. large multimedia production industry) need to be accomplished by a collaborative web service in which several constituent services interact with each other and work in dynamic cycle [8, 9]. By using the service-oriented concept, it also allows resources to be accessed collaboratively guided by some guidelines or privacy rules [9].

In this paper, we describe the development of temporal based multimedia data archiving application under cloud computing platform. The paper is organized as follows. The developed of temporal based multimedia data archive application is discussed in Sect. 2. This section explains the concept of how time elements are embedded into multimedia data archive model. Section 3 presents the design and deployment of the application in cloud environment. Last section will conclude the research work.

2 Temporal Based Multimedia Data Archive Model

The developed temporal-based multimedia data archive model consists of two main components: *archiving object* and *data manager*. Here, the model applied *relational database management system* (RDBMS) and *extensible markup language*-based data storage (XML).

2.1 The Service Design

The concept that we term as *application as a service* is applied to all primitive operations involved in multimedia data archiving process. Particularly, the elements of SaaS that are used with the dynamic alignment and binding of elementary services at the time of archiving process. The context of the proposed service model is classified into two main operations and can be defined as: *service* and *query*. If a service set S containing the services and an ontology domain T then,

- Each service $s \in S$ can be defined as:
A service $s \in S$ is composed of a set of essential input $I_s \subseteq T$, a set of generated outputs $T_s \subseteq T$, a set of native data used by the operation of s $D_s \subseteq T$ and a set of primitive operations, O^s offered by s to be performed on the distributed multimedia archive resources over the Internet, where each primitive operation, O^s denotes an interaction between the service and query.
- Submitted query, q can be defined as:
 q consists of a set of provided inputs $I_q \subseteq T$, a set of preferred outputs $T_q \subseteq T$ and a set of required primitive operations O^q .

By adapting the common multimedia data archive application, a web services manager has been developed which manage four main services {*web respond, transaction log, crawler engine, data integration*}.

2.2 Time Granularities in Multimedia Data Archive

In the proposed model, the relational data scheme for multimedia data archive is extended by embedding time elements as additional attributes. The conventional data scheme will be associated with a temporal context to common events such as period of valid data and transaction time of ordering. The organization with huge transaction of multimedia data requires an explicit or implicit temporal context, and it can be expressed in terms of an appropriate time granularity.

A temporal based information system will associate with real-world facts, events and be able for query reasoning [10, 11]. Therefore, information system engineer needs to develop a formal characteristic of time granularities for defining its

relations. The theory of time-element is divided into two categories: intervals and points [3]. If T is denoted a nonempty set of time-elements and d is denoted a function from T to \mathbb{R}^+ which is the set of nonnegative real numbers then: t is classified as an interval if $d(t) > 0$ and otherwise, t is classified as a point. Hence, the set of time-elements, T , can be expressed as $T = I \cup P$, where I is the set of intervals and P is the set of points. Temporal management aspects of any transaction in database are as follows [3, 12]:

- The capability to detect change in a specific object over a certain period of time.
- The use of data to conduct analysis of past events such as the change of valid time for the object due to any event.
- To keep track of all the transactions status on the object life cycle.

In this proposed model, two (2) elements of time is considered: transaction time (tt) and valid time (vt). Time element unit is considered in the format of [day/month/year]. In this context, transaction time represents the transaction date when multimedia data is recorded (write) into the database. The transaction date is recorded during read, edit and delete operation. Valid time represents the valid period for the multimedia objects stored in the database. This valid period is changed when the object stored in a database is modified or edited. In the model, valid time involves a time interval, and can be categorized into two (2) different attributes known as valid-from and valid-until, $vt = [vt\text{-from}, vt\text{-until}]$.

In a dynamic multimedia data archive, the transaction of objects can be viewed as: Each multimedia data that gone through a set of modification process, $U = \{u_1, u_2, \dots, u_n\}$ will have a set of versions which can be noted as:

$$M = \{m_1, m_2, \dots, m_n\} \text{ then} \quad (1)$$

$$\forall m_i \in M \Leftrightarrow \exists (u_i \in U) \text{ and} \quad (2)$$

$$\forall m_i \in M \Rightarrow \exists (vt\text{-from} \in I \cup vt\text{-until} \in I) \quad (3)$$

where $vt\text{-from} < vt\text{-until}$, meaning that each multimedia data stored in the temporal data archive would have a valid time (which is classified as a time interval), and:

$$\forall u_i \in U \Rightarrow \exists (tt \in P) \quad (4)$$

meaning that each updating process of multimedia data in a data archive would have a transaction time (which is classified as a time point).

If we have a multimedia data archive that contains a set of multimedia data signed as, $M = \{m_1, m_2, \dots, m_n\}$ then the complete model for temporal based multimedia data archive is:

$$\text{TEMPORAL}(m_i \in M) \subseteq (tt \cap vt) \quad (5)$$

where tt is transaction time and vt is valid time.

Thus, if the multimedia data archive has a set of features attributes A_i then a complete scheme for a temporal based multimedia data management can be signed as:

$$R = (A_1, A_2, \dots, A_n, tt, vt\text{-from}, vt\text{-until}) \quad (6)$$

3 The Application Deployment

A public cloud platform is used to deploy the prototype of the proposed model. The *CodeIgnitor* PHP framework and Java API for web services (JAX-RS) are employed to develop and execute the functions in the prototype.

3.1 The Architecture

The proposed application model consists of three layers that are connected and interacted each other: *user layer*, *task layer* and *technology layer*. The user layer consists of dashboard and interface for cloud manager and consumers to interact with the prototype services. Cloud manager performs service management, device management and authorization. The task layer is referred to modules of the application including primitive operation, select transaction, searching process, data formatting, data integration and reporting. Meanwhile, technology layer represents a platform for deploying the application which consists of devices, cloud infrastructure and network devices. Figure 1 shows the layers of application architecture.

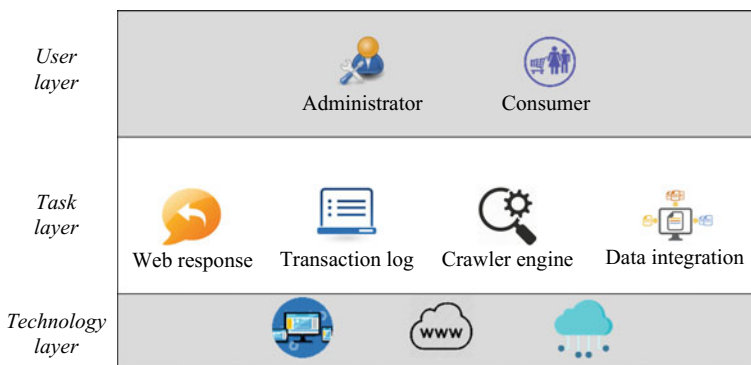


Fig. 1 The layers of application architecture

3.2 The Prototype

The prototype is developed based on the concept of service-oriented architecture. Service oriented architecture offers benefits of dynamic discovery and binding to services on a per-use basis and creation of sub-services based on existing services. Figure 2 depicts the architecture of application prototype. In this concept, there are three roles can be involved in the application: *<service provider, service consumer, service broker>* [13]. In this model, the developed temporal-based multimedia data archive engine is considered as a service broker. The broker has the ability to register any services provided by multimedia data provider and provide the integration repository multimedia data (the archive) to consumer. Service providers are the sites that offer multimedia datasets and its transaction to be registered in service broker database. Service consumers can be considered as a group of users that registered with service broker and allowed to use all primitive operations that are defined in broker’s application. The services including document, data and functions offered by service broker and service providers are described by web services description language (WSDL).

Cloud computing technology offers several layers of services including physical resource layer, data resource layer and application service layer [14]. In the cloud-based implementation phase, the model application is developed under the *application service layer*. This layer provides the software service platform including other supporting software, data integration module and database. The operations of updating and reporting are based on *keyword-search* and *similarity-search* that registered in service provider database. The consumer can explore multimedia data based on keyword. A simple object access protocol

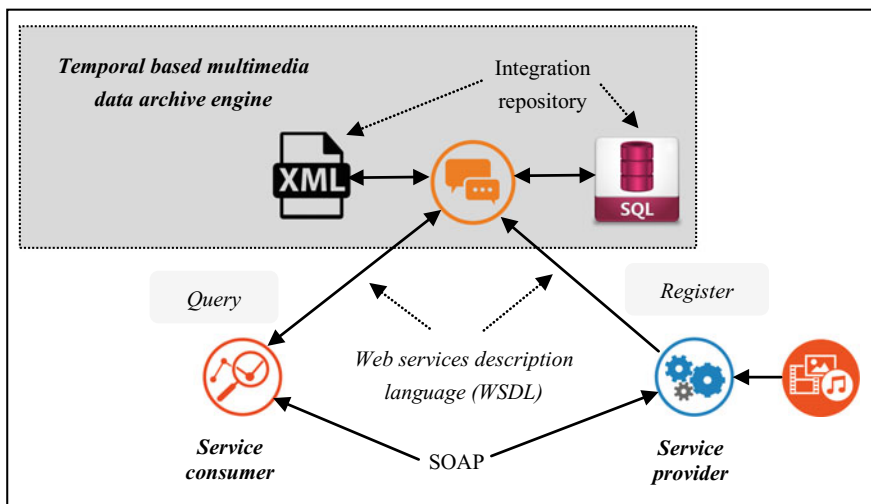


Fig. 2 Service oriented application architecture

(SOAP) is developed for communication purposes between service consumer and service provider. Technically, the service communication uses XML format for messages exchange and HTTP as application layer protocol.

4 Conclusions

The main functional components in multimedia data archive discussed in this paper are web response, transaction log, crawler engine and data integration. These functional components are propositioned as a service concept in the developed model. A temporal based data management is embedded into the model for providing an effective data management and archiving. The prototype of the model is developed based on the cloud computing platform.

Acknowledgements This study is funded by University of Sultan Zainal Abidin (UniSZA), Code: R0004-R274. Special thanks to Faculty of Informatics & Computing, UniSZA for all the supports in this research project.

References

1. Klepac, G.: Integration of different analytical concepts on multimedia contents in service of intelligent knowledge extraction. In: *Artificial Intelligence: Concepts, Methodologies, Tools and Applications*, vol. 4, pp. 2493–2522 (2016)
2. Hung, J.C., Takaziwa, M., Chen, S-C.: Large-scale multimedia data management: techniques and applications. *Multimed. Tools Appl.* **75**(23), 15341–15346 (2016)
3. Farham, M., Rahman, M.N.A., Yuzarimi, M.L., Saiful, B.M.: Managing multimedia data: a temporal-based approach. *Int. J. Multimed. Ubiquitous Eng.* **7**(4), 73–85 (2012)
4. Al-Qurishi, M., Al-Rakhami, M., Al-Rubaian, M., Alamri, A.: A framework of knowledge management as a service over cloud computing platform. In: *Proceedings of the Intelligent Information Processing, Security and Advanced Communication* (2015)
5. Duggal, S., Sharma, M.K.: Proposed framework of e-learning services through cloud. In: *Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies* (2016)
6. Purohit, L., Kumar, S.: Web service selection using semantic matching. In: *Proceedings of the International Conference on Advances in Information Communication Technology and Computing* (2016)
7. Khabou, I., Rouached, M., Abid, M., Bouaziz, R.: Enhancing web services compositions with privacy capabilities. In: *Proceedings of the 17th International Conference on Information Integration and Web-based Applications and Service* (2015)
8. Vathsala, A.V., Mohanty, H.A.: Survey on checkpointing web services. In: *Proceedings of the 6th International Workshop on Principles of Engineering Service-Oriented and Cloud Systems*, vol. 2014, pp. 11–17 (2014)
9. Huang, C-Y., Wu, C-H.: A web service protocol realizing interoperable internet of things tasking capability. *Sensors* **16**(9), 1–23 (2016)
10. Hagedon, S., Rath, T.: Efficient spatio-temporal event processing with SPARK. In: *Proceedings of the 20th International Conference on Extending Database Technology*, pp. 570–573 (2017)

11. Petkovic, D.: Temporal data in relational database systems: a comparison. In: Rotcha, A. et al. (eds.) *New Advances in Information Systems and Technologies, Advances in Intelligence Systems and Computing*, pp. 13–23 (2016)
12. Wieland, M., Pittore, M.: A spatio-temporal building exposure database and information life-cycle management solution. *Int. J. Geo-Inf.* **6**(4), 1–20 (2017)
13. Lin, M., Cheung, W.: Automatic tagging web services using machine learning techniques. In: *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pp. 258–265 (2014)
14. Nordin, M., Rahman, A., Abdullahi, N.S., Fadzil, M.A.K., Syadiah, N.S., Syarilla, I.A.S.: A gamification model for resource sharing in Malaysian Schools using cloud computing platform. In: Maria, A. et al. (eds.) *Intelligent Systems Design and Applications*, pp. 406–416 (2017)

Chapter 9

Thalas Test Android Application



Iliana Mohd Ali and Nooraida Samsudin

Abstract Thalassemia is a genetic blood disorder. People with Thalassemia disease are not able to make enough hemoglobin, which causes severe anemia. The public knowledge about thalassemia disease is still lack. Therefore, Thalas Test is necessary to ensure that users are more aware of the symptoms of the disease. It provides knowledge and give basic diagnose of thalassemia disease. The objectives of the study are to develop Thalas Test Android Application and to construct online questionnaires regarding thalassemia symptoms by adding the technique of rule-based system that is used in Artificial Intelligence (AI). Thalas Test will display a series of questions related to the disease. The user will answer the questions. Thalas Test will then produce the analysis of results. The methodology used is waterfall. Thalas Test can help the user check before seeing the doctor. If affected by the disease symptoms, patients can refer to the nearest hospital for further treatment.

Keywords Thalassemia · Questionnaires · Rule-based system · Artificial intelligence

1 Introduction

Gartner, Inc. has identified ‘Mobile Health Monitoring’ as the fifth among the ‘Top 10 Consumer Mobile Applications for 2012’ [1]. This list is supported by the ‘Top 10 Strategic Technology Trends for 2014’ which includes ‘Mobile Apps and Applications’ [2]. Analysts expect global mobile health market’s value will increase to \$11.8 billion by 2018 [3].

I. M. Ali (✉) · N. Samsudin
Faculty of Computer, Media and Technology Management,
TATI University College, 24000 Teluk Kalong, Kemaman, Terengganu, Malaysia
e-mail: iliana@tatiuc.edu.my

N. Samsudin
e-mail: nooraida@tatiuc.edu.my

Thalassemia is a genetic blood disorder. People with Thalassemia disease are not able to make enough hemoglobin, which causes severe anemia. The public knowledge about thalassemia disease is still lack. Therefore, Thalas Test is necessary to ensure that users are more aware of the symptoms of the disease. The objectives of the study are to develop Thalas Test Android Application and to construct online questionnaires regarding symptoms by adding the technique of rule-based system that is used in Artificial Intelligence (AI). Thalas Test will display a series of questions related to the disease. The user will then answer the questions. Thalas Test will then produce the analysis of results.

This paper is organized into several sections. Section 2 discusses the related work focusing on various applications. Section 3 presents the design and methodology involved including the flowchart. Section 4 presents the user interfaces and evaluation for Thalas Test. Finally, the work of this paper is summarized in the final section.

2 Related Work

In order to develop Thalas Test Android Application, few related works have been reviewed. The related works are summarized as in Table 1, starting from the earlier published work to the most recent ones.

Table 1 Past researches on health questionnaire

Source	Technique	Feature used	Domain	Disadvantage/ Advantage	Future direction
Kelly, et al. [4]	Accelerometer, gyroscope, sensor, SVM regression models	Current activity, daily activity, health status questionnaire	Health	Inexpensive, can track health status on an ongoing basis	Prediction accuracy should be further improved before use in clinical setting
Levin, et al. [5]	Online questionnaire	Sophisticated prompting, secure database integration, push notification, reminder	Health behavior	Thorough conduct of questionnaire	

3 Methodology

This research adopts the steps of waterfall model which progress from one phase to another linearly, as illustrated in Fig. 1 [6].

Waterfall model was chosen because parts of the application are generally well understood. It can be observed from Fig. 1 that the study commenced with stage 1—the requirements specification. User and application requirements need to be gathered in order to obtain clear picture pertaining to the specific features of the application. Table 2 shows the relationship between user and the application requirements.

Next, it is followed by the second stage which involves designing the workings Thalass Test Android Application and the software.

Then, the third stage of Thalass Test application development (Coding) was the implementation stage. Testing was performed concurrently with programming of the application. The overview of Thalass Test operation is displayed in Fig. 2. Ten inputs are demanded from the users which are questions related to users and parents. Those questions are gathered from a doctor of Hospital Sultanah Nur Zahirah, Kuala Terengganu. Once this information is inserted in the application, user’s total of yes response will be calculated. As a result, either user is suffering from Thalassaemia disease or not will be displayed.

Afterwards, stage four involves testing Thalass Test Application before it can be fully utilized by users. Thalass Test function must be tested to ensure that it is error

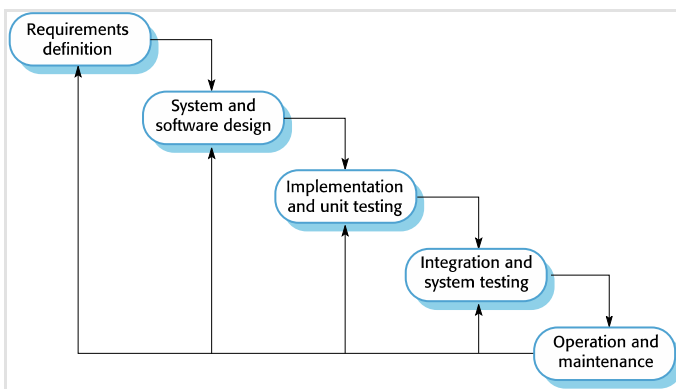


Fig. 1 Waterfall model steps

Table 2 User requirements and application requirements

No	User requirements	Application requirements
1.	Key-in questions response	Questions response either YES or NO
2.	View results	Either suffering from Thalassaemia or not

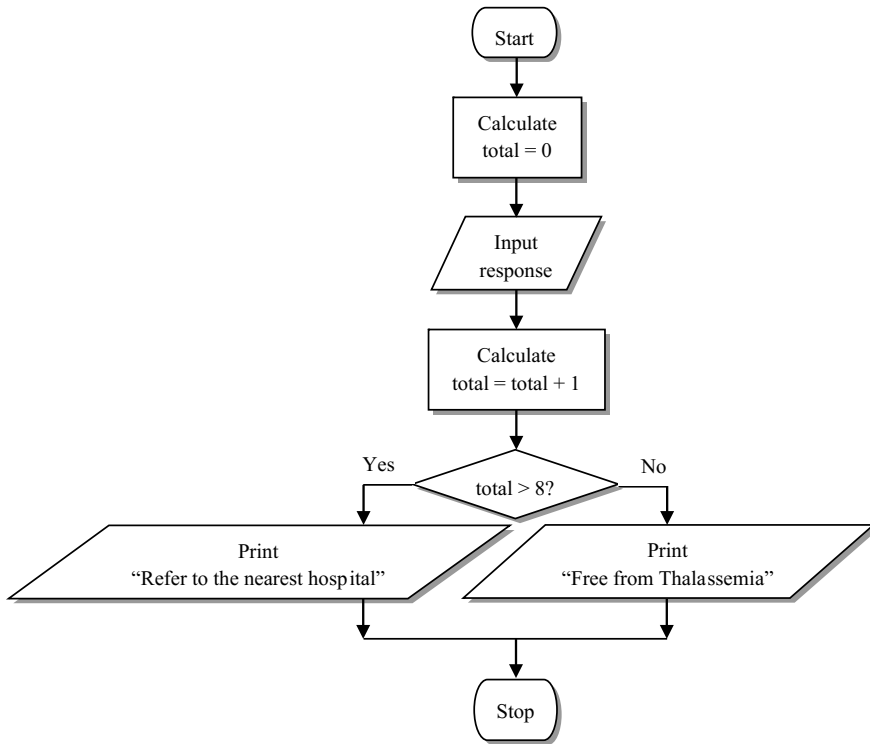


Fig. 2 Overview of Thalas Test Operation

free and the end result meets user requirements as determined earlier in the study. Finally, stage five requires that the developer to perform frequent operation and maintenance so that Thalas Test keeps on functioning as its best ability.

4 Results

As shown in Fig. 3, the overall test operation of Thalas Test application had indicated a successful outcome in designing and developing the application. Figure 3a shows Thalas Test interface requesting questions responses regarding user's and parents. In Fig. 3b and c, once the RESULT button was pressed, Thalas Test displayed either user is free from Thalassemia or the user may be suffering from Thalassemia and need assistance from the nearest hospital.

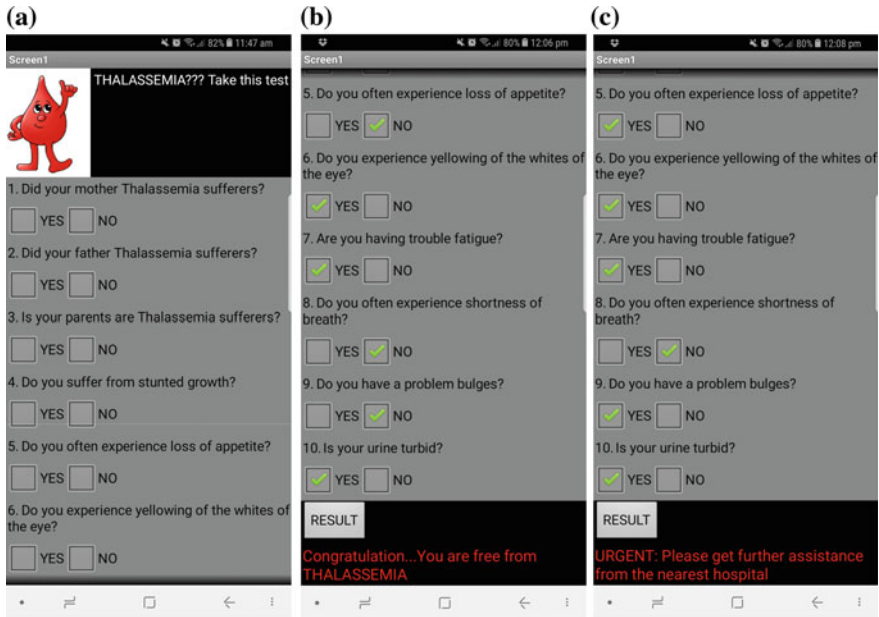


Fig. 3 a–c Thalas Test interfaces

4.1 User Evaluation

To obtain data on users’ opinion about the application’s ease of use aspect, the instrument—questionnaire was administered. A total of 30 respondents participated in users’ evaluation process. The respondents of the study were the users who had

Table 3 Demographic profile

Question	Item	Frequency	Percentage (%)
Age	18–20	16	53
	21–23	14	47
	24–26	0	0
	27 and above	0	0
Gender	Male	22	73
	Female	8	27
Level of education	Diploma	15	50
	Degree	15	50
	Master	0	0
	Ph.D.	0	0
Experience using similar app	Yes	23	77
	No	7	23

tested the application for the first time. These users were requested to fill out the questionnaire upon the completion of the application testing and rate their experience in using the application. Data were then analyzed and tabulated as in Table 3.

Table 4 and Fig. 4 show the overall results on the evaluation for Thalaz Test Application from the aspect of ease of use. It was discovered that mode for user feedback regarding the ease of use is 3 and 4 and most users are agreeing for ease of use for this application. The range for average based on user feedback is 3.20–3.60 with the standard deviation range from 0.606 to 0.802. The highest average is for item a based on user feedback is 3.60 (SD = 0.675). The lowest average is for item e which is 3.20 (SD = 0.714).

Table 4 Results of Ease of Use Construct

No	Item	Min	Max	Mode	Average	Standard deviation
a.	The app is easy to use	1	4	4	3.60	0.675
b.	It is simple to use	1	4	4	3.54	0.681
c.	It is user friendly	1	4	4	3.40	0.770
d.	It requires the fewest steps possible to accomplish what I want to do	1	4	3	3.33	0.802
e.	It is flexible to use	1	4	3	3.20	0.714
f.	I don't notice any inconsistencies when I used it	2	4	3	3.33	0.606

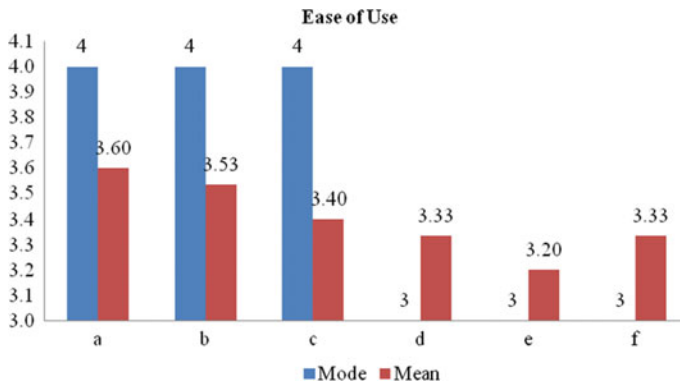


Fig. 4 Ease of Use Graph

5 Conclusion

An android application that calculates percentage of Thalassemia possibility is needed by the health conscious community. With the development of Thalass Test Android Application, the researchers are able to assist particularly people as they are either free from Thalassemia or need assistance from the nearest hospital.

References

1. Gartner Identifies the Top 10 Consumer Mobile Applications for 2012. <http://www.gartner.com/newsroom/id/1230413>
2. Gartner Identifies the Top 10 Strategic Technology Trends for 2014. <http://www.gartner.com/newsroom/id/2603623>
3. Greenspun, H., Coughlin, S.: mHealth in an mWorld How mobile technology is transforming health care. Deloitte Development LLC, Washington (2012)
4. Kelly, D., Curran, K., Caulfield, B.: Automatic prediction of health status using smartphone derived behaviour profiles. *IEEE J. Biomed. Health Inform.* **99**, 1–10 (2016). <https://doi.org/10.1109/JBHI.2017.2649602>
5. Levin, M.E., Pierce, B., Schoendorff, B.: The acceptance and commitment therapy matrix mobile app: A pilot randomized trial on health behaviors. *J. Context. Behav. Sci.* **6**, 268–275. <https://doi.org/10.1016/j.jcbs.2017.05.003>
6. Sommerville, I.: *Software Engineering*. Pearson Education, Essex (2016)

Chapter 10

Dimensions of Mobile Information Behavior



Zuraidah Arif, Abd Latif Abdul Rahman
and Asmadi Mohamed Ghazali

Abstract The aim of this study was to develop and validate new dimension of mobile information behavior for Small Medium Enterprise (SME) workers in Malaysia. The items of this new developed scale were derived from previous study on information behavior. SME managers were participating in this study. A Kaiser-Meyer-Olkin (KMO) show that samples met the factor analysis criteria adequacy was 0.79 which indicates the appropriateness of the input in factor analysis. In addition Bartlett test was statistically significant with 0.000. The principal components method with varimax rotation was conducted to extract common factors. Item with factor loading >0.5 were selected to ensure a stable factor structure with adequate sample size. Factor extracted based on eigen-values greater than 1. Item with factor loading less than 0.5 and cross loading were removed. All items group into four dimensions are acquired based on the result of the rotated component matrix. Validity for the new dimension of mobile information behavior was found. Four dimensions were categorized as experiential need, social functional need, physical intellectual access, and social access. Result support validation of mobile information behavior dimensions for Malaysia SME workers.

Keywords Mobile information behavior · Factor analysis

Z. Arif (✉) · A. L. A. Rahman · A. M. Ghazali
Universiti Teknologi MARA Caw. Bedong, Kedah, Malaysia
e-mail: zuraidah_arif@kedah.uitm.edu.my

A. L. A. Rahman
e-mail: ablatif@kedah.uitm.edu.my

A. M. Ghazali
e-mail: asmadi615@kedah.uitm.edu.my

1 Introduction

Information behavior caused by needs that exists from people work task, or curiosity on uncertainty things or matter. Information behavior may affected by the information user and provider context (i.e. environmental and personal). In environmental context includes location, culture and social influences, activity-related, financial constrain, and technology [1]. With the increasing number of mobile device user for instance in year 2016 almost 1500 million units of Smartphone were sold to consumers [2].

However, there is limited study was done in studying dimension structure of mobile information behavior. Popularity of mobile device (i.e. Smartphone and tabs) is increased since the proliferating of mobile social application. Searching information using mobile device is the faster's way to fulfill information need. The action of searching information using mobile may related with principal of least effort by Carter, 1962 as stated in [3]. Needs to Smartphone may derived from various function and services offered from the Smartphone itself. As stated by [4] "user attachment to Smartphone contributes a great deal to shaping the experiential value derived from using value-added mobile services". Information need may originate with the need in seeking answers. According to [5] information seeking behavior is the intention and action of seeking answers for accurate information in order to stratify some goals, which user may interact with a database or World Wide Web.

The use of mobile phones creates social stimulation for the user [6]. Currently, with the function and services offered by Smartphone, user can easily socialize virtually. On the other hand, functional value is the perceived utility acquired from an alternative's capacity for functional and practical performance [7]. Functional value in the context of Smartphone can be interpreted as the value gained from being able to use mail, banking or gaining access to public transportation information in Smartphone amongst others [8]. The characteristics (i.e. customization (to attain personal information), convenience (easy access to information and services [9], and content (to retrieve updated and in-depth information)) of application on mobile device not only enabling its user to socialize, but also to perform other activity seamlessly including searching, access and retrieval, and use information.

In the remote area, mobile phone is the tool for exchanging information in supporting livelihood. As stated, [10] that phone-based application also important in facilitating access to information and resources, for example phone-based radio allow user to listen to the news or memory card to store data. Moreover, Smartphone can act as an intermediary between potential users and the information itself [11]. The term access itself refers to a "source pertinent to an inquiry, to comply with the evidence that result in acquiring the knowledge desired" [12]. In accessing information, user need physical device that can hold and provide ways to retrieve the information [13]. According to [14] in his study stated the feature of a Smartphone that regularly in use by respondents are 97% use it for text messaging, 92% use for voice or video calls, 89% use the internet, 88% use for emailing, and 75% use for social networking. It is important to well categorized, organized,

displayed, and represented information for easy searching and access. Searching skill is an advantage in enabling information from various formats able to be access. Therefore, by learning intellectual access, is enable to disclose the appropriate ways of making information accessible whenever users retrieved the information and it brings the request and the information in an efficient manner through the representation of available information sources [13]. Since intellectual access dimensions include the quantity and readability of information provided, as well as possible connections between documents, knowing how to request records and how to pursue adjudication are also mechanism of intellectual access [15, 16].

2 Aim of Study

The aims of this study were to develop and validate a new scale for mobile information behavior.

3 Method

3.1 Design

The instrument contains 35 items with 7-point likert response scale. The highest scores [4] demonstrated the highest level of agreement of the item rated. The 35-item instrument consisted of four items for experiential need; nine items for social need; six items for functional need; five items for physical access; five items for intellectual access; and six items for social access.

3.2 Research Instrument

A survey using questionnaires had been administered in this research. The total subject in this study was 99 participants from SME industry. The questionnaire covers four dimensions. Firstly, experiential need which is related to user preferences to used Smartphone in fulfilling need. Secondly, social functional need is the encouragement to socialize and to perform other activity effectively. Thirdly, physical intellectual access is the ability of getting access to the computer system or mobile system in accessing information. Lastly, social access is based on the theory of normative behavior by Burnett et al., in which within specific contexts, information behavior is like day-to-day activities.

3.3 Data Collection

Participant working in SME industry were approached to participate in the study. Participants are ensuring that the data would only be used for educational purpose only. The participants are selected from Klang Valley in Malaysia. The questionnaire was provided in English version only. All questionnaires were completed by the participant that made up 100% response rate. The responses were coded according to scale measurement.

3.4 Data Analysis

Data analysis was performed using SPSS 23.0 for windows statistical software package. Descriptive analysis was used to summarize sample characteristics. To determine the construct validity of the mobile information behavior, content and construct validity was also performed to determined instrument validity. Cronbach's alpha coefficient, split-half coefficient, and item analysis verified the reliability of the instrument.

4 Results

4.1 Sample Characteristics

Total numbers of 99 participants working in SME industry were involved in this study. The participants involved in his study were in the age group 20–50 years old. 63% were male, and 37% female. 40% of the participants were in the executive position.

4.2 Validity of Mobile Information Behavior

Result from Kaiser-Meyer-Olkin (KMO) and Bartlett test shows that samples met the factor analysis criteria. The KMO measure on sampling adequacy was 0.79 which indicates the appropriateness of the input in factor analysis. In addition Bartlett test was statistically significant with 0.000. EFA established the construct validity of the mobile information behavior; it's to ensure that the scale could evaluate interaction efficiently. The principal components method with varimax rotation was conducted to extract common factors. Item with factor loading >0.5 were selected to ensure a stable factor structure with adequate sample size. Factor

extracted based on eigen-values greater than 1. Item with factor loading less than 0.5 and cross loading were removed.

4.3 Factor Analysis Result of Mobile Information

Four dimensions are acquired based on the result of the rotated component matrix. Item were arranging into dimensions based on the size of loading with respect to statistical analysis (Table 1). The first dimension consist of fourteen items that mostly associated with dimension entitle social functional need (SFN). Most items were initially from the group social need and six items from functional need. The

Table 1 Correlations of each dimension item with their own scale and with other scales

Item	Dimensions			
	1	2	3	4
<i>Experiential Need (EN)</i>				
I need mobile information to fulfill daily need				0.753
I need mobile information to solve problem in daily routine				0.728
<i>Social Functional Need (SFN)</i>				
I need mobile information that could escape myself from loneliness	0.769			
I need mobile information that could escape myself from alienation	0.782			
I need mobile information that could give myself love from others	0.788			
I need mobile information that could stabilize myself	0.799			
I need mobile information that could give me self-respect of others	0.786			
I need mobile information that could give me the location of myself	0.836			
I need mobile information that could give me the direction of my ways	0.795			
I need mobile information that could give me the location of my friends	0.731			
I need mobile information to fulfill their daily entertainment need (Enjoyment)	0.726			
I always find out at least one mobile information that's needed	0.761			
I am aware of mobile information needs	0.748			
I will seek for mobile information	0.813			
My organization motivates staff to use mobile devices for seeking information to fulfill their daily need	0.852			
My organization encourages staff to use the mobile device as a fast way to seek answers for their information need	0.846			

(continued)

Table 1 (continued)

Item	Dimensions			
	1	2	3	4
<i>Physical Intellectual Access</i>				
I have no problem buying mobile information devices		0.856		
I have no problem to access mobile information		0.911		
My mobile devices are always with me		0.900		
My mobile devices are always connected to the Internet		0.861		
I have access to a mobile device most of the times went to seek answers to my question		0.835		
I have the knowledge to access mobile information		0.764		
I have the skills to access mobile information		0.780		
I have all capabilities needed to access mobile information		0.719		
<i>Social Access</i>				
My friends share mobile information to be used in my daily life			0.506	
My officemates share mobile information to be used in my daily life			0.764	
My Superiors share mobile information to be used in my daily life			0.786	
Extraction Method: Principal Component Analysis				
Rotation Method: Varimax with Kaiser Normalization				
a. Rotation converged in 6 iterations				

second dimension consist of eight items associated with dimension entitle physical intellectual access (PIA). Most items were initially from the group physical access and three items from intellectual access. The third dimension identified as social access (SA). The third consist three items. All item from group social access. The fourth dimension consists of two items and was name experiential need (EN).

4.4 Reliability

Cronbach’s alpha coefficients and split-half coefficients were conducted to express the internal reliability of the instrument. The alpha’s for each experiential need, social functional need, physical intellectual access, and social access was 0.97,

Table 2 Reliability test

Variable	Cronbach’s alpha
Experiential needs	0.977
Social and functional needs	0.986
Physical and intellectual access	0.980
Social access	0.995

0.98, 0.98, and 0.99. Coefficients alpha reach the internal consistency estimates of reliability for these four subscales. Based on the item analysis one item need to be delete in social access dimensions “my friends share mobile information to be used in my daily life” to achieved alpha 0.99. All items were relevant to construct the content of the mobile information behavior. The result was shown in Table 2 and supports the measure’s reliability.

5 Conclusion

Result from this study indicated that the mobile information behavior dimension has high-quality reliability and validity to assess the core value for mobile information behavior. Furthermore, regarding the purpose of identifying dimension structure of set of variable, mobile information behavior offers opportunities to generalize the scale other population [17].

In previous study information need [18, 19] and information access [20] are dimensions in information behavior, this study proved otherwise, and respondent responded differently regarded mobile information behavior.

Current study discovers that experiential need, social functional need, physical intellectual access, and social access are dimensions in mobile information behavior. The identification of mobile information behavior dimensions contributes to the body of knowledge.

Acknowledgements This research was funded by the Research Acculturation Grant-RAGS-600-RMI/RAGS/5/3 (103/2014) managed by the Research Management Institute (RMI), UiTM. The authors wish to thank the Ministry of Education, Malaysia and Universiti Teknologi MARA (UiTM) Kedah, Malaysia for the support in this study.

References

1. Robson, A., Robinson, L.: Building on models of information behavior: linking information seeking and communication. *J. Doc.* **69**(2), 169–193 (2013)
2. The Statistic Portal (2017)
3. Faibisoff, S.G., Ely, D.P.: Information and information need. *Information Report and Bibliographies*, 5(5) (1974)
4. Tojib, D., Tsarenka, Y., Sembada, A.Y.: The facilitating role of smartphones in increasing use of value-added mobile services. *New Media Soc.* **17**(8), 1220–1240 (2015)
5. Ellis, D., et al.: Information seeking and mediated searching. Part 5. User–intermediary interaction. *J. Am. Soc. Inf. Sci. Technol.* **53**(11), 883–893 (2002)
6. Grant, I., O’Donohoe, S.: Why young consumers are not open to mobile marketing communications. *Int. J. Advert.* **26**(2), 223–246. <http://dx.doi.org/10.1080/10803548.2007.11073008> (2007)
7. Sheth, J.N., Newman, B.I., Gross, B.L.: Why we buy what we buy: a theory of consumption values. *J. Bus. Res.* **22**, 159–170 (1991)

8. Jarenfors, O.A., Stureson, S.H.: Value creation through smartphones: an ethnographic study about consumer value and social interaction through smartphones. Master thesis-marketing and consumption (2012)
9. Balasubramanian, S., Peterson, R.A., Jarvenpaa, S.L.: Exploring the implications of m-commerce for markets and marketing. *J. Acad. Mark. Sci.* **30**(4), 348–361 (2002)
10. Baird, T.D., Hartter, J.: Livelihood diversification, mobile phones and information diversity in Northern Tanzania. *Land Use Policy* (2017) <http://dx.doi.org/10.1016/j.landusepol.2017.05.031>
11. Mathiesen, K., Fallis, D.: Information ethics and the library profession. In: Himma, K.E., Tavanni, H.T. (eds.) *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Hoboken, NJ (2008)
12. Buckland, M.K.: Information as thing. *JASIS* **42**(5), 351–360 (1991)
13. Jaeger, P.T., Bowman, C.A.: *Understanding Disability: Inclusion, Access, Diversity, and Civil Rights*. Praeger, Westport, CT (2005)
14. Smith, A.: U.S. smartphone use in 2015. Pew Research Center (2015)
15. Grunewald, M.H.: E-FOIA and the “mother of all complaints:” Information delivery and delay reduction. *Adm. Law Rev.* **50**(2), 345–369 (1998)
16. Tankersley, M.E.: How the electronic freedom of information act amendments of 1996 update public access for the information age. *Adm. Law Rev.* **50**(2), 421–458 (1998)
17. Hair, J.F., et al.: *Multivariate data analysis: a global perspective*. Pearson, Upper Saddle River (1998)
18. Wilson, T.D.: On studies and information needs. *J. Doc.* **37**(1), 3–15 (1981)
19. Leckie, G.J., Pettigrew, K.E., Sylvain, C.: Modeling the information seeking of professionals: a general model derived from research on engineers, health care professionals and lawyers. *Libr. Q.* **66**(2), 161–193 (1996)
20. Oliver, R.: Interactive information systems: information access and retrieval. *Electron. Libr.* **13**(3), 187–194 (1995)

Part II
Mathematics

Chapter 11

A Promising Method to Approximate Fractional Derivatives Under Uncertainty



Ali Ahmadian, Norazak Senu, Fudziah Ismail and Soheil Salahshour

Abstract In this work, we apply a new and promising method base on tau method for solving a variety of differential equations of fractional order under fuzzy concept. We employ a linearization method to approximate the fractional derivative of the Caputo-type under uncertainty, then, we get to a fuzzy algebraic linear system and we solve it using any type of numerical technique to achieve the solution. The algorithm handles the problem in a direct manner without any need to restrictive assumptions. We emphasize the power of the method by applying it to an example.

Keywords Fractional differential equations · Caputo derivative · Tau method · Fuzzy settings theory

1 Introduction

Fractional calculus seamlessly generalizes the notion of standard integer-order calculus to its fractional-order counter- part, leading to a broader class of mathematical models. Spectral methods have been discovered exceptionally powerful devices for tackling numerous sorts of fractional differential equations (FDEs), which have roused many authors to employ them for these types of equations. The operational

A. Ahmadian (✉) · N. Senu · F. Ismail
Institute for Mathematical Research, Universiti Putra Malaysia,
43400, Selangor, Serdang, Malaysia
e-mail: ahmadian.hosseini@gmail.com

N. Senu
e-mail: norazak@upm.edu.my

F. Ismail
e-mail: fudziah@upm.edu.my

S. Salahshour
Young Research and Elite Club, Mobarakeh Branch, Islamic Azad University,
Mobarakeh, Iran
e-mail: soheilsalahshour@yahoo.com

matrix of the Legendre polynomials was investigated by [1, 2] and considered for numerical solution of different types of FDEs subject to initial conditions, which was extended by Kazem et al. [3] who utilized fractional Legendre orthogonal polynomials for solving FDEs. Afterwards, Doha et al. [4] presented robust Chebyshev spectral algorithms for solving linear and nonlinear FDEs. Subsequently, Bhrawy et al. [5] employed a technique to approximate multi-term FDEs with variable coefficients using a quadrature shifted Legendre tau method. Ultimately, this procedure was continued by several authors [6, 7].

The motivation of this research is to extract an explicit algorithm for fuzzy fractional-order derivative of shifted Jacobi polynomials of any degree in terms of shifted Jacobi polynomials themselves; in the fuzzy Caputo sense with order ($0 < \nu \leq 1$). Also, we focus on the simplified solution technique for solving the fuzzy fractional differential equation (FFDE).

In this paper, the discussed FFDE is approximated based on the shifted Jacobi polynomials presented in [7], then matrix operations are replaced with the fractional derivatives in the residual form of the problem in which we utilize the shifted Jacobi spectral tau (JST) method to build the spectral solution for FDEs under uncertain conditions.

2 Basic Definitions

Definition 2.1 Let us specify by \mathfrak{R}_F the class of fuzzy subsets of the real axis satisfying the following properties:

- i. q is normal, i.e., there exists $p_0 \in \mathfrak{R}$ such that $q(p_0) = 1$,
- ii. q is a convex fuzzy set (i.e.)

$$q(\vartheta a + (1 - \vartheta)b) \geq \min\{q(a), q(b)\},$$

$$\forall \vartheta \in [0, 1], a, b \in \mathfrak{R}.$$

- iii. q is upper semi-continuous on \mathfrak{R}_F .
- iv. $\text{cl}\{a \in \mathfrak{R} | q(a) > 0\}$ is compact where cl denotes the closure of a subset.

Then \mathfrak{R}_F is called the space of fuzzy numbers. Clearly $\mathfrak{R} \subset \mathfrak{R}_F$. $[q]^\alpha = \{a \in \mathfrak{R} | q(a) \geq \alpha\}$ is specified for $0 < \alpha \leq 1$ and $[q]^0 = \{a \in \mathfrak{R} | q(a) > 0\}$

The notation $[q]^\alpha = \left[\begin{matrix} q^+ \\ - \\ q^- \end{matrix} \right]$ denotes the α -level set of q .

Definition 2.2 Let $\Pi : I \rightarrow \mathfrak{R}_F$, $t_0 \in I$. We say Π is differentiable at ρ_0 , if there exists an element $\Pi'(\rho_0) \in \mathfrak{R}_F$ such that

- (1) For all $h > 0$ sufficiently close to 0, there exist $\Pi(\rho_0 + h)\Theta\Pi(\rho_0)$ and $\Pi(\rho_0)\Theta\Pi(\rho_0 - h)$ and the limits (in metric D)

$$\lim_{h \rightarrow 0^+} \frac{\Pi(\rho_0 + h)\Theta\Pi(\rho_0)}{h} = \lim_{h \rightarrow 0^+} \frac{\Pi(\rho_0)\Theta\Pi(\rho_0 - h)}{h} = \Pi'(\rho_0),$$

or

- (2) For all $h > 0$ sufficiently close to 0, there exist $\Pi(\rho_0)\Theta\Pi(\rho_0 + h)$ and $\Pi(\rho_0 - h)\Theta\Pi(\rho_0)$ and the limits (in metric D)

$$\lim_{h \rightarrow 0^-} \frac{\Pi(\rho_0)\Theta\Pi(\rho_0 + h)}{-h} = \lim_{h \rightarrow 0^-} \frac{\Pi(\rho_0 - h)\Theta\Pi(\rho_0)}{-h} = \Pi'(\rho_0).$$

Jacobi polynomials: The analytic form of the shifted Jacobi polynomials $P_i^{(\gamma, \xi)}(z)$ of degree i is acquired by

$$P_i^{(\gamma, \xi)}(x) = \sum_{k=0}^i (-1)^{i-k} \frac{\Gamma(i + \xi + 1)\Gamma(i + k + \gamma + \xi + 1)}{\Gamma(k + \xi + 1)\Gamma(i + \gamma + \xi + 1)(i - k)!k!} x^k$$

that

$$P_{0,i}^{(\gamma, \xi)}(0) = (-1)^i \frac{\Gamma(i + \xi + 1)}{\Gamma(\xi + 1)i!}, P_{1,i}^{(\gamma, \xi)}(1) = \frac{\Gamma(i + \gamma + 1)}{\Gamma(\gamma + 1)i!}.$$

A function g belonging to $L_w^2(\Omega)$ can be expanded by

$$g(x) = \sum_{i=0}^{+\infty} \mu_i P_i^{(\gamma, \xi)}(x),$$

where

$$\mu_i = \frac{1}{v_i^{\gamma, \xi}} \int_0^1 P_i^{(\gamma, \xi)}(x) g(x) w^{(\gamma + \xi)}(x) dx, \quad i = 0, 1, \dots$$

Practically, only the first $m + 1$ -terms shifted Jacobi polynomials are taken in counter. Hence, we have

$$\begin{aligned} g(x) &\cong g_m(x) = \sum_{i=0}^m \mu_i P_i^{(\gamma, \xi)}(x) = \Upsilon^T \Phi(x), \\ \Upsilon &= [\mu_0, \mu_1, \dots, \mu_m]^T, \\ \Phi(x) &= [P_0^{(\gamma, \xi)}(x), P_1^{(\gamma, \xi)}(x), \dots, P_m^{(\gamma, \xi)}(x)]^T. \end{aligned} \tag{1}$$

3 Solution Method

Lemma 3.1 Let $\Phi(x)$ be shifted Jacobi vector described in Eq. (1) and also suppose that $v > 0$. Then $D^v\Phi(x) \simeq D^{(v)}\Phi(x)$, in which $D^{(v)}$ is $(m + 1) \times (m + 1)$ plays role as the operational matrix of the Caputo-type derivative assumed in the problem and is characterized by:

$$D^{(v)} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \\ A_v(\lceil v \rceil, 0) & A_v(\lceil v \rceil, 1) & A_v(\lceil v \rceil, 2) & \dots & A_v(\lceil v \rceil, N) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ A_v(i, 0) & A_v(i, 1) & A_v(i, 2) & \dots & A_v(i, N) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ A_v(m, 0) & A_v(m, 1) & A_v(m, 2) & \dots & A_v(m, m) \end{pmatrix}$$

where

$$A_v(i, j) = \sum_{k=\lceil v \rceil}^i a_{ijk},$$

and a_{ijk} is given by

$$a_{ijk} = \frac{(-1)^{i-k} \Gamma(j + \xi + 1) \Gamma(i + \xi + 1) \Gamma(i + k + \xi + 1)}{v_i \Gamma(j + k + \gamma + \xi + 1) \Gamma(k + \xi + 1) \Gamma(i + \gamma + \xi + 1) \Gamma(k - v + 1) (i - k)!} \times \sum_{l=0}^j \frac{(-1)^{j-l} \Gamma(j + l + \gamma + \xi + 1) \Gamma(\gamma + 1) \Gamma(l + k + \xi - v + 1)}{\Gamma(l + \xi + 1) \Gamma(l + k + \gamma + \xi - v + 2) (j - l)!}$$

It is worthy to note that in $D^{(v)}$, the first $\lceil v \rceil$ rows, are all zeros.

In this study, we aim to find the numerical solution of the following mathematical model constructed based on FDE by employing the Jacobi polynomials with tau technique. Therefore, we have

$$\frac{d^v M}{dt} + \varpi_2 M = \varpi_1 H \exp(-\varpi_1 t), \quad 0 < v \leq 1 \tag{2}$$

where M is a fuzzy function and H is a fuzzy or crisp constant. Also, ϖ_1 and ϖ_2 are non-fuzzy constant

Now, by utilization of the spectral tau method prescribed in [1–3, 8] and Eq. (1) and Lemma 3.1 we approximate the solution of the Eq. (2) as follows:

$$\sum_{j=0}^m a_j \left[\left(D^{(\nu)} P_j^{(\gamma, \xi)}(x), P_k^{(\gamma, \xi)}(x) \right)_{w^{(\gamma, \xi)}} + \varpi_2 \left(P_j^{(\gamma, \xi)}(x), P_k^{(\gamma, \xi)}(x) \right)_{w^{(\gamma, \xi)}} \right] = \left(\varpi_1 M e^{-\varpi_1 t}, P_k^{(\gamma, \xi)}(x) \right)_{w^{(\gamma, \xi)}}, \quad k = 0, 1, \dots, m - 1, \tag{3}$$

where is the multiplication with respect to fuzzy notion. It is clear that the unknown fuzzy coefficients (a_j) are determined by solving this non-crisp linear algebraic system using an iterative method.

4 Numerical Results

In this section the technique prescribed in the last section is applied to extract the numerical solution of Eq. (2). Here, we assume that the constant ϖ_1 and ϖ_2 are 0.05 and the fuzzy initial value is $M[0; r] = [M_{01}^r, M_{02}^r] = [0.5 + 0.5r, 1.5 - 0.5r]$ (Figs 1 and 2).

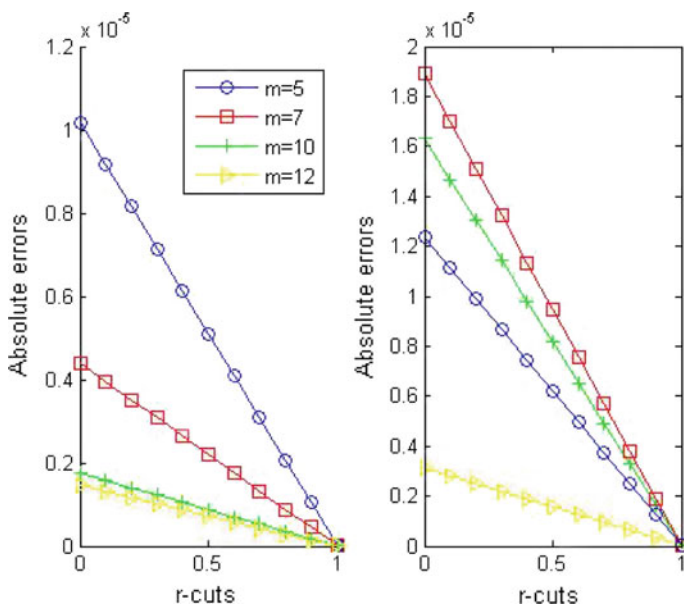


Fig. 1 (Left) Absolute errors of the approximate solution of Eq. (2) using proposed method with different number of Jacobi functions ($\gamma = 0, \xi = 0$); $\nu = 0.85$. (Right) Absolute errors of the approximate solution with different values of γ and ξ , red: $\gamma = 0.5, \xi = 0.5$; green: $\gamma = 0.5, \xi = 0$; blue: $\gamma = 0.5, \xi = 0.5$; yellow: $\gamma = 0, \xi = 0$; $\nu = 0.85$

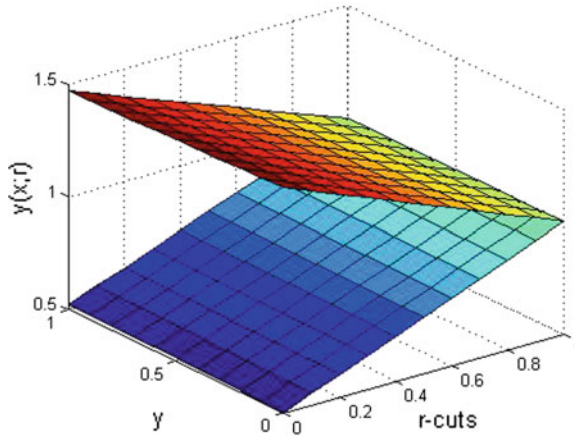


Fig. 2 Approximate solution of Eq. (2) using the proposed method with $m = 8$, $\nu = 0.85$ and $\gamma = 0$, $\xi = 0$

5 Conclusion

In this study an uncertain FDE was approximated by an effective numerical technique using the Jacobi polynomials. The results prove that the algorithm is valid and reliable to the similar problems arising in the real-world systems under fuzzy notion

References

1. Doha, E.H., Bhrawy, A.H., Ezz-Eldien, S.S.: Efficient Chebyshev spectral methods for solving multi-term fractional orders differential equations. *Appl. Math. Model.* **35**, 5662–5672 (2011)
2. Bhrawy, A.H., Alofi, A.S., Ezz-Eldien, S.S.: A quadrature tau method for fractional differential equations with variable coefficients. *Appl. Math. Lett.* **24**, 2146–2152 (2011)
3. Kazem, S., Abbasbandy, S., Kumar, S.: Fractional-order Legendre functions for solving fractional- order differential equations. *Appl. Math. Model.* **37**, 5498–5510 (2013)
4. Esmaili, S., Shamsi, M.: A pseudo-spectral scheme for the approximate solution of a family of fractional differential equations. *Commun. Nonlinear Sci. Numer. Simul.* **16**, 3646–3654 (2011)
5. Doha, E.H., Bhrawy, A.H., Ezz-Eldien, S.S.: A new Jacobi operational matrix: An application for solving fractional differential equations. *Appl. Math. Model.* **36**, 4931–4943 (2012)
6. Bede, B., Gal, S.G.: Generalizations of the differentiability of fuzzy number value functions with applications to fuzzy differential equations. *Fuzzy Sets Syst.* **151**, 581–599 (2005)
7. Ahmadian, A., Suleiman, M., Salahshour, S., Baleanu, D.: A Jacobi operational matrix for solving fuzzy linear fractional differential equation. *Adv. Differ. Equations* **2013**, 104 (2013)
8. Doha, E.H., Bhrawy, A.H., Ezz-Eldien, S.S.: A Chebyshev spectral method based on operational matrix for initial and boundary value problems of fractional order. *Comput. Math Appl.* **62**, 2364–2373 (2011)

Chapter 12

Energy Dissipation of Free Convection Boundary Layer Flow in a Jeffrey Fluid Across a Horizontal Circular Cylinder with Suspended Nanoparticles



Syazwani Mohd Zokri, Nur Syamilah Arifin,
Abdul Rahman Mohd Kasim, Nurul Farahain Mohammad
and Mohd Zuki Salleh

Abstract An investigation has been executed to study the influence of energy dissipation on free convection boundary layer flow in a Jeffrey fluid over a horizontal circular cylinder with suspended nanoparticles. Mathematical formulation is modelled in terms of partial differential equations with some physical conditions. The suitable non-dimensional and non-similarity variables are introduced to transform the system of equations and then the resulting equations are solved using the Keller-box method. The graphical results for the temperature profile and Nusselt number are plotted and explained for several values of Brownian motion parameter. This study reveals that the Brownian motion parameter has caused the increment in the temperature profile while decrement in the Nusselt number. The theoretical results generated in this study are important to the researchers and engineers as they can be used as a reference or basis for comparison purposes especially for validating data or experiments in the future.

S. M. Zokri (✉) · N. S. Arifin · A. R. M. Kasim · M. Z. Salleh
Faculty of Industrial Sciences and Technology,
Universiti Malaysia Pahang (UMP), 26300 Kuantan, Pahang, Malaysia
e-mail: syazwanizokri@gmail.com

N. S. Arifin
e-mail: nursyamilaharifin@gmail.com

A. R. M. Kasim
e-mail: rahmanmohd@ump.edu.my

M. Z. Salleh
e-mail: zuki@ump.edu.my

N. F. Mohammad
Department of Computational Theoretical Science, Kulliyyah of Science,
International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia
e-mail: farahain@iium.edu.my

Keywords Energy dissipation · Horizontal circular cylinder · Suspended nanoparticles · Free convection

1 Introduction

The investigation of heat transfer on free convection boundary layer flow from a horizontal circular cylinder with several effects has been studied by many researchers, for instance Merkin [1], Azim and Chowdhury [2] and Mohamed et al. [3]. This attraction is stimulated from the industrial applications such as the coating of wires, polymer processing, fiber technology and geothermal energy extraction.

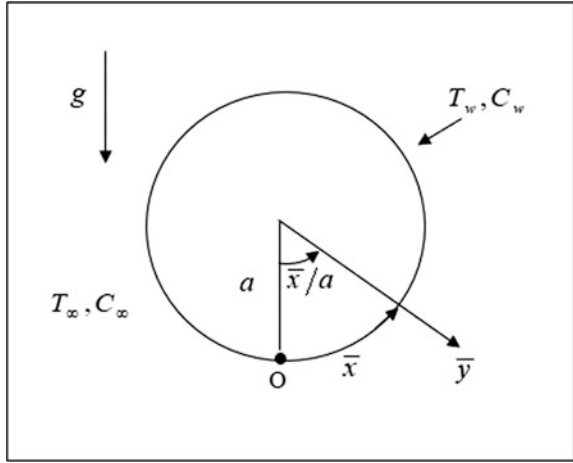
In this study, the Jeffrey nanofluid is considered, where the Jeffrey fluid is regarded as a base fluid while the suspended nanoparticles that consists of carbide, nitride, metal (Al, Ag, Au, etc.) and metal oxide (Al_2O_3 , CuO, TiO_2 , etc.) are functioned to enhance the thermal properties of the base fluid. Eastman et al. [4] and Choi et al. [5] reported the increment of thermal conductivity for about 40 and 150% as a result of small dispersion of Cu nanoparticles or carbon nanotubes in ethylene glycol or oil, respectively. Recent analysis of this fluid model is performed by Abbasi et al. [6] and Ramzan et al. [7] on the stretching sheet and inclined stretching cylinder, respectively.

Inspired by the previous studies, the present investigation aims to tackle the effect of energy dissipation on free convection boundary layer flow in Jeffrey nanofluid due to a horizontal circular cylinder. This effect cannot be ignored as the generation of heat due to internal friction is vital especially when dealing with highly viscous fluid.

2 Mathematical Formulation

A steady two-dimensional free convection boundary layer flow from a horizontal circular cylinder of radius a is taken into consideration. In the presence of nanoparticles, the cylinder is embedded in Jeffrey fluid with constant and ambient temperature, i.e. T_w and T_∞ as well as constant and ambient concentration i.e. C_w and C_∞ . The coordinate of \bar{x} and \bar{y} are measured from the lower stagnation point and normal to the cylinder surface. The geometry of this problem is sketched in Fig. 1. By integrating the energy dissipation effect, the governing equations become:

$$\frac{\partial \bar{u}}{\partial \bar{x}} + \frac{\partial \bar{v}}{\partial \bar{y}} = 0, \quad (1)$$

Fig. 1 Geometry of the problem

$$\bar{u} \frac{\partial \bar{u}}{\partial \bar{x}} + \bar{v} \frac{\partial \bar{u}}{\partial \bar{y}} = \frac{\nu}{1 + \lambda} \left[\frac{\partial^2 \bar{u}}{\partial \bar{y}^2} + \lambda_1 \left(\bar{u} \frac{\partial^3 \bar{u}}{\partial \bar{x} \partial \bar{y}^2} - \frac{\partial \bar{u}}{\partial \bar{x}} \frac{\partial^2 \bar{u}}{\partial \bar{y}^2} + \frac{\partial \bar{u}}{\partial \bar{y}} \frac{\partial^2 \bar{u}}{\partial \bar{x} \partial \bar{y}} + \bar{v} \frac{\partial^3 \bar{u}}{\partial \bar{y}^3} \right) \right] + g[\beta_T(T - T_\infty) + \beta_C(C - C_\infty)] \sin\left(\frac{\bar{x}}{a}\right), \quad (2)$$

$$\bar{u} \frac{\partial T}{\partial \bar{x}} + \bar{v} \frac{\partial T}{\partial \bar{y}} = \alpha \frac{\partial^2 T}{\partial \bar{y}^2} + \frac{\nu}{C_p(1 + \lambda)} \left[\left(\frac{\partial \bar{u}}{\partial \bar{y}} \right)^2 + \lambda_1 \left(\bar{u} \frac{\partial \bar{u}}{\partial \bar{y}} \frac{\partial^2 \bar{u}}{\partial \bar{x} \partial \bar{y}} + \bar{v} \frac{\partial \bar{u}}{\partial \bar{y}} \frac{\partial^2 \bar{u}}{\partial \bar{y}^2} \right) \right] + \tau \left[D_B \frac{\partial C}{\partial \bar{y}} \frac{\partial T}{\partial \bar{y}} + \frac{D_T}{T_\infty} \left(\frac{\partial T}{\partial \bar{y}} \right)^2 \right], \quad (3)$$

$$\bar{u} \frac{\partial C}{\partial \bar{x}} + \bar{v} \frac{\partial C}{\partial \bar{y}} = D_B \frac{\partial^2 C}{\partial \bar{y}^2} + \frac{D_T}{T_\infty} \frac{\partial^2 T}{\partial \bar{y}^2} \quad (4)$$

with boundary conditions

$$\begin{aligned} \bar{u}(\bar{x}, 0) = 0, \bar{v}(\bar{x}, 0) = 0, T(\bar{x}, 0) = T_w, C(\bar{x}, 0) = C_w \text{ at } \bar{y} = 0 \\ \bar{u}(\bar{x}, \infty) \rightarrow 0, \bar{v}(\bar{x}, \infty) \rightarrow 0, T(\bar{x}, \infty) \rightarrow T_\infty, C(\bar{x}, \infty) \rightarrow C_\infty \text{ as } \bar{y} \rightarrow \infty \end{aligned} \quad (5)$$

where \bar{u} and \bar{v} are the velocity components along the \bar{x} and \bar{y} axes, respectively. Let λ , λ_1 , β_T , β_C , α , ν , g , T , C , C_p , D_B and D_T be the respective ratio of relaxation to retardation times, retardation time, thermal expansion, concentration expansion, thermal diffusivity, kinematic viscosity, gravity acceleration, local temperature, local concentration, specific heat capacity at constant pressure, Brownian diffusion and thermophoretic diffusion coefficients. Further, $\tau = (\rho c)_p / (\rho c)_f$ is the ratio of

heat capacity of the nanoparticle to fluid. The following non-dimensional variables are implemented to reduce (1)–(5) into dimensionless form:

$$u = \frac{a}{v} Gr^{-1/2} \bar{u}, \quad v = \frac{a}{v} Gr^{-1/4} \bar{v}, \quad x = \frac{\bar{x}}{a}, \quad y = Gr^{1/4} \frac{\bar{y}}{a}, \quad \theta = \frac{T - T_\infty}{T_w - T_\infty}, \quad \phi = \frac{C - C_\infty}{C_w - C_\infty} \quad (6)$$

Now, the dimensionless form of the above equations can be written as follows:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad (7)$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \frac{1}{1+\lambda} \left[\frac{\partial^2 u}{\partial y^2} + \lambda_2 \left(u \frac{\partial^3 u}{\partial x \partial y^2} - \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial y} \frac{\partial^2 u}{\partial x \partial y} + v \frac{\partial^3 u}{\partial y^3} \right) \right] + [\theta + N\phi] \sin x, \quad (8)$$

$$u \frac{\partial \theta}{\partial x} + v \frac{\partial \theta}{\partial y} = \frac{1}{Pr} \frac{\partial^2 \theta}{\partial y^2} + \frac{Ec}{(1+\lambda)} \left[\left(\frac{\partial u}{\partial y} \right)^2 + \lambda_2 \left(u \frac{\partial u}{\partial y} \frac{\partial^2 u}{\partial x \partial y} + v \frac{\partial u}{\partial y} \frac{\partial^2 u}{\partial y^2} \right) \right] + Nb \frac{\partial \phi}{\partial y} \frac{\partial \theta}{\partial y} + Nt \left(\frac{\partial \theta}{\partial y} \right)^2, \quad (9)$$

$$u \frac{\partial \phi}{\partial x} + v \frac{\partial \phi}{\partial y} = \frac{1}{Le} \left(\frac{\partial^2 \phi}{\partial y^2} + \frac{Nt}{Nb} \frac{\partial^2 \theta}{\partial y^2} \right) \quad (10)$$

$$\begin{aligned} u(x, 0) = 0, \quad v(x, 0) = 0, \quad \theta(x, 0) = 1, \quad \phi(x, 0) = 1 \quad \text{at } y = 0 \\ u(x, \infty) \rightarrow 0, \quad v(x, \infty) \rightarrow 0, \quad \theta(x, \infty) \rightarrow 0, \quad \phi(x, \infty) \rightarrow 0 \quad \text{as } y \rightarrow \infty \end{aligned} \quad (11)$$

Introducing the following functions to solve Eqs. (7)–(10):

$$\psi = xf(x, y), \quad \theta = \theta(x, y), \quad \phi = \phi(x, y), \quad u = \frac{\partial \psi}{\partial y} \quad \text{and} \quad v = -\frac{\partial \psi}{\partial x} \quad (12)$$

Here, Eq. (7) is automatically satisfied, while Eqs. (8)–(11) become:

$$\begin{aligned} \frac{1}{1+\lambda} f'''' + ff'' - (f')^2 + \frac{\lambda_2}{1+\lambda} \left[(f'')^2 - ff^{(iv)} \right] + \frac{\sin x}{x} [\theta + N\phi] = x \left[f' \frac{\partial f'}{\partial x} \right. \\ \left. - f'' \frac{\partial f}{\partial x} + \frac{\lambda_2}{1+\lambda} \left(f''' \frac{\partial f'}{\partial x} + f^{iv} \frac{\partial f}{\partial x} - f' \frac{\partial f'''}{\partial x} - f'' \frac{\partial f''}{\partial x} \right) \right], \end{aligned} \quad (13)$$

$$\begin{aligned} \frac{1}{Pr} \theta'' + f\theta' + Nb\theta'\phi' + Nt(\theta')^2 = x \left[f' \frac{\partial \theta}{\partial x} - \theta' \frac{\partial f}{\partial x} \right. \\ \left. - x \frac{Ec}{(1+\lambda)} \left((f'')^2 - \lambda_2 \left(ff''f''' + xf''f''' \frac{\partial f}{\partial x} - xf'f'' \frac{\partial f''}{\partial x} - f'(f'')^2 \right) \right) \right], \end{aligned} \quad (14)$$

$$\phi'' + Le f\phi' + \frac{Nt}{Nb} \theta'' = xLe \left[f' \frac{\partial \phi}{\partial x} - \phi' \frac{\partial f}{\partial x} \right] \quad (15)$$

$$\begin{aligned}
 f(x, 0) = 0, f'(x, 0) = 0, \theta(x, 0) = 1, \phi(x, 0) = 1 \text{ at } y = 0 \\
 f'(x, \infty) \rightarrow 0, f''(x, \infty) \rightarrow 0, \theta(x, \infty) \rightarrow 0, \phi(x, \infty) \rightarrow 0 \text{ as } y \rightarrow \infty
 \end{aligned}
 \tag{16}$$

where $\lambda_2 = \frac{\lambda_1 Gr^{1/2} \nu}{a^2}$, $Gr = \frac{g \beta_T (T_w - T_\infty) a^3}{\nu^2}$, $Pr = \frac{\mu}{\alpha \rho}$, $Nb = \frac{\tau D_B (C_w - C_\infty)}{\nu}$, $Nt = \frac{\tau D_T (T_w - T_\infty)}{\nu T_\infty}$, $Le = \frac{\nu}{D_B}$, $Ec = \frac{\nu^2 Gr}{a^2 C_p (T_w - T_\infty)}$ and $N = \frac{\beta_C (C_w - C_\infty)}{\beta_T (T_w - T_\infty)}$ are the Deborah number, Grashof number, Prandtl number, Brownian motion parameter, thermophoresis parameter, Lewis number, Eckert number and concentration buoyancy parameter while f' represents the derivative of $f(x, y)$ with respect to y . The physical quantities of interests are the Nusselt number and Sherwood number:

$$Nu_x Gr^{-1/4} = -(\partial \theta / \partial y)_{\bar{y}=0} \text{ and } Sh_x Gr^{-1/4} = -(\partial \phi / \partial y)_{\bar{y}=0}
 \tag{17}$$

3 Result and Discussions

This study utilizes the Keller-box method to solve the governing Eqs. (13)–(15) with boundary conditions (16). Generally, this method entails the following steps: (a) decomposition of the N th order of partial differential equations to N first order equations, (b) finite difference discretization, (c) linearization of non-linear equations and (d) Block-tridiagonal elimination solution of linear equations. Validation of the results is done by comparing the present result with existing publications and a great agreement between them is perceived as written in Table 1.

The graphical results of this study are presented through Figs. 2, 3 and 4. In Fig. 2, as Brownian motion parameter, Nb increases, the temperature profile is boosted. An increase in Nb has accelerated the random movement of nanoparticles, which is triggered by the collision of the molecules in the fluid. This will generate more heat and the elevation of the temperature profile is predicted. Also, at $y = 1$ and $Nb = 0.5-1.0$, it is detected that the percentage increment of the temperature is 5.34%. Meanwhile, the Nusselt number is decreased owing to the increment in Nb as sketched in Fig. 3. This signifies that the heat transfer rate is considerably reduced. Figure 4 is plotted to examine the effect of Prandtl number, Pr on the

Table 1 Comparison between the present numerical results and the existing publications when $\lambda = N = Nb = Nt = Ec = 0$, $Pr = 1.0$ and $\lambda_2 \rightarrow 0$

x	$Nu_x Gr^{-1/4}$			
	Merkin [1]	Azim and Chowdhury [2]	Mohamed et al. [3]	Present
0.0	0.4214	0.4216	0.4214	0.4214
$\pi/3$	0.4007	0.4006	0.4008	0.4009
$2\pi/3$	0.3364	0.3356	0.3364	0.3363
π	0.1945	0.1912	0.1939	0.1931

Fig. 2 Effect of Nb on $\theta(y)$

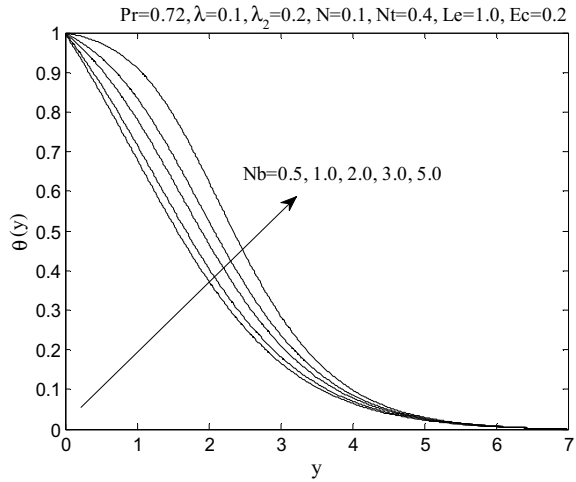
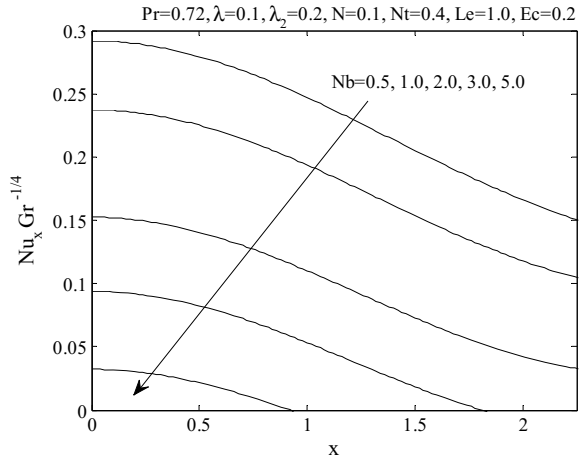
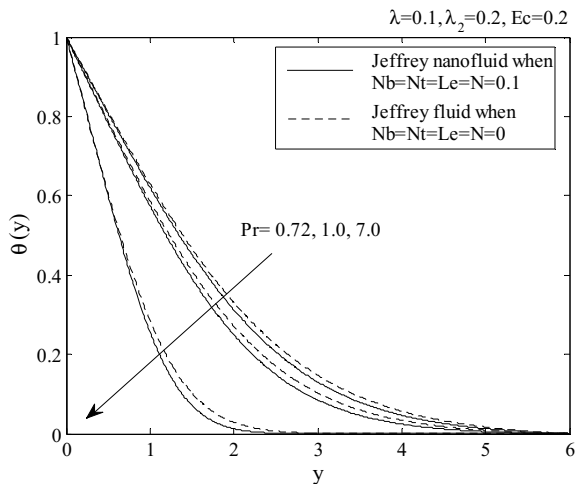


Fig. 3 Effect of Nb on $Nu_x Gr^{-1/4}$



temperature profile for both Jeffrey fluid and Jeffrey nanofluid, in which the temperature profile for Jeffrey nanofluid is lower than Jeffrey fluid. The reason is that, larger values of Pr imply the declination of conduction while enhancing pure convection. That being the case, the rate of heat transfer is increased, and both temperature and thermal boundary layer thickness are diminished. In comparison to Jeffrey fluid, it can be concluded that the thermal boundary layer thickness of Jeffrey nanofluid is the thinnest.

Fig. 4 Comparison between the Jeffrey fluid and Jeffrey nanofluid on $\theta(y)$ for different values of Pr



4 Conclusion

In general, increasing Brownian motion parameter has increased the temperature profile whereas the Nusselt number is decreased. Besides, for dissimilar Prandtl number, the temperature profile for Jeffrey nanofluid is found to be lower than Jeffrey fluid.

Acknowledgements The authors are grateful to the funding provided by Universiti Malaysia Pahang (UMP) through grants PGRS1703100, RDU170358 and RDU150101.

References

1. Merkin, J.H.: Free convection boundary layer on an isothermal horizontal cylinder. In: American Society of Mechanical Engineers and American Institute of Chemical Engineers, Heat Transfer Conference, St. Louis, USA pp. 1–4 (1976)
2. Azim, N., Chowdhury, M.: MHD-conjugate free convection from an isothermal horizontal circular cylinder with joule heating and heat generation. *J. Comput. Methods Phys.* 2013 (2013). <https://doi.org/10.1155/2013/180516>
3. Mohamed, M.K.A., Noar, N.A.Z.M., Salleh, M.Z., Ishak, A.: Free convection boundary layer flow on a horizontal circular cylinder in a nanofluid with viscous dissipation. *Sains Malays.* **45** (2), 289–296 (2016)
4. Eastman, J.A., Choi, S., Li, S., Yu, W., Thompson, L.: Anomalously increased effective thermal conductivities of ethylene glycol-based nanofluids containing copper nanoparticles. *Appl. Phys. Lett.* **78**(6), 718–720 (2001). <https://doi.org/10.1063/1.1341218>
5. Choi, S., Zhang, Z., Yu, W., Lockwood, F., Grulke, E.: Anomalous thermal conductivity enhancement in nanotube suspensions. *Appl. Phys. Lett.* **79**(14), 2252–2254 (2001). <https://doi.org/10.1063/1.1408272>

6. Abbasi, F., Shehzad, S., Hayat, T., Alhuthali, M.: Mixed convection flow of Jeffrey Nanofluid with thermal radiation and double stratification. *J. Hydrodyn., Ser. B* **28**(5), 840–849 (2016). [https://doi.org/10.1016/s1001-6058\(16\)60686-8](https://doi.org/10.1016/s1001-6058(16)60686-8)
7. Ramzan, M., Bilal, M., Chung, J.D.: Effects of thermal and solutal stratification on Jeffrey magneto-nanofluid along an inclined stretching cylinder with thermal radiation and heat generation/absorption. *Int. J. Mech. Sci.* **131**, 317–324 (2017). <https://doi.org/10.1016/j.ijmecsci.2017.07.012>

Chapter 13

Generalized Half-Step Hybrid Block for Solving Second Order Ordinary Differential Equations Directly



Kamarun Hizam Mansor, Zurni Omar and Azizah Rohni

Abstract This paper proposes a new generalized half-step hybrid block method with one off-step point to find the direct solution of second order ordinary differential equations. In developing this method, a power series adopted as an approximate solution is interpolated at x_n and x_{n+p} points while its second derivatives collocated at all points in the interval i.e. x_n , x_{n+p} and $x_{n+\frac{1}{2}}$ to obtain the main continuous scheme, where $0 < p < \frac{1}{2}$. The analysis of the method such as order, zero stability, consistency and convergence is also discussed. The derived method is then compared with the existing methods in terms of accuracy. The numerical results suggest that the new method can be served as a viable alternative to solve the initial value problems of second order.

Keywords Half-step hybrid block method · Second order ordinary differential equations · One off-step point · Direct solution

1 Introduction

Mathematical formulation for numerous problems in engineering, biology, economics and business often leads to the solution of ordinary differential equations (ODEs). In some cases, the problems can be represented by a second order initial value problem (IVP) in the form

$$y'' = f(x, y, y'), y(a) = \eta_0, y'(a) = \eta_1 \text{ and } x \in [a, b] \quad (1)$$

K. H. Mansor (✉) · Z. Omar · A. Rohni
School of Quantitative Sciences, Universiti Utara Malaysia, Sintok, Malaysia
e-mail: hizam@uum.edu.my

Z. Omar
e-mail: zurni@uum.edu.my

A. Rohni
e-mail: r.azizah@uum.edu.my

Generally, conventional numerical methods such as Euler and Runge-Kutta methods will be used to solve the equivalent first order system of (1). Besides enlarging the number of equations in (1), this strategy also approximates numerical solutions only at one point at a time. Block methods were then introduced with the aim of approximating numerical solutions at many points simultaneously and proven to be cost effective and produced better approximation (refer to [1–4]). Nevertheless, block methods have zero stability barrier. In order to overcome this setback, hybrid block method was introduced (see [5–7]). In 2013, James et al. [8] introduced a half-step continuous block method for first ODEs using the interpolation and collocation approach. In this study, we will extend their work to solve second order ODEs by developing half-step hybrid block methods with generalized one off-step point.

2 Development of the Method

In this section, a half-step block hybrid method with one generalized off step (hybrid) point for solving (1) is described. Let the approximate solution of (1) be the power series polynomial given by:

$$y(x) = \sum_{j=0}^{i+c-1} a_j \left(\frac{x-x_n}{h} \right)^j \quad (2)$$

where $x \in \left[x_n, x_{n+\frac{1}{2}} \right]$ for $n = 0, 1, 2, \dots, N-1$, i is the number of interpolation points which is equal to the order of differential equation and c is the number of collocation points. Meanwhile the constant step size, $h = x_n - x_{n-1}$ of partition of interval $[a, b]$ is given by $a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$.

Now, differentiating (2) twice yields

$$y''(x) = f(x, y, y') = \sum_{j=2}^{i+c-1} \frac{j(j-1)}{h^2} a_j \left(\frac{x-x_n}{h} \right)^{j-2} \quad (3)$$

Interpolating (2) at x_n, x_{n+p} ($i = 2$) and collocating (3) at all points, i.e x_n, x_{n+p} and $x_{n+\frac{1}{2}}$ ($c = 3$) in that interval gives five equations which can be written in the following matrix form

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & p & p^2 & p^3 & p^4 \\ 0 & 0 & 2/h^2 & 0 & 0 \\ 0 & 0 & 2/h^2 & 6/h^2 & 12/h^2 \\ 0 & 0 & 2/h^2 & 3/h^2 & 3/h^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} y_n \\ y_{n+p} \\ f_n \\ f_{n+p} \\ f_{n+\frac{1}{2}} \end{pmatrix} \quad (4)$$

Solving (4) using Gaussian elimination gives the values of a_0, a_1, a_2, a_3 and a_4 as follows

$$\begin{aligned} a_0 &= y_n, a_2 = \frac{h^2 f_n}{2}, \\ a_1 &= \frac{h^2 p^2 (2 - 5p + 2p^2) f_n - 2h^2 p^4 f_{n+\frac{1}{2}} + h^2 p^2 (1-p) f_{n+p} + (6 - 12p) y_n + (12p - 6) y_{n+p}}{6p(2p-1)}, \\ a_3 &= \frac{h^2 \left\{ (4p^2 - 1) f_n - 4p^2 f_{n+\frac{1}{2}} + f_{n+p} \right\}}{6p(2p-1)}, a_4 = \frac{h^2 \left\{ (2p-1) f_n - 2p f_{n+\frac{1}{2}} + f_{n+p} \right\} f_n}{6p(2p-1)} \end{aligned} \quad (5)$$

All values of $a'_j, j = 0(1)(4)$ are then substituted back into Eq. (2) to give approximate solution equation

$$\begin{aligned} y(x) &= y_n + \frac{\left\{ h^2 p^2 (2 - 5p + 2p^2) f_n - 2h^2 p^4 f_{n+\frac{1}{2}} + h^2 p^2 (1-p) f_{n+p} + (6 - 12p) y_n + (12p - 6) y_{n+p} \right\}}{6p(2p-1)h} (x - x_n) \\ &+ \left(\frac{h^2 f_n}{2h^2} \right) (x - x_n)^2 - \frac{\left\{ (4p^2 - 1) f_n - 4p^2 f_{n+\frac{1}{2}} + f_{n+p} \right\}}{6ph(2p-1)} (x - x_n)^3 \\ &+ \frac{\left\{ (2p-1) f_n - 2p f_{n+\frac{1}{2}} + f_{n+p} \right\}}{6ph^2(2p-1)} (x - x_n)^4 \end{aligned} \quad (6)$$

Evaluating Eq. (6) at the non-interpolating point, i.e at $x_{n+\frac{1}{2}}$ leads to

$$\begin{aligned} y_{n+\frac{1}{2}} &= \left(1 - \frac{1}{2p} \right) y_n + \left(\frac{1}{2p} \right) y_{n+p} + h^2 \left(\frac{8p^3 - 16p^2 + 8p - 1}{96p} \right) f_n \\ &+ h^2 \left(\frac{1 + 2p - 4p^2}{96p} \right) f_{n+p} - h^2 \left(\frac{4p^2 + 2p - 1}{48} \right) f_{n+\frac{1}{2}} \end{aligned} \quad (7)$$

Differentiating (6) once, we have

$$\begin{aligned} y'(x) &= -\frac{1}{hp} y_n + \frac{1}{hp} y_{n+p} + \left\{ \frac{1}{6} hp(p-2) + (x - x_n) \right\} f_n + \frac{hp(p-1)}{6-12p} f_{n+p} \\ &+ \frac{hp^3}{3-6p} f_{n+\frac{1}{2}} - \left\{ \frac{(1+2p)}{2hp} f_n + \frac{1}{2hp(2p-1)} f_{n+p} - \frac{2p}{h(2p-1)} f_{n+\frac{1}{2}} \right\} (x - x_n)^2 \\ &+ \left\{ \frac{2}{3h^2 p} \left(f_n + \frac{f_{n+p}}{2p-1} - \frac{2p}{2p-1} f_{n+\frac{1}{2}} \right) \right\} (x - x_n)^3 \end{aligned} \quad (8)$$

The following equation are obtained after evaluating (8) at all points, i.e x_n, x_{n+p} and $x_{n+\frac{1}{2}}$.

$$y'_n = -\frac{1}{hp}y_n + \frac{1}{hp}y_{n+p} + \frac{1}{6}hp(p-2)f_n + \left(\frac{hp(p-1)}{6-12p}\right)f_{n+p} + \left(\frac{hp^3}{3-6p}\right)f_{n+\frac{1}{2}} \quad (9)$$

$$y'_{n+p} = -\frac{1}{hp}y_n + \frac{1}{hp}y_{n+p} - \frac{1}{6}hp(p-1)f_n + \left(\frac{hp(3p-2)}{12p-6}\right)f_{n+p} + \left(\frac{hp^3}{6p-3}\right)f_{n+\frac{1}{2}} \quad (10)$$

$$y'_{n+\frac{1}{2}} = -\frac{1}{hp}y_n + \frac{1}{hp}y_{n+p} + h\left(\frac{4p^3-8p^2+6p-1}{24p}\right)f_n + \left(\frac{h-4hp^2+4hp^3}{24p-48p^2}\right)f_{n+p} + h\left(\frac{2p^3-3p+1}{6-12p}\right)f_{n+\frac{1}{2}} \quad (11)$$

Combining Eqs. (7) and (9) produces a block of the form

$$A^{[1]_2}Y_m^{[1]_2} = B_1^{[1]_2}R_1^{[1]_2} + B_2^{[1]_2}R_2^{[1]_2} + h^2\left[D^{[1]_2}R_3^{[1]_2} + E^{[1]_2}R_4^{[1]_2}\right] \quad (12)$$

where

$$A^{[1]_2} = \begin{bmatrix} -\frac{1}{2p} & 1 \\ -\frac{1}{hp} & 0 \end{bmatrix}, Y_m^{[1]_2} = \begin{bmatrix} y_{n+p} \\ y_{n+\frac{1}{2}} \end{bmatrix}, B_1^{[1]_2} = \begin{bmatrix} 0 & 1-\frac{2}{p} \\ 0 & -\frac{1}{hp} \end{bmatrix}, R_1^{[1]_2} = \begin{bmatrix} y_{n-\frac{1}{2}} \\ y_n \end{bmatrix},$$

$$B_2^{[1]_2} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, R_2^{[1]_2} = \begin{bmatrix} y'_{n-\frac{1}{2}} \\ y'_n \end{bmatrix}, D^{[1]_2} = \begin{bmatrix} 0 & \frac{8p^3-16p^2+8p-1}{96p} \\ 0 & \frac{p(p-2)}{6h} \end{bmatrix}, R_3^{[1]_2} = \begin{bmatrix} f_{n-\frac{1}{2}} \\ f_n \end{bmatrix},$$

$$E^{[1]_2} = \begin{bmatrix} \frac{1+2p-4p^2}{96p} & \frac{1-2p-4p^2}{48} \\ \frac{p(p-1)}{h(6-12p)} & \frac{p^3}{h(3-6p)} \end{bmatrix}, R_4^{[1]_2} = \begin{bmatrix} f_{n+p} \\ f_{n+\frac{1}{2}} \end{bmatrix}.$$

Now, multiplying Eq. (12) by the inverse of $A^{[1]_2}$ gives

$$I^{[1]_2}Y_m^{[1]_2} = \bar{B}_1^{[1]_2}R_1^{[1]_2} + \bar{B}_2^{[1]_2}R_2^{[1]_2} + h^2\left[\bar{D}^{[1]_2}R_3^{[1]_2} + \bar{E}^{[1]_2}R_4^{[1]_2}\right] \quad (13)$$

which leads to the following equations

$$y_{n+p} = y_n + hpy'_n + \frac{1}{6}h^2(p-2)p^2f_n + h^2\left(\frac{(p-1)p^2}{12p-6}\right)f_{n+p} + h^2\left(\frac{p^4}{6p-3}\right)f_{n+\frac{1}{2}} \quad (14)$$

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2}y'_n + h^2\left(\frac{8p-1}{96p}\right)f_n + \left(\frac{h^2}{96p-192p^2}\right)f_{n+p} + h^2\left(\frac{4p-1}{96p-48}\right)f_{n+\frac{1}{2}} \tag{15}$$

Substituting Eqs. (14) and (15) into (10) and (11) respectively gives the derivative of the block as below

$$y'_{n+p} = y'_n - \frac{1}{6}h(2p-3)pf_n + hp\left(\frac{4p-3}{12p-6}\right)f_{n+p} + 2h\left(\frac{p^3}{6p-3}\right)f_{n+\frac{1}{2}} \tag{16}$$

$$y'_{n+\frac{1}{2}} = y'_n + h\left(\frac{6p-1}{24p}\right)f_n + \left(\frac{h}{24p-48p^2}\right)f_{n+p} + h\left(\frac{1-3p}{6-12p}\right)f_{n+\frac{1}{2}} \tag{17}$$

3 Analysis of the Method

3.1 Order of the Method

The linear difference operator L associated with (13) is defined as

$$L[y(x); h] = I^{[1]_2} Y_m^{[1]_2} - \bar{B}_1^{[1]_2} R_1^{[1]_2} - \bar{B}_2^{[1]_2} R_2^{[1]_2} - h^2 \left[\bar{D}^{[1]_2} R_3^{[1]_2} + \bar{E}^{[1]_2} R_4^{[1]_2} \right] \tag{18}$$

where $y(x)$ is an arbitrary test function continuously differentiable on $[a, b]$. Y_m and $R_3^{[1]_2}$ components are expanded in Taylors series respectively and its terms are collected in powers of h to give

$$L[y(x); h] = \bar{C}_0 y(x) + \bar{C}_1 h y'(x) + \bar{C}_2 h y''(x) + \dots \tag{19}$$

Definition 3.1 Hybrid block method (13) and associated linear operator (18) are said to be of order d if $\bar{C}_0 = \bar{C}_1 = \bar{C}_2 = \dots = \bar{C}_{d+1} = 0$ and $\bar{C}_{d+2} \neq 0$ with error vector constants \bar{C}_{d+2} .

Expanding (13) in Taylor series about x_n gives

$$\left(\begin{array}{l}
 \sum_{j=0}^{\infty} \frac{(p)^j h^j y_n^j}{j!} - y_n - hp y_n' + \frac{1}{6} h^2 (p-2) p^2 y_n'' \\
 - \sum_{j=0}^{\infty} \left\{ \frac{p^2 (p-1)}{12p-6} \frac{(p)^j h^{j+2} y_n^{j+2}}{j!} + \frac{p^4}{6p-3} \left(\frac{1}{2}\right) \frac{j h^{j+2} y_n^{j+2}}{j!} \right\} \\
 \sum_{j=0}^{\infty} \left(\frac{1}{2}\right) \frac{j h^j y_n^j}{j!} - y_n - \frac{h}{2} y_n' - \frac{h^2 (8p-1)}{96p} y_n'' \\
 - \sum_{j=0}^{\infty} \left\{ \frac{1}{96p-192p^2} \frac{(p)^j h^{j+2} y_n^{j+2}}{j!} + \frac{4p-1}{96p-48} \left(\frac{1}{2}\right) \frac{j h^{j+2} y_n^{j+2}}{j!} \right\} \\
 \sum_{j=0}^{\infty} \frac{(p)^j h^j y_n^{j+1}}{j!} - y_n' + \frac{1}{6} hp (2p-3) y_n'' \\
 - \sum_{j=0}^{\infty} \left\{ \frac{ph(4p-3)}{12p-6} \frac{(p)^j h^{j+1} y_n^{j+2}}{j!} + \frac{2p^3 h}{6p-3} \left(\frac{1}{2}\right) \frac{j h^{j+1} y_n^{j+2}}{j!} \right\} \\
 \sum_{j=0}^{\infty} \left(\frac{1}{2}\right) \frac{j h^j y_n^{j+1}}{j!} - y_n' - \frac{1}{24p} h (6p-1) y_n'' \\
 - \sum_{j=0}^{\infty} \left\{ \frac{h}{24p-48p^2} \frac{(p)^j h^{j+1} y_n^{j+2}}{j!} + \frac{h(1-3p)}{6-12p} \left(\frac{1}{2}\right) \frac{j h^{j+1} y_n^{j+2}}{j!} \right\}
 \end{array} \right) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{20}$$

By comparing the coefficient of h , we obtain the order of the method to be $[2, 2, 2, 2]^T$ with error constant

$$\bar{C}_4 = \left[\frac{1}{144} p^4 (2p-1), \frac{1-2p}{2304}, \frac{p^3}{72} (4p-1), \frac{1-p}{288} \right]^T \tag{21}$$

3.2 Zero Stability

The hybrid block method (13) is said to be *zero-stable* if no root of the first characteristic polynomial $\rho(r) = |rI - \bar{B}_1^{[1]_2}|$ is having a modulus greater than one and every root of modulus one is simple, where I is identity matrix and $\bar{B}_1^{[1]_2}$ is the coefficients matrix of y - function.

$$\pi(r) = |rI - B| = \left| r \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right| = r(r-1) \tag{22}$$

which implies $r = 0, 1$. Hence, our method is zero stable for all $p \in (0, \frac{1}{2})$.

3.3 Consistency and Convergence

The one step hybrid block method (13) is said to be consistent if its order is greater than or equal one, i.e. $d \geq 1$. This proves that our method is consistent for all $p \in (0, \frac{1}{2})$.

Theorem 3.1 (Henrici [9]) *Consistency and zero stability are sufficient conditions for a linear multistep method to be convergent.*

Since the method (13) is consistent and zero stable, this implies that it is convergent for all $p \in (0, \frac{1}{2})$.

3.4 Numerical Results

To determine the accuracy and stability of our methods, the following second order ODEs problems are tested for $0 < x < 10$. However, for the sake of comparison, we have to choose the same interval as adopted existing methods.

Problem 1 $y'' + x(y')^2 = 0, y(0) = 1, y'(0) = \frac{1}{2}$, for $0 \leq x \leq 1$ with $h = \frac{1}{100}$

Exact solution: $y(x) = 1 + \frac{1}{2} \ln(\frac{2+x}{2-x})$. Source: [5] (Table 1).

Problem 2 $y'' - y = 0, y(0) = 1, y'(0) = 1$, for $0 \leq x \leq 1$ with $h = \frac{1}{10}$

Exact solution: $y(x) = e^x$. Source: [10] (Table 2).

Problem 3 $y'' - y = 0, y(0) = 0, y'(0) = -1$, for $0 \leq x \leq 1$ with $h = \frac{1}{10}$

Exact solution: $y(x) = 1 - e^x$. Source: [11] (Table 3).

Table 1 Exact solution and new method error for Problem 1

x	Exact solution	Error new method	Adesanya et al.
0.10	1.050041729278491400	3.543832E-12	2.3314E-14
0.20	1.100335347731075800	1.418576E-11	1.8918E-13
0.30	1.151140435936467000	3.150724E-11	6.5836E-13
0.40	1.202732554054082300	5.461565E-11	1.6406E-12
0.50	1.255412811882995700	8.180456E-11	3.4350E-12
0.60	1.309519604203112100	1.098912E-10	6.4921E-12
0.70	1.365443754271396400	1.330058E-10	1.1529E-11
0.80	1.423648930193602200	1.402813E-10	1.9728E-11
0.90	1.484700278594052200	1.113387E-10	3.3111E-11
1.00	1.549306144334055400	7.104317E-12	5.5286E-11

Table 2 Exact solution and new method error for Problem 2

x	Exact solution	Error new method	Sagir
0.10	1.105170918075647700	1.039463E-09	–
0.20	1.221402758160169900	2.735561E-09	–
0.30	1.349858807576003200	5.173235E-09	5.7600E-10
0.40	1.491824697641270300	8.451993E-09	1.6413E-09
0.50	1.648721270700128200	1.268766E-08	1.7001E-09
0.60	1.822118800390508900	1.801438E-08	2.3905E-09
0.70	2.013752707470476600	2.458685E-08	3.4705E-09
0.80	2.225540928492467900	3.258291E-08	4.4925E-09
0.90	2.459603111156950300	4.220642E-08	4.1569E-09
1.00	2.718281828459046000	5.369056E-08	4.4590E-09

Table 3 Exact solution and new method error for Problem 3

x	Exact solution	Error new method	Mohammed
0.10	-0.105170918095000010	7.284528E-07	2.198000E-05
0.20	-0.221402758151111160	1.591683E-06	6.070400E-06
0.30	-0.349858797968125750	2.609982E-06	1.005100E-05
0.40	-0.491824677364233350	3.806420E-06	1.402530E-05
0.50	-0.648721226946174510	5.207203E-06	1.799340E-05
0.60	-0.822118730616258730	6.842080E-06	2.161620E-05
0.70	-1.013752594666440500	8.744796E-06	2.799300E-05
0.80	-1.225540768042424400	1.095361E-05	3.456100E-05
0.90	-1.459602880616953900	1.351185E-05	4.111400E-05
1.00	-1.718281520346842900	1.646859E-05	4.765600E-05

4 Conclusion

A new generalized hybrid block method with one off-step point in half step method to solve the second order ordinary differential equations directly has been successfully developed. The new method which is of order two possesses good properties of numerical method. Furthermore, the performance of the developed method has proven to be compatible, or even better if compared to the existing methods when solving the same problems.

References

1. Ademiluyi, R., Duromola, M., Bolaji, B.: Modified block method for the direct solution of initial value problems of fourth order ordinary differential equations. *Aust. J. Basic Appl. Sci.* **8**(10), 389–394 (2014)
2. James, A., Adesanya, A., Joshua, S.: Continuous block method for the solution of second order initial value problems of ordinary differential equation. *Int. J. Pure Appl. Math.* **83**(3), 405–416 (2013)
3. Kuboye, J., Omar, Z.: New zero-stable block method for direct solution of fourth order ordinary differential equations. *Indian J. Sci. Technol.* **8**(12), 1–8 (2015)
4. Kuboye, J., Omar, Z.: Numerical solution of third order ordinary differential equations using a seven-step block method. *Int. J. Math. Anal.* **9**(15), 743–754 (2015)
5. Adesanya, A., Ibrahim, Y., Abdulkadi, B., Anake, T.: Order four continuous hybrid block method for the solutions of second order ordinary differential equations. *J. Math. Comput. Sci.* **4**(5), 817–825 (2014)
6. Anake, T.A., Awoyemi, D.O., Adesanya, A.: One-step implicit hybrid block method for the direct solution of general second order ordinary differential equations. *IAENG Int. J. Appl. Math.* **42**(4), 224–228 (2012)
7. Anake, T.A., Awoyemi, D.O., Adesanya, A.: A one step method for the solution of general second order ordinary differential equations. *Int. J. Sci. Technol.* **2**(4), 159–163 (2012)
8. James, A.A., Adesanya, A.O., Sunday, J., Yakubu, D.G.: Half-step continuous block method for the solutions of modeled problems of ordinary differential equations. *Am. J. Comput. Math.* **3**, 261–269 (2013)
9. Henrici, P.: *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York (1962). <https://doi.org/10.1126/science.136.3511.143-a>
10. Sagir, A.: An accurate computation of block hybrid method for solving stiff ordinary differential equations. *J. Math.* **4**, 18–21 (2012)
11. Mohammed, U.: A class of implicit five—step block method for general second order ordinary differential equations. *J. Nigeria Math. Soc. (JNMS)* **30**, 25–39 (2011)

Chapter 14

Mixed Convection Boundary Layer Flow on a Solid Sphere in a Viscoelastic Micropolar Fluid



Laila Amera Aziz, Abdul Rahman Mohd Kasim, Mohd Zuki Salleh
and Sharidan Shafie

Abstract This study considers the mixed convection boundary layer flow of a viscoelastic micropolar fluid past a solid sphere with aligned MHD effect. The governing equations are first transformed into dimensionless form using dimensionless variables before the application of stream function which then produced a set of partial differential equations. The equations are solved numerically using a finite difference method known as the Keller-box scheme in Fortran program. Validations of present results are performed by comparing the present work with previous publications and the results show excellent agreement. Results on the effects of the viscoelastic parameter, K , material parameter, K_1 , mixed convection parameter, λ and the magnetic parameter, M on the distribution of velocity, temperature as well as microrotation were discussed together with the graphical representation.

Keywords Viscoelastic micropolar · Solid sphere · MHD effect ·
Boundary layer · Numerical solution

L. A. Aziz (✉) · A. R. M. Kasim · M. Z. Salleh
Applied & Industrial Mathematics Research Group,
Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang,
26300 Kuantan, Pahang, Malaysia
e-mail: laila@ump.edu.my

A. R. M. Kasim
e-mail: rahmanmohd@ump.edu.my

M. Z. Salleh
e-mail: zuki@ump.edu.my

S. Shafie
Faculty of Science, Department of Mathematical Sciences,
Universiti Teknologi Mara, 81310 Johor Bahru, Johor, Malaysia
e-mail: sharidan@utm.my

1 Introduction

Viscoelastic micropolar fluid is a complex non-Newtonian fluid which displays both characteristics of viscoelastic and micropolar fluid. This type of fluid deformed when force is exerted on them and return to their original state when the force is removed. Besides the semi-deformation characteristics, viscoelastic micropolar fluid also contains microstructures in the form of bar-like elements [1]. Example of viscoelastic micropolar fluid includes animal blood and anisotropic liquid such as liquid crystal.

The aligned MHD effect on the flow of the viscoelastic micropolar fluid as it passes a solid sphere will be considered in this study. The MHD effect on the fluid flow when the fluid are considered separately has been published in [2, 3] where micropolar fluid is investigated while these studies [4–6] considered the MHD effect for the flow of viscoelastic fluid on various geometrics. The proposed mathematical model in this study will be a generalised model of mixed convection boundary layer flow past a solid sphere with MHD effect for three different types of fluids, namely viscoelastic, micropolar as well as viscoelastic micropolar fluid. Although a similar study has been conducted in [7], the study only focused on the flow at the stagnation point and the equations involved are ordinary differential equations. However, for this study, partial differential equations are solved and the solutions obtained are not only limited to the stagnation point at the boundary layer.

2 Mathematical Formulation

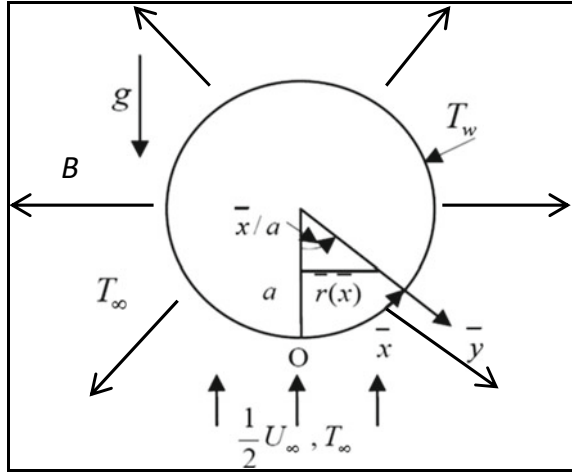
With reference to the mathematical model proposed by [8, 9], the governing boundary layer flow equation of this problem is given by Eqs. (1)–(4) which consists of continuity, momentum, micropolar and energy equation, respectively, subject to the boundary conditions in (5) (Fig. 1).

$$\frac{\partial}{\partial \bar{x}}(\bar{r}\bar{u}) + \frac{\partial}{\partial \bar{y}}(\bar{r}\bar{v}) = 0 \quad (1)$$

$$\begin{aligned} \bar{u}\frac{\partial \bar{u}}{\partial \bar{x}} + \bar{v}\frac{\partial \bar{u}}{\partial \bar{y}} = \bar{u}_e\frac{d\bar{u}_e}{d\bar{x}} + \left(\frac{\mu + \kappa}{\rho}\right)\frac{\partial^2 \bar{u}}{\partial \bar{y}^2} + \frac{k_0}{\rho}\left[\frac{\partial}{\partial \bar{x}}\left(\bar{u}\frac{\partial^2 \bar{u}}{\partial \bar{y}^2}\right) + \bar{v}\frac{\partial^3 \bar{u}}{\partial \bar{y}^3} + \frac{\partial \bar{u}}{\partial \bar{y}}\frac{\partial^2 \bar{v}}{\partial \bar{y}^2}\right] \\ + g\beta(T - T_\infty)\sin\left(\frac{\bar{x}}{a}\right) + \frac{\kappa}{\rho}\frac{\partial \bar{H}}{\partial \bar{y}} - \frac{\sigma}{\rho}(\bar{u} - \bar{u}_e)B^2\sin^2\alpha \end{aligned} \quad (2)$$

$$\rho j\left(\bar{u}\frac{\partial \bar{H}}{\partial \bar{x}} + \bar{v}\frac{\partial \bar{H}}{\partial \bar{y}}\right) = -\kappa\left(2\bar{H} + \frac{\partial \bar{u}}{\partial \bar{y}}\right) + \gamma\frac{\partial^2 \bar{H}}{\partial \bar{y}^2} \quad (3)$$

Fig. 1 Physical model and coordinate system



$$\bar{u} \frac{\partial T}{\partial \bar{x}} + \bar{v} \frac{\partial T}{\partial \bar{y}} = \alpha \frac{\partial^2 T}{\partial \bar{y}^2} \tag{4}$$

$$\begin{aligned} \bar{u} = \bar{v} = 0, \quad T = T_w, \quad \bar{H} = -\frac{1}{2} \frac{\partial \bar{u}}{\partial \bar{y}} \quad \text{on} \quad \bar{y} = 0, \\ \bar{u} = \bar{u}_e(\bar{x}), \quad \frac{\partial \bar{u}}{\partial \bar{y}} = 0, \quad T = T_\infty, \quad \bar{H} = 0 \quad \text{as} \quad \bar{y} \rightarrow \infty \end{aligned} \tag{5}$$

In these equations, the radial distance, $\bar{r}(x)$, the velocity outside boundary layer, $\bar{u}_e(x)$, microinertia per unit mass, j , and the spin gradient, γ , are defined as $\bar{u}_e(\bar{x}) = \frac{3}{2} U_\infty \sin(\frac{\bar{x}}{a})$ and $\bar{r}(\bar{x}) = a \sin(\frac{\bar{x}}{a})$, $j = \frac{a^2 \nu}{U_\infty}$ and $\gamma = (\mu + \frac{\kappa}{2})j$, respectively. The dynamic viscosity, micropolar vortex viscosity, density, viscoelastic vortex viscosity, gravitational acceleration, coefficient of thermal expansion, electrical conductivity and magnetic field are represented by $\mu, \kappa, \rho, k_0, g, \beta, \sigma$ and B , respectively.

Then, the above equations are converted to dimensionless equations as shown in Eqs. (7)–(10) by using the non-dimensional variables introduced in Eq. (6).

$$\begin{aligned} x = \bar{x}/a, \quad y = \text{Re}^{1/2}(\bar{y}/a), \quad u = \bar{u}/U_\infty, \quad v = \text{Re}^{1/2}\bar{v}/U_\infty \\ u_e(x) = \bar{u}_e(\bar{x})/U_\infty, \quad r(x) = \bar{r}(\bar{x})/a, \quad H = (a/U_\infty)\text{Re}^{-1/2}\bar{H} \quad \theta = \frac{T-T_\infty}{T_w-T_\infty} \end{aligned} \tag{6}$$

$$\frac{\partial}{\partial x}(ru) + \frac{\partial}{\partial y}(rv) = 0 \tag{7}$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = u_e \frac{du_e}{dx} + (1 + K_1) \frac{\partial^2 u}{\partial y^2} + K \left[\frac{\partial}{\partial x} \left(u \frac{\partial^2 u}{\partial y^2} \right) + v \frac{\partial^3 u}{\partial y^3} + \frac{\partial u \partial^2 v}{\partial y \partial y^2} \right] + \lambda \theta \sin(x) + K_1 \frac{\partial H}{\partial y} - M(u - u_e) \sin^2 \alpha \quad (8)$$

$$u \frac{\partial H}{\partial x} + v \frac{\partial H}{\partial y} = -K_1 \left(2H + \frac{\partial u}{\partial y} \right) + \left(1 + \frac{K_1}{2} \right) \frac{\partial^2 H}{\partial y^2} \quad (9)$$

$$u \frac{\partial \theta}{\partial x} + v \frac{\partial \theta}{\partial y} = \frac{1}{Pr} \frac{\partial^2 \theta}{\partial y^2} \quad (10)$$

$$u = v = 0, \quad \theta = 1, \quad H = -\frac{1}{2} \frac{\partial u}{\partial y} \quad \text{on } y = 0, \quad (11)$$

$$u_e = \frac{3}{2} \sin x, \quad \frac{\partial u}{\partial y} = 0, \quad \theta = 0, \quad H = 0 \quad \text{as } y \rightarrow \infty$$

Next, the following similarity variables are introduced to solve the equations with the stream function as defined in Eq. (13).

$$\psi = xr(x)f(x, y), \quad H = xh(x, y), \quad \theta = \theta(x, y) \quad (12)$$

$$u = \frac{1}{r} \frac{\partial \psi}{\partial y}, \quad v = -\frac{1}{r} \frac{\partial \psi}{\partial x} \quad (13)$$

$$(1 + K_1) \frac{\partial^3 f}{\partial y^3} - \left(\frac{\partial f}{\partial y} \right)^2 + \frac{9 \sin x \cos x}{4x} + \lambda \frac{\sin x}{x} \theta + K_1 \frac{\partial h}{\partial y} - M \left(\frac{\partial f}{\partial y} - \frac{3 \sin x}{2x} \right) \sin^2 \alpha + \left(1 + \frac{x}{\sin x} \cos x \right) f \frac{\partial^2 f}{\partial y^2} + K \left\{ 2 \frac{\partial f \partial^3 f}{\partial y \partial y^3} - \left(1 + \frac{x}{\sin x} \cos x \right) \left(f \frac{\partial^4 f}{\partial y^4} + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) \right\} = x \left(\frac{\partial f}{\partial y} \frac{\partial^2 f}{\partial x \partial y} - \frac{\partial f \partial^2 f}{\partial x \partial y^2} \right) + Kx \left\{ \frac{\partial f \partial^4 f}{\partial x \partial y^4} - \frac{\partial^3 f \partial^2 f}{\partial y^3 \partial x \partial y} - \frac{\partial f \partial^4 f}{\partial y \partial x \partial y^3} + \frac{\partial^2 f \partial^3 f}{\partial y^2 \partial y^2 \partial x} \right\} \quad (14)$$

$$\left(1 + \frac{K_1}{2} \right) \frac{\partial^2 h}{\partial y^2} + \left(1 + \frac{x}{\sin x} \cos x \right) f \frac{\partial h}{\partial y} - K_1 \left(2h + \frac{\partial^2 f}{\partial y^2} \right) - h \frac{\partial f}{\partial y} = x \left(\frac{\partial f}{\partial y} \frac{\partial h}{\partial x} - \frac{\partial f \partial h}{\partial x \partial y} \right) \quad (15)$$

$$\frac{1}{Pr} \frac{\partial^2 \theta}{\partial y^2} + \left(1 + \frac{x}{\sin x} \cos x \right) f \frac{\partial \theta}{\partial y} = x \left(\frac{\partial f}{\partial y} \frac{\partial \theta}{\partial x} - \frac{\partial f \partial \theta}{\partial x \partial y} \right) \quad (16)$$

$$f(0) = f'(0) = 0, \quad \theta(0) = 1, \quad h(0) = -\frac{1}{2} f''(0) \quad \text{on } y = 0, \quad (17)$$

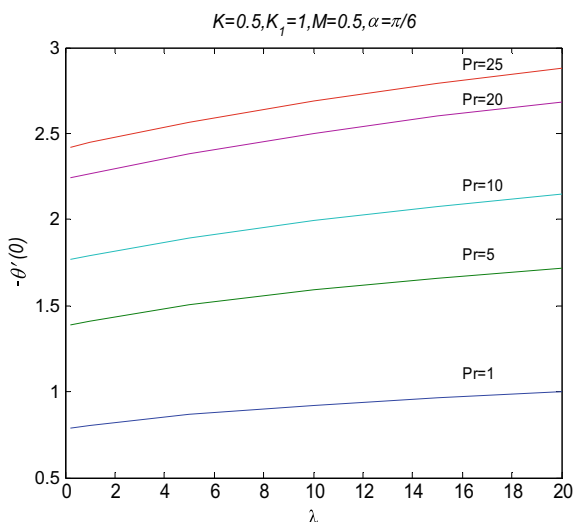
$$f' = \frac{3 \sin x}{2x}, \quad f'' = 0, \quad \theta = 0, \quad h = 0 \quad \text{as } y \rightarrow \infty$$

The set of partial differential equations (14)–(17) are then solved numerically by using the Keller-box [10] scheme in Fortran program.

Table 1 Values of $f''(0)$ and $-\theta'(0)$ for various values of λ at $K = 0, K_1 = 0, M = 0, Pr = 0.7$

λ	$f''(0)$		$-\theta'(0)$	
	Nazar et al. [9]	Present values	Nazar et al. [9]	Present values
0	2.4151	2.532705	0.8162	0.81985
1	2.8064	2.801357	0.8463	0.840638
5	4.2257	4.217931	0.923	0.921762
10	5.7995	5.78745	0.9981	0.996776
20	8.5876	8.565406	1.1077	1.106159

Fig. 2 Variation of heat transfer with λ at various Pr values

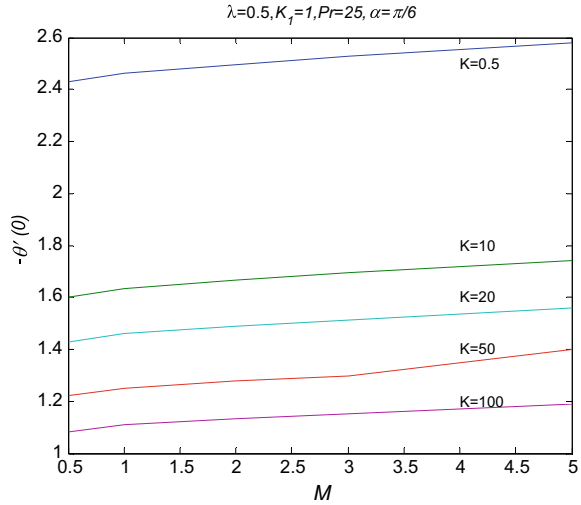


3 Results and Discussion

For the purpose of result verification, the finding from this study is compared to a limiting case ($M = K = K_1 = 0$) by Nazar [9] with both results showing excellent agreement as displayed in Table 1.

Figure 2 shows the variation of the heat transfer with the mixed convection parameter, λ at different Prandtl numbers. From the figure, it is evident that the heat transfer coefficient increases along with the increment of the values of mixed convection parameter and Pr. The result obtained meets the expectation since for liquid with high Pr, the momentum diffusivity is more dominant than thermal diffusivity thus being more efficient at transferring energy through convection. As for Fig. 3, it is observed that the heat transfer coefficient increases as the magnetic parameter, M increases. However, the opposite effect occurs when the viscoelastic parameter is increased where such behavior is caused by the friction force in the viscoelastic fluid.

Fig. 3 Variation of heat transfer with M at various K values



4 Conclusion

The convective boundary layer flow of viscoelastic micropolar fluid past a sphere with aligned MHD effect is investigated. The governing equations are transformed into a set of partial differential equations which are solved numerically in Fortran by the Keller-box scheme. From the results, the following behaviours are observed.

- The increase of mixed convection parameter, Prandtl number and the magnetic parameter increases the heat transfer coefficient.
- As the viscoelastic parameter increases, the heat transfer coefficient decreases.

Acknowledgements The authors would like to express our gratitude for the financial support received from Universiti Malaysia Pahang for RDU 161106 and RDU 1703258.

References

1. Eringen, A.C., Aeronautics, P.U.L.I.S.O., Astronautics: Theory of Micropolar Fluids. Defense Technical Information Center (1965)
2. Sandeep, N., Sulochana, C.: Dual solutions for unsteady mixed convection flow of MHD micropolar fluid over a stretching/shrinking sheet with non-uniform heat source/sink. *Eng. Sci. Technol. Int. J.* **18**(4), 738–745 (2015)
3. Rashad, A.M., Abbasbandy, S., Chamkha, A.J.: Mixed convection flow of a micropolar fluid over a continuously moving vertical surface immersed in a thermally and solutally stratified medium with chemical reaction. *J. Taiwan Inst. Chem. Eng.* **45**(5), 2163–2169 (2014). <https://doi.org/10.1016/j.jtice.2014.07.002>

4. Aziz, L.A., Kasim, A.R.M., Al-Sharifi, H., Salleh, M.Z., Mohammad, N.F., Shafie, S., Ali, A.: Influence of aligned MHD on convective boundary layer flow of viscoelastic fluid. In: AIP Conference Proceedings, vol. 1, p. 030005. AIP Publishing (2017)
5. Kasim, A., Mohammad, N., Anwar, I., Shafie, S.: MHD effect on convective boundary layer flow of a viscoelastic fluid embedded in porous medium with Newtonian heating. *Recent Adv. Math.* **4**, 182–189 (2013)
6. Rashidi, M., Ali, M., Freidoonimehr, N., Rostami, B., Hossain, M.A.: Mixed convective heat transfer for MHD viscoelastic fluid flow over a porous wedge with thermal radiation. *Adv. Mech. Eng.* **6**, 735939 (2014)
7. Aziz, L.A., Kasim, A.R.M., Salleh, M.Z., Yusoff, N.S., Shafie, S.: Magnetohydrodynamics effect on convective boundary layer flow and heat transfer of viscoelastic micropolar fluid past a sphere. *J. Phys. Conf. Ser.* **890**(1), 012003 (2017)
8. Dasman, A., Kasim, M., Rahman, A., Mohammad, N.F., Mangi, A., Shafie, S.: Mixed convection boundary layer flow of viscoelastic fluids past a sphere. In: *Defect and Diffusion Forum*, pp. 57–63. Trans Tech Publications (2013)
9. Nazar, R., Amin, N., Pop, I.: Mixed convection boundary layer flow about an isothermal sphere in a micropolar fluid. *Int. J. Therm. Sci.* **42**(3), 283–293 (2003). [https://doi.org/10.1016/S1290-0729\(02\)00027-3](https://doi.org/10.1016/S1290-0729(02)00027-3)
10. Cebeci, T., Bradshaw, P.: Finite-difference solution of boundary-layer equations. In: *Physical and Computational Aspects of Convective Heat Transfer*, pp. 385–428. Springer (1984)

Chapter 15

Mathematical Modelling of Bank Financial Management in Malaysia with Goal Programming Approach



Chen Jia Wai, Lam Weng Siew and Lam Weng Hoe

Abstract Financial management is important to the companies such as financial institutions to manage their assets and liabilities. The banks or financial institutions have to achieve various goals in optimizing the financial management such as asset accumulation, liability reduction, equity, earning, profitability and optimum management items. Therefore, goal programming has been introduced to optimize the financial management since it involves multiple goals. Goal programming is a mathematical model which aims to solve multiple objectives decision problems. The objective of this study is to develop a mathematical model to optimize the financial management of RHB Bank in Malaysia with goal programming approach. Six goals from the financial statements, namely total asset, total liability, total equity, profit, earning and optimum management items are investigated for the period from year 2011 to 2015. The results of this study show that all goals are fully achieved based on the mathematical model with goal programming approach. Besides that, the potential improvements for asset, liability and equity have been identified in this study. This study is significant because it helps to develop a mathematical model to examine the financial strengths and determine the potential improvements for RHB Bank in Malaysia.

Keywords Financial management • Goal • Potential improvement • LINGO

C. J. Wai (✉) · L. W. Siew · L. W. Hoe
Centre for Mathematical Sciences, Faculty of Science, Department of Physical and Mathematical Science, Universiti Tunku Abdul Rahman, Kampar Campus, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
e-mail: jiawai_chen@hotmail.com

L. W. Siew
e-mail: lamws@utar.edu.my

L. W. Hoe
e-mail: whlam@utar.edu.my

1 Introduction

Financial performance of a company is always a popular issue because there are various goals to be accomplished. Asset and liability management is one of the strategies to supervise the company's investments while controlling its liabilities [1]. Naderi et al. [1] denotes that asset and liability management supports the quality of asset and liability in premeditating the risk in future. Halim et al. [2] indicates a desired liquidity and outcomes rely on the asset and liability of a company. Kosmidou and Zopounidis [3] mention that asset and liability management is the planning of company's asset as well as liability regarding different criteria such as managerial and market constraints in order to increase the company's value. Goal Programming (GP) model was introduced by Charnes et al. [4] to identify the optimal solution among all feasible solutions under multiple objectives [5]. GP model is commonly used for optimization in many organizations. This is because GP model can help to achieve multiple objectives simultaneously.

Based on the past studies, GP model has been applied in bank asset and liability management [1–3]. Arewa et al. [6] investigated financial statements of a Nigerian Bank with GP model. Mohammadi and Sherafati [7] hybrid GP model with Fuzzy Analytic Hierarchy Process (FAHP) to examine the liquidity of Parsian Bank. Mean-while, GP model is applicable in other fields such as incineration plants [8], resource allocation [9], bakery production [10], residents' scheduling [11] as well as personnel management [12].

The objective of this study is to develop a GP model in optimizing the financial management of the RHB bank in Malaysia. Potential improvements can be identified and recommended for further improvement in this study. Since RHB Bank incorporated with Kwong Yik Banking Corporation in 1913, it has become the country's oldest bank and it was the first local bank in Malaysia. RHB Bank has involved in various finance related business and has over 380 delivery channels in eight countries [13]. Six common goals are identified, which are total asset, total liability, total equity, profitability, earnings as well as optimum management items.

2 Data and Methodology

2.1 Goal Programming Model

Goal programming is a mathematical model which aims to solve multi-objective decision problem [14, 15]. In this study, preemptive method is applied to solve the goal programming model. Preemptive method satisfies the highest priority of the goal followed by the second priority of the goal and so on. The goal programming model is formulated as follow.

$$\text{Minimize } z = \rho_i \text{ where } i = 1, 2, 3, \dots, n. \tag{1}$$

Subject to

$$\sum_{j=1}^m (a_{ij}x_j + d_i^- - d_i^+) = g_i \tag{2}$$

$$x_j, d_i^-, d_i^+ \geq 0 \tag{3}$$

where

z is objective function;

ρ_i is deviation variable for $i = 1, 2, 3, \dots, n$;

d_i^- is negative deviation variable for $i = 1, 2, 3, \dots, n$;

d_i^+ is positive deviation variable for $i = 1, 2, 3, \dots, n$;

x_j is decision variable for $j = 1, 2, 3, \dots, m$;

a_{ij} is parameter for decision variable; and

g_i is aspiration level for $i = 1, 2, 3, \dots, n$.

2.2 Data

The financial strength of RHB Bank is investigated in this study. Annual financial reports from year 2011 until 2015 are downloaded from Bursa Malaysia. The data of the financial statement is given in Table 1.

Decision variables x_j refer to amount of financial statement in year j . In goal programming model, the goals are formulated as soft constraints below.

Asset constraint:

$$0.1524x_1 + 0.1891x_2 + 0.1911x_3 + 0.2194x_4 + 0.2307x_5 \geq 0.9826$$

Table 1 RHB financial statement from year 2011 until 2015

Goal	Year (RM' trillion)					Total
	2011	2012	2013	2014	2015	
Asset	0.1524	0.1891	0.1911	0.2194	0.2307	0.9826
Liability	0.1409	0.1737	0.1741	0.2005	0.2076	0.8969
Equity	0.0115	0.0153	0.0169	0.0189	0.0231	0.0857
Profit	0.0015	0.0018	0.0018	0.0021	0.0015	0.0087
Earnings	0.0043	0.0048	0.0060	0.0062	0.0062	0.0275
Total	0.3106	0.3848	0.3900	0.4470	0.4692	2.0015

Liability constraint:

$$0.1409 x_1 + 0.1737 x_2 + 0.1741 x_3 + 0.2005 x_4 + 0.2076 x_5 \leq 0.8969$$

Equity constraint:

$$0.0115 x_1 + 0.0153 x_2 + 0.0169 x_3 + 0.0189 x_4 + 0.0231 x_5 \geq 0.0857$$

Profit constraint:

$$0.0015 x_1 + 0.0018 x_2 + 0.0018 x_3 + 0.0021 x_4 + 0.0015 x_5 \geq 0.0087$$

Earning constraint:

$$0.0043 x_1 + 0.0048 x_2 + 0.0060 x_3 + 0.0062 x_4 + 0.0062 x_5 \geq 0.0275$$

Financial Statement managing constraint:

$$0.3106 x_1 + 0.3848 x_2 + 0.3900 x_3 + 0.4470 x_4 + 0.4692 x_5 \geq 2.0015$$

$$x_1, x_2, x_3, x_4, x_5, d_1^+, d_2^+, d_3^+, d_4^+, d_5^+, d_6^+, d_1^-, d_2^-, d_3^-, d_4^-, d_5^-, d_6^- \geq 0$$

Next, preemptive GP model is formulated by forming the objective function and adding deviation variables into the goal constraints.

Objective function:

$$\text{Minimize: } P1(d_1^-) + P2(d_2^+) + P3(d_3^-) + P4(d_4^-) + P5(d_5^-) + P6(d_6^-)$$

Subject to

$$0.1524 x_1 + 0.1891 x_2 + 0.1911 x_3 + 0.2194 x_4 + 0.2307 x_5 + d_1^- - d_1^+ = 0.9826$$

$$0.1409 x_1 + 0.1737 x_2 + 0.1741 x_3 + 0.2005 x_4 + 0.2076 x_5 + d_2^- - d_2^+ = 0.8969$$

$$0.0115 x_1 + 0.0153 x_2 + 0.0169 x_3 + 0.0189 x_4 + 0.0231 x_5 + d_3^- - d_3^+ = 0.0857$$

$$0.0015 x_1 + 0.0018 x_2 + 0.0018 x_3 + 0.0021 x_4 + 0.0015 x_5 + d_4^- - d_4^+ = 0.0087$$

$$0.0043 x_1 + 0.0048 x_2 + 0.0060 x_3 + 0.0062 x_4 + 0.0062 x_5 + d_5^- - d_5^+ = 0.0275$$

$$0.3106 x_1 + 0.3848 x_2 + 0.3900 x_3 + 0.4470 x_4 + 0.4692 x_5 + d_6^- - d_6^+ = 2.0015$$

In this study, the GP model is solved with LINGO software. LINGO is an optimization modeling software for solving linear programming, non-linear programming, integer programming and goal programming [16–18].

Table 2 Goal achievement

Goals	Output value	Goals achievement
G1	$d_1^- = 0$	Fully achieved
G2	$d_2^+ = 0$	Fully achieved
G3	$d_3^- = 0$	Fully achieved
G4	$d_4^- = 0$	Fully achieved
G5	$d_5^- = 0$	Fully achieved
G6	$d_6^- = 0$	Fully achieved

Table 3 Deviation variables

Goals	$d_i^- = 0$	$d_i^+ = 0$
G1	0	1.6862×10^{-4}
G2	2.0415×10^{-3}	0
G3	0	2.2025×10^{-3}
G4	0	0
G5	0	0
G6	0	0

3 Result and Discussion

Tables 2 and 3 present the empirical results of goal achievement and deviation variables respectively based on the optimal solution obtained.

All zero output values in Table 2 denote that all six goals are fully achieved. This result is in line with the past studies on bank financial management [2, 19]. Based on Table 3, the values in d_1^+ indicates that the goal for total asset of RHB Bank can be increased RM 0.00017 trillion in future. Besides that, the total liability can be decreased further by RM 0.00204 trillion. Amount of total equity can be increased as well by RM 0.00220 trillion. For the profitability, the value remained RM 8724 million because both deviation variables are zero. Both positive and negative deviation variables for Goal 5 and Goal 6 illustrate zero values also. It shows that total earnings for the bank does not change for five years which is equal to RM 27514 million while ratio for the items' value is equal to RM 2001505 million.

4 Conclusion

This study examines six common goals for the financial management of RHB Bank in Malaysia. GP model is developed in optimize the financial management of RHB Bank. The results indicate that RHB Bank is able to achieve all six goals based on GP model. Besides that, there are potential improvements identified on assets, liability, equity and profit. Therefore, this study can help to identify the new target values for the bank's goal for further improvement.

References

1. Naderi, S., Minouei, M., Gashti, H.P.: Asset and liability optimal management mathematical modeling for bank. *J. Basic Appl. Sci. Res.* **3**(1), 484–493 (2013)
2. Halim, B.A., Karim, H.A., Fahami, N.A., Mahad, N.F., Nordin, S.K.S., Hassan, N.: Bank financial statement management using a goal programming model. *Procedia Soc. Behav. Sci.* **211**, 498–504 (2015). <https://doi.org/10.1016/j.sbspro.2015.11.066>
3. Kosmidou, K., Zopounidis, C.: An optimization scenario methodology for bank asset liability management. *Oper. Res. Int. J.* **2**(2), 279–287 (2002)
4. Charnes, A., Cooper, W.W., Ferguson, R.O.: Optimal estimation of executive compensation by linear programming. *Manag. Sci.* **1**(2), 138–151 (1995). <https://doi.org/10.1287/mnsc.1.2.138>
5. Yahia-Berrouiguet, A., Tissourassi, K.: Application of goal programming model for allocating time and cost in project management: a case study from the company of construction seror. *Yugosl. J. Oper. Res.* **25**(2), 283–289 (2015). <https://doi.org/10.2298/YJOR131010010Y>
6. Arewa, A., Owoputi, J.A., Torbira, L.L.: Financial statement management, liability reduction and asset accumulation: an application of goal programming model to a nigerian bank. *Int. J. Financ. Res.* **4**(4), 83–90 (2013). <https://doi.org/10.5430/ijfr.v4n4p83>
7. Mohammadi, R., Sherafati, M.: Optimization of bank liquidity management using goal programming and fuzzy ahp. *Res. J. Recent Sci.* **4**(6), 53–61 (2015)
8. Petridis, K., Dey, P.K.: A DEA/goal programming model for incineration plants performance in the UK. *Procedia Environ. Sci.* **35**, 257–264 (2016). <https://doi.org/10.1016/j.proenv.2016.07.006>
9. Jayaraman R., Liuzzi D., Colapinto C., Malik T.: A fuzzy goal programming model to analyze energy, environmental and sustainability goals of the United Arab Emirates. In: Boros, E. (ed.) *Annals of Operations Research 2015*, vol 226. Springer (2015). <https://doi.org/10.1007/s10479-015-1825-5>
10. Hassan, N., Pazil, A.H.M., Idris, N.S., Razman, N.F.: A goal programming model for bakery production. *Adv. Environ. Biol.* **7**(1), 187–190 (2013)
11. Güler, M.G., Idin, K., Güler, E.Y.: A goal programming model for scheduling residents in an anesthesia and reanimation department. *Expert Syst. Appl.* **40**, 2117–2126 (2013). <https://doi.org/10.1016/j.eswa.2012.10.030>
12. Sen, N.: Goal programming model for personnel management in tea industry: a case study of barak valley of assam (India). *Am. J. Math. Stat.* **3**(6), 312–314 (2013). <https://doi.org/10.5923/j.ajms.20130306.03>
13. RHB Bank Berhad Corporate Information. <http://www.rhbgroup.com/about-us/who-we-are/corporate-information>
14. Lam, W.S., Lam, W.H.: Strategic decision making in portfolio management with goal programming model. *Am. J. Oper. Manag. Inf. Syst.* **1**(1), 34–38 (2016)
15. Lam, W.S., Jaaman, S.H., Ismail, H.: Portfolio optimization in enhanced index tracking with goal programming approach. In: *The 2014 UKM FST Postgraduate Colloquium, AIP Conference Proceedings 1614*, pp. 968–972 (2014). <https://doi.org/10.1063/1.4895332>
16. Lam, W.S., Liew, K.F., Lam, W.H.: An empirical comparison on the efficiency of healthcare companies in Malaysia with data envelopment analysis model. *Int. J. Serv. Sci. Manag. Eng.* **4**(1), 1–5 (2017)

17. Lam, W.S., Jaaman, S.H., Ismail, H.: Portfolio optimization for index tracking modelling in Malaysia stock market. In: 2nd International Conference on Mathematical Sciences and Statistics 2016, AIP Conference Proceedings 1739, pp. 020025 (2016). <https://doi.org/10.1063/1.4952505>
18. Lam, W.S., Liew, K.F., Lam, W.H.: An empirical investigation on the efficiency of the financial companies in Malaysia with DEA model. *Am. J. Inf. Sci. Comput. Eng.* **3**(3), 32–38 (2017)
19. Chen, J.W., Lam, W.S., Lam, W.H.: Optimization on the financial management of the bank with goal programming model. *J. Fundam. Appl. Sci.* **9**(6S), 442–451 (2017)

Chapter 16

Numerical Solutions on Boundary Layer of Casson Micropolar Fluid Over a Stretching Surface



Abdul Rahman Mohd Kasim, Hussein Ali Mohammed Al-Sharifi,
Nur Syamilah Arifin, Mohd Zuki Salleh and Sharidan Shafie

Abstract This paper aims to discuss the numerical solutions on the problem of Casson Micropolar fluid moving over a stretching sheet. The boundary layer and Boussinesq approximation has been consider in the formulation on the respective fluid model. The mathematical model of the fluid is first transform into a set of ordinary differential equation using similarity transformation before the computation process is carried out via Bvp4c method which embedded in Matlab program. The comparison results between present computation and previous published output show a strong agreement in the limiting case where the present model is revert back as previous model by setting some dimensionless parameters into constant values. The present result in term of velocity, temperature as well as angular velocity of the fluid is presented graphically.

Keywords Boundary layer · Numerical solution · Casson micropolar fluid · Stretching surface

A. R. M. Kasim (✉) · H. A. M. Al-Sharifi · N. S. Arifin · M. Z. Salleh
Applied & Industrial Mathematics Research Group,
Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang,
26300 Kuantan, Pahang, Malaysia
e-mail: rahmanmohd@ump.edu.my

H. A. M. Al-Sharifi
e-mail: PSE14001@stdmail.ump.edu.my

N. S. Arifin
e-mail: nursyamilaharifin@gmail.com

M. Z. Salleh
e-mail: zuki@ump.edu.my

H. A. M. Al-Sharifi
Department of Mathematics, College of Education for Pure Sciences,
University of Karbala, Karbala, Iraq

S. Shafie
Faculty of Science, Department of Mathematical Sciences, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia
e-mail: sharidan@utm.my

1 Introduction

The Navier-Stokes equations have been observed to fall short when it comes to describing fluids with a lofty molecular weight. The wide diversity of Newtonian fluids renders their properties difficult to describe by way of a single constitutive equation. This circumstance led to the development of collection of non-Newtonian fluid models. Some of them have been discovered by [1–5]. Despite the difference on its constitutive equations, the model of non-Newtonian fluid is more complicated and need extra effort in computation since the equations is more complex. Previous literatures are mostly focus on the problem of non-Newtonian fluid without considering the impact of microrotation acting on the flow. Therefore, the focus of this endeavour is to tackle the problem on the steady two dimensional convective flow of a Casson fluid together with the microrotation effect over a stretching surface. The Newtonian heating boundary condition was considers in the model where the mixed convection is taking into account as the source of heat transfer.

2 Mathematical Formulation

The problem under two-dimensional steady stagnation point of a Casson micropolar type of fluid flowing over a geometry stretching surface is considered. The uniform ambient temperature denoted as T_∞ is deliberate on the system of fluid flow. The velocity distribution is assumed to be in the forms $u_e(x) = 0$ and the velocity of the stretching surface is $u_w(x) = cx$, where c is positive constants. The physical configuration is captured in Fig. 1.

In few publications (see [6–8]), the constitutive equations which represent Casson fluid are once introduced as

$$\tau_{ij} = \begin{cases} 2\left(\mu_{\beta c} + \frac{p_v}{\sqrt{2\pi_c}}\right)e_{ij}, & \pi > \pi_c \\ 2\left(\mu_{\beta c} + \frac{p_v}{\sqrt{2\pi}}\right)e_{ij}, & \pi < \pi_c \end{cases} \quad (1)$$

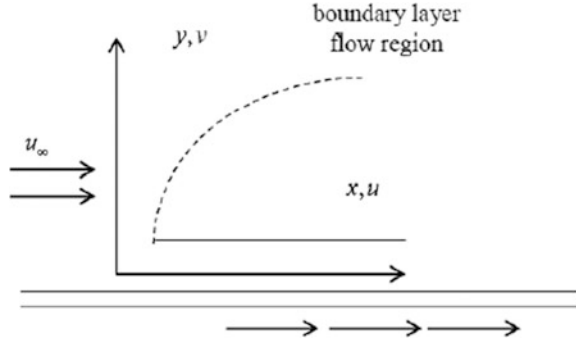
where $e_{ij} = \frac{1}{2} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right)$.

There are four main equations cover in this problem where it is adopted in [3, 9]. Continuity equation:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad (2)$$

Momentum equation:

Fig. 1 Physical geometry of fluid flow



$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \left(\nu \left(1 + \frac{1}{\beta_c} \right) + \frac{\kappa}{\rho} \right) \frac{\partial^2 u}{\partial y^2} + \frac{\kappa}{\rho} \frac{\partial N}{\partial y} + g_c \beta_T (T - T_\infty) \quad (3)$$

Energy equation:

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \frac{\kappa}{\rho c_p} \frac{\partial^2 T}{\partial y^2} \quad (4)$$

Micropolar equation:

$$u \frac{\partial N}{\partial x} + v \frac{\partial N}{\partial y} = \frac{\gamma}{\rho j} \frac{\partial^2 N}{\partial y^2} - \frac{\kappa}{\rho j} \left(2N + \frac{\partial u}{\partial y} \right) \quad (5)$$

The Eqs. (2)–(5) are subjected to the following boundary conditions

$$u = cx, v = 0, N = -m_0 \frac{\partial N}{\partial y}, \frac{\partial T}{\partial y} = -h_s T \text{ at } y = 0 \quad (6)$$

$$u \rightarrow 0, N \rightarrow 0, T \rightarrow T_\infty \text{ as } y \rightarrow \infty$$

where N is angular velocity while m_0 is constant parameter.

Introducing the similarity transformations (7) into Eqs. (2)–(5)

$$u = cx f'(\eta), v = -\sqrt{cx} f(\eta), N = cx \sqrt{\frac{c}{\nu}} g(\eta), \eta = y \sqrt{\frac{c}{\nu}}, \theta(\eta) = \frac{T - T_\infty}{T_\infty} \quad (7)$$

led to transform the respective equations to the following ordinary differential equations,

$$\left(1 + \frac{1}{\beta} + K \right) f''' + ff'' - (f')^2 + Kg' + \lambda\theta = 0 \quad (8)$$

$$\left(1 + \frac{K}{2}\right)g''(\eta) + f(\eta)g'(\eta) - f'(\eta)g(\eta) - K(2g(\eta) + f''(\eta)) = 0 \tag{9}$$

$$\theta'' + \text{Pr}f\theta' = 0. \tag{10}$$

The boundary conditions (6) also transformed to

$$\begin{aligned} f(0) = 0, f'(0) = 1, g(0) = -m_0f''(0), \theta'(0) = -\gamma(1 + \theta(0)) \text{ at } \eta = 0 \\ f'(\infty) = 0, g(\infty) = 0, \theta(\infty) = 0 \text{ as } \eta \rightarrow \infty \end{aligned} \tag{11}$$

where, Casson parameter β_c , material parameter K , mixed parameter λ , Prandtl number Pr , Conjugate parameter γ_0 , are defined as:

$$\beta_c = \frac{p_y}{\sqrt{2\pi}\mu\beta_c}, K = \frac{\kappa}{\mu}, \lambda = \frac{g_c\beta_T(T_f - T_\infty)}{c^2x}, \text{Pr} = \frac{\mu c_p}{k}, \gamma_0 = \frac{h_s}{\sqrt{\epsilon}_v} \tag{12}$$

3 Results and Discussion

The numerical scheme named Bvp4c was applied to the set of ordinary differential Eqs. (8)–(10) with respect to boundary condition (11). The Bvp4c is actually the solver from the Matlab software which applied the finite different approach where the solution starts from initial guess that supplied at an initial mesh point. Table 1 portrays a comparison between the results from this study and the documented results from a related study by Qasim et al. [9]. The comparison outcomes of $\theta(0)$ and $-\theta'(0)$ verified that the results from both sources are concurring.

Tables 2 and 3 represent the values of $f''(0)$ for different values of β_c and K respectively. The increment on the values of β_c led to give the lower value of $f''(0)$. Meanwhile, the opposite trend is found for value of K . These phenomena led to a conclusion that, the bigger values of Casson parameter gave the lesser resistance forces between the motion of fluid and the surface. This is true since the function represent the Casson parameter is in the form of rational function (Table 4).

Table 1 Comparison value of $\theta(0)$ and $-\theta'(0)$ for different values of Pr when $K = 0$ and $\gamma_0 = 1$

Pr	$\theta(0)$		$-\theta'(0)$	
	Qasim et al. [9]	Present	Qasim et al. [9]	Present
3	6.05168	6.0514	7.05168	7.0514
5	1.76039	1.7603	2.76039	2.7603
7	1.11682	1.1168	2.11682	2.1168
10	0.76452	0.76452	1.76452	1.7645
100	0.14781	0.1478	1.14780	1.1478

Table 2 Values of $f''(0)$ at $m_0 = 0.5$, $\lambda = 0.6$, $Pr = 0.72$, $K = 0.09$ and $\gamma_0 = 0.001$ for various β_c

β_c	$f''(0)$
1.0	-0.72722
1.3	-0.76930
1.5	-0.79075
1.8	-0.81634
2.0	-0.83017

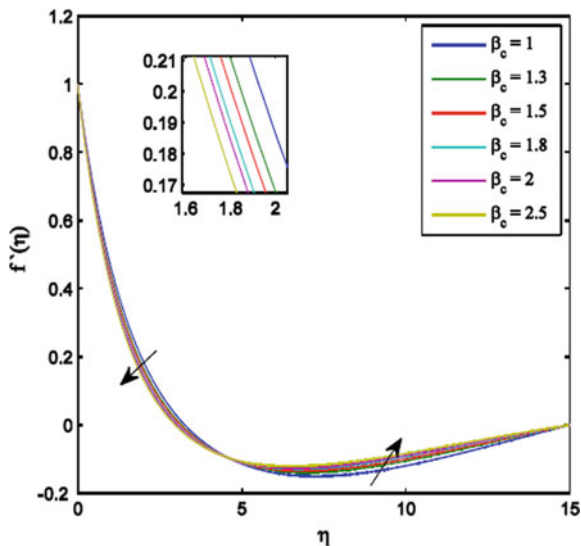
Table 3 Values of $f''(0)$ at $m_0 = 0.5$, $\lambda = 0.9$, $Pr = 1$, $\beta_c = 1$ and $\gamma_0 = 0.1$ for various K

K	$f''(0)$
0.1	-0.67657
0.2	-0.67069
0.3	-0.66470
0.4	-0.65871
0.5	-0.65280

Table 4 Values of $-\theta'(0)$ at $m_0 = 0.5$, $K = 0.1$, $\lambda = 0.5$, $\beta_c = 1$ and $\gamma_0 = 0.1$ for various Pr

Pr	$f''(0)$
0.6	0.12876
0.7	0.12504
0.8	0.12236
0.9	0.12034
1.0	0.11874

Fig. 2 Velocity profile at $m_0 = 0.5$, $K = 0.09$, $\lambda = 0.6$, $Pr = 0.72$ and $\gamma_0 = 0.001$ for various β_c



Figures 2, 3 and 4 captured the distribution of velocity, microrotation and temperature of fluid respectively. Those figures were fulfilled the boundary conditions considered which led to support the correctness of the solutions. The bigger values of β_c improve the fluid velocity ($\eta > 5$), microrotation as well as temperature. This is the fact that the bigger values of Casson parameter led to reduce the thickness of momentum and thicken the thermal boundary layer.

Fig. 3 Microrotation profile at $m_0 = 0.5$, $K = 0.09$, $\lambda = 0.6$, $Pr = 0.72$ and $\gamma_0 = 0.001$ for various β_c

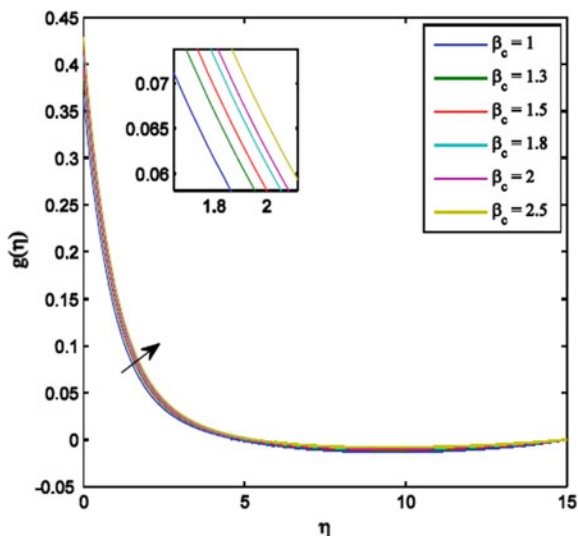
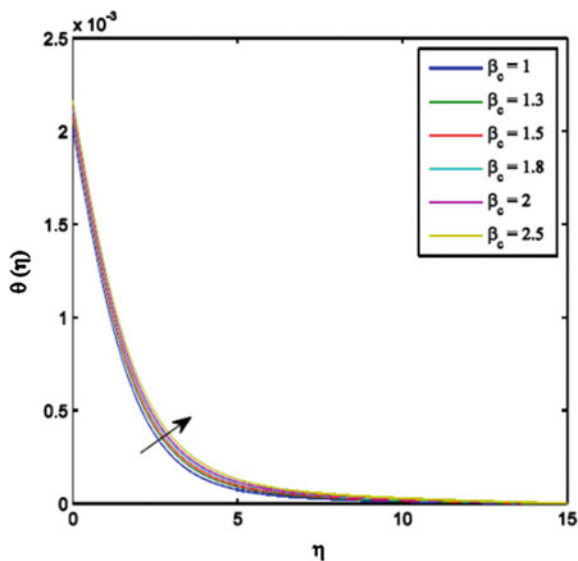


Fig. 4 Temperature profile at $m_0 = 0.5$, $K = 0.09$, $\lambda = 0.6$, $Pr = 0.72$ and $\gamma_0 = 0.001$ for various β_c



4 Conclusion

The Problem of Casson micropolar fluid with Newtonian heating boundary condition is investigated. It is revealed the parameter Casson, Material, and Prandtl number affected the flow characteristic of the fluid.

Acknowledgements The authors gratefully acknowledge the financial support received from Universiti Malaysia Pahang for (PGRS170397, RDU 160330, & RDU 170328).

References

1. Kasim, A.R.M., Mohammad, N.F., Anwar, I., Shafie, S.: MHD effect on convective bound-ary layer flow of a viscoelastic fluid embedded in porous medium with Newtonian heating. *Recent Adv. Math.* **4**, 182–189 (2013)
2. Kasim, A.R.M., Mohammad, N.F., Shafie, S.: Effect of heat generation on free convection boundary layer flow of a viscoelastic fluid past a horizontal circular cylinder with constant surface heat flux. In: 5th International Conference on Research and Education in Mathematics (ICREM5), Bandung, Indonesia, vol. 1450, pp. 286–292 (2011). <https://doi.org/10.1063/1.4724156>
3. Aurangzaib, Kasim, A.R.M., Mohammad, N.F., Shafie, S.: Unsteady MHD mixed convec-tion flow with heat and mass transfer over a vertical plate in a micropolar fluid-saturated po-rous medium. *J. Appl. Sci. Eng.* **2**, 141–150 (2013). <https://doi.org/10.6180/jase.2013.16.2.05>
4. Kasim, A.R.M., Jiann, L.Y., Shafie, S., Ali, A.: The effects of heat generation or absorption on MHD stagnation point of jeffrey fluid. In: Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21), vol. 1605, pp. 404–409. AIP Publishing (2014). <https://doi.org/10.1063/1.4887623>
5. Zokri, S.M., Arifin, N.S., Mohamed, M.K.A., Salleh, M.Z., Kasim, A.R.M., Mohammad, N.F.: Numerical solutions on mixed convection boundary layer and heat transfer of jeffrey fluid over a horizontal circular cylinder by using keller-box method. In: National Conference for Postgraduate Research (NCON-PGR) 2016, pp. 913–926 (2016)
6. El-Dabe, N.T., Ghaly, A.Y., Rizkallah, R.R., Ewis, K.M., Al-Bareda, A.S.: Numerical solution of MHD boundary layer flow of non-newtonian casson fluid on a moving wedge with heat and mass transfer and induced magnetic field. *J. Appl. Math. Phys.* **3**, 649–663 (2015). <https://doi.org/10.4236/jamp.2015.36078>
7. Hussanan, A., Salleh, M.Z., Khan, I.: Effects of newtonian heating and inclined magnetic field on two dimensional flow of a casson fluid over a stretching sheet. In: 5th World Con-ference on Applied Sciences, Engineering & Technology, pp 251–255 (2016)
8. Vajravelu, K., Prasad, K.V., Vaidya, H., Basha, N.Z., Ng, C.-O.: Mixed convective flow of a casson fluid over a vertical stretching sheet. *Int. J. Appl. Comput. Math.* **3**, 1619–1638 (2016). <https://doi.org/10.1007/s40819-016-0203-6>
9. Qasim, M., Khan, I., Shafie, S.: Heat transfer in a micropolar fluid over a stretching sheet with newtonian heating. *PLoS One* **8**, e59393 (2013). <https://doi.org/10.1371/journal.pone.0059393>

Chapter 17

One-Step Third-Derivative Block Method with Two-Hybrid Points for Solving Non-linear Dirichlet Second Order Boundary Value Problems



Mohammad Alkasassbeh and Zurni Omar

Abstract The introduction of higher derivative in the development of numerical methods for solving second order boundary value problems of ordinary differential equations has been explored by few researchers. Taking the advantages of both hybrid block methods and the presence of higher derivative in deriving numerical methods, this study proposes a new one-step hybrid block method using two-hybrid (off-step) points for solving directly initial value problems of second order ordinary differential equations with the introduction of the third derivative. To derive this method, the approximate power series solution is interpolated at $\left\{x_n, x_{n+\frac{1}{3}}\right\}$ while its second and third derivatives are collocated at all points $\left\{x_n, x_{n+\frac{1}{3}}, x_{n+\frac{2}{3}}, x_{n+1}\right\}$ on the given interval. The basic properties such as zero stability, order, consistency and convergence are also investigated in this study. In order to solve non-linear Dirichlet second-order boundary value problems, we first convert them to their equivalent initial value problems by using non-linear shooting method. Then the proposed method is employed to solve the resultant initial value problems. The numerical results indicate that the new derived method outperforms the existing methods in solving the same problems.

Keywords Hybrid block method · Dirichlet boundary value problems · Interpolation · Collocation · Third derivative

M. Alkasassbeh (✉) · Z. Omar
School of Quantitative Sciences, Universiti Utara Malaysia, Changlun, Malaysia
e-mail: mfkasassbeh@yahoo.com

Z. Omar
e-mail: zurni@uum.edu.my

© Springer Nature Singapore Pte Ltd. 2019
L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_17

1 Introduction

In this study, we are interested in solving the following second-order Dirichlet boundary value problems (BVPs)

$$y'' = (x, y, y'), y(a) = \alpha, y(b) = \beta \quad (1)$$

using hybrid block methods. The reason behind introducing hybrid block methods is to overcome the zero-stability condition encountered in block methods. Several researchers such as [1–5] proposed hybrid block methods which combined step and off-step point(s) to solve ODEs [5–7].

In order to further improve the accuracy of the solution, [8] proposed continuous third derivative formulas with block extensions for solving second order ODEs. In this article, a new one-step hybrid block method with two-hybrid points in the presence of third derivative is proposed to solve initial value problems (IVPs) of second order ordinary differential equations directly using interpolation and collocation approach. In order to solve (1) using the newly derived method, we must first transform (1) to the equivalent IVPs by employing the nonlinear shooting method. This can be achieved by approximating the solution to the BVP (1) using the solutions to the following sequence of IVP

$$y'' = (x, y, y'), y(a) = \alpha, y'(a) = t_k, k = 0, 1, 2, \dots, x \in [a, b] \quad (2)$$

which satisfies the following condition

$$\lim_{n \rightarrow \infty} y(b, t_k) = y(b) = \beta. \quad (3)$$

To start the shooting method, we first choose an initial guess t_0 that produces $y(b, t_0)$. To arrive at a solution to the BVP (1), the parameter t has to be chosen such that $y(b, t) = \beta$; i.e., we have to solve the equation

$$F(t) = 0,$$

where the function $F : R \rightarrow R$ is defined by

$$F(t) = y(b, t) - \beta.$$

For finding a zero of F , the iterative method with three-step proposed by [9] is employed instead of Newton method because the former method converges faster. For the computation of the derivative $F'(t)$ which is required for the three-step iterative method, we assume that the solution y to the IVP (2) depends on a continuously differentiable manner on the parameter t . By setting $v(x, t) = \frac{\partial y(x, t)}{\partial t}$, Eq. (2) can now be written as

$$y''(x, t) = f(x, y(x, t), y'(x, t)), \text{ for } x \in [a, b], \text{ with } y(a, t) = \alpha, y'(a, t) = t. \quad (4)$$

Differentiating Eq. (4) once with respect to t , we obtain

$$\frac{\partial y''(x, t)}{\partial t} = f_x(x, y, y') \frac{\partial x}{\partial t} + f_y(x, y, y') \frac{\partial y(x, t)}{\partial t} + f_{y'}(x, y, y') \frac{\partial y'(x, t)}{\partial t}.$$

Since both x and t are independent variables in the above equation, then $\frac{\partial x}{\partial t} = 0$. The above equation is reduced to

$$\frac{\partial y''(x, t)}{\partial t} = f_y(x, y, y') \frac{\partial y(x, t)}{\partial t} + f_{y'}(x, y, y') \frac{\partial y'(x, t)}{\partial t} \quad (5)$$

after substituting $\frac{\partial y(x, t)}{\partial t} = v(x, t)$. Now, differentiating the corresponding initial conditions of Eq. (4) with respect to the variable t we obtain the initial conditions

$$v(a, t) = 0, v'(a, t) = 1.$$

Since $F'(t) = v(b, t)$, computing the derivative of F requires solving the additional initial value problem (4) and (5) for $v(x, t)$, where $y(x, t)$ is known from solving (4). Note that from a numerical approximation, $y(x, t)$ is known only at grid points. It can be summarized the algorithm for the shooting method with the three-step iterative method in the following steps:

Step 1: Choose an initial slope $t \in R$ for the second order BVP Eq. (4).

Step 2: Solve numerically the obtained IVP and its derivative with respect, i.e.

$$y''(x, t) = f(x, y, y')$$

and

$$\frac{\partial y''(x, t)}{\partial t} = f_y(x, y, y') \frac{\partial y(x, t)}{\partial t} + f_{y'}(x, y, y') \frac{\partial y'(x, t)}{\partial t}.$$

Step 3: If $|F(t)| = |y(b, t) - \beta| < \epsilon$ for some small value $\epsilon > 0$ then stop; otherwise, update t_{k+1} using the following equation

$$t_{k+1} = T_k + U_k - \frac{F(T_k + U_k)}{F'(t_k)}$$

where

$$T_k = t_k - \frac{F(t_k)}{F'(t_k)}, U_k = -\frac{F(T_k)}{F'(t_k)}$$

and go back to Step 2 [10].

2 Development of the Method

In order to approximate the solution of Eq. (1) above, the following function is employed

$$y(x) = \sum_{j=0}^{2s+r-1} a_j \left(\frac{x-x_n}{h}\right)^j \tag{6}$$

where r and s refer to the interpolation and collocation points respectively. On differentiating (6) twice and thrice the below two equations are produced

$$y''(x) = \sum_{j=0}^{2s+r-1} \frac{j!a_j}{h^2(j-2)!} \left(\frac{x-x_n}{h}\right)^{j-2} = f(x, y, y'), \tag{7}$$

$$y'''(x) = \sum_{j=0}^{2s+r-1} \frac{j!a_j}{h^3(j-3)!} \left(\frac{x-x_n}{h}\right)^{j-3} = g(x, y, y'). \tag{8}$$

Now, Eq. (6) is interpolated at the points $x_{n+\hat{r}}$, for $\hat{r} = \{0, \frac{1}{3}\}$, i.e. $r = 2$, and Eqs. (7) and (8) are collocated at all points $x_{n+\hat{s}}$, for $\hat{s} = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, i.e. $s = 4$, and then combining the resulted equations yields a block in matrix form

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & \frac{1}{3} & \frac{1}{9} & \frac{1}{27} & \frac{1}{81} & \frac{1}{243} & \frac{1}{729} & \frac{1}{2187} & \frac{1}{6561} & \frac{1}{19683} \\ 0 & 0 & \frac{2}{h^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{h^2} & \frac{2}{h^2} & \frac{4}{3h^2} & \frac{20}{27h^2} & \frac{10}{27h^2} & \frac{14}{81h^2} & \frac{56}{729h^2} & \frac{8}{243h^2} \\ 0 & 0 & \frac{2}{h^2} & \frac{4}{h^2} & \frac{16}{3h^2} & \frac{160}{27h^2} & \frac{160}{27h^2} & \frac{448}{81h^2} & \frac{3584}{729h^2} & \frac{1024}{243h^2} \\ 0 & 0 & \frac{2}{h^2} & \frac{6}{h^2} & \frac{12}{h^2} & \frac{20}{h^2} & \frac{30}{h^2} & \frac{42}{h^2} & \frac{56}{h^2} & \frac{72}{h^2} \\ 0 & 0 & 0 & \frac{6}{h^3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{6}{h^3} & \frac{8}{h^3} & \frac{20}{3h^3} & \frac{40}{9h^3} & \frac{70}{27h^3} & \frac{112}{81h^3} & \frac{56}{81h^3} \\ 0 & 0 & 0 & \frac{6}{h^3} & \frac{16}{h^3} & \frac{80}{3h^3} & \frac{320}{9h^3} & \frac{1120}{27h^3} & \frac{3584}{81h^3} & \frac{3584}{81h^3} \\ 0 & 0 & 0 & \frac{6}{h^3} & \frac{24}{h^3} & \frac{60}{h^3} & \frac{120}{h^3} & \frac{210}{h^3} & \frac{336}{h^3} & \frac{504}{h^3} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \end{pmatrix} = \begin{pmatrix} y_n \\ y_{n+\frac{1}{3}} \\ f_n \\ f_{n+\frac{1}{3}} \\ f_{n+\frac{2}{3}} \\ f_{n+1} \\ g_n \\ g_{n+\frac{1}{3}} \\ g_{n+\frac{2}{3}} \\ g_{n+1} \end{pmatrix} \tag{9}$$

Solving system (9) for the unknown values a'_j s using Gaussian manipulation and substituting these values back into Eq. (6) produces a linear multi-step continuous hybrid block method in the form

$$y(x) = \sum_{i=0, \frac{1}{3}} \alpha_i y_{n+i} + h^2 \sum_{i=0, \frac{1}{3}, \frac{2}{3}, 1} \beta_i f_{n+i} + h^3 \sum_{i=0, \frac{1}{3}, \frac{2}{3}, 1} \gamma_i g_{n+i} \tag{10}$$

where the coefficients α_i , β_i and γ_i for $i = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ are listed below

$$\alpha_0 = \frac{3}{h} \left(\frac{h}{3} - x - x_n \right)$$

$$\alpha_1 = \frac{3}{h} (x - x_n)$$

$$\begin{aligned} \beta_0 = & \frac{(x - x_n)^2}{2} - \frac{97}{16h^2} (x - x_n)^4 + \frac{1691}{80h^3} (x - x_n)^5 - \frac{141}{4h^4} (x - x_n)^6 \\ & + \frac{129}{4h^5} (x - x_n)^7 - \frac{3483}{224h^6} (x - x_n)^8 + \frac{99}{32h^7} (x - x_n)^9 \\ & - \frac{7536502154204791h}{78812993478983680} (x - x_n), \end{aligned}$$

$$\begin{aligned} \beta_{\frac{1}{3}} = & \frac{243}{36028797018963968h^2} (x - x_n)^4 + \frac{243}{20h^3} (x - x_n)^5 \\ & - \frac{81}{2h^4} (x - x_n)^6 + \frac{2997}{56h^5} (x - x_n)^7 - \frac{3645}{112h^6} (x - x_n)^8 \\ & + \frac{243}{32h^7} (x - x_n)^9 - \frac{1301h}{30240} (x - x_n), \end{aligned}$$

$$\begin{aligned} \beta_{\frac{2}{3}} = & -\frac{81}{16h^2} (x - x_n)^4 - \frac{2187}{80h^3} (x - x_n)^5 + \frac{243}{4h^4} (x - x_n)^6 \\ & - \frac{1863}{28h^5} (x - x_n)^7 + \frac{8019}{224h^6} (x - x_n)^8 - \frac{243}{32h^7} (x - x_n)^9 \\ & - \frac{181h}{7560} (x - x_n), \end{aligned}$$

$$\begin{aligned} \beta_1 = & \frac{(x - x_n)^4}{h^2} - \frac{119}{20h^3} (x - x_n)^5 + \frac{15}{h^4} (x - x_n)^6 - \frac{1077}{56h^5} (x - x_n)^7 \\ & + \frac{1377}{112h^6} (x - x_n)^8 - \frac{99}{32h^7} (x - x_n)^9 - \frac{3329h}{816480} (x - x_n), \end{aligned}$$

$$\begin{aligned} \gamma_0 = & \frac{9(x - x_n)(\frac{h}{3} - x + x_n)}{1120h^6} \left(-\frac{2597h^7}{2187} - \frac{2597h^6x}{729} + \frac{2597h^7x_n}{729} \right. \\ & + \frac{12523h^5x^2}{243} - \frac{25046h^5xx_n}{243} + \frac{12523h^5x_n^2}{243} - \frac{15197h^4x_n^3}{81} \\ & + \frac{15197h^4x^2x_n}{27} - \frac{15197h^4xx_n^2}{27} + \frac{15197h^4x_n^3}{81} + \frac{9121h^3x^4}{27} \\ & - \frac{36484h^3x^3x_n}{27} + \frac{18242h^3x^2x_n^2}{9} - \frac{36484h^3xx_n^2}{27} + \frac{9121h^3x_n^4}{27} \\ & - \frac{2975h^2x^5}{9} + \frac{14875h^2x^4x_n}{9} - \frac{29750h^2x^3x_n^2}{9} + \frac{29750h^2x^2x_n^3}{9} \\ & - \frac{14875h^2xx_n^4}{9} + \frac{2975h^2x_n^5}{9} + \frac{505hx^6}{3} - 1010hx^5x_n + 2525hx^4x_n^2 \\ & - 1010hx_n^2 - \frac{10100hx^3x_n^3}{3} + 2525hx^2x_n^4 + 1225x^4x_n^3 + \frac{505hx_n^6}{3} \\ & - 35x^7 + 245x^6x_n - 735hx^5x_n^2 - 1225x^4x_n^3 + 735hx^2x_n^5 - 1225x^3x_n^4 \\ & \left. + 735hx^2x_n^5 - 245xx_n^6 + \frac{70}{2}hx_n^7 \right), \end{aligned}$$

$$\begin{aligned}\gamma_{\frac{1}{3}} &= \frac{54(x-x_n)^5}{5h^2} - \frac{9(x-x_n)^4}{4h} - \frac{873(x-x_n)^6}{40h^3} + \frac{1269(x-x_n)^7}{56h^4} \\ &\quad - \frac{2673(x-x_n)^8}{224h^5} + \frac{81(x-x_n)^9}{32h^6} + \frac{313(x-x_n)}{22680}, \\ \gamma_{\frac{2}{3}} &= \frac{513(x-x_n)^5}{80h^2} - \frac{9(x-x_n)^4}{8h} - \frac{153(x-x_n)^6}{10h^3} + \frac{513(x-x_n)^7}{28h^4} \\ &\quad - \frac{1215(x-x_n)^8}{112h^5} + \frac{81(x-x_n)^9}{32h^6} + \frac{89(x-x_n)}{18144}, \\ \gamma_1 &= \frac{(x-x_n)^5}{2h^2} - \frac{(x-x_n)^4}{12h} - \frac{51(x-x_n)^6}{40h^3} + \frac{93(x-x_n)^7}{56h^4} \\ &\quad - \frac{243(x-x_n)^8}{224h^5} + \frac{9(x-x_n)^9}{32h^6} + \frac{137(x-x_n)}{408240}.\end{aligned}$$

On deriving Eq. (10) once we obtain

$$y'(x) = \frac{1}{h} \sum_{i=0, \frac{1}{3}} \alpha'_i y_{n+i} + h \sum_{i=0, \frac{1}{3}, \frac{2}{3}, 1} \beta'_i f_{n+i} + h^2 \sum_{i=0, \frac{1}{3}, \frac{2}{3}, 1} \gamma'_i g_{n+i} \quad (11)$$

Interpolating (10) at the points $\{x_{n+\frac{2}{3}}, x_{n+1}\}$ and collocating (11) at the points $\{x_n, x_{n+\frac{1}{3}}, x_{n+\frac{2}{3}}, x_{n+1}\}$, a hybrid block method is formed as shown below

$$A^{(0)}Y_{m+1} = A^{(1)}Y_m + \sum_{i=0}^1 B^{(i)}F_{m+i} + \sum_{i=0}^1 D^{(i)}G_{m+i} \quad (12)$$

where

$$A^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, A^{(1)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & \frac{h}{3} \\ 0 & 0 & 1 & 0 & 0 & \frac{2h}{3} \\ 0 & 0 & 1 & 0 & 0 & h \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$B^{(1)} = \begin{pmatrix} \frac{1301 h^2}{90720} & \frac{181 h^2}{22680} & \frac{3329 h^2}{2449440} \\ \frac{296 h^2}{2835} & \frac{7575499373 1993 h^2}{1970324836 974592} & \frac{344 h^2}{76545} \\ \frac{243 h^2}{1120} & \frac{1823957849 0849715 h^2}{1261007895 66373888} & \frac{2955487255 46181 h^2}{1576259869 5796736} \\ \frac{313 h}{2016} & \frac{89 h}{2016} & \frac{397 h}{54432} \\ \frac{20 h}{63} & \frac{4065749663 59773 h}{1970324836 974592} & \frac{20 h}{1701} \\ \frac{20 h}{63} & \frac{81 h}{224} & \frac{3895412624 12043 h}{2814749767 106560} \end{pmatrix},$$

$$D^{(0)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \frac{371 h^3}{349920} \\ 0 & 0 & 0 & 0 & 0 & \frac{206 h^3}{76545} \\ 0 & 0 & 0 & 0 & 0 & \frac{h^3}{224} \\ 0 & 0 & 0 & 0 & 0 & \frac{1283 h^3}{272160} \\ 0 & 0 & 0 & 0 & 0 & \frac{43 h^3}{8505} \\ 0 & 0 & 0 & 0 & 0 & \frac{19 h^3}{3360} \end{pmatrix}, D^{(1)} = \begin{pmatrix} \frac{-313 h^3}{68040} & \frac{-89 h^3}{54432} & \frac{-137 h^3}{1224720} \\ \frac{-20 h^3}{1701} & \frac{-52 h^3}{8505} & \frac{-4 h^3}{10935} \\ \frac{-9 h^3}{1120} & \frac{-9 h^3}{1120} & \frac{-h^3}{840} \\ \frac{-851 h^2}{30240} & \frac{-269 h^2}{30240} & \frac{-163 h^2}{272160} \\ \frac{-16 h^2}{945} & \frac{-19 h^2}{945} & \frac{-8 h^2}{8505} \\ \frac{-9 h^2}{1120} & \frac{81 h^2}{1120} & \frac{-19 h^2}{3360} \end{pmatrix},$$

$$Y_{m+1} = \left[y_{n+\frac{1}{3}}, y_{n+\frac{2}{3}}, y_{n+1}, y'_{n+\frac{1}{3}}, y'_{n+\frac{2}{3}}, y'_{n+1} \right]^T,$$

$$Y_m = \left[y_{n-\frac{2}{3}}, y_{n-\frac{1}{3}}, y_n, y'_{n-\frac{1}{3}}, y'_{n-\frac{2}{3}}, y'_n \right]^T,$$

$$F_{m+1} = \left[f_{n+\frac{1}{3}}, f_{n+\frac{2}{3}}, f_{n+1} \right]^T,$$

$$F_m = \left[f_{n-\frac{2}{3}}, f_{n-\frac{1}{3}}, f_{n-1}, f_{n-\frac{2}{3}}, f_{n-\frac{1}{3}}, f_n \right]^T,$$

$$G_{m+1} = \left[g_{n+\frac{1}{3}}, g_{n+\frac{2}{3}}, g_{n+1} \right]^T, \text{ and}$$

$$G_m = \left[g_{n-\frac{2}{3}}, g_{n-\frac{1}{3}}, g_{n-1}, g_{n-\frac{2}{3}}, g_{n-\frac{1}{3}}, g_n \right]^T.$$

3 Analysis of the Method

3.1 Zero Stability

Definition 3.1 The zero stability property of the hybrid block method formula (12) is satisfied if all roots (R_z) of the first characteristic equation $\rho(R)$ are inside the unit circle and if $R_z = 1$ then the multiplicity of (R_z) must not exceed two.

Now,

$$\det[RA^{(0)} - A^{(1)}] = \left| R \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right| = 0$$

implies that

$$R^4(R - 1)^2 = 0$$

whose solutions are $R = 0, 0, 0, 0, 1, 1$. Hence (12) is zero stable.

3.2 Order of the Method

The block method formula (12) according to [11, 12] possess an order p if the linear operator ∇ associated with the block can be expressed

$$\nabla\{y(x), h\} = A^{(0)}Y_{m+1} - A^{(1)}Y_m - \sum_{i=0}^1 B^{(i)}F_{m+i} - \sum_{i=0}^1 D^{(i)}G_{m+i} \quad (13)$$

Using Taylor series expansion and gathering similar terms

$$\nabla\{y(x), h\} = \sum_{i=0}^{\infty} C_i h^i y^{(i)} = 0 \quad (14)$$

where

$$\hat{C}_0 = \hat{C}_1 = \hat{C}_2 = \dots = \hat{C}_{p+1} = 0 \text{ and } \hat{C}_{p+2} \neq 0.$$

Comparing the coefficients of $y^{(i)}$ and h^i produces $\hat{C}_0 = \hat{C}_1 = \hat{C}_2 = \dots = \hat{C}_9 = 0$ with vector of error constants

$$\hat{C}_{10} = [1.1967e^{-10}, 3.6268e^{-10}, 7.2903e^{-10}, 6.2603e^{-10}, 8.3203e^{-10}, 1.4581e^{-10}]^T$$

which concludes that the order p of this algorithm is 8.

3.3 Consistency

Definition 3.2 A hybrid block method is said to be consistent if ($p \geq 1$) i.e., the order is greater than one.

From the analysis shown above for the hybrid block method, we conclude that the order is greater than one. Thus, the hybrid block method consistent.

3.4 Convergence

Theorem 3.1 Referring to [13] a linear multi-step method is convergent if it is consistent and zero stable.

The hybrid block method formula (12) is convergent since it fulfills both the consistency and zero stability conditions.

4 Numerical Examples

This section, consider the performance and accuracy of the hybrid third-derivative one-step implicit block method (12). A three experimental Dirichlet non-linear boundary value problems are solved simultaneously using this method. It should be noted that the tolerance used for both Problems 2 and 3 is $1.00E(-05)$ and the step size is $h = \frac{1}{20}$ for example 2 and $h = \frac{1}{100}$ for Problem 3. The starting value $t_0 = 0, 0.5$ and $\frac{4}{3}$ used for the three problems respectively.

The following notations are used in the tables:

- * The (*) sign refers to maximum error.
- Tol Tolerance.
- TS Total Step.
- HA Method proposed by Ha (2001) [16].
- DMS Method suggested by Jafri et al. (2009) [16].
- 2PDAM4 Direct two-step Adams Moulton method of order four [16].
- 2PDAM5 Direct two-step Adams Moulton method of order five [16].

Table 1 Comparison of the developed method with [14, 15]

Tol	TS	Error [14]	Error [15]	Error(new method)
1E(-2)	11	2.46E(-4)	7.66E(-5)	9.66E(-06)
1E(-4)	16	3.94E(-5)	2.47E(-6)	9.70E(-06)
1E(-6)	26	4.94E(-6)	3.55E(-8)	3.64E(-14)
1E(-8)	32	6.17E(-7)	1.16E(-8)	1.56E(-14)
1E(-10)	57	5.00E(-8)	1.07E(-10)	5.34E(-15)

Problem (1)

$$f(x, y, y') = -y^2 + \sin^2(\pi x) - \pi^2 \sin(\pi x); y(0) = 0, y(1) = 0, 0 \leq x \leq 1$$

Exact Solution: $y(x) = \sin(\pi x)$. Source: [9] (Tables 1, 2 and 3).

Problem (2)

$$f(x, y, y') = \frac{3}{2}y^2; y(0) = 4, y(1) = 1, 0 \leq x \leq 1$$

Exact Solution: $(x) = \frac{4}{(1+x)^3}$, Source: [16].

Problem (3)

$$f(x, y, y') = \frac{32 + 2x^3 - yy'}{8}; y(1) = 17, y(3) = \frac{43}{3}, 1 \leq x \leq 3$$

Exact Solution: $(x) = x^2 + \frac{16}{x}$. Source: [16].

Table 2 Comparison of the developed method with HA, DMS and 2PDAM4

x	Error(HA)	Error(DMS)	Error(2PDAM4)	Error(new method)
0.10	3.0E(-06)	3.19E(-04)*	4.17E(-6)	9.35E(-10)
0.20	4.0E(06)*	2.62E(-04)	4.65E(-6)*	1.96E(-09)
0.30	4.0E(-06)	2.11E(-04)	4.62E(-06)	3.14E(-09)
0.40	4.0E(-06)	1.68E(-04)	4.31E(-06)	4.56E(-09)
0.50	4.0E(-06)	1.32E(-04)	3.84E(-06)	6.25E(-09)
0.60	3.0E(-06)	1.00E(-04)	3.26E(-06)	8.27E(-09)
0.70	3.0E(-06)	7.21E(-05)	2.58E(-06)	1.06E(-08)
0.80	3.0E(-06)	4.62E(-05)	1.82E(-06)	1.35E(-08)
0.90	3.0E(-06)	2.17E(-05)	9.64E(-07)	1.69E(-08)
1.00	3.0E(-06)	2.40E(-07)	1.75E(-10)	2.08E(-08)*

Table 3 Comparison of the developed method with HA, 2PDAM4 and 2PDAM5

x	Error(HA)	Error(2PDAM4)	Error(2PDAM5)	Error (new method)
1.00	0.00	0.00	0.00	0.00
1.20	2.00E(-06)	1.88E(-08)	5.86E(-09)*	6.75E(-14)
1.40	3.00E(-06)	2.28E(-08)*	4.03E(-09)	7.46E(-14)
1.60	1.00E(-06)	2.09E(-08)	2.84E(-09)	6.92E(-14)
1.80	1.00E(-06)	1.71E(-08)	1.98E(-09)	7.28E(-14)
2.00	0.00E(-06)	1.30E(-08)	1.35E(-09)	5.68E(-14)
2.20	2.00E(-06)	9.14E(-09)	8.80E(-10)	3.55E(-14)
2.40	4.00E(-06)	5.89E(-09)	5.34E(-10)	3.73E(-14)
2.60	6.00E(-06)	3.31E(-09)	2.86E(-10)	4.79E(-14)
2.80	9.00E(-06)	1.37E(-09)	1.14E(-10)	6.03E(-14)
3.00	1.10E(05)*	3.48E(-15)	3.94E(-15)	7.81E(-14)*

5 Conclusion

A new hybrid block method is successfully derived. The numerical results indicate that the developed method outperform other existing methods regarding error for solving the same second-order boundary value problems with Dirichlet conditions. Thus, the developed method should be highly considered as a better option in solving the given problems due to its superior performance and possessing good properties of a numerical method as well.

References

1. Abdelrahim, R.F., Omar, Z.: Direct solution of second-order ordinary differential equation using a single-step hybrid block method of order five. *Math. Comput. Appl.* **21**, 1–12 (2016). <https://doi.org/10.3390/mca21020012>
2. Sahi, R.K., Jator, S.N., Khan, N.A.: A Simpson's-type second derivative method for stiff systems. *Int. J. Pure Appl. Math.* **81**, 619–633 (2012)
3. Gear, C.W.: Hybrid methods for initial value problems in ordinary differential equations. *J. Soc. Ind. Appl. Math. Ser. B: Numer. Anal.* **2**, 69–86 (1965). <https://doi.org/10.1137/0702006>
4. Dahlquist, G.: Convergence and stability in the numerical integration of ordinary differential equations. *Mathematica Scandinavica* **4**, 33–53 (1956). <https://doi.org/10.7146/math.scand.a-10454>
5. Jator, S.N.: Solving second order initial value problems by a hybrid multistep method without predictors. *Appl. Math. Comput.* **217**, 4036–4046 (2010). <https://doi.org/10.1016/j.amc.2010.10.010>
6. Enright, W.H.: Second derivative multistep methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.* **11**, 321–331 (1974). <https://doi.org/10.1137/0711029>
7. Gragg, W.B., Stetter, H.J.: Generalized multistep predictor-corrector methods. *J. ACM (JACM)* **11**, 188–209 (1964). <https://doi.org/10.1145/321217.321223>
8. Jator, S., Akinfenwa, A., Okunuga, S.A., Sofoluwe, A.B.: High-order continuous third derivative formulas with block extensions for $y'' = f(x; y; y')$. *Int. J. Comput. Math.* **90**, 1899–1914 (2013). <https://doi.org/10.1080/00207160.2013.766329>
9. Yun, J.H.: A note on three-step iterative method for nonlinear equations. *Appl. Math. Comput.* **202**, 401–405 (2008). <https://doi.org/10.1016/j.amc.2008.02.002>
10. Kress, R.: *Numerical Analysis*. Springer, New York (1998). https://doi.org/10.1007/978-1-4612-0599-9_9
11. Lambert, J.: *Computational Methods in Ordinary Differential Equations*. Wiley, Chichester (1973). <https://doi.org/10.1137/0709052>
12. Fatunla, S.: Block methods for second order odes. *Int. J. Comput. Math.* **41**, 55–63 (1991). <https://doi.org/10.1080/00207169108804026>
13. Henrici, P.: *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York (1962). <https://doi.org/10.1126/science.136.3511.143-a>
14. Chen, X., Chen, Z., Wu, B., Xu, Y.: Fast multilevel augmentation methods for nonlinear boundary integral equations. *SIAM J. Numer. Anal.* **49**, 2231–2255 (2011). <https://doi.org/10.1137/100807600>

15. Phang, P.S., Abdul Majid, Z., Ismail, F., Othman, K.I., Suleiman, M.: New algorithm of two-point block method for solving boundary value problem with Dirichlet and Neumann boundary conditions. *Math. Probl. Eng.* **2013**, 1–10 (2013). <https://doi.org/10.1155/2013/917589>
16. Phang, P.S., Majid, Z.A., Suleiman, M.: Solving nonlinear two point boundary value problem using two step direct method. *J. Qual. Meas. Anal.* **7**, 129–140 (2011). <https://doi.org/10.1063/1.4915691>

Chapter 18

Pricing Asian Option by Solving Black–Scholes PDE Using Gauss–Seidel Method



W. S. Koh, R. R. Ahmad, S. H. Jaaman and J. Sulaiman

Abstract The main purpose of this paper is to study the pricing of the Asian option by using Gauss–Seidel iterative method via the finite difference approximation equation. Actually, Asian option is an option that is taking the average price of the underlying asset over the lifetime of the option. To solve the proposed problem numerically, a two-dimensional Black–Scholes Partial Differential Equation (PDE) governing the Asian option is discretized by using the second-order Crank–Nicolson discretization scheme. Then, the linear system generated from the discretization process is solved by using Gauss–Seidel iterative method. The results of the numerical computation were shown and discussed.

Keywords Asian option · Black–Scholes partial differential equation · Crank–Nicolson approximation scheme · Gauss–Seidel iterative method

1 Introduction

Option is a financial derivative which gives the holder the right to trade the underlying asset by a certain date for a certain price. The call option gives the holder the rights to buy while the put option gives the holder the right to sell. Asian option are path dependent options whose payoff functions depend on the average of

W. S. Koh (✉)
INTI International University, Nilai, Malaysia
e-mail: weisin.koh@newinti.edu.my

W. S. Koh · R. R. Ahmad · S. H. Jaaman
Universiti Kebangsaan, Bangi, Malaysia
e-mail: rozy@ukm.edu.my

S. H. Jaaman
e-mail: shj@ukm.edu.my

J. Sulaiman
Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: jumat@ums.edu.my

underlying asset over life of the option. It is called ‘Asian’ because it was started in Tokyo, Japan.

Generally, there are plenty approaches to determine Asian option pricing namely, Monte Carlo approach [1, 2], solving the Partial Differential Equation (PDE) approach [3–5] and other models such as Variance Gamma model [6] and Levy model [7, 8]. In the PDE approaches, there are closed form [3–5, 11–13] and numerical solutions [9, 10].

Certain transformations were used to transform the PDE into different equations to obtain closed form solutions such as the Laplace transform [11, 12]. Rogers and Shi used general transformation techniques to reduce the PDE to a one dimensional PDE and gave the analytical solutions [3]. However, Dubois and Lelièvre [10] provided numerical solutions for Roger and Shi’s PDE [3]. Then, Vecer [4] and Barucci et al. [13] introduced a change of variables techniques and solved it analytically respectively. Recently, Elshegmani [5] used the general stochastic differential equations to obtain a modified PDE with four different cases.

Nevertheless, Lee and Chin [9] solved the two-dimensional Black–Scholes PDE by using simple Crank–Nicolson finite difference method. The discretization is based on a four-point stencil approximation scheme on each time level. But, they form a tridiagonal linear system by setting the option’s price on average space a constant. In other words, the computation is performed on each average stock’s price interval at each time interval with direct method.

In this paper, however, we discretized the two dimensional Black–Scholes PDE and a five-point stencil approximation scheme was formed on each time level. Then, the linear system was solved by using Gauss–Seidel iterative method on each time interval. Finally, the numerical results were shown and discussed.

2 The Model

Suppose A is the running sum of a stock price or underlying asset within time t is

$$A_t = \int_0^t s(u)du \quad (1)$$

And the Black–Scholes PDE governing the arithmetic Asian options is

$$\frac{\partial v}{\partial t} + \frac{\sigma^2 s^2}{2} \frac{\partial^2 v}{\partial s^2} + rs \frac{\partial v}{\partial s} + s \frac{\partial v}{\partial A} - rv = 0 \quad (2)$$

where v is the option price, σ is the volatility, s is the stock or underlying asset’s price, r is the risk free interest rate. The payoff function is

$$V(s, A_T, T) = \max\left(\frac{A_T}{T} - K, 0\right), \text{ for fixed strike call option;}$$

$$V(s, A_T, T) = \max\left(K - \frac{A_T}{T}, 0\right), \text{ for fixed strike put option;}$$

$$V(s_T, A_T, T) = \max\left(s_T - \frac{A_T}{T}, 0\right), \text{ for floating strike call option;}$$

$$V(s_T, A_T, T) = \max\left(\frac{A_T}{T} - s_T, 0\right), \text{ for floating strike put option;}$$

where K is the strike price. Based on certain transformation technique, Rogers and Shi [3] showed that when $A \geq KT$,

$$V(s, A, t) = S\left(\frac{1 - e^{-r(T-t)}}{rT}\right) + e^{-r(T-t)}\left(\frac{A}{T} - K\right).$$

Thus, the option pricing when $A < KT$ is solved by numerical methods.

3 Crank–Nicolson Discretization Scheme

The Crank–Nicolson approximation equations can be derived from Eq. (1) as

$$\frac{v_{i,j,k+1} - v_{i,j,k}}{\Delta t} = \frac{1}{2}(L_{k+1} + L_k) \quad (3)$$

where,

$$L_k = -rs_i D_s - \frac{\sigma^2 s_i^2}{2} D_{ss} - s_i D_A + r v_{i,j,k},$$

$$D_s = \frac{v_{i+1,j,k} - v_{i-1,j,k}}{2\Delta s},$$

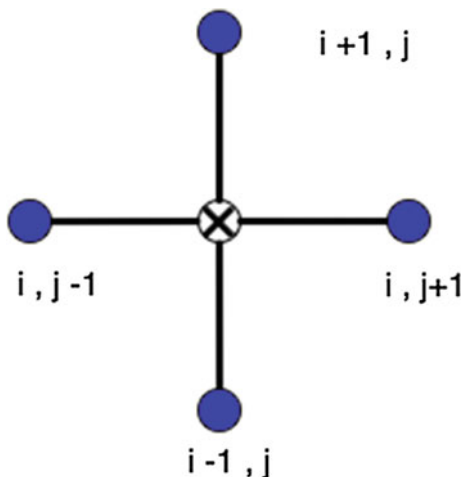
$$D_{ss} = \frac{v_{i-1,j,k} - 2v_{i,j,k} + v_{i+1,j,k}}{(\Delta s)^2},$$

$$D_A = \frac{v_{i,j+1,k} - v_{i,j-1,k}}{2\Delta A},$$

i , j , and k denote the nodes for stock price s , average stock price A , and time t respectively.

The five-point approximation scheme for Crank–Nicolson approximation scheme at each time level, k can be illustrated in Fig. 1.

Fig. 1 Computational nodes at each time level, k



4 Gauss–Seidel Iterative Method

The approximation equation in (3) generated large sparse linear system of form

$$B \tilde{v} = \tilde{f} \tag{4}$$

where, B is a pentadiagonal coefficient matrix, \tilde{f} is the known column vector, computed from the previous level with the boundary condition and \tilde{v} is the column vector for the option price v at time t . In general, the linear system in (4) was solved by using Algorithm 1 (Gauss–Seidel) method.

Algorithm 1 (GS method):

- i. Initializing all the parameters. Set $h = 0$.
- ii. General iteration

$$v_i^{(h+1)} = \frac{1}{B_{ii}} \left(f_i - \sum_{j=1}^{i-1} B_{ij} v_j^{(h+1)} - \sum_{j=i+1}^{i-n} B_{ij} v_j^{(h)} \right)$$

- iii. Convergence test.

If the error tolerance is fulfilled, the value option at that time level is $v_i^{(h+1)}$ and the algorithm stops.

Else, set $h = h + 1$ and go to step ii.

Table 1 Numerical results of 5-point Crank–Nicolson scheme solved with Gauss–Seidel (GS) iterative method and Levy’s approximation formula with $K = 20$, $T_{max} = 1$

R	σ	s = 20			s = 40			s = 80		
		CNGS	Levy	Relative error	CNGS	Levy	Relative error	CNGS	Levy	Relative error
0.05	0.25	1.31	1.38	4.98E-2	19.86	19.99	6.86E-2	58.86	59.01	2.44E-3
	0.50	2.45	2.45	1.78E-2	19.61	20.01	1.71E-2	58.29	59.01	1.23E-2
	0.70	3.09	3.42	9.64E-2	19.69	20.18	2.45E-2	58.02	59.01	1.68E-2
0.10	0.25	1.54	1.62	4.75E-2	19.98	19.97	6.20E-4	57.92	58.03	1.92E-3
	0.50	2.51	2.68	6.33E-2	19.62	19.98	1.82E-2	57.68	58.03	1.36E-2
	0.70	3.21	3.56	9.79E-2	19.72	20.13	2.03E-2	57.03	58.03	1.73E-2
0.15	0.25	1.54	1.62	4.75E-2	19.98	19.97	6.99E-4	56.52	57.08	2.05E-3
	0.50	2.80	2.86	2.13E-2	19.65	19.94	1.46E-2	56.52	57.08	9.82E-3
	0.70	3.71	3.70	9.73E-4	19.95	20.07	5.72E-3	56.69	57.54	6.95E-3

5 Numerical Results

For the numerical experiments, we computed based on the fixed strike call Asian option. Table 1 presents the numerical results of Asian option values. The numerical solutions are compared with Levy’s approximation closed solution [7].

6 Conclusion

From Table 1, we can see that the results computed by the Gauss–Seidel method based on the standard Crank–Nicolson approximation equation is near to the Levy’s formula. Moreover, the accuracy of this 5-point stencil approximation Crank–Nicolson scheme is better than the accuracy of the 4-point stencil approximation Crank–Nicolson scheme [9].

Thus, we have shown that the two-dimensional PDE can be approximated by using standard Crank–Nicolson approximation scheme and be solved using Gauss–Seidel iterative method. In future, a more efficient iterative method can be investigated to enhance the computational time.

References

1. Kemna, A.G., Vorst, A.C.: A pricing method for options based on average asset values. *J. Bank. Financ.* **14**(1), 113–129 (1990). [https://doi.org/10.1016/0378-4266\(90\)90039-5](https://doi.org/10.1016/0378-4266(90)90039-5)
2. Boyle, P., Broadie, M., Glasserman, P.: Monte Carlo methods for security pricing. *J. Econ. Dyn. Control.* **21**(8), 1267–1321 (1997). [https://doi.org/10.1016/S0165-1889\(97\)00028-6](https://doi.org/10.1016/S0165-1889(97)00028-6)

3. Rogers, L.C.G., Shi, Z.: The value of an Asian option. *J. App. Probab.* **32**(4), 1077–1088 (1995). <https://doi.org/10.1017/S0021900200103559>
4. Vecer, J.: A new PDE approach for pricing arithmetic average Asian options. *J. Comput. Financ.* **4**(4), 105–113 (2001). <https://doi.org/10.21314/JCF.2001.064>
5. Elshegmani, Z.A., Ahmad, R.R., Jaaman, S.H.: On the modified arithmetic Asian option equation and its analytical solution. *App. Math. Sci.* **5**(25–28), 1217–1227 (2011)
6. Avramidis, A.N., L'Ecuyer, P.: Efficient Monte Carlo and quasi-Monte Carlo option pricing under the variance gamma model. *Manag. Sci.* **52**(12), 1930–1944 (2006). <https://doi.org/10.1287/mnsc.1060.0575>
7. Levy, E.: Pricing European average rate currency options. *J. Int. Money Financ.* **11**(5), 474–491 (1992). [https://doi.org/10.1016/0261-5606\(92\)90013-N](https://doi.org/10.1016/0261-5606(92)90013-N)
8. Fusai, G., Meucci, A.: Pricing discretely monitored Asian options under Lévy processes. *J. Bank. Financ.* **32**(10), 2076–2088 (2008). <https://doi.org/10.1016/j.jbankfin.2007.12.027>
9. Lee, T.Y., Chin, S.T.: A simple Crank-Nicolson scheme for Asian option. In: Proceedings of the 6th IMT-GT Conference on Mathematics, Statistics and its Applications (ICMSA 2010). UTAR, Kuala Lumpur, pp. 381–294 (2010)
10. Dubois, F., Lelièvre, T.: Efficient pricing of Asian options by the PDE approach. *J. Comput. Financ.* **8**(2), 55–64 (2004). <https://doi.org/10.21314/JCF.2005.138>
11. Geman, H., Yor, M.: Bessel processes, asian options and perpetuities. *Math. Financ.* **3**(4), 349–375 (1993). <https://doi.org/10.1111/j.1467-9965.1993.tb00092.x>
12. Elshegmani, Z.A., Ahmad, R.R.: Solving an Asian option PDE via the Laplace transform. *ScienceAsia* **39**(SUPPL. 1), 67–69 (2013). <https://doi.org/10.2306/scienceasia1513-1874.2013.39S.067>
13. Barucci, E., Polidoro, S., Vespri, V.: Some results on partial differential equations and Asian options. *Math. Model. Methods Appl. Sci.* **11**(03), 475–497 (2001). <https://doi.org/10.1142/S0218202501000945>

Chapter 19

Reinforcement Learning for Computing Power Grid Network Operating Functions



Shivam Sharma, Pinki Gupta and Laxminarayan Das

Abstract High voltage electric power grid physical nature learning is rarely a deterministic process rather it is a stochastic process in nature. Markov decision process and reinforcement learning algorithms are available to learn the quantitative and numerical estimation of the electric power generation, transmission, transformation, and distributing line physical measurement. Power grid managers use reinforcement learning process to regulate or control the parameters within a coded programming and computerized instrumentation. In this paper, we emphasize reinforcement algorithms to simplify medium level power grid problems.

Keywords Reinforcement learning · Power grid physical nature learning · Electric power physical measurement · Laplace transform and reinforcement learning

1 Introduction

The power grid is a system of systems, or a complex of different components, namely, Generators, loads, Transmission lines, power-system gear, switches, relays, transformers, computers, PLC, supervisory control and sensors, supply and demands.

The power grid networking system computation with reinforcement learning classification includes the grid management tasks such as maintenance schedule,

S. Sharma (✉)
United Health Group, Bangalore, India
e-mail: shivam.sharma15@gmail.com

P. Gupta
Atria, Gurgaon, India
e-mail: gupta3361@gmail.com

L. Das
Department of Applied Mathematics, Delhi Technological University, New Delhi, Delhi, India
e-mail: Indas@dce.ac.in

transmission schedule, power exchange instruments' functional unit behavior, subsystem dependency, utility cost calculation by sensing the connecting phase line load and distribution line consumed power, depreciate components' placement decision through computation, Electrical equipment and material service site or store house inventory regulation.

2 List of Electrical Equipment and Material Used in Transmission Grid Support System

The power grids meant for industrial electric energy transmission and control process consists of different equipment namely, Anchoring and structure foundations, Arresters, Bushings, cable arresters, connectors, power drain system, Enclosures, fiber glass, dropout switching fuse line material (wire, plate), insulators, pole-line hardware, track accessories.

The power grid for train engine power service system consists of overhead switch gear, protective devices, train engine switch gear, basement fixer insulators, structured line pole with catenary shaped connecting wires, reinforced cemented concrete basement, transformer units, wire stretcher, vibrator damper, insulators, suspension assembler, end-jumper accessories, anchors, center break switches, etc.

The transmission system fitted in hydroelectric power generator source consists of equipment dam-side overhead or underground switch gears, vibration shock absorber basement or pole insulators.

The transmission system fitted to thermoelectric power generator consists of heat transforming protective material.

The electronic communication microwave transmission tower consists of metallic trusses fitted with sensitive painted metallic antenna, transponder, booster, fixer screw, knot bolt, power cables etc.

The networked computer electronics transmission system is also a developed version topic for reinforcement learning. The signal formatted message transmission is done through channelizing electrons repeatedly and oscillating frequency-time-propagation time medium path schedule are organized with digital numbers bit string forms and analog instruments transmits to the destination computing device, this process is named as transmission schedule. The computing instrument senses whether there is a signal to be transmitted or not, and if transmitted what is the signal's patterns sequence transmitted and should the signal be transformed to piecewise continuity or discontinuity and delay phase difference learning are performed by tools measuring multidisciplinary transmission route line or radio wave formatted signal communicating system.

3 Reinforcement Learning to Computationally Determine or Estimate Transmission Line

The solution of certain mathematical differential equation's model is overall used to encapsulate the electrical, electromagnetic or electromechanical functions representing the power grid activities. The scalable models to grasp macroscopic trends are graph networks with or without ring topology for indexing continuum model or tree topology. The graph theory notations usually represent the transmission line symbolic or deterministic computational purpose.

Sometimes, the linear and integer programming branch and bound, heuristic methods (Tabu search) are used to decide which line to be assigned with effective load. This is also a reinforcement learning module for those machines which were not trained with decision making prior knowledge. Linear program simplex method, ellipsoid method (Khachiyan 1979), interior point method (Karmarkar 1984) are some of the methods used for interpreting or justifying the decision of replacing the effective equipment for separating depreciated items.

Power flow in electricity networks which are renewable in nature, produce independent power and consume uninterrupted smart energy. To analyze the network with regards to demand satisfaction, optimal energy production and fault tolerance it is necessary to compute a flow of energy in such an electricity network that can effectively implement decisions made by advanced computation. Energy flow in an electricity network obeys elemental laws of physics. To calculate the amount of energy flowing through each edge, traditionally the power flow (PF) method is used. The Optimal power flow (OPF) method is also used to calculate electrical flow by minimizing the production costs. Both the PF and OPF methods are nonlinear optimization problems.

Transmission line parameter, Properties of transmission lines, Power flow path-direction, Direct current approximation are topics of soft computing for solving networked computer power system. The extension of this study with specific input data is also applicable for corporations and industries using power supply system.

In the corporate industry or train engine, power grid network functional data and equipment health standards are mechanically learned either by deterministic power scaling or computation or by analyzing the power flow-based approaches. The mathematical transformations relating space and time graphs such as Laplace transform is a form of reinforcement learning method. The initial or bounded value for time and space variables associating with Laplace transform helps in learning and scaling the digital-analog interface. The algorithmic framework is as follows:

1. Determination of discrete points from the network flow metric or the line capacity, node or grid capacity, dynamically demarcate source line and feeder line (demarcation with time variable).
2. Determine the all possible interpolating polynomial using numerical methods.

3. Determine the corresponding polynomials approximate functions.
4. Determine the Laplace transform of all the approximate functions and trace the graph with the computer programming.
5. Determine the inverse Laplace transform of a function of s , the resulting function is a function with variable t , trace the patterned graph and compare with the time functional value graph, (The time functional value was the charge or electric potential real value of the component in the power transmission grid system whose performance was estimated or determined by assigning a numeric value, and the data was saved during transmission phase).
6. If the patterns match, then the machine can provide conclusive decision based on the inference logical statements.

Synthesizing this reinforcement learning process, we conclude the power-grid network computing is a vast topic with the multistep programming with different class of computation units. We briefly narrated how a machine can test whether the learned information is correct or with least error. The professional software is available in the software computing farms. The original manufacturers of the electric power grid network computational unit rarely describe the machine learning program codes. Therefore, the Power grid managers have set up research and development units to frame programmable logic codes for regulating their power-grid-network corporation data through both the advance computation as well as post-operation computation. The above algorithm is a part of the following algorithm.

4 A Sequential Reinforcement Learning

Step1. Represent the power grid network in Graph theory notation.

Step2. Use Linear and integer programming (branch and bound, heuristic methods (Tabu search), also read linear program simplex method, ellipsoid method (Khachiyan 1979), interior point method (Karmarkar 1984).

Step3. Power flow in electricity network function constitutes of the variables, the state renewable energy sources, which produces independent power and consumes uninterrupted smart energy, etc. To analyze the network with regards to demand satisfaction, optimal energy production and fault tolerance it is necessary to compute a flow of energy in such an electricity network. Energy flow in an electricity network obeys elemental laws of physics. To calculate the amount of energy flowing through each edge, traditionally the power flow (PF) method is used. The Optimal power flow (OPF) method is also used to calculate electrical flow by minimizing the production costs. Both the PF and OPF methods are nonlinear optimization problems.

1. Transmission line parameter
2. Properties of transmission lines
3. Power flow
4. Direct current approximation

Step4. Flow based approaches: Transformation to an s-t graph, Standard flow model, Balanced flow model, Bottleneck flow model, Minimum cost flow model, Combination of cost minimization and load balancing.

Step5. Hybrid models: Mathematical model, Mathematical properties, Structural findings.

Step6. Case study and Conclusion.

5 Reinforcement Learning: A New Perspective

The concept of Reinforcement Learning consists of an agent interacting with its environment, with every action a that the agent takes, it results in a change of its state from s to s' . There is a reward associated with transition from a state s to s' . Therefore, it is a constant learning process which is taking place here, and the result of every action, leads to a cumulative learning experience for the agent. So in the case of the power grid, the way we see reinforcement learning is that we know the commercial product produced which we call S during the energy transmission $f(t)$ in discrete time, so when we do the Laplace transformation of the function $p(t)$ which was the function of flow of energy with respect to continuous time that we arrived at after interpolation and approximation and then taking its Inverse Laplace Transformation and thus obtaining $q(s)$, subsequently we compare our $q(s)$ to the commercial product produced S (the exact space variable output in discrete time) that we already know. Thus, the difference in the value of $q(s)$ and S will determine how effective our interpolation and approximation were in determining the value of flow of energy as a function of continuous time. Thus the reward function is seen here as the inverse of the difference in the value of S and $q(s)$ and our task is to optimize our interpolation and approximation (which we define as the action a that our agent takes) in such a way so as to maximize our reward function. In a similar light, the state s in our case is the value of the function $p(t)$ that we obtain after interpolation and approximation, and this change of states after successive interpolation and approximation and the subsequent learning as a result of this is what makes our reinforcement learning paradigm. Thus, our objective is to find a policy P (which is a mapping from state to action), which will determine what action to take if you are in a particular state s , to as to maximize our long term reward. So, we have constant data from the power grid about the flow of energy or charges in discrete time, and we are constantly interpolating and approximating to arrive at a state s , which is the value of the function of flow of energy with respect to continuous time. Thus, our policy will guide us to the correct action a (the interpolation and approximation to be performed) so as to maximize our long-term reward. Thus, with this we conclude our reinforcement learning methodology.

6 Conclusion

Reinforcement learning is less descriptive and more diagnostic process. The example of reinforcement learning is the learning of a beginner's smart grid electronics computer operating system and drive card internal process logical transition and task assignment learning. The beginner has limited knowledge of electrical system formula such as mathematically estimation of current, voltage or charge potential electromotive force, transformer, capacitor, electromagnetic flux through different geometrical or topological shapes but interpretation in logical digital and analogue system in real time computer process machines are sequential learning. The Laplace transform is a measure of space with time variable and inverse transformation is the diagnosis of space properties.

Chapter 20

Similarity Measure for Fuzzy Number Based on Distances and Geometric Shape Characteristics



Nur Amira Mat Saffie, Khairul A. Rasmani
and Nor Hashimah Sulaiman

Abstract Fuzzy similarity measure is a very useful technique in the area of fuzzy decision making. Although various techniques have been developed to measure fuzzy similarity, it can be observed that several limitations still exist. In particular, there are methods that produce inconsistent degrees of similarity with regard to the compared fuzzy numbers which do not reflect the geometric characteristic of the fuzzy numbers. In this paper, a new fuzzy similarity measure for generalized trapezoidal fuzzy number (GTFN) is developed based on distance and geometric shape characteristics like height, area, centre, and perimeter to overcome the problems. Three existing fuzzy similarity measures were selected and compared with the proposed measure. The analysis on eight sets of GTFNs using the fuzzy similarity measures shows that the proposed measure produced a reasonable interpretation of similarity degree based on the graphical representation. Hence, this measure can serve as an alternative method in calculating the degree of similarity of compared GTFNs. The introduced concept in the proposed measure will enable a broader implementation of fuzzy similarity measures in real-world decision making.

Keywords Similarity measure · Fuzzy similarity measure · Generalized trapezoidal fuzzy number · Geometric distance · Geometric shape characteristics

N. A. M. Saffie (✉) · K. A. Rasmani · N. H. Sulaiman
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Malaysia
e-mail: amira_saffie@yahoo.com

K. A. Rasmani
e-mail: khairulanwar@ns.uitm.edu.my

N. H. Sulaiman
e-mail: nhashima@tmsk.uitm.edu.my

1 Introduction

Fuzzy number is a very useful concept since it can represent linguistic expressions mathematically, commonly in the form of triangular fuzzy number or trapezoidal fuzzy number. The fuzzy number can well describe human cognition in interval numbers because in real-world situations, there often exists conditions or states that are difficult to express by crisp numbers [1]. In the area of decision making, artificial intelligence, and data analysis under fuzzy environment, comparison of fuzzy numbers are commonplace which is normally done using the concept of similarity measure.

The concept of similarity is essentially important in most real-world situations since it is a basic concept in human cognition [2]. Similarity measure can be described as a function that computes the degree of similarity between two compared objects. There are numerous factors that affect the similarity degree of fuzzy numbers such as distances or positions, overlapping midpoint, heights, perimeters, shapes, areas, and so on. Based on these factors, various techniques to identify the degree of similarity had been introduced by many researchers.

Although various techniques had been established to measure the similarity degree of fuzzy numbers, it can be observed that several limitations for the existing fuzzy similarity measures still exist. According to Khorshidi [3] and Patra [4], the existing measures fail to determine degrees of similarity properly, while Chen and Chen [5] and Chen [6] claimed that the degrees of similarity are not correctly obtained in certain selected situations. Therefore, in this paper, a new fuzzy similarity measure is developed to overcome the said problems. This study aims to propose a fuzzy similarity measure for Generalized Trapezoidal Fuzzy Number (GTFN) based on distances and geometric shape characteristics like height, area, centre, and perimeter.

The rest of the paper is planned as follows: Sect. 2 briefly reviews the existing fuzzy similarity measures for GTFN. The proposed fuzzy similarity measure is described in Sect. 3. The comparison of the proposed technique with the selected existing fuzzy similarity measures is presented in Sect. 4, and the conclusions are finally discussed in Sect. 5.

2 Existing Fuzzy Similarity Measures for Generalized Trapezoidal Fuzzy Numbers

In this section, some basic definition of Generalized Trapezoidal Fuzzy Number (GTFN) and several existing fuzzy similarity measures are presented.

Definition 1 ([5, 6]) A GTFN is denoted as $M = (m_1, m_2, m_3, m_4; w_M)$, where $m_1, m_2, m_3, m_4 \in R$ such that $0 \leq m_1 \leq m_2 \leq m_3 \leq m_4 \leq 1$ and $0 < w_M \leq 1$.

Definition 2 ([3, 5, 6]) M is called a normal GTFN if $w_M = 1$ and denoted as $M = (m_1, m_2, m_3, m_4)$. Noted that, M is called a generalized triangular fuzzy number if $m_2 = m_3$, M is a crisp interval if $m_1 = m_2$ and $m_3 = m_4$, and M is a singleton fuzzy number if $m_1 = m_2 = m_3 = m_4$ and $w_M = 1$.

The following definitions are the descriptions of three selected existing fuzzy similarity measures. Suppose that $M = (m_1, m_2, m_3, m_4; w_M)$ and $N = (n_1, n_2, n_3, n_4; w_N)$ are two GTFNs and the calculation of M and N are conducted analogously.

- (a) Fuzzy similarity measure using arithmetic operators of the GTFNs based on the combined concept of geometric distance and centre of gravity (COG) points [7]

$$S(M, N) = 1 - \frac{\sum_{i=1}^4 |m_i - n_i|}{8} - \frac{d(M, N)}{2}, \tag{1}$$

where the distance COG of M and N , $d(M, N)$ is calculated as

$$d(M, N) = \frac{\sqrt{(x_M - x_N)^2 + (y_M - y_N)^2}}{\sqrt{1.25}}, \tag{2}$$

such that

$$y_M = \begin{cases} \frac{w_M \left(\frac{m_3 - m_2}{m_4 - m_1} + 2 \right)}{6} & \text{if } m_1 \neq m_4, \\ \frac{w_M}{2} & \text{if } m_1 = m_4, \end{cases} \tag{3}$$

$$x_M = \begin{cases} \frac{y_M^* (m_3 + m_2) + (m_4 + m_1)(w_M - y_M^*)}{2w_M} & \text{if } w_M \neq 0, \\ \frac{m_4 + m_1}{2} & \text{if } w_M = 0. \end{cases} \tag{4}$$

- (b) Fuzzy similarity measure that is based on the combination of geometric distances, areas and heights [4]

$$S(M, N) = \left(1 - \frac{\sum_{i=1}^4 |m_i - n_i|}{4} \right) \left(1 - \frac{1}{2} \{ |Ar(M) - Ar(N)| + |w_M - w_N| \} \right), \tag{5}$$

where the area, Ar is given by

$$Ar(M) = \frac{1}{2}w_M(m_4 + m_3 - m_2 - m_1). \tag{6}$$

(c) Fuzzy similarity measure proposed by Khorshidi [3] that combined the concept of geometric distances, distance COG, areas, heights and perimeters

$$S(M, N) = \left(1 - \frac{\sum_{i=1}^4 |m_i - n_i|}{4} X d(M, N) \right) \left(1 - \frac{|Ar(M) - Ar(N)| + |w_M - w_N| + \frac{|P(M) - P(N)|}{\max(P(M), P(N))}}{3} \right), \tag{7}$$

where the distance COG, $d(M, N)$ is calculated based on Eqs. (2), (3), and (4). Then, the area, Ar is defined based on Eq. (6) and the perimeter, P is calculated as

$$P(M) = \sqrt{(m_1 - m_2)^2 + w_M^2} + \sqrt{(m_3 - m_4)^2 + w_M^2} + (m_3 - m_2) + (m_4 - m_1). \tag{8}$$

Noted that $\frac{|P(M) - P(N)|}{\max(P(M), P(N))} = 0$ when $\max(P(M), P(N)) = 0$.

3 Proposed Fuzzy Similarity Measure

In this section, a new formulation of fuzzy similarity measure based on geometric distance, centre of gravity (COG) points (x^*, y^*) , height (w), area (Ar), and perimeter (P) is proposed in order to calculate the degree of similarity between two generalized trapezoidal fuzzy numbers (GTFNs), M and N . The formula is presented as follows:

$$S(M, N) = \left(1 - \frac{\sum_{i=1}^4 |m_i - n_i|}{4} \right) (1 - |x_M^* - x_N^*|)^{B(S_M, S_N)} \left(1 - \frac{|w_M - w_N| + \frac{|y_M^* - y_N^*|}{\max(y_M^*, y_N^*)} + |Ar(M) - Ar(N)| + \frac{|P(M) - P(N)|}{\max(P(M), P(N))}}{4} \right), \tag{9}$$

where the COG points consist of points at x -axis (x_M^*) and y -axis (y_M^*)

$$y_M^* = \begin{cases} w_M \left(\frac{m_3 - m_2}{m_4 - m_1} + 2 \right) & \text{if } m_1 \neq m_4 \\ \frac{6}{2} \frac{w_M}{2} & \text{if } m_1 = m_4 \end{cases} \tag{10}$$

$$x_M^* = \begin{cases} \frac{y_M^*(m_3 + m_2) + (m_4 + m_1)(w_M - y_M^*)}{2w_M} & \text{if } w_M \neq 0 \\ \frac{m_4 + m_1}{2} & \text{if } w_M = 0 \end{cases} \tag{11}$$

$$B(S_M, S_N) = \begin{cases} 1 & \text{if } S_M + S_N > 0 \\ 0 & \text{if } S_M + S_N = 0 \end{cases} \tag{12}$$

such that $S_M = m_4 - m_1$ and $S_N = n_4 - n_1$.

The perimeter, P and area, Ar of M are calculated based on Eqs. (6) and (8). The closer the calculated degree of similarity $S(M, N)$ is to 1, the higher the degree of similarity between M and N . For comparison between two sets of GTFNs (M, N_a) and (M, N_b) where $a \neq b$, the set with higher degree of similarity indicates that the pair of GTFNs in the set is more similar as compared to the pair of GTFNs in the other set. The proposed fuzzy similarity measure satisfies the following properties:

Property 1 $S(M, N) = S(N, M)$;

Property 2 Two GTFNs M and N are identical if and only if $S(M, N) = 1$;

Property 3 If $M = (m_1, m_2, m_3, m_4; w_M)$ and $N = (n_1, n_2, n_3, n_4; w_N)$ are two GTFNs with the same geometric shape, the same height and the same offset, d , where $d = m_1 - n_1 = m_2 - n_2 = m_3 - n_3 = m_4 - n_4$, then $S(M, N) = 1 - |d|$;

Property 4 If $M = (m, m, m, m; 1.0)$ and $N = (n, n, n, n; 1.0)$ are two real numbers, then $S(M, N) = 1 - |a - b|$.

4 Comparison of Fuzzy Similarity Measures

The performance of the proposed fuzzy similarity measure is compared with Xu [7], Patra [4], and Khorshidi [3] over eight sets of GTFNs based on Patra [4] and Chen [6]. The graphical representation of the eight sets are presented in Fig. 1.

The results from the implementation of the proposed fuzzy similarity measure and the selected existing fuzzy similarity measures are shown in Table 1.

Both fuzzy numbers N in Set 1 and Set 2 obviously have different heights. It can also be seen that the overlapping area in Set 1 is twice the overlapping area in Set 2. However, Xu [7] obtained quite similar degrees of similarity for both sets which are 0.9627 and 0.8434, respectively as compared with other methods. The proposed measure produces reasonable degrees of membership with regard to Set 1, 0.9191, and Set 2, 0.3738.

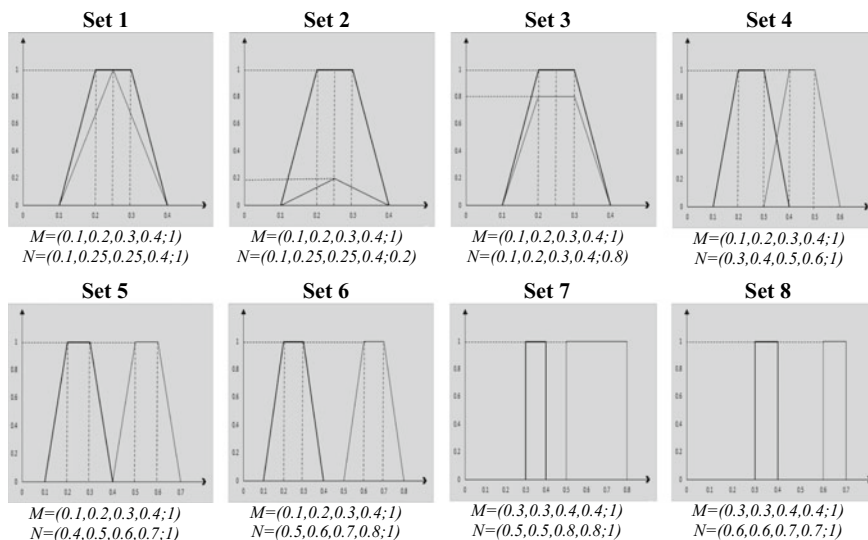


Fig. 1 Graphical representations of eight sets of fuzzy numbers

Table 1 Comparison of calculated similarity degree between the proposed measure and the existing measures

Set	Xu [7]	Patra [4]	Khorshidi [3]	Proposed method
1	0.9627	0.9506	0.9700	0.9191
2	0.8434	0.5021	0.4507	0.3738
3	0.9652	0.8800	0.8650	0.8488
4	0.8106	0.8000	0.9642	0.6400
5	0.7158	0.7000	0.9195	0.4900
6	0.6211	0.6000	0.8569	0.3600
7	0.7158	0.6300	0.8110	0.4467
8	0.7158	0.7000	0.9195	0.4900

Set 3 and Set 4 are compared due to the different height and positions of N in both sets. Based on Fig. 1, the overlapping areas of M and N in Set 3 are greater than the overlapping areas of M and N in Set 4. Intuitively, M and N in Set 3 are more similar than M and N in Set 4. However, Khorshidi [3] rates Set 4 with a higher degree of similarity than Set 3, which are 0.9642 and 0.865, respectively. The proposed measure assigns degrees of similarity for Sets 3 and 4 as 0.8488 and 0.64, respectively, which is in line with human intuition.

For Sets 5 and 6, the proposed measure is consistent with Xu [7], Patra [4], and Khorshidi [3] in the sense that all the measures produce higher degrees of similarity for Set 5 as compared to Set 6. In fact, the proposed measure produces the lowest

degrees amongst all the measures in both sets partly due to the non-overlapping area in the compared fuzzy numbers.

Fuzzy numbers N in Set 7 and Set 8 which both have different areas and perimeters are compared with the same fuzzy number M . Xu [7] produced the same similarity degrees for both sets even though M and N have different positions, areas and perimeters. However, the proposed method produced a slightly different similarity degree for both sets as 0.4467 and 0.49, respectively, which can be considered more reasonable.

From the discussions, it can be concluded that the results obtained from the proposed technique produced a more reasonable interpretation and are consistent with the results produced by other selected existing fuzzy similarity measures. In summary, the proposed technique overcame the drawbacks of some of the existing fuzzy similarity measures.

5 Conclusion

This paper has presented a new fuzzy similarity measure for Generalized Trapezoidal Fuzzy Numbers (GTFNs). The proposed similarity measure combined the concepts of geometric distance, centre of gravity (COG) points, height, area, and perimeter in order to calculate the degree of similarity between two GTFNs. A comparison of the proposed method was made with three existing fuzzy similarity measures over eight sets of compared fuzzy numbers. The results showed that the proposed method produced a reasonable interpretation and can serve as an alternative method in calculating the degree of similarity of compared GTFNs. To show that the proposed measure is very useful in real-world situations, work is currently carried out to implement the proposed measure in decision making to estimate excessive domestic water usage using combinative algorithms.

Acknowledgements This research work is funded by the Ministry of Education, Malaysia under the Fundamental Research Grant Scheme (FRGS) with reference number 600-RMI/FRGS TD 5/3 (1/2015). The authors would also like to thank the Research Management Institute (RMI), Universiti Teknologi MARA, Malaysia.

References

1. Luo, L., Ren, H.: A new similarity measure-based MADM method under dynamic interval-valued intuitionistic fuzzy environment. *AMSE J.* **59**(1), 84–92 (2016)
2. Beg, I., Ashraf, S.: Similarity measures for fuzzy sets. *Appl. Comput. Math.* **8**(2), 192–202 (2009)
3. Khorshidi, H.A., Nikfalazar, S.: An improved similarity measure for generalized fuzzy numbers and its application to fuzzy risk analysis. *Appl. Soft Comput.* **52**, 478–486 (2017). <https://doi.org/10.1016/j.asoc.2016.10.020>

4. Patra, K., Mondal, S.K.: Fuzzy risk analysis using area and height based similarity measure on generalized trapezoidal fuzzy numbers and its application. *Appl. Soft Comput.* **28**, 276–284 (2015). <https://doi.org/10.1016/j.asoc.2014.11.042>
5. Chen, S.-J., Chen, S.-M.: A new method to measure the similarity between fuzzy numbers. In: 2001 The 10th IEEE International Conference on Fuzzy Systems, Melbourne, Australia, pp. 1123–1126. IEEE (2001)
6. Chen, S.-J.: A new similarity measure of generalized fuzzy numbers based on geometric-mean averaging operator. In: 2006 IEEE International Conference on Fuzzy Systems, Vancouver, Canada, pp. 1879–1886. IEEE (2006)
7. Xu, Z., Shang, S., Qian, W., Shu, W.: A method for fuzzy risk analysis based on the new similarity of trapezoidal fuzzy numbers. *Expert Syst. Appl.* **37**(3), 1920–1927 (2010). <https://doi.org/10.1016/j.eswa.2009.07.015>

Chapter 21

Solving Fourth Order Linear Initial and Boundary Value Problems Using an Implicit Block Method



Oluwaseun Adeyeye and Zurni Omar

Abstract The introduction of new approaches to numerically approximate higher order ordinary differential equations (ODEs) is vastly being explored in recent literature. The reason for adopting these numerical approaches is because some of these higher order ODEs fail to have an approximate solution or the current numerical approach being adopted has less accuracy. The application of an implicit block method for solving fourth order ordinary differential equations (ODEs) is considered in this article. The solution encompasses both initial and boundary value problems of fourth order ODEs. The implicit block method is developed for a set of six equidistant points using a new linear block approach (LBA). The LBA produces the required family of six-step schemes to simultaneously evaluate the solution of the fourth order ODEs at individual grid points in a self-starting mode. The basic properties of the implicit block method are investigated, and the block method is seen to satisfy the property of convergence which is displayed in the numerical results obtained. Furthermore, in comparison to works of past authors the implicit block method gives more impressive results.

Keywords Implicit block method · Six-step · Linear block approach · Fourth order · Ordinary differential equations

1 Introduction

Block methods for the solution of higher order ordinary differential equations (ODEs) have become a common approach in literature due to the shortcomings experienced by the initial numerical approaches for solving higher order ODEs [1]. These initial approaches include the approach of reducing the higher order ODEs to

O. Adeyeye (✉) · Z. Omar

School of Quantitative Sciences, Universiti Utara Malaysia, Changlun, Malaysia
e-mail: adeyeye_oluwaseun@ahsgs.uum.edu.my

Z. Omar

e-mail: zurni@uum.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_21

a system of first order ODEs and then suitable numerical methods such as Euler's method, Trapezoidal rule, Runge Kutta methods, amongst others are then applied to solve the system of ODEs [2–5]. The approach of reduction is able to obtain the required solution, however, there is too much rigor involved in the process of reduction and also the implementation process. Other initial numerical approaches that bypassed the reduction stress was a direct approximation of the higher order ODEs using Taylor series expansions or predictor-corrector methods [6, 7]. Similar to the concept of reduction, these other initial approaches likewise gave good solutions but when considering the higher order ODEs having exact solutions, the Taylor series expansions and predictor-corrector methods give low accuracy when compared to the exact solution of the ODE [8]. Therefore, the introduction of block methods which has floored these initial rigorous and less accurate initial numerical approaches [9].

Conventionally, a linear multistep method of the form

$$\sum_{j=0}^k \alpha_j y_{n+j} = h^m \sum_{j=0}^k \beta_j f_{n+j} \quad (1)$$

is said to be implicit when $\beta_k = 0$ and explicit otherwise [10]. The advantage of adopting implicit linear multistep methods over its explicit counterpart is because the order of the method (p) obtained by the implicit method is higher which implies better accuracy. This is because, the higher the order of the numerical method, the better the accuracy [11]. Therefore, this article will be considering implicit block methods for the solution of fourth order ODEs.

Numerical solutions to fourth order ODEs encompassing initial value problems (IVPs) or boundary value problems (BVPs) have been considered by authors in literature. These includes the works of [5] where the authors adopted embedded pairs of Runge-Kutta type methods for direct solution to fourth order ODEs, [12] adopting seven-step method to directly solve fourth order ODEs, [13] also solving fourth order IVPs using modified implicit hybrid block method, amongst others [14–16]. However, the motivation to introduce methods with better accuracy informs the development of the implicit block method in this article.

To develop the required block method, a new approach coined as the linear block approach (LBA) is adopted. This approach is easy to implement and less rigorous to adopt. Details of the algorithm of the LBA and its adoption to develop the required implicit block method is discussed in the next section.

2 Methodology

This section describes the derivation of the implicit block method using LBA. The following algorithm shows the steps involved in developing the specific six-step implicit block method. LBA follows the concept of considering the general form of the block method while implementing a step-by-step wise mode to obtain the expected block method for solving fourth order ODEs.

Algorithm 2.1

START

Step 1: Obtain the block method from the given expression

$$y_{n+\xi} = \sum_{i=0}^3 \frac{(\xi h)^i}{i!} y_n^{(i)} + \sum_{i=0}^6 \phi_{\xi i} f_{n+i}, \xi = 1, 2, \dots, 6 \tag{2}$$

Step 2: Obtain the 1st, 2nd and 3rd derivative schemes of the block method from

$$y_{n+\xi}^{(a)} = \sum_{i=0}^{3-a} \frac{(\xi h)^i}{i!} y_n^{(i+a)} + \sum_{i=0}^6 \omega_{\xi ia} f_{n+i}, a = 1_{(\xi=1,2,\dots,6)}, 2_{(\xi=1,2,\dots,6)}, 3_{(\xi=1,2,\dots,6)} \tag{3}$$

$\phi_{\xi i} = A^{-1}B$ and $\omega_{\xi ia} = A^{-1}D$ where

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & h & 2h & 3h & 4h & 5h & 6h \\ 0 & \frac{(h)^2}{2!} & \frac{(2h)^2}{2!} & \frac{(3h)^2}{2!} & \frac{(4h)^2}{2!} & \frac{(5h)^2}{2!} & \frac{(6h)^2}{2!} \\ 0 & \frac{(h)^3}{3!} & \frac{(2h)^3}{3!} & \frac{(3h)^3}{3!} & \frac{(4h)^3}{3!} & \frac{(5h)^3}{3!} & \frac{(6h)^3}{3!} \\ 0 & \frac{(h)^4}{4!} & \frac{(2h)^4}{4!} & \frac{(3h)^4}{4!} & \frac{(4h)^4}{4!} & \frac{(5h)^4}{4!} & \frac{(6h)^4}{4!} \\ 0 & \frac{(h)^5}{5!} & \frac{(2h)^5}{5!} & \frac{(3h)^5}{5!} & \frac{(4h)^5}{5!} & \frac{(5h)^5}{5!} & \frac{(6h)^5}{5!} \\ 0 & \frac{(h)^6}{6!} & \frac{(2h)^6}{6!} & \frac{(3h)^6}{6!} & \frac{(4h)^6}{6!} & \frac{(5h)^6}{6!} & \frac{(6h)^6}{6!} \end{pmatrix}, B = \begin{pmatrix} \frac{(\xi h)^4}{4!} \\ \frac{(\xi h)^5}{5!} \\ \frac{(\xi h)^6}{6!} \\ \frac{(\xi h)^7}{7!} \\ \frac{(\xi h)^8}{8!} \\ \frac{(\xi h)^9}{9!} \\ \frac{(\xi h)^{10}}{10!} \end{pmatrix}, D = \begin{pmatrix} \frac{(\xi h)^{(4-a)}}{(4-a)!} \\ \frac{(\xi h)^{(5-a)}}{(5-a)!} \\ \frac{(\xi h)^{(6-a)}}{(6-a)!} \\ \frac{(\xi h)^{(7-a)}}{(7-a)!} \\ \frac{(\xi h)^{(8-a)}}{(8-a)!} \\ \frac{(\xi h)^{(9-a)}}{(9-a)!} \\ \frac{(\xi h)^{(10-a)}}{(10-a)!} \end{pmatrix}$$

STOP

To implement individual steps of Algorithm 2.1, the following procedure is as highlighted below.

Considering Step 1, Eq. (2) takes the form

$$\begin{aligned}
 y_{n+1} &= y_n + hy'_n + \frac{(h)^2}{2!}y''_n + \frac{(h)^3}{3!}y'''_n + (\phi_{10}f_n + \phi_{11}f_{n+1} \\
 &\quad + \phi_{12}f_{n+2} + \phi_{13}f_{n+3} + \phi_{14}f_{n+4} + \phi_{15}f_{n+5} + \phi_{16}f_{n+6}) \\
 y_{n+2} &= y_n + 2hy'_n + \frac{(2h)^2}{2!}y''_n + \frac{(2h)^3}{3!}y'''_n + (\phi_{20}f_n + \phi_{21}f_{n+1} \\
 &\quad + \phi_{22}f_{n+2} + \phi_{23}f_{n+3} + \phi_{24}f_{n+4} + \phi_{25}f_{n+5} + \phi_{26}f_{n+6}) \\
 y_{n+3} &= y_n + 3hy'_n + \frac{(3h)^2}{2!}y''_n + \frac{(3h)^3}{3!}y'''_n + (\phi_{30}f_n + \phi_{31}f_{n+1} \\
 &\quad + \phi_{32}f_{n+2} + \phi_{33}f_{n+3} + \phi_{34}f_{n+4} + \phi_{35}f_{n+5} + \phi_{36}f_{n+6}) \\
 y_{n+4} &= y_n + 4hy'_n + \frac{(4h)^2}{2!}y''_n + \frac{(4h)^3}{3!}y'''_n + (\phi_{40}f_n + \phi_{41}f_{n+1} \\
 &\quad + \phi_{42}f_{n+2} + \phi_{43}f_{n+3} + \phi_{44}f_{n+4} + \phi_{45}f_{n+5} + \phi_{46}f_{n+6}) \\
 y_{n+5} &= y_n + 5hy'_n + \frac{(5h)^2}{2!}y''_n + \frac{(5h)^3}{3!}y'''_n + (\phi_{50}f_n + \phi_{51}f_{n+1} \\
 &\quad + \phi_{52}f_{n+2} + \phi_{53}f_{n+3} + \phi_{54}f_{n+4} + \phi_{55}f_{n+5} + \phi_{56}f_{n+6}) \\
 y_{n+6} &= y_n + 6hy'_n + \frac{(6h)^2}{2!}y''_n + \frac{(6h)^3}{3!}y'''_n + (\phi_{60}f_n + \phi_{61}f_{n+1} \\
 &\quad + \phi_{62}f_{n+2} + \phi_{63}f_{n+3} + \phi_{64}f_{n+4} + \phi_{65}f_{n+5} + \phi_{66}f_{n+6})
 \end{aligned} \tag{4}$$

Moving on to Step 2, Eq. (3) takes the following expressions

$$\begin{aligned}
 y'_{n+1} &= y'_n + hy''_n + \frac{(h)^2}{2!}y'''_n + (\omega_{101}f_n + \omega_{111}f_{n+1} + \omega_{121}f_{n+2} \\
 &\quad + \omega_{131}f_{n+3} + \omega_{141}f_{n+4} + \omega_{151}f_{n+5} + \omega_{161}f_{n+6}) \\
 y'_{n+2} &= y'_n + 2hy''_n + \frac{(2h)^2}{2!}y'''_n + (\omega_{201}f_n + \omega_{211}f_{n+1} + \omega_{221}f_{n+2} \\
 &\quad + \omega_{231}f_{n+3} + \omega_{241}f_{n+4} + \omega_{251}f_{n+5} + \omega_{261}f_{n+6}) \\
 y'_{n+3} &= y'_n + 3hy''_n + \frac{(3h)^2}{2!}y'''_n + (\omega_{301}f_n + \omega_{311}f_{n+1} + \omega_{321}f_{n+2} \\
 &\quad + \omega_{331}f_{n+3} + \omega_{341}f_{n+4} + \omega_{351}f_{n+5} + \omega_{361}f_{n+6}) \\
 y'_{n+4} &= y'_n + 4hy''_n + \frac{(4h)^2}{2!}y'''_n + (\omega_{401}f_n + \omega_{411}f_{n+1} + \omega_{421}f_{n+2} \\
 &\quad + \omega_{431}f_{n+3} + \omega_{441}f_{n+4} + \omega_{451}f_{n+5} + \omega_{461}f_{n+6}) \\
 y'_{n+5} &= y'_n + 5hy''_n + \frac{(5h)^2}{2!}y'''_n + (\omega_{501}f_n + \omega_{511}f_{n+1} + \omega_{521}f_{n+2} \\
 &\quad + \omega_{531}f_{n+3} + \omega_{541}f_{n+4} + \omega_{551}f_{n+5} + \omega_{561}f_{n+6})
 \end{aligned} \tag{5}$$

and

$$\begin{aligned}
y''_{n+1} &= y''_n + hy'''_n + (\omega_{102}f_n + \omega_{112}f_{n+1} + \omega_{122}f_{n+2} \\
&\quad + \omega_{132}f_{n+3} + \omega_{142}f_{n+4} + \omega_{152}f_{n+5} + \omega_{162}f_{n+6}) \\
y''_{n+2} &= y''_n + 2hy'''_n + (\omega_{202}f_n + \omega_{212}f_{n+1} + \omega_{222}f_{n+2} \\
&\quad + \omega_{232}f_{n+3} + \omega_{242}f_{n+4} + \omega_{252}f_{n+5} + \omega_{262}f_{n+6}) \\
y''_{n+3} &= y''_n + 3hy'''_n + (\omega_{302}f_n + \omega_{312}f_{n+1} + \omega_{322}f_{n+2} \\
&\quad + \omega_{332}f_{n+3} + \omega_{342}f_{n+4} + \omega_{352}f_{n+5} + \omega_{362}f_{n+6}) \\
y''_{n+4} &= y''_n + 4hy'''_n + (\omega_{402}f_n + \omega_{412}f_{n+1} + \omega_{422}f_{n+2} \\
&\quad + \omega_{432}f_{n+3} + \omega_{442}f_{n+4} + \omega_{452}f_{n+5} + \omega_{462}f_{n+6}) \\
y''_{n+5} &= y''_n + 5hy'''_n + (\omega_{502}f_n + \omega_{512}f_{n+1} + \omega_{522}f_{n+2} \\
&\quad + \omega_{532}f_{n+3} + \omega_{542}f_{n+4} + \omega_{552}f_{n+5} + \omega_{562}f_{n+6}) \\
y''_{n+6} &= y''_n + 6hy'''_n + (\omega_{602}f_n + \omega_{612}f_{n+1} + \omega_{622}f_{n+2} \\
&\quad + \omega_{632}f_{n+3} + \omega_{642}f_{n+4} + \omega_{652}f_{n+5} + \omega_{662}f_{n+6})
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
y'''_{n+1} &= y'''_n + (\omega_{103}f_n + \omega_{113}f_{n+1} + \omega_{123}f_{n+2} + \omega_{133}f_{n+3} \\
&\quad + \omega_{143}f_{n+4} + \omega_{153}f_{n+5} + \omega_{163}f_{n+6}) \\
y'''_{n+2} &= y'''_n + (\omega_{203}f_n + \omega_{213}f_{n+1} + \omega_{223}f_{n+2} \\
&\quad + \omega_{233}f_{n+3} + \omega_{243}f_{n+4} + \omega_{253}f_{n+5} + \omega_{263}f_{n+6}) \\
y'''_{n+3} &= y'''_n + (\omega_{303}f_n + \omega_{313}f_{n+1} + \omega_{323}f_{n+2} \\
&\quad + \omega_{333}f_{n+3} + \omega_{343}f_{n+4} + \omega_{353}f_{n+5} + \omega_{363}f_{n+6}) \\
y'''_{n+4} &= y'''_n + (\omega_{403}f_n + \omega_{413}f_{n+1} + \omega_{423}f_{n+2} + \omega_{433}f_{n+3} \\
&\quad + \omega_{443}f_{n+4} + \omega_{453}f_{n+5} + \omega_{463}f_{n+6}) \\
y'''_{n+5} &= y'''_n + (\omega_{503}f_n + \omega_{513}f_{n+1} + \omega_{523}f_{n+2} + \omega_{533}f_{n+3} \\
&\quad + \omega_{543}f_{n+4} + \omega_{553}f_{n+5} + \omega_{563}f_{n+6}) \\
y'''_{n+6} &= y'''_n + (\omega_{603}f_n + \omega_{613}f_{n+1} + \omega_{623}f_{n+2} + \omega_{633}f_{n+3} \\
&\quad + \omega_{643}f_{n+4} + \omega_{653}f_{n+5} + \omega_{663}f_{n+6})
\end{aligned} \tag{7}$$

To obtain the unknown ϕ coefficients, it is defined that $\phi_{\xi_i} = A^{-1}B$ where A and B are as stated above. Therefore, with reference to Eq. (4) and Algorithm 2.1,

$$\begin{aligned}
& (\phi_{10}, \phi_{11}, \phi_{12}, \phi_{13}, \phi_{14}, \phi_{15}, \phi_{16})^T \\
&= \left(\frac{95929h^4}{3628800}, \frac{4001h^4}{129600}, \frac{23033h^4}{725760}, \frac{811h^4}{30240}, \frac{10693h^4}{725760}, \frac{4219h^4}{907200}, \frac{2323h^4}{3628800} \right)^T \\
& (\phi_{20}, \phi_{21}, \phi_{22}, \phi_{23}, \phi_{24}, \phi_{25}, \phi_{26})^T \\
&= \left(\frac{4127h^4}{14175}, \frac{8782h^4}{14175}, \frac{199h^4}{405}, \frac{388h^4}{945}, \frac{127h^4}{567}, \frac{998h^4}{14175}, \frac{137h^4}{14175} \right)^T \\
& (\phi_{30}, \phi_{31}, \phi_{32}, \phi_{33}, \phi_{34}, \phi_{35}, \phi_{36})^T \\
&= \left(\frac{49239h^4}{44800}, \frac{34263h^4}{11200}, \frac{3159h^4}{1792}, \frac{261h^4}{160}, \frac{8019h^4}{8960}, \frac{3159h^4}{11200}, \frac{1737h^4}{44800} \right)^T \\
& (\phi_{40}, \phi_{41}, \phi_{42}, \phi_{43}, \phi_{44}, \phi_{45}, \phi_{46})^T \\
&= \left(\frac{7808h^4}{2835}, \frac{124928h^4}{14175}, \frac{10112h^4}{2835}, \frac{4096h^4}{945}, \frac{928h^4}{405}, \frac{2048h^4}{2835}, \frac{1408h^4}{14175} \right)^T \\
& (\phi_{50}, \phi_{51}, \phi_{52}, \phi_{53}, \phi_{54}, \phi_{55}, \phi_{56})^T \\
&= \left(\frac{807125h^4}{145152}, \frac{701875h^4}{36288}, \frac{790625h^4}{145152}, \frac{59375h^4}{6048}, \frac{653125h^4}{145152}, \frac{7625h^4}{5184}, \frac{29375h^4}{145152} \right)^T \\
& (\phi_{60}, \phi_{61}, \phi_{62}, \phi_{63}, \phi_{64}, \phi_{65}, \phi_{66})^T \\
&= \left(\frac{1719h^4}{175}, \frac{6318h^4}{175}, \frac{243h^4}{35}, \frac{684h^4}{35}, \frac{243h^4}{35}, \frac{486h^4}{175}, \frac{9h^4}{25} \right)^T
\end{aligned} \tag{8}$$

Likewise, to obtain the unknown ω coefficients, it is defined that $\omega_{\xi ia} = A^{-1}D$ where A and D are as stated above. Therefore, with reference to Eqs. (5) and (6) and Algorithm 2.1,

$$\begin{aligned}
& (\omega_{101}, \omega_{111}, \omega_{121}, \omega_{131}, \omega_{141}, \omega_{151}, \omega_{161})^T \\
&= \left(\frac{343801h^3}{3628800}, \frac{6031h^3}{43200}, \frac{32981h^3}{241920}, \frac{5177h^3}{45360}, \frac{15107h^3}{241920}, \frac{5947h^3}{302400}, \frac{9809h^3}{3628800} \right)^T \\
& (\omega_{201}, \omega_{211}, \omega_{221}, \omega_{231}, \omega_{241}, \omega_{251}, \omega_{261})^T \\
&= \left(\frac{6887h^3}{14175}, \frac{5996h^3}{4725}, \frac{233h^3}{270}, \frac{416h^3}{567}, \frac{379h^3}{945}, \frac{596h^3}{4725}, \frac{491h^3}{28350} \right)^T \\
& (\omega_{301}, \omega_{311}, \omega_{321}, \omega_{331}, \omega_{341}, \omega_{351}, \omega_{361})^T \\
&= \left(\frac{52893h^3}{44800}, \frac{43173h^3}{11200}, \frac{14499h^3}{8960}, \frac{9h^3}{5}, \frac{8829h^3}{8960}, \frac{3483h^3}{11200}, \frac{1917h^3}{44800} \right)^T \\
& (\omega_{401}, \omega_{411}, \omega_{421}, \omega_{431}, \omega_{441}, \omega_{451}, \omega_{461})^T \\
&= \left(\frac{30904h^3}{14175}, \frac{37312h^3}{4725}, \frac{1808h^3}{945}, \frac{2176h^3}{567}, \frac{248h^3}{135}, \frac{2752h^3}{4725}, \frac{1136h^3}{14175} \right)^T \\
& (\omega_{501}, \omega_{511}, \omega_{521}, \omega_{531}, \omega_{541}, \omega_{551}, \omega_{561})^T \\
&= \left(\frac{505625h^3}{145152}, \frac{162125h^3}{12096}, \frac{85625h^3}{48384}, \frac{66875h^3}{9072}, \frac{119375h^3}{48384}, \frac{1625h^3}{1728}, \frac{18625h^3}{145152} \right)^T \\
& (\omega_{601}, \omega_{611}, \omega_{621}, \omega_{631}, \omega_{641}, \omega_{651}, \omega_{661})^T \\
&= \left(\frac{891h^3}{175}, \frac{3564h^3}{175}, \frac{81h^3}{70}, \frac{432h^3}{35}, \frac{81h^3}{35}, \frac{324h^3}{175}, \frac{9h^3}{50} \right)^T
\end{aligned} \tag{9}$$

must be simple or less than one.

The resulting expression from (12) is $\rho(z) = z^5(z - 1)$ which has roots simple and less than one. Therefore, the implicit block method is zero-stable, hence convergent.

4 Numerical Results and Discussion

To display the accuracy and convergence of the implicit block method, certain linear initial and boundary value problems and the results are graphically shown.

Problem 1 Consider the initial value problem

$$y^{iv} = -y'', y(0) = 0, y'(0) = -\frac{1.1}{72 - 50\pi}, y''(0) = \frac{1}{144 - 100\pi}, y'''(0) = \frac{1.2}{144 - 100\pi}$$

Exact Solution: $y = \frac{1-x-\cos x-1.2\sin x}{144-100\pi}$.

Problem 2 Consider the initial value problem

$$y^{iv} = -xy - (8 + 7x + x^3)e^x, y(0) = 0, y'(0) = 1, y''(0) = 0, y'''(0) = -3$$

Exact Solution: $y = x(1 - x)e^x$.

Problem 3 Consider the boundary value problem that involves a cantilever beam of length L with both ends fixed, distributed load $k(x)$, modulus of elasticity E and the moment of inertia I . The problem is solved for $k(x) = x$, $L = 1$ and $EI = 1$.

$$EIy^{iv} = k(x), y(0) = 0, y'(0) = 0, y''(L) = 0, y'''(L) = 0$$

with exact solution: $y = \frac{1}{120}(20x^2 - 10x^3 + x^5)$. Source: [16]

Problem 4 Consider the boundary value problem

$$y^{iv} = (x^4 + 14x^3 + 49x^2 + 32x - 12)e^x, y(0) = y'(0) = y(1) = y'(1) = 0$$

corresponding to the bending of a thin beam clamped at both ends.

Exact Solution: $y = x^2(1 - x^2)e^x$.

Problems 1–4 have considered both initial and boundary value problems. The implicit block method was adopted to solve these problems in comparison to previously existing authors. In Problems 1 and 3, the convergence and high accuracy of the implicit block method is observed as seen in Tables 1 and 3. Likewise, in considering the maximum error encountered whilst solving Problem 2, it is observed from Table 2 that as the number of steps increased over the interval

Table 1 Comparison of error for Problem 1

x	Error	Error (implicit block method)
0.1	6.51E-19	6.51E-19
0.2	1.30E-18	8.67E-19
0.3	4.77E-18	4.34E-19
0.4	1.73E-17	1.73E-18
0.5	4.34E-17	0.00E+00
0.6	9.54E-17	8.67E-19
0.7	1.81E-16	0.00E+00
0.8	3.16E-16	0.00E+00
0.9	5.19E-16	0.00E+00
1.0	8.05E-16	0.00E+00

Table 2 Comparison of maximum error for Problem 2

Number of steps for $[0, 2]$	Max error	Max error (implicit block method)
83	8.25E-07	2.61E-13
145	1.75E-07	2.84E-14
234	1.75E-07	3.02E-14

Table 3 Comparison of error for Problem 3

x	Error [16]	Error (implicit block method)
0.13	2.39E-17	4.34E-19
0.25	3.47E-17	0.00E+00
0.38	1.11E-16	0.00E+00
0.50	2.29E-16	0.00E+00
0.63	3.12E-16	0.00E+00
0.75	3.68E-16	0.00E+00
0.88	4.44E-16	0.00E+00
1.00	5.13E-16	0.00E+00

$[0, 2]$, the implicit block method showed impressive convergence as the error is so close to the exact solution. In addition, the ability to properly get an accurate result at the boundary for the boundary value problems, Problem 4 considered documenting the error at the boundary point $x = 1$. Table 4 also shows the high accuracy of the implicit block method with accurate prediction at the boundary point tending faster to the $y(1) = 0$.

Table 4 Comparison of error at the boundary $x = 1$ for Problem 4

h	Max error	Max error (implicit block method)
1/5	4.70E-07	2.54E-16
1/7	3.28E-08	1.04E-16
1/9	8.83E-09	2.53E-19

5 Conclusion

This article has presented an implicit six-step block method for solving both fourth order initial and boundary value problems. The development of the implicit block method adopted the linear block approach which is quite straightforward to adapt. In addition, the properties required to ensure convergence of the block method was satisfied and these was affirmed from the numerical results presented. Therefore, this article has given a better convergent and highly accurate numerical approach for solving fourth order initial and boundary value problems.

References

1. Fatunla, S.O.: Block methods for second order ODEs. *Int. J. Comput. Math.* **41**(1–2), 55–63 (1991). <https://doi.org/10.1080/00207169108804026>
2. Bulirsch, R., Stoer, J.: Numerical treatment of ordinary differential equations by extrapolation methods. *Numer. Math.* **8**(1), 1–13 (1966). <https://doi.org/10.1007/BF02165234>
3. Fang, Y., You, X., Ming, Q.: Exponentially fitted two-derivative runge-kutta methods for the schrodinger equation. *Int. J. Mod. Phys.* **24**(10), 1350073 (2013). <https://doi.org/10.1142/S0129183113500733>
4. Kalogiratou, Z., Monovasilis, T., Psihoyios, G., Simos, T.E.: Runge-kutta type methods with special properties for the numerical integration of ordinary differential equations. *Phys. Rep.* **536**(3), 75–146 (2014). <https://doi.org/10.1016/j.physrep.2013.11.003>
5. Hussain, K., Ismail, F., Senu, N.: Two embedded pairs of runge-kutta type methods for direct solution of special fourth-order ordinary differential equations. *Math. Prob. Eng.* **2015** (196595), 1–12 (2015). <https://doi.org/10.1155/2015/196595>
6. Awoyemi, D.O.: Algorithmic collocation approach for direct solution of fourth-order initial-value problems of ordinary differential equations. *Int. J. Comput. Math.* **82**(3), 321–329 (2005). <https://doi.org/10.1080/00207160412331296634>
7. Adeyeye, O., Kayode, S.J.: Two-step two-point hybrid methods for general second order differential equations. *Afr. J. Math. Comput. Sci. Res.* **6**(10), 191–196 (2013). <https://doi.org/10.5897/AJMCSR2013.0502>
8. Jator, S.N., Lee, L.: Implementing a seventh-order linear multistep method in a predictor-corrector mode or block mode: which is more efficient for the general second order initial value problem. *SpringerPlus* **3**(1), 1–8 (2014). <https://doi.org/10.1186/2193-1801-3-447>
9. Jator, S.N.: Solving second order initial value problems by a hybrid multistep method without predictors. *Appl. Math. Comput.* **217**(8), 4036–4046 (2010). <https://doi.org/10.1016/j.amc.2010.10.010>
10. Lambert, J.D.: *Computational methods in ordinary differential equations*. London (1973)

11. Hull, T.E., Enright, W.H., Fellen, B.M., Sedgwick, A.E.: Comparing numerical methods for ordinary differential equations. *SIAM J. Numer. Anal.* **9**(4), 603–637 (2006). <https://doi.org/10.1137/0709052>
12. Omar, Z., Kuboye, J.O.: New seven-step numerical method for direct solution of fourth order ordinary differential equations. *J. Math. Fundam. Sci.* **48**(2):94–105 (2016). <http://dx.doi.org/10.5614%2Fj.math.fund.sci.2016.48.2.1>
13. Kayode, S.J., Duromola, M.K., Bolaji, B.: Direct solution of initial value problems of fourth order ordinary differential equations using modified implicit hybrid block method. *J. Sci. Res. Rep.* **3**, 2790–2798 (2014). <https://doi.org/10.9734/JSRR/2014/11953>
14. Omar, Z., Abdelrahim, R.: Direct solution of fourth order ordinary differential equations using a one step hybrid block method of order five. *Int. J. Pure Appl. Math.* **109**(4), 763–777 (2016). <https://doi.org/10.12732/ijpam.v109i4.1>
15. Papakostas, S.N., Tsitmidelis, S., Tsitouras, C.: Evolutionary generation of 7th order runge–kutta–nyström type methods for solving $y(4) = f(x,y)$. In: *International Conference of Computational Methods in Sciences and Engineering*, Athens, March 2015. AIP Conference Proceedings, vol. 1702, no. 1, pp. 190018-1–190018-4. AIP Publishing, New York (2015). <https://doi.org/10.1063/1.4938985>
16. Jator, S.N.: Numerical integrators for fourth order initial and boundary value problems. *Int. J. Pure Appl. Math.* **47**(4), 563–576 (2008). <https://doi.org/10.12732/ijpam.v47i4.8>

Chapter 22

Stability Analysis of Explicit and Semi-implicit Euler Methods for Solving Stochastic Delay Differential Equations



Norhayati Rosli, Noor Amalina Nisa Ariffin, Yeak Su Hoe and Arifah Bahar

Abstract This paper dealt with the stability analysis of explicit and semi implicit Euler methods in approximating the solutions of linear stochastic delay differential equations (SDDEs). It has been proved that the methods are convergent with strong order 0.5 and are numerically stable in general mean square (GMS) and mean square (MS) sense for certain conditions. A comparative study of the stability explicit and semi implicit Euler methods in approximating the solutions of SDDEs are performed to visualize the theoretical results. Numerical experiments are conducted by applying both methods to linear SDDEs.

Keywords Stochastic delay differential equations · Mean square stable Explicit method · Implicit method · General mean square stable

N. Rosli (✉) · N. A. N. Ariffin
Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang,
26300 Gambang, Pahang, Malaysia
e-mail: norhayati@ump.edu.my

N. A. N. Ariffin
e-mail: amalinanisa1188@gmail.com

Y. S. Hoe (✉) · A. Bahar
Faculty of Science, Department of Mathematical Sciences, UTM Centre for Industrial and Applied Mathematics (UTM-CIAM), Universiti Teknologi Malaysia,
81310 Johor Bahru, Malaysia
e-mail: s.h.yeak@utm.my

A. Bahar
e-mail: arifah@utm.my

1 Introduction

Most of the natural systems around us are subjected to the presence of delayed feedback and are influenced by the uncontrolled environmental noise. For instance, the growth of the cancer cells is non-instantaneous but responds only after some time lag, $r > 0$. Cancer cells also subject to the uncontrolled factors such as blood pressures variations and individual characteristics like genes and stress impacts [1]. The most suitable mathematical equations describing such system is stochastic delay differential equations (SDDEs). In SDDEs, time delay and noisy behaviour are incorporating to its deterministic counterpart. Due to the presence of both effects, solving SDDEs is not an easy task. Analytical solutions of SDDEs are often unavailable. Thus, numerical methods provide a way to solve problem. The development of numerical methods for SDDEs is now among of the research interest. Amongst of the recent works are [2–4]. They proposed explicit numerical methods for solving SDDEs. The convergence and stability analysis of the semi-implicit method for linear SDDEs has been presented in [5]. It is the aimed of this paper to investigate the performance of explicit and semi-implicit Euler methods in approximating the solution of SDDEs. This paper is arranged as follows; Sect. 2 presents the mean-square stability properties of explicit and semi implicit Euler methods. Numerical experiment is conducted in Sect. 3. Concluding remarks are given in Sect. 4.

2 Stochastic Delay Differential Equations

Consider SDDEs of Ito form

$$\begin{aligned} dX(t) &= f(X(t), X(t-r))dt + g(X(t), X(t-r))dW(t), \quad t \in [-r, T] \\ X(t) &= \Phi(t), \quad t \in [-r, 0] \end{aligned} \quad (1)$$

where $\{W_t : t \in \mathbb{R}\}$ is a standard Wiener process with $W_0 = 0$ and the increments $W(t) - W(s) \sim N(0, t - s)$, $0 \leq s \leq t$. $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ are drift and diffusion functions, respectively and $\Phi(t)$ is an initial function defined on interval $[-r, 0]$ which is F_0 -measurable and right continuous, $E\|\Phi\|^2 < \infty$, where $\|\Phi\| = \sup_{-r \leq s \leq 0} |\Phi(s)|$. $\Phi(t)$ does not depends on $W(t)$ and $r > 0$ is a positive fixed delay.

Linear SDDEs is written by

$$\begin{aligned} dX(t) &= [aX(t) + bX(t-1)]dt + [cX(t) + dX(t-1)]dW(t), \quad t \in [-r, T] \\ X(t) &= 1 + t, \quad t \in [-r, 0] \end{aligned} \quad (2)$$

2.1 Euler Scheme for SDDEs

Let consider

$$X_{n+1} = X_n + [\alpha(aX_{n+1} + bX_{n+1-m}) + (1 - \alpha)(aX_n + bX_{n-m})]h + (cX_n + dX_{n-m})\Delta W_n \tag{3}$$

where α is a parameter with $0 \leq \alpha \leq 1$ and $h > 0$ satisfies $r = mh$, for positive integer, m and $t_n = nh$. X_n is an approximate solution to $X(t_n)$. If $t_n \leq 0$, $X_n = \Phi(t_n)$. The increment $\Delta W_n = W(t_{n+1}) - W(t_n)$ are independent and normally distributed with mean zero and variance, Δt .

A method is said to be explicit if $\alpha = 0$, hence we have

$$X_{n+1} = X_n + (aX_n + bX_{n-m})h + (cX_n + dX_{n-m})\Delta W_n \tag{4}$$

A semi-implicit Euler scheme is given by (2) for $0 < \alpha \leq 1$.

2.2 Convergence and Mean Square Stability of Euler Scheme

The following results are cited from [5].

Theorem 1 ([5]) *Assume that $ah\alpha < 1$. The numerical solution produced by (3) is convergent to the exact solution of (1) in the mean square sense with order of 0.5, i.e. there exists a positive constant C such that*

$$\max_{1 \leq n \leq N} \left(E(\varepsilon_n)^2 \right)^{\frac{1}{2}} \leq Ch^{\frac{1}{2}} \quad \text{as } h \rightarrow 0 \tag{5}$$

where $\varepsilon_n = X(t_n) - X_n$ is defined as a global error.

Lemma 1 ([5]) *If*

$$a < -|b| - \frac{1}{2}c^2 \tag{6}$$

then the solution of (2) is mean square stable, that is

$$\lim_{t \rightarrow \infty} E|X(t)|^2 = 0 \tag{7}$$

Definition 1 ([5]) Under condition (6) a numerical method is said to be general mean square stable (GMS), if there exists $h_0(a,b,c,d) > 0$, such that any application of the method to problem (1) generates numerical approximations, X_n which satisfy

$$\lim_{n \rightarrow \infty} E|X_n|^2 = 0 \tag{8}$$

for every stepsize $h = \frac{\tau}{m}$.

Definition 2 ([5]) Under condition (6) a numerical method is said to be mean square stable (MS), if there exists $h_0(a,b,c,d) > 0$, such that any application of the method to problem (1) generates numerical approximations, X_n which satisfy

$$\lim_{n \rightarrow \infty} E|X_n|^2 = 0 \tag{9}$$

for all $h \in (0, h_0(a,b,c,d))$ with $h = \frac{\tau}{m}$.

Theorem 2 ([5]) Under condition (6) and let

$$K = \frac{|a| + |b|}{2|a|} + \frac{2a + 2|b| + (|c| + |d|)^2}{2|a|(|a| + |b|)} \tag{10}$$

If $K < 0$, then for all $\alpha \in [0, 1]$, the Euler method (3) is GMS stable. If $K \geq 0$, then for $\alpha \in (K, 1]$, the Euler method (3) is GMS stable and for $\alpha \in [0, K]$, it is MS-stable and $h_0(a,b,c,d) = \min\{h', h''\}$, where $h' = \max\{h_1, h_2\}$, $h'' = \max\left\{\frac{1}{|a|}, h_2\right\}$ and $h_1 = \min\left\{\frac{1}{|a|}, \frac{-(2a + 2|b| + (|c| + |d|)^2)}{(a + |b|)^2}\right\}$, $h_2 = \frac{-(2a + 2|b| + (|c| + |d|)^2)}{(a + |b|)^2}$.

3 Numerical Experiments

Let consider linear SDDE (2) with sets of coefficients are given in Table 1.

By Theorem 2, SDDE (2) is GMS-stable for set of coefficients C1 if $0.1693 < \alpha \leq 1$ and it is MS-stable for $0 \leq \alpha \leq 0.1693$. For C2, a linear SDDE (2) is GMS-stable for $0 \leq \alpha \leq 1$ when $h \in (0, 1.250)$. For C3, the solution obtained is GMS-stable for $\alpha \in (0, 1]$ and it is MS-stable if $\alpha = 0$ when $h \in (0, 1.0)$.

Table 1 Coefficients of linear SDDEs and the corresponding values of K and h_0

Coefficients	a	b	c	d	K	$h_0(a,b,c,d)$
C1	-2	0.2	0.5	0.0	0.1693	0.6921
C2	-0.8	0.2	0.2	0.2	-0.0250	1.2500
C3	-1.0	0.2	0.2	0.2	0.0000	1.0000

Theoretically, it shows that a set of coefficients C1 produce unstable solution if the explicit Euler method is applied, but it is GMS-stable and MS-stable for semi-implicit method under certain values of α . Moreover, the solution is GMS-unstable for $\alpha < 0.1693$. Linear SDDE generates from C2 is GMS-stable for both methods. Linear SDDE generates from C3 is MS-stable for both methods and GMS-unstable for explicit method. The theoretical results are confirmed by applying explicit Euler method (4) and semi implicit (3) with fixed parameter $\alpha = 0, 0.1$ and 1.0. Table 2 shows the corresponding methods for each α .

In Figs. 1, 2 and 3, the stepsize is fixed to $h = \frac{1}{10}$ and the parameter α is changed according to Table 2.

Table 2 Explicit and semi-implicit Euler methods

α	Method	Formula
0	Explicit	$X_{n+1} = X_n + (aX_n + bX_{n-m})h + (cX_n + dX_{n-m})\Delta W_n$
0.1	Semi-implicit	$X_{n+1} = X_n + [0.1(aX_{n+1} + bX_{n+1-m}) + 0.9(aX_n + bX_{n-m})]h + (cX_n + dX_{n-m})\Delta W_n$
1.0	Semi-implicit	$X_{n+1} = X_n + (aX_{n+1} + bX_{n+1-m})h + (cX_n + dX_{n-m})\Delta W_n$

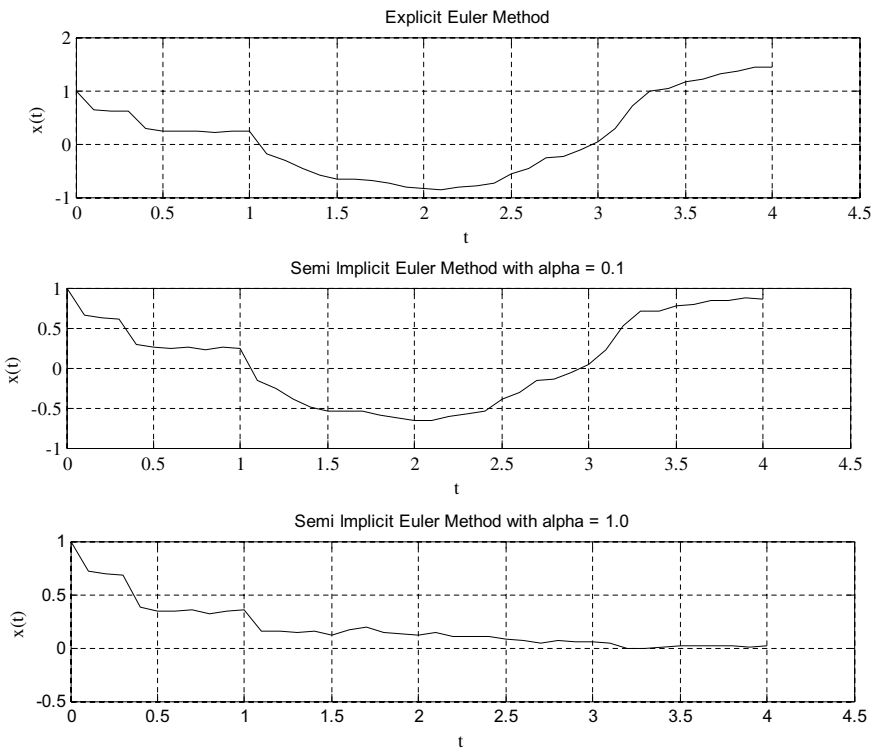


Fig. 1 Simulation result for C1 with fixed stepsize $h = \frac{1}{10}$ for $\alpha = 0, 0.1$ and 1.0

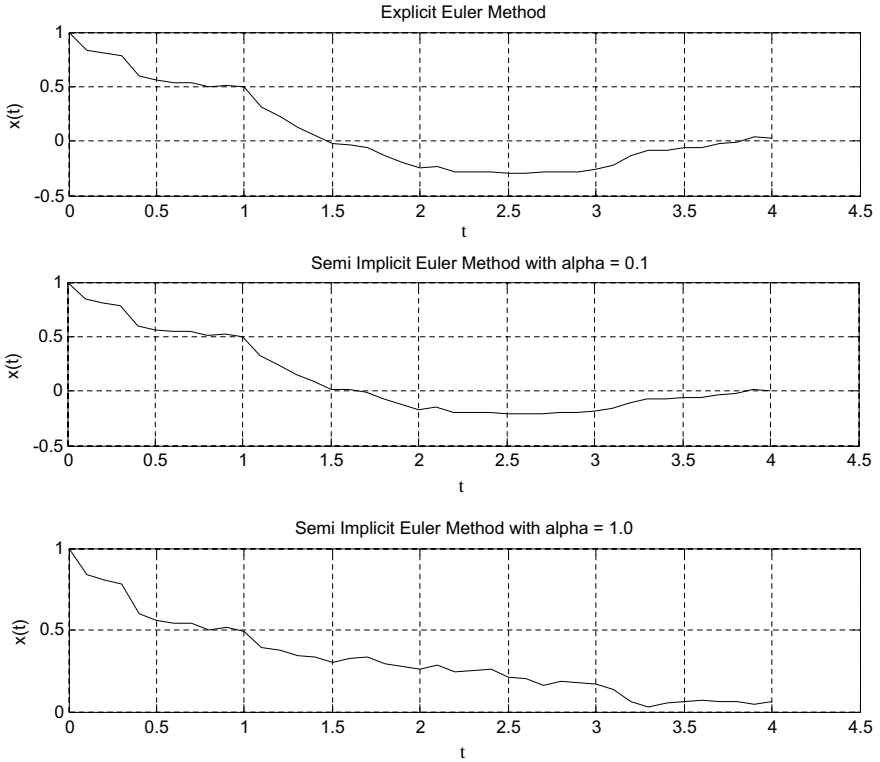


Fig. 2 Simulation result for C2 with fixed stepsize $h = \frac{1}{10}$ for $\alpha = 0, 0.1$ and 1.0

For $\alpha = 0$ and 0.1 , the simulated results for C2 are mean-square unstable. This indicates that for small values of α (when the method is reduced to explicit scheme) the results become unstable for C2. However, for C1 and C3, the results tend to negative values for $\alpha = 0$ and 0.1 , hence indicates instability of the solution. However, the simulated result produced by semi-implicit method with $\alpha = 1.0$ possess the stability in mean-square for C1, C2 and C3.

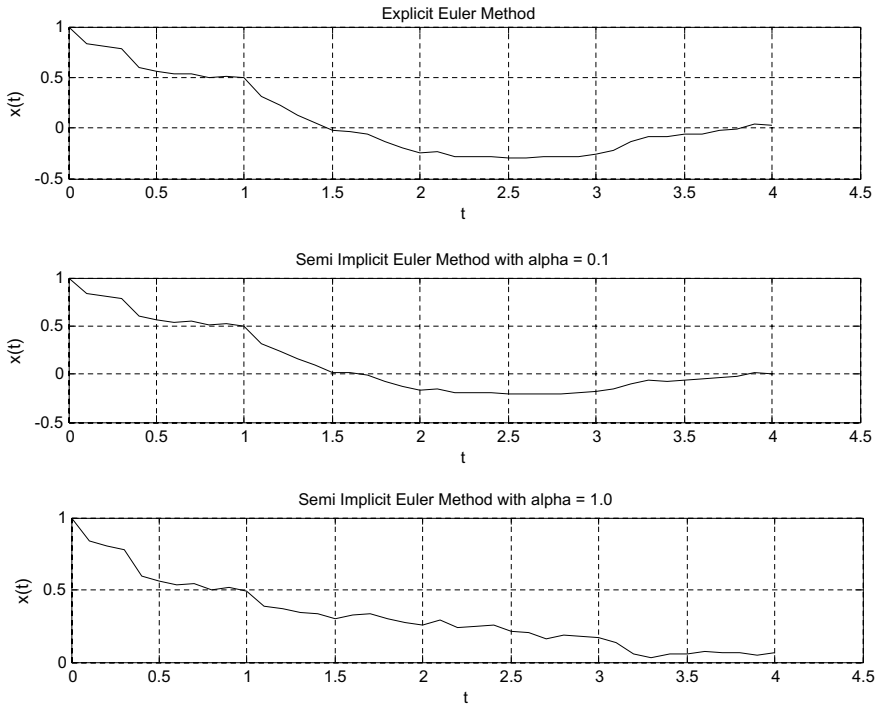


Fig. 3 Simulation result for C3 with fixed stepsize $h = \frac{1}{10}$ for $\alpha = 0, 0.1$ and 1.0

4 Conclusions

It can be concluded that the stability of explicit and semi-implicit methods are influenced by the values of h and α . Small values of α produce instable results compare than large value of α (for $\alpha = 1.0$).

Acknowledgements We would like to thank the Ministry of Education (MOE) and Research and Innovation Department, Universiti Malaysia Pahang (UMP) for their financial supports through FRGS Vote No: RDU130122 and Internal UMP Grant RDU1703190.

References

1. Ditlevsen. S., Samson, A.: Stochastic Biomathematical Models, pp. 3–35. Springer, New York (2013)
2. Baker, C.T.H., Buckwar, E.: Numerical analysis of explicit one-step methods for stochastic delay differential equations. *J. Comput. Math.* **3**, 315–335 (2000)
3. Kuchler, U., Platen, E.: Strong discrete time approximation of stochastic differential equations with time delay. *Math. Comput. Simul.* **54**, 189–205 (2000)

4. Rosli, N., Bahar, A., Yeak, S.H., Mao, X.: A systematic derivation of stochastic Taylor methods for stochastic delay differential equations. *Bull. Malays. Math. Sci. Soc.* **36**(3), 555–576 (2013)
5. Liu, M., Cao, W., Fan, Z.: Convergence and stability of the semi implicit Euler method for a linear stochastic differential delay equation. *J. Comput. Appl. Maths.* **170**, 255–268 (2004)

Chapter 23

Stability Analysis of 4-Stage Stochastic Runge-Kutta Method (SRK4) and Specific Stochastic Runge-Kutta Method (SRKS1.5) for Stochastic Differential Equations



Noor Amalina Nisa Ariffin, Norhayati Rosli
and Abdul Rahman Mohd Kasim

Abstract This paper is devoted to investigate the mean-square stability of 4-stage stochastic Runge-Kutta (SRK4) and specific stochastic Runge-Kutta (SRKS1.5) methods for linear stochastic differential equations (SDEs). The mean-square stability functions of SRK4 and SRKS1.5 are derived. The regions in which the methods are stable in the mean-square sense are plotted. Numerical experiments are performed to verify the stability properties of both methods.

Keywords Stochastic differential equations · Stochastic Runge-kutta · Mean square stable · Explicit method · General mean square stable

1 Introduction

Nowadays, stochastic differential equations (SDEs) have been extensively used to model diverse areas in neural networks, ecosystem, population dynamics, genetics and macroeconomic systems. However, due to the presence of the Wiener process in SDEs the analytical solutions of SDEs is hard to be found [1]. Hence solving these SDEs numerically are required. Latest work on numerical solution for SDEs has been done by [2] where the independent internal stages of stochastic

N. A. N. Ariffin (✉) · N. Rosli · A. R. M. Kasim
Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang,
26300 Gambang, Pahang, Malaysia
e-mail: amalinanisa1188@gmail.com

N. Rosli
e-mail: norhayati@ump.edu.my

A. R. M. Kasim
e-mail: rahmanmohd@ump.edu.my

Runge-Kutta (SRK) method have been introduced. The new proposed method in [2] has been proven can overcome the complexity in deriving the SRK method for SDEs. Crucial question that might be asked while developing the numerical methods for SDEs is whether the method possess the stability property or not? There are a few researchers who investigated the stability of the numerical methods for SDEs (see [3–8]). The investigation of the stability of 4-stage stochastic Runge-Kutta method (SRK4) with strong order 1.5 and also specific stochastic Runge-Kutta method (SRKS1.5) method with the same order of convergence will be carried out in this paper. Mean square stability of numerical methods for SDEs is performed by applying both SRK4 and SRKS1.5 schemes to the linear SDEs model. Then, the stability functions for both methods will be obtained and visualized in the (u, v) -plane. This stability region will assist us identifying the step size to be used that can guarantee the stability of the approximate solution.

2 Stochastic Runge-Kutta (SRK) Methods for SDEs

The general form of SRK method can be written as [7]

$$\begin{aligned}
 Y_i(t) &= Y_n(t_0) + \Delta \sum_{j=1}^{i-1} a_{ij} f(y_j(t)) + \sum_{j=1}^{i-1} (b_{ij}^{(1)} J_1 + b_{ij}^{(2)} \frac{J_{10}}{h}) g(y_j(t)) \\
 y_{n+1}(t) &= y_n + \Delta \sum_{j=1}^s \alpha_j f(y_j(t)) + \sum_{j=1}^s (\gamma_j^{(1)} J_1 + \gamma_j^{(2)} \frac{J_{10}}{h}) g(y_j(t))
 \end{aligned}
 \tag{1}$$

where $A = (a_{ij})$ and $B = (b_{ij})$ are $s \times s$ matrices of real elements, $\alpha^T = (\alpha_1, \dots, \alpha_s)$ and $\gamma^T = (\gamma_1, \dots, \gamma_s)$ are row vectors $\in \mathbb{R}^s$. J_1 and J_{10} integrals are both stochastic component where $J_1 = \int_{t_n}^{t_{n+1}} \circ dW(t)$ and $J_{10} = \int_{t_n}^{t_{n+1}} \int_{s_n}^s \circ dW(t) d(s)$ respectively.

SRK4 is developed based on (1) by Burrage [7] and the numerical scheme of SRK4 method in Butcher’s tableau form is shown as follows

0		0		0	
0.5	0	$B^{(1)}$	-0.72429	$B^{(2)}$	2.70200
0	0.5		0.42373		1.75726
0	0	1	-1.57847	0.84010	1.73837
0	0	1	0	0	0
α	1	1	1	$\gamma^{(1)}$	-0.78007
				$\gamma^{(2)}$	0.07363
					1.48652
					0.21992
					1.69395
					1.63610
					-3.02400
					-0.30604

(2)

SRK method (1) suffers from the complexity of investigating the order conditions due to the large numbers of equations to be solved. The new explicit SRK method with several groups of independent internal stages have been introduced by Aiguo and Xiao in 2015 to overcome that difficulty [2]. The general form of new independent $s -$ stage specific SRK method is given by

$$\begin{aligned}
 Y_{i_0}^0(t) &= Y_n(t_0) + \Delta \sum_{j=1}^{s_0} a_{i_0j}^{(0)} f(y_j^{(0)}(t)) + J_1 \sum_{j=1}^{s_0} b_{i_0j}^{(0)} g(y_j^{(0)}(t)) \\
 Y_{i_1}^1(t) &= Y_n(t_0) + \Delta \sum_{j=1}^{s_1} a_{i_1j}^{(1)} f(y_j^{(1)}(t)) + \frac{J_{10}}{\Delta} \sum_{j=1}^{s_1} b_{i_1j}^{(1)} g(y_j^{(1)}(t)) \\
 y_{n+1}(t) &= y_n + \Delta \sum_{i_0=1}^{s_0} \alpha_{i_0}^{(0)} f(y_{i_0}^{(0)}(t)) + J_1 \sum_{i_0=1}^{s_0} \gamma_{i_0}^{(0)} g(y_{i_0}^{(0)}(t)) \\
 &\quad + \Delta \sum_{i_1=1}^{s_1} \alpha_{i_1}^{(1)} f(y_{i_1}^{(1)}(t)) + \frac{J_{10}}{\Delta} \sum_{i_1=1}^{s_1} \gamma_{i_1}^{(1)} g(y_{i_1}^{(1)}(t))
 \end{aligned} \tag{3}$$

where $i_0 = 1, 2, \dots, s_0$ and $i_1 = 1, 2, \dots, s_1$ with $s_0 = 4$ and $s_1 = 3$ are the independent internal stages for this specific SRK method (SRKS1.5). In the Butcher's tableau form it can be written as shown below.

$$\begin{array}{c|ccc}
 A^{(0)} & 0.5 & & \\
 & 1 & 0 & \\
 & 3 & 0 & 0 \\
 \hline
 \alpha^{(0)} & -1 & -1 & 0 & 0
 \end{array}
 \quad
 \begin{array}{c|ccc}
 B^{(0)} & 0.5 & & \\
 & 0 & 0.5 & \\
 & 0 & 0 & 1 \\
 \hline
 \gamma^{(0)} & 1/6 & 1/3 & 1/3 & 1/6
 \end{array}
 \quad
 \begin{array}{c|cc}
 A^{(1)} & 0 & \\
 & 1 & 0 \\
 \hline
 \alpha^{(1)} & -2/3 & 2/3 & 0
 \end{array}
 \quad
 \begin{array}{c|cc}
 B^{(1)} & 1.5 & \\
 & 1.5 & 0 \\
 \hline
 \gamma^{(1)} & 0 & 1 & -1
 \end{array} \tag{4}$$

3 Stability Analysis of Stochastic Runge-Kutta (SRK) Methods for SDEs

Among various numbers of definition for numerical stability which currently appear in the literature are such as A -stability, β -stability, T -stability and mean-square exponential (MSE) stability [8–11]. The concept of mean square stability as defined by [5] is one of the most feasible ones. Let consider a linear SDE of the form

$$dX(t) = aX(t)dt + bX(t)dW(t), \quad t \in [0, T] \tag{5}$$

The exact solution of (5) is given by $X(t) = \Phi_{t,t_0}(X_0)$, where $\Phi_{t,t_0} = \exp((a)(t - t_0) + b(W(t) - W(t_0)))$.

Definition 1 (*Mean–square Stability* [7]) The numerical solution of linear SDE (5), y_{n+1} is said to be mean square stable if

$$\lim_{n \rightarrow \infty} (E|y_{n+1}|^2) = 0. \tag{6}$$

Next, the condition that guarantees the stability of the numerical methods to linear SDEs (5) is presented. By applying the SRK4 scheme (2) and SRKS 1.5

scheme (4) to the linear test in Eq. (5), the approximate solutions for both processes, y at time, t_{n+1} can be written in the form (7) and (8) respectively.

$$\begin{aligned}
 y_{n+1} = & y_n \left(1 - \frac{0.5534719789J_{10}^2 a J_1 b^3}{\Delta} + 0.6066682049\Delta a^2 J_1 b^2 J_{10} + 0.05522579906J_1^4 b^4 \right. \\
 & - 0.4131967610J_{10}^2 a^2 b^2 - 0.0005991230abJ_{10} + \frac{1}{2}\Delta^2 a^2 + \Delta a + 0.999987bJ_1 + \frac{1}{24}\Delta^4 a^4 \\
 & + 0.4999948469b^2 J_1^2 + \frac{1}{6}\Delta^3 a^3 + 0.1666640206J_1^3 b^3 + 0.9999884934\Delta abJ_1 + 4.034544850aJ_1 b^2 J_{10} \\
 & + \frac{0.000207138J_1^2 b^3 J_{10}}{\Delta} - \frac{4.620510630J_{10}^2 ab^2}{\Delta} - \frac{0.000627241J_{10}^2 J_1 b^3}{\Delta^2} - \frac{0.2827392065J_1^3 J_{10} b^4}{\Delta} \\
 & + \frac{0.2865172139J_{10}^2 J_1^2 b^4}{\Delta^2} + 0.5798840050\Delta^2 a^2 bJ_1 - 0.1661142185\Delta a^2 bJ_{10} - 0.3419882133\Delta aJ_1^2 b^2 \\
 & - 0.1029520247\Delta aJ_1^3 b^3 + 0.05043435167\Delta^3 a^3 J_1 b + 0.1485044200\Delta^2 a^3 bJ_{10} - 0.1157387149\Delta^2 a^2 J_1^2 b^2 \\
 & \left. - \frac{0.0001581892J_{10} J_1 b^2}{\Delta} + 0.4431985843aJ_1^2 b^3 J_{10} + \frac{0.000001J_{10} b}{\Delta} - \frac{0.0029505522J_{10}^2 b^2}{\Delta^2} \right)
 \end{aligned} \tag{7}$$

$$\begin{aligned}
 y_{n+1} = & y_n \left(1 - 2\Delta a - \frac{1}{2}\Delta^2 a^2 + \frac{1}{2}J_1 b \Delta a + J_1 b + \frac{1}{2}J_1^2 b^2 + \frac{1}{4}J_1^2 b^2 \Delta a + \frac{1}{6}J_1^3 b^3 + \frac{1}{24}J_1^3 b^3 \Delta a + \frac{1}{24}J_1^4 b^4 \right)
 \end{aligned} \tag{8}$$

By squaring both sides of (7) and (8) and taking its expectation, Eqs. (7) and (8) can be written in the form of $E(y_{n+1}^2) = E(y_n^2)R(\Delta, a, b)$. Then, by applying the changes of variables of $u = \Delta(a + b^2)$ and $v = \Delta b^2$ [7] to the $R(\Delta, a, b)$, the stability functions for SRKS1.5 and SRK4 can be written as in Eqs. (9) and (10).

$$\begin{aligned}
 R_{SRKS1.5}(u, v) = & 1 - \frac{5}{4}(u - v)^2 v + (u - v)v^2 - \frac{1}{4}(u - v)^3 v + \frac{3}{16}(u - v)^2 v^2 + \frac{25}{48}(u - v)v^3 \\
 & + \frac{5}{192}(u - v)^2 v^3 - 4u + 8v + 3(u - v)^2 - \frac{1}{2}(u - v)v + 2(u - v)^3 + \frac{1}{4}(u - v)^4 + \frac{25}{24}v^3 + \frac{35}{192}v^4,
 \end{aligned} \tag{9}$$

$$\begin{aligned}
 R_{SRK4}(u, v) = & 1 + 2u + 0.101922069v^4 + 0.270309581v^3 + 1.71057106v^2 + 2(u - v)^2 + \frac{4}{3}(u - v)^3 \\
 & + \frac{2}{3}(u - v)^4 + \frac{1}{4}(u - v)^5 + \frac{5}{72}(u - v)^6 + \frac{1}{72}(u - v)^7 + \frac{1}{576}(u - v)^8 + 0.23985221(u - v)^3 v^2 \\
 & - 0.00553023635(u - v)^2 v^3 + 0.1478030732(u - v)^5 v + 0.02157131671(u - v)^6 v \\
 & + 1.925252077(u - v)v^2 + 1.645042404(u - v)^3 v + 2.863798646(u - v)^2 v + 1.3704558(u - v)^2 v^2 \\
 & + 3.269893207(u - v)v + 0.6354256573(u - v)^4 v
 \end{aligned} \tag{10}$$

The corresponding stability functions are plotted in Figs. 1 and 2.

The white regions correspond to the regions where the functions are stable. Based on Figs. 1 and 2, it is clear that the SRKS1.5 method shows better stability property compare to SRK4. It can be confirmed by performing numerical experiments that are presented in the next section.

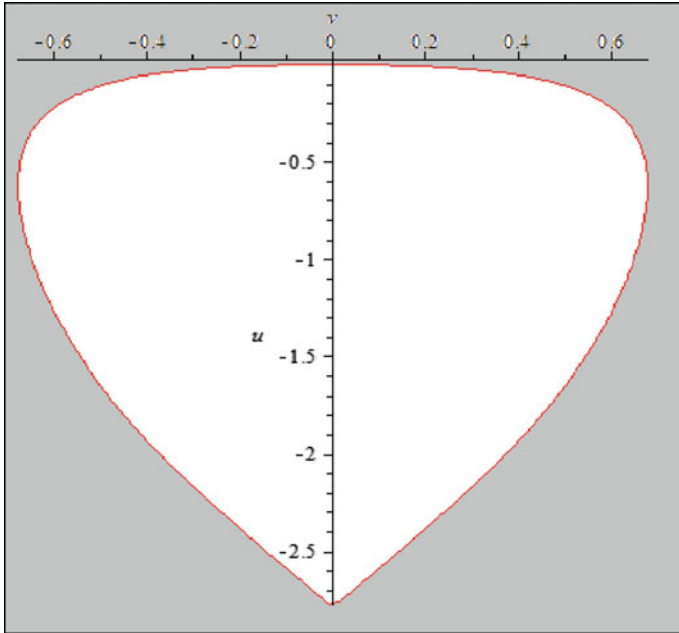


Fig. 1 The stability region of the stability function (8) by SRK4

4 Results and Discussions

We carried out the numerical experiment to verify the stability properties of SRK4 and SRKS1.5 which have been presented in Sect. 3. Numerical experiments show that the step size, Δ influences the mean-square stability of the corresponding methods. Let the value of $a = -4.4$ and $b = 0.6325$. The value of resulting u and v for $\Delta = 0.125, 0.25, 0.5, 1.0$ are listed in Table 1.

The second moment of y_T for $T \in [0,20]$ of Eq. (5) are estimated and the expectation for $N = 1000$ sample paths with 5 batches are computed as $E|y_n|^2 = \frac{1}{5(1000)} \sum_{i=1}^{1000} |y_n(\varpi_i)|^2$. The results are illustrated in Figs. 3 and 4.

Based on Figs. 3 and 4, the results obtained indicate the stability of the numerical solution if the chosen step size is in the stability region. However, if the chosen step size is out of the stability region, the numerical solutions for both methods are not stable.

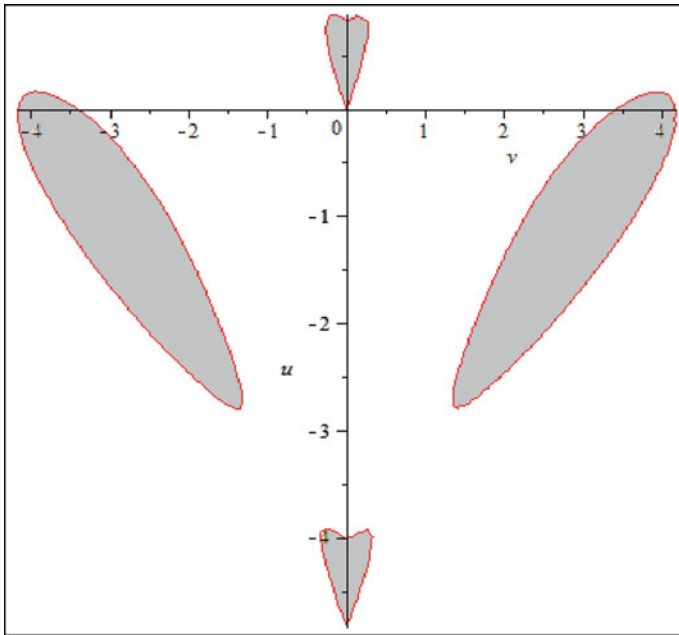


Fig. 2 The stability region of the stability function (9) by SRKS1.5

Table 1 Value of parameter u and v

Δ	0.125	0.25	0.5	1.0
u	-0.5	-1	-2	-4
v	0.05	0.1	0.2	0.4

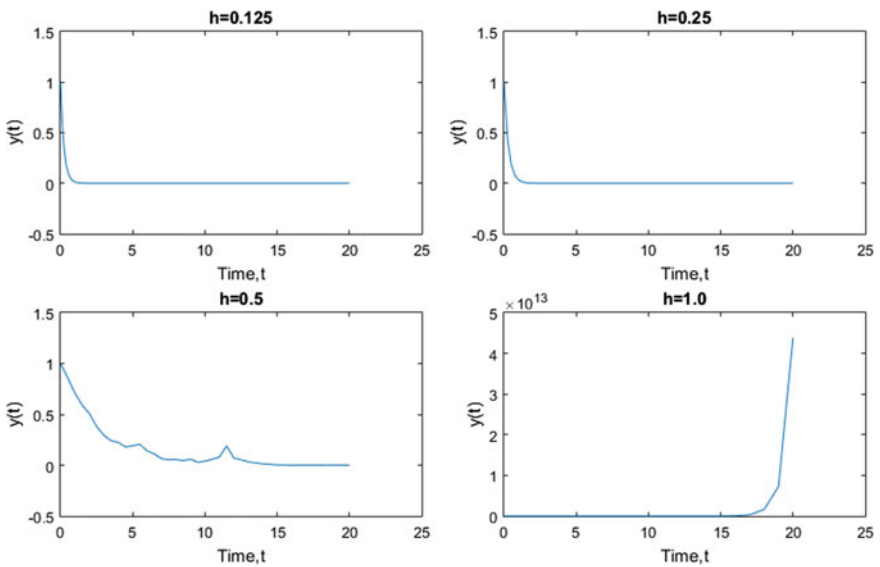


Fig. 3 Numerical solution of SDE (12) via SRK4

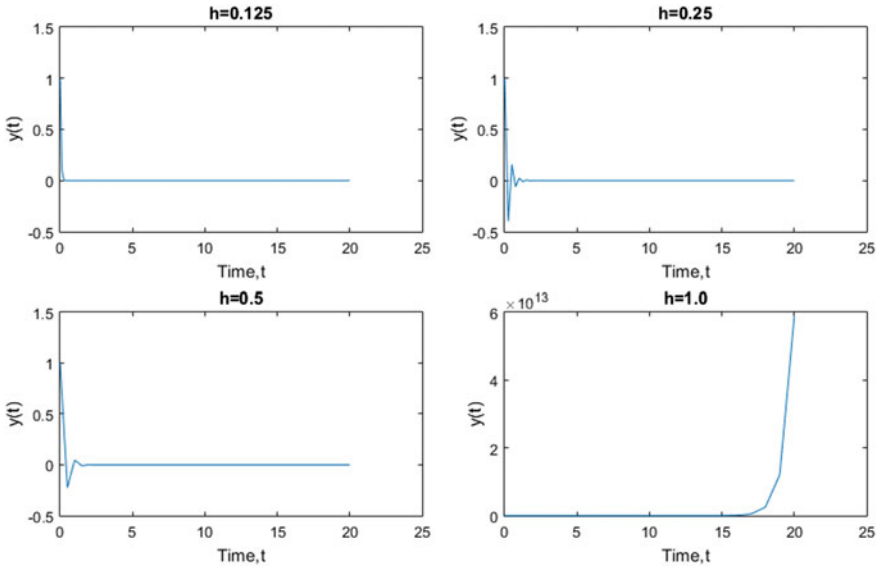


Fig. 4 Numerical solution of SDE (12) via SRKS1.5

5 Conclusion

The stability regions for SRK4 and SRKS1.5 for linear SDE (3) have been plotted. Figures 1 and 2 show that the specific SRK method SRKS1.5 have better stability property compared to SRK4. The numerical experiment proved that the numerical stability of the numerical solutions are influenced by step size, Δ .

Acknowledgements We would like to thank the Ministry of Education (MOE) and Research and Innovation Department, Universiti Malaysia Pahang (UMP) for their financial supports through FRGS Vote No: RDU130122 and Internal UMP Grant RDU1703190.

References

1. Norhayati, R., Arifah, B., Yeak, S.H., Haliza, A.R., Madidah, M.S.: Performance of Euler-Maruyama, 2-stage SRK and 4- stage SRK in approximating the strong solution of stochastic model. *Sains Malaysiana* **39**, 851–857 (2010)
2. Xiao, A., Tang, X.: High strong order stochastic Runge-kutta method for stratonovich stochastic differential equations with scalar noise. Springer, New York (2015)
3. Milstein, G.N.: Numerical Integration of Stochastic Differential Equations. Springer, Berlin (1995)
4. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1992)

5. Burrage, K., Burrage, P.M.: High strong order explicit Runge-Kutta methods for stochastic ordinary differential equations. *Appl. Numer. Math.* **22**(1–3), 81–101 (1996)
6. Saito, Y., Mitsui, T.: Stability analysis of numerical schemes for stochastic differential equations. *SIAM J. Num. Anal.* **33**(6), 2254–2267 (1996)
7. Burrage, P.M.: Runge-Kutta methods for stochastic differential equations. Ph.D. thesis, University of Queensland, Australia (1999)
8. Higham, D.J.: Mean-square and asymptotic stability of the stochastic theta method. *SIAM J. Numer. Anal.* **38**(3), 753–769 (2000)
9. Saito, Y., Mitsui, T.: Mean-square stability of numerical schemes for stochastic differential systems. *Vietnam J. Math.* **30**, 551–560 (2002)
10. Liu, M.Z., Spijker, M.N.: The stability of the θ -methods in the numerical solution of delay differential equations. *IMA J. Numer. Anal.* **10**(1), 31–48 (1990)
11. Ryashko, L.B., Schurz, H.: Mean square stability analysis of some linear stochastic systems. *Dyn. Syst. Appl.* **6**, 165–190 (1997)

Chapter 24

The VIKOR Method with Pythagorean Fuzzy Sets and Their Applications



Wan Rosanisah Wan Mohd and Lazim Abdullah

Abstract Pythagorean fuzzy sets (PFS) is proposed by Yager, which characterized by a membership, nonmembership and hesitation degree. The condition that the square sum of its membership and nonmembership degree is less than or equal to one and is very useful for decision makers (DMs) to depict the fuzzy character of data comprehensively. It is hypothesized that the PFS is also capable to model uncertainty and impreciseness in the practical decision making problems. A handful of research focused on Vlsekriterijumska Optimizacija I Kompromisno Resenje (VIKOR) method based on IFS but no one pay attention to propose the new VIKOR based on PFS. Therefore, the purpose of this paper is to propose the PFS for VIKOR method. This study uses Pythagorean fuzzy sets to handle the linguistic uncertainty and imprecision of human beings judgment. Finally, the compromise solution can be obtained. An illustrative example is demonstrated to show their practicality and effectiveness of the proposed VIKOR based on PFS.

Keywords Decision making · Linguistic · Pythagorean fuzzy set · Uncertainty · VIKOR

1 Introduction

Uncertainty is a major issue in the multi criteria decision making (MCDM) process. A large number of MCDM methods have been proposed to tackle the issues in, for instance, TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) [1], ELECTRE (Elimination and Choice Expressing REality) [2], PROMETHEE [3] and VIKOR (Vlsekriterijumska Optimizacija I Kompromisno

W. R. W. Mohd (✉) · L. Abdullah
Pusat Pengajian Informatik Dan Matematik Gunaan, Universiti Malaysia Terengganu,
21030 Kuala Nerus, Terengganu, Malaysia
e-mail: wanrosanisahwm@gmail.com

L. Abdullah
e-mail: lazim_m@umt.edu.my

Resenje) [4]. MCDM is the process of finding the best option from all the feasible alternatives that provides multiple and conflicting criteria. It deals with the real life situations that surround with the imprecise and vague in nature. It cannot be avoided as the decisions or opinion are coming from the decision makers' (DMs).

The DMs are unable to give the appropriate evaluation for the alternatives due to the time limitations, lack of data and lack of knowledge. To overcome this limitation, the only way to overcome the uncertainty, vagueness and ill-defined is fuzzy set (FS) theory [5]. Since it has performed successfully, a numerous researchers have been applied in MCDM [5–7]. However, in many cases, it cannot be used because it only consider the membership function (single-valued function). As a generalization to the FS, intuitionistic fuzzy set (IFS) [8] has been introduced which additional membership, so called non-membership and appropriate to manage MCDM problems [9–11]. Despite of its effectiveness in solving MCDM problems, the constraint of this IFS is the sum of their memberships degree are not greater than 1. In certain condition, these sets are failing to express the idea of the DMs because sometimes the summation of the degrees are exceeding one. Apparently, IFS are not able to address these situations.

In order to solve these limitation, Yager [12] introduced Pythagorean fuzzy set (PFS) as an extension to the IFS. The PFS is similar to the IFS which is characterized by the membership and non membership as well, but it is restricted to the square sum of the membership degrees are less than or equal to 1. The advantage of PFS is more general as compared to the IFS. Obviously, it can be said that all the IFS is a part of PFS. In other words, PFS can solve the problems of IFS, but IFS cannot solve the all problem of PFS [13]. To address this issue, in this paper, we propose a new VIKOR outlook by considering the Pythagorean fuzzy information. Specifically, this paper aims to propose VIKOR based on PFS.

The outline of the work is structured as follows. In Sect. 2 discusses the step by step of Pythagorean fuzzy VIKOR. Accordingly, Sect. 3 is the numerical example. Finally, Sect. 4 is a conclusion and future work.

2 Proposed Pythagorean Fuzzy VIKOR

In this section, the general steps of VIKOR is presented. The VIKOR is aiming to find a compromise solution to rank or select from various alternatives in the conflicting criteria [4]. The proposed algorithm of PF-VIKOR as shown below.

Step 1: Using the linguistic rating of Pythagorean fuzzy numbers (PFNs) to defuzzify the criteria weight and alternatives using the Eq. (1):

$$a = \frac{a_1 + 4a_2 + a_3}{6} \quad (1)$$

Step 2: Calculate f_{ij} and find the best and worst values that denoted by f_j^* and the f_j^- respectively for all criteria/attribute $j = 1, 2, \dots, n$ and alternatives $i = 1, 2, \dots, m$

$$\begin{aligned} f_j^* &= \max [f_{ij} | i = 1, 2, \dots, n] \\ f_j^- &= \min [f_{ij} | i = 1, 2, \dots, n] \end{aligned} \tag{2}$$

Step 3: Calculate S_i and R_i with the below equations:

$$\begin{aligned} S_i &= \sum_{j=1}^n w_j \frac{f_j^* - f_{ij}}{f_j^* - f_j^-} \\ R_i &= \max \left[w_j \frac{f_j^* - f_{ij}}{f_j^* - f_j^-} \right] \end{aligned} \tag{3}$$

Step 4: Compute $Q_i, i = 1, 2, \dots, n$ values with the following equation.

$$Q_i = v \frac{S_i - S^*}{S^- - S^*} + (1 - v) \frac{R_i - R^*}{R^- - R^*} \tag{4}$$

$$S^* = \min S_i; S^- = \max S_i; R^* = \min R_i; R^- = \max R_i; \tag{5}$$

Step 6: The best alternatives ranked based on the Q . If the values are minimum then, it has to check the two conditions either satisfied or not.

Condition 1 The acceptable advantage:

$$Q(A_2) - Q(A_1) \geq DQ \tag{6}$$

A_1 is referring to the best alternative or first score meanwhile A_2 is referring to second score of the Q_i . And M denote the number of threshold.

$$DQ = 1 / (M - 1) \tag{7}$$

Condition 2 Acceptable stability in decision making

Alternative with the first position considered as the best.

If these alternative could not satisfied with those conditions, then another compromise solutions are offered including:

- If the Condition 2 is not fulfilled, then A_1 and A_2 is considered.
- If the Condition 1 is not fulfilled, then A_1, A_2, \dots, A_N is considered. The A_1 and A_N is determined by the equation $Q_{A_N} - Q_{A_1} < DQ$ with max N .

3 Numerical Example

This part demonstrate the proposed PF-VIKOR method work using an illustrative example from [14] is about the green supplier development program selection was selected as the case of the study. Subsequently, the DMs are requested to represent their judgment based on the linguistic rating of Pythagorean fuzzy number for the criteria and alternatives (see Table 1).

3.1 Implementation

Step 1: The criteria weights \tilde{w}_j that has been aggregated (refer Table 2) are obtained. Refers to C1 (L, VL, M), the fuzzy weights are denoted by $w_j = (w_{j1}, w_{j2}, w_{j3})$:

$$w_{j1} = \min_k(0.7, 0.6, 0.8), \quad w_{j2} = \frac{1}{3}(0.6 + 0.4 + 0.6), \quad \tilde{w} = (0.6, 0.533, 0.3)$$

$$w_{j3} = \max_k(0.1, 0.2, 0.3)$$

Using Eq. (1) to defuzzify the weights \tilde{w}_j . For example, $\tilde{w}_j = (0.6, 0.533, 0.3)$ then $w_j = \frac{0.6 + (0.533 \times 4) + 0.3}{6} = 0.506$. Similarly, remaining criteria weight are calculated. The overall result will be obtained as shown in the last column in the Table 2.

Table 1 Linguistic ratings for alternatives and criteria [14]

Linguistic term (alternatives)	Linguistic term (criteria)	Pythagorean fuzzy number (PFN)
Very poor (VP)	Very low (VL)	(0.6,0.4,0.2)
Poor (P)	Low (L)	(0.7,0.6,0.1)
Fair (F)	Medium (M)	(0.8,0.6,0.3)
Good (G)	High (H)	(0.8,0.7,0.4)
Very good (VG)	Very high (VH)	(0.9,0.6,0.5)

Table 2 Linguistic and aggregated fuzzy weight ratings

Criteria	Linguistic variables			Aggregate fuzzy rating	Crisp
	D1	D2	D3		
C1	L	VL	M	(0.6,0.533,0.3)	0.506
C2	H	L	M	(0.7,0.467,0.4)	0.494
...
C16	L	L	L	(0.7,0.467,0.1)	0.444

Table 3 The f_j^* and f_j^- values

Criteria	Crisp rating			f_j^*	f_j^-
	P1	P2	P3		
C1	0.6	0.494	0.628	0.494	0.628
C2	0.444	0.461	0.544	0.444	0.544
...
C16	0.667	0.506	0.506	0.506	0.667

Table 4 The S_i , R_i and Q_i values

	P1	P2	P3
S_i	5.858	3.805	5.255
R_i	0.667	0.628	0.628
Q_i	1	0	0.353

Table 5 Alternatives rankings (ascending order)

S_i	P2	P3	P1
R_i	P3	P2	P1
Q_i	P2	P3	P1

Step 2: The decision makers are asking to rate for the three alternatives using linguistic ratings as presented in Table 1. Then, convert into Pythagorean fuzzy numbers. In a similar way, the criteria are generated. Using Eq. (2), the values of f_j^* and the f_j^- for entire criteria are calculated. The result obtained such in Table 3.

Step 3: Compute values of S_i , R_i and Q_i using Eqs. (3) and (4). The result obtained as presented in Table 4.

Step 4: The S_i , R_i and Q_i of the three alternatives are ranked in ascending order as Table 5.

Table 5 shows that the alternative P2 is the best ranked as compared to others because it has the minimum value of Q_i . Then, it has to check the two conditions. Using Eq. (7), $DQ = 0.5$. Then, condition 1 is checked using Eq. (6), $Q(P3) - Q(P2) = 0.353 - 0 = 0.353 < 0.5$. The result show it fails. Then, Condition 2 is checked and it satisfied. As the ranking $P2 = P3 > P1$, both P2 and P3 signify the best alternative.

4 Conclusions and Future Works

This work has proposed a new VIKOR to overcome MCDM problems using Pythagorean fuzzy sets. Pythagorean fuzzy number (PFN) is used to the VIKOR method in handling the linguistic uncertainty and imprecision of experts' opinion. Through this study, the linguistic variables of PFN are used to present the criteria of weight. The criteria of weight are important as it might influence the final result.

A numerical example showed the applicability and evaluation of the proposed method. Based on the result, it shows that P2 and P3 has emerged and selected as the appropriate alternatives in the green supplier development program (s). However, in this study it has a limitation where the method to determine the weight of criteria is general. In further action, it may be required another method to determine the criteria of weight using PFN to evaluate the selection problems.

References

1. Hwang, C.L., Yoon, K.S.: *Multiple Attribute Decision Methods and Applications*. Springer, Berlin, Germany (1981)
2. Roy, B., Bertier, P.: *La méthode ELECTRE II: une méthode de classement en présence de critères multiples* (1971)
3. Mareschal, B., Brans, J.P., Vincke, P.: *PROMETHEE: A New Family of Outranking Methods in Multicriteria Analysis*. Université Libre de Bruxelles, ULB (1984)
4. Opricovic, S.: *Multicriteria optimization of civil engineering systems*. Fac. Civ. Eng., Belgrade (1998)
5. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
6. Bellman, R.E., Zadeh, L.A.: Decision-making in a fuzzy environment. *Manag. Sci.* **17**, 141–161 (1970)
7. Yager, R.R.: Fuzzy decision making including unequal objectives. *Fuzzy Sets Syst.* **1**, 87–95 (1978)
8. Nakamura, K.: Preference relations on a set of fuzzy utilities as a basis for decision making. *Fuzzy Sets Syst.* **20**, 147–162 (1986)
9. Atanassov, K.T.: Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **20**, 87–96 (1986)
10. Xu, Z.S., Yager, R.R.: Dynamic intuitionistic fuzzy multi-attribute decision making. *Int. J. Approx. Reason.* **48**, 246–262 (2008)
11. Liu, H.W., Wang, G.J.: Multi-criteria decision-making methods based on intuitionistic fuzzy sets. *Eur. J. Oper. Res.* **179**, 220–233 (2007)
12. Boran, F.E., Genc, S., Kurt, M.: A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method. *Exp. Syst. Appl.* **36**, 11363–11368 (2009)
13. Yager, R.R.: Pythagorean fuzzy subsets. In: *Proceeding of the Joint IFSA World Congress and NAFIPS Annual Meeting*, Edmonton, Canada, pp. 57–61 (2013)
14. Zeng, S., Chen, J., Li, X.: A hybrid method for Pythagorean fuzzy multiple-criteria decision making. *Int. J. Inf. Technol. Decis. Mak.* **14**, 1–16 (2015). <https://doi.org/10.1142/S0219622016500012>

Chapter 25

Topological Properties of Flat Electroencephalography



**Tan Lit Ken, Tahir Ahmad, Nor Azwadi Che Sidik,
Chuan Zun Liang, Lee Kee Quen, Gan Yee Siang, Goh Chien Yong
and Tey Wah Yen**

Abstract Electroencephalograph is one of the useful and favored instruments in diagnosing various brain disorders especially in epilepsy due to its non-invasive characteristic and ability in providing wealthy information about brain functions. While epileptic foci localization is possible with the aid of EEG signals, it relies greatly on the ability to ex-tract hidden information or pattern within electroencephalography signals. Flat electroencephalography being an enhancement of

T. L. Ken (✉) · N. A. C. Sidik · L. K. Quen

Department of Mechanical Precision Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
e-mail: tlken@utm.my

N. A. C. Sidik

e-mail: azwadi@utm.my

L. K. Quen

e-mail: lkquen@utm.my

T. Ahmad

Department of Mathematical Sciences and Centre for Sustainable Nanomaterials, Universiti Teknologi Malaysia, Johor Darul Takzim, 81310 Skudai, Malaysia
e-mail: tahir@ibnusina.utm.my

C. Z. Liang

Science Programme, Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang Darul Makmur, Malaysia
e-mail: chuan@utm.edu.my

G. Y. Siang · G. C. Yong

Department of Mathematics, Xiamen Universiti Malaysia, 43900 Sepang, Selangor Darul Ehsan, Malaysia
e-mail: ysgan@xmu.edu.my

G. C. Yong

e-mail: gcyong@xmu.edu.my

T. W. Yen

Faculty of Engineering, Technology & Built Environment, UCSI University, North Wing Campus, 56000 Cheras Kuala Lumpur, Malaysia
e-mail: teywy@ucsiuniversiti.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_25

electroencephalography carries affluent information about seizure process. In the perspective of topological dynamical systems, epileptic seizure and Flat EEG are two equivalent object, hence, findings attained from Flat EEG can be implied on epileptic seizure. In other words, Flat EEG serves as a great alternative platform to study epileptic seizure. Although there exists various researches on Flat EEG utilizing various mathematical models, topological study on its states connectivity has yet to exist. Since both events of epileptic seizure and Flat EEG are continuous processes, topological studies on its states connectivity can provides great in-sight into seizure process. In this paper, structures of the events will be modelled and explored topologically. In addition, the extracted topological properties will also be interpreted physically. Based on the theorems derived, Flat EEG is found to be well-behaved from topological viewpoint.

Keywords Topology · Flat Electroencephalograph · Epilepsy

1 Introduction

Epilepsy is a brain disorder in which the sufferer experience recurrent seizures. It usually results in unusual behavior, sensations and even loss of consciousness. Formally, it is defined by the Commission on Epidemiology and Prognosis and International League Against Epilepsy 1993 as “the occurrence of at least two unprovoked seizure” [1]. Epilepsy is usually happen without any warning, unprompted and occur more than once. Although it can occur to anyone at any age regardless of gender, it is not contagious i.e. it is not inherited. Statistically, the number of people with epilepsy in the world is about 1.5% of the world populations [2].

EEG has been used extensively in diagnosing, classifying the type of seizure occurring and locating the source of electrical activity [3]. It is also one of the most significant laboratory tests in detecting epilepsies [4–6]. Its wide acceptance is because EEG allows neurologists to examine and pinpoint damaged brain tissue and also to make preparation prior to surgery to circumvent or decrease the risk of injury on important parts of the brain. Recently, gaining the graphic electrical activity within the brain has in general become a essential part of surgical [7].

2 Literature Review

Based on the literature surveyed, there exist various approaches and procedures to localize epileptic foci with each having their own advantages and weaknesses. For examples, presurgical method via multimodality neuroimaging [8], by using large-area magnetometer and functional brain anatomy [9], exploring the signal structure by studying the correlations between electrodes captured by linear,

nonlinear and multi linear data analysis using kernel functions [10], 3-D source localization of epileptic foci by combining EEG and MRI data [11] and statistical approaches such as Bayesian technique [12] and maximum likelihood estimation (MLE) method by Jan et al. [13].

Epilepsy can be studied more effectively via a novel technique called Flat electroencephalography (Flat EEG). This approach enable EEG signals to be viewed on the first component of Fuzzy Topographic Topological Mapping (FTTM). An immediate consequence from this is that, by FTTM model, EEG signals can be depicted in 3-dimension space. There exist various research utilizing different mathematical concepts to visualize and extract “hidden” information within EEG signals via Flat EEG. Most of these researches focuses on a particular setting, in this context known as time. As an example, with the application of Fuzzy C-Means (FCM) on Flat EEG, one can compute the cluster centers of electrical activities within the brain during seizure. This made brainstorm tracking during epileptic seizure to be possible [14].

Besides that, an algebraic study on Flat EEG has shown that it is possible to transform Flat EEG from topological structure to algebraic structure. It also shows that Flat EEG can be decomposed into a semigroup of upper triangular matrices under matrix multiplication. Hence, signifying epileptic seizure is non-chaotic [15].

Also, the topological dynamic study on the structure of Flat EEG, proves that Flat EEG in the presence of artifacts can still offer important descriptions of electrical goings-on in the brain during seizure attack [16]. All this signifies that Flat EEG is a worthy alternative platform to study epileptic seizure.

The event of epileptic seizure and Flat EEG are continuous processes which is apparent from the embedment of time parameter. In fact, it had been modelled as dynamical systems [17]. However, to date, the connectivity of the states in the events has not been studied. Thusly, in this study, the events will be studied topologically, anticipating several properties of the events to be exposed.

3 Materials and Method

Consider the flow of Flat EEG modelled in [17]:

Flow of Flat EEG

The flow is $\psi_t(y)$ where $\psi_t(y) : \mathfrak{R} \times Y \rightarrow Y$ and such that the following two properties are fulfilled:

- i. $\psi_0(y) = y \quad \forall y \in Y = \mathfrak{R}^n$, and
- ii. for all t and $s \in \mathfrak{R}$

$$\psi_t \circ \psi_s = \psi_{t+s}$$

Here, for any $y_k \in Y = \mathfrak{R}^n$, $\psi_t(y)$ is generally defined as $\psi_t(y_k) = y_i$ i.e., the state of the system which initiate from y_k at time i is y_i .

Basically, for each $y_k \in Y = \mathfrak{R}^n$, the flow of Flat EEG defines an event of Flat EEG (EoFE).

In general, an EoFE can be pictured geometrically as a trajectory in the state space. This line is stretched on both directions and without endpoint. Undoubtedly, such distinctive property are also possessed by the real number line, \mathfrak{R} . In actual fact, an event can be viewed as the topological deformation of real number line, \mathfrak{R} . In point of fact, an event being topologically equivalent to \mathfrak{R} is apparent from the point that each states is a unique consequence of a time, t . For that reason, there exists a homeomorphism which maps from the trajectory (with order topology) to the real number line, \mathfrak{R} (with order topology). Basically, the homeomorphism is

$$h : O_{\psi_t(y_k)} \rightarrow \mathfrak{R}$$

and can be defined as

$$h(y_i) = i \text{ where } i \in \mathfrak{R}$$

where $O_{\psi_t(y_k)}$ is an EoFE.

Theorem 1 *The function $h : O_{\psi_t(y_k)} \rightarrow \mathfrak{R}$ defined as $h(y_i) = i$ where $i \in \mathfrak{R}$ is a homeomorphism.*

Proof The proof is trivial. Notice that each of the elements in the set $O_{\psi_t(y_k)}$ is a unique consequence of a time, t . Therefore, the function h can be seen as an identity function. Identity function is bijective and bicontinuous, hence it is a homeomorphism. For details, refer to the proof of theorem 7 in [17].

Before going any further, we introduce the following notations

- $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ - An EoFE with order topology
- (\mathfrak{R}, τ_S) - Real number line with standard topology
- (\mathfrak{R}, τ_O) - Real number line with order topology

Corollary 1 $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is homeomorphic to (\mathfrak{R}, τ_O) .

Proof By Theorem 1.

It is well known that the order topology (or interval topology) on \mathfrak{R} corresponds with the standard topology (topology induced by the usual metric) on \mathfrak{R} . Equivalently, they comprise of the same set of open sets though the open sets are generated differently. Consequently, all the topological properties on \mathfrak{R} equipped with the standard metric are also possessed by \mathfrak{R} when endowed with order topology because topological properties are defined via open sets. Metrizable is one of the various topological properties and the standard topology on \mathfrak{R} is an example of metrizable space. This is clear because the topology is generated by the

standard distance metric on \mathfrak{R} . Since, the standard topology on \mathfrak{R} corresponds with the order topology on \mathfrak{R} , the order topology on \mathfrak{R} is then metrizable space as well.

Proposition 1 ([18]) τ_S and τ_O on \mathfrak{R} coincides (homeomorphic).

Corollary 2 (\mathfrak{R}, τ_O) is metrizable.

Proof Note that (\mathfrak{R}, τ_S) is metrizable. Since τ_S and τ_O on \mathfrak{R} coincides (Proposition 1). Thus, (\mathfrak{R}, τ_O) is also metrizable.

Since an event (with order topology) is topologically same to real number line, \mathfrak{R} (with order topology), thus it is also topologically equivalent to a real number line, \mathfrak{R} (with standard topology). As a result, topological properties on \mathfrak{R} (with standard topology) are preserved to the event with order topology.

Corollary 3 $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is homeomorphic to (\mathfrak{R}, τ_S) .

Proof By Corollary 1 and Proposition 1.

An immediate implication of the metrizability of (\mathfrak{R}, τ_S) and Corollary 3 is $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ too become metrizable. Thusly, an event with the order topology is also a metrizable space.

Corollary 4 $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is metrizable.

Proof Since (\mathfrak{R}, τ_S) is metrizable and (\mathfrak{R}, τ_S) is topologically equivalent to $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ (Corollary 3).

Being metrizable, it shows that there exist a distance between any two states during the events. As time is the quantity that contributes to the metrizability, the quantification then gives the meaning of temporal distance between the states. That is, any two states can be assigned to a real number which represents the temporal distance between the two moments.

Corollary 5 $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is Hausdorff.

Proof Every metrizable space is Hausdorff [18]. Since $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is metrizable (Corollary 4), thus it is Hausdorff.

Corollary 6 $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is normal.

Proof Every metrizable space is normal [18]. Since $(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}})$ is metrizable (Corollary 4), thus it is normal.

On the other hand, being Hausdorff implies that, any two states are distinguishable topologically no matter how closely and similar they are. For instance, two Flat EEG frames which may have equivalent electrical potentials (and hence number of cluster centers) are distinguishable topologically (due to the Hausdorff

property). Furthermore, this holds not only for any two states in the events, but also for any disjoint closed interval of states because the events are normal spaces.

Corollary 7 $\left(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}}\right)$ is connected.

Proof Since $\left(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}}\right)$ is homeomorphic to (\mathfrak{R}, τ_S) , whereby (\mathfrak{R}, τ_S) is a connected space [18].

EEG signals are often tagged with segments of similar traits (e.g. frequency by clinicians for the assessment by neurophysiologists [3]. For instance, an EEG signal of an epileptic patient can be divided into three major segments called preictal, ictal and postictal. However, according to Asano et al. [19], the definition of ictal onset zones is frequently a subjective matter and varies among electroencephalographers.

Since the event of Flat EEG is connected, topologically, Corollary 7 confirms Asano et al. [19] claims. The reason is because in a connected space, any two sub events which union forms the whole event must intersect somewhere. It is the existence of this intersection that makes it subjective to determine the border line for segmentation.

Corollary 8 $\left(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}}\right)$ is separable.

Proof Since $\left(O_{\psi_t(y_k)}, \tau_{O_{\psi_t(y_k)}}\right)$ is homeomorphic to (\mathfrak{R}, τ_S) , whereby (\mathfrak{R}, τ_S) is a separable space [20].

This shows that the events are separable. Topologically, it implies that there exists a countable dense subset within the events i.e., a countable subset which closure is the events. Physically, it means that Flat EEG can be sampled at least countably infinite number of times because for any two states within the events, one can always find another state (from that countable dense subset). However, since there also exists an uncountable dense subset (formed by the set of irrational numbers), it shows that Flat EEG can be sampled infinite number of times because for any two states within the events, one can always find another state (from that uncountable dense subset).

Equivalently, it also means that for any two states in the events, there are infinitely many other states in between them. In short Corollary 8, says that Flat EEG can be sampled at any time. To date, Flat EEG can only be sampled (or digitized) at milliseconds. Such limitations is due to the available technology i.e., limited temporal resolution. However, Corollary 8 illustrates that in reality, infinitely many Flat EEG frames can be sample in between any two moments and it is just a matter of time to produce such technology. In general, this characteristic can also be realized from the completeness property of the real numbers.

Basically, any topologically properties of real number line with standard topology were also topological properties of event with order topology. Others topological properties are like first countable, second countable, Lindelof and etc.

4 Results and Discussions

From topological viewpoint, the connectivity of states in an EoFE are equivalent to the connectivity of real numbers in \mathfrak{R} . Consequently, they have equivalent topological properties. In other words, EoFE can be view as the real number line topologically. Topological properties established on an EoFE can be implied on an event of epileptic seizure (EoES) since their respective flows are topologically conjugated. Also, these events are well-behaved from topological viewpoint since it exhibit one of the significant topological properties namely, metrizability.

5 Conclusions

Albeit EoES and EoFE may seem complex and dissimilar, they share many equivalent qualitative properties. In particular, they have equivalent topological properties. Generally, this study demonstrates not only that topological properties established on EoFE can be implied on EoES but topological properties of the real number line \mathfrak{R} can be implied on EoES and EoFE as well.

Acknowledgement The authors would like to thank their family members and friends, for their continuous support and assistance. The authors would also like to express his appreciation to Universiti Teknologi Malaysia. This research is supported by the university GUP Tier 1 grant Vot No. 15H42.

References

1. Panayiotopoulos, C.P.: Atlas of Epilepsies. Springer, London (1992)
2. Jahnecke, C.A.N., Schwarz, L., Sovierzoski, M.A., Azevedo, D.F.M., Argoud, F.I.M.: C++ Video-EEG processing system with sights to the epileptic seizure detection. In: World Congress on Medical Physics and Biomedical Engineering. IFMBE Proceedings, pp 1052–1055 (2007)
3. Sanei, S., Chambers, J.: EEG Signals Processing. Wiley, England (2007)
4. Popp, A.J., Deshaies, E.M.: A Guide to the Primary Care of Neurological Disorders. American Associations of Neurosurgeons, Thieme (2007)
5. Yudofsky, S.C., Hales, R.E.: The American Psychiatric Publishing Textbook of Neuropsychiatry and Behavioral Neurosciences. American Psychiatric Publishing, USA (2008)
6. Gilhus, N.E., Barnes, M.R., Brainin, M.: European Handbook of Neurological Management. Wiley, England (2011)
7. Miller, J.W., Cole, A.J.: Is it necessary to define the ictal onset zone with EEG prior to performing respective epilepsy surgery? Elsevier Epilepsy Behav. **20**(2), 178–181 (2011)
8. Desco, M., Pascau, J., Pozo, M.A., Santos, A., Reig, S., Gispert, J., Garcia, B.P.: Multimodality localization of epileptic foci. In: Proceedings of SPIE, Medical Imaging Physiology and Function from Multidimensional Images, pp 362–370 (2001)

9. Tiihonen, J., Hari, R., Kjolha, M., Nousiainen, U., Vapalahti, M.: Localization of epileptic foci using large-area magnetometer and functional brain anatomy. *Ann. Neurol.* **27**(3), 283–290 (2004)
10. Evim, A., Canan, A.B., Haluk, B., Bulent, Y.: Computational analysis of epileptic focus localization. In: ACTA Press Anaheim, BioMed 2006 Proceedings of the 24th IASTED International Conference on Biomedical Engineering, pp 317–322 (2006)
11. Natasa, M., Malek, D., Ilker, Y., Prasanna, J.: 3-d source localization of epileptic foci integrating EEG and MRI data. *Brain Topogr.* **16**(2), 111–119 (2003)
12. Toni, A., Aopa, N., Matti, S.H., Liro, P.J., Jouko, L., Aki, V., Mikko, S.: Bayesian analysis of the neuromagnetic inverse problem with l^p -norm priors. *Neuroimage* **26**(3), 870–884 (2005)
13. Jan, C.D.M., Fetsje, B., Pawel, G., Cezary, A.S., Maria, I.B., Heethaar, R.M.: A maximum-likelihood estimator for trial-to-trial variations in noisy MEG/EEG data sets. *IEEE Trans. Biomed. Eng.* **51**(2), 2123–2138 (2004)
14. Fauziah, Z.: Dynamic profiling of electroencephalography data during seizure using fuzzy information space. Ph.D. Thesis, Universiti Teknologi Malaysia (2008)
15. Faisal, A.M.B., Tahir, A.: EEG signals during epileptic seizure as a semigroup of upper triangular matrices. *Am. J. Appl. Sci.* **7**(4), 540–544 (2010)
16. Tan, L.K., Tahir, A.: Structural stability of flat electroencephalography. *Life Sci. J.* **11**(8), 165–170 (2014)
17. Tahir, A., Tan, L.K.: Topological Conjugacy between seizure and flat electroencephalography. *Sci. Publ. Am. J. Appl. Sci.* **7**(3), 1470–1476 (2010)
18. Munkres, J.R.: *Topology*. Pearson Prentice Halls, USA (2000)
19. Asano, E., Otto, M., Aashit, S., Csaba, J., Diane, C.C., Jean, G., Harry, T.C.: Qualitative visualization of ictal subdural EEG changes in children with neocortical focal seizures. *Clin. Neurophysiol.* **115**(12), 2718–2727 (2004)
20. Davis, S.W.: *Topology*. McGraw-Hill, Singapore (2005)

Chapter 26

Two-Phase Mixed Convection Flow of Dusty Williamson Fluid with Aligned Magnetic Field over a Vertical Stretching Sheet



Nur Syamilah Arifin, Syazwani Mohd Zokri,
Abdul Rahman Mohd Kasim, Mohd Zuki Salleh
and Nurul Farahain Mohammad

Abstract The mixed convection flow of dusty Williamson fluid over a vertical stretching sheet with the influence of aligned magnetic field is investigated. The mathematical model on two-phase flows of dust particles embedded in Williamson fluid has been considered under Newtonian heating boundary condition by initially applying the similarity transformation to its governing equations. The transformed ordinary differential equations are solved numerically using Runge-Kutta Fehlberg (RKF45) method. Several pertinent parameters such as aligned magnetic field, Williamson parameter, mixed convection parameter, fluid particle interaction parameter, Prandtl number and conjugate parameter on the flow and heat transfer are visualized in graphical form. The results revealed that the fluid particle interaction parameter influencing the fluid velocity which resulted to decrease the fluid motion. The two-phase fluid flow model presented herein can be transformed into classical problem of single phase fluid flow under the condition of the fluid particle interaction parameter is neglected. Therefore, the present mathematical model can be offered as the generalized model of complex fluid with suspended particles

N. S. Arifin (✉) · S. M. Zokri · A. R. M. Kasim · M. Z. Salleh
Applied & Industrial Mathematics Research Group, Faculty of Industrial Sciences &
Technology, Universiti Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia
e-mail: nursyamilarifin@gmail.com

S. M. Zokri
e-mail: Ssyazwanizokri@gmail.com

A. R. M. Kasim
e-mail: rahmanmohd@ump.edu.my

M. Z. Salleh
e-mail: zuki@ump.edu.my

N. F. Mohammad
Department of Computational and Theoretical Sciences, Kulliyah of Science,
International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia
e-mail: farahain@iium.edu.my

Keywords Mixed convection • Dusty Williamson fluid • Stretching sheet • Aligned magnetic field • Newtonian heating

1 Introduction

The shear thinning (pseudoplastic) fluid is regards to those fluids that viscosity reduc-es with the increasing rate of shear stress. Such fluids are polymer solution, ketchup, whipped cream and paint which provide another non-Newtonian fluid model. One of the models that have been given little discussion about is Williamson fluid where chyme in small intestines and blood are the closed fit to portray its behavior, thus become fundamental in investigating the physiological fluid. A model of Williamson has been discovered by Williamson [1]. Several literary respected to this circumstance has been published [2–4]. Previous research has been mostly restricted on fluid (single phase) flow [5, 6]. However, the two-phase flow of which the mixture of dust particle into fluid is an interesting flow model that needs to explore more. There have been few studies into dusty non-Newtonian fluid by considering various aspects [7–9]. The main purpose of this study is to propose a two-phase model of dusty Williamson fluid for mixed convection flow with Newtonian heating (NH).

2 Mathematical Formulation

The steady, two dimensional incompressible flow of dusty Williamson fluid is con-sidered over a stretching sheet with linear velocity $u_w(x) = ax$. An acute angle α_1 with magnetic field is applied to the flow as shown in Fig. 1. Supposed, the dust particles are in spherical shape, uniform size and number density are taken as constant throughout the flow. The governing boundary layer equations for two-phase flow [10, 11] can be written as follow

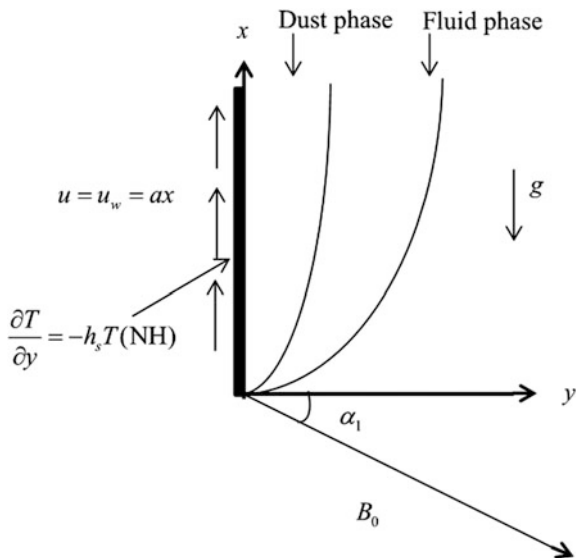
Fluid phase:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad (1)$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = v \frac{\partial^2 u}{\partial y^2} + \sqrt{2}v\Gamma \frac{\partial u}{\partial y} \frac{\partial^2 u}{\partial y^2} + \frac{\rho_p}{\rho\tau_v} (u_p - u) - \frac{\sigma u B_0^2}{\rho} \sin^2 \alpha_1 + g\beta^* (T - T_\infty), \quad (2)$$

$$\rho c_p \left(u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} \right) = k \left(\frac{\partial^2 T}{\partial y^2} \right) + \frac{\rho_p c_s}{\gamma_T} (T_p - T), \quad (3)$$

Fig. 1 Flow configuration



Dust phase:

$$\frac{\partial u_p}{\partial x} + \frac{\partial v_p}{\partial y} = 0, \quad (4)$$

$$\rho_p \left(u_p \frac{\partial u_p}{\partial x} + v_p \frac{\partial u_p}{\partial y} \right) = \frac{\rho_p}{\tau_v} (u - u_p), \quad (5)$$

$$\rho_p c_s \left(u_p \frac{\partial T_p}{\partial x} + v_p \frac{\partial T_p}{\partial y} \right) = -\frac{\rho_p c_s}{\gamma_T} (T_p - T) \quad (6)$$

where (u, v) and (u_p, v_p) are the velocities components of the fluid and particle phases along x and y axes, respectively. μ is the coefficient of fluid's viscosity, ρ and ρ_p are the density of fluid and dust phases, $\tau_v = 1/k$ is the relaxation time of particles phase, k is the Stoke's resistance, c_p and c_s are specific heat of fluid and dust particle, T and T_p are the temperature of fluid and particle phases, γ_T is the thermal relaxation time, B_0 is the magnetic-field strength, $\Gamma > 0$ is the time constant, g is the gravity acceleration and β^* is the thermal expansion coefficient.

The proposed equations are respect to the following boundary conditions

$$u = u_w(x) = ax, v = 0, \frac{\partial T}{\partial y} = -h_s T \text{ at } y = 0$$

$$u \rightarrow 0, u_p \rightarrow 0, v_p \rightarrow v, T \rightarrow T_\infty, T_p \rightarrow T_\infty \text{ at } y \rightarrow \infty \quad (7)$$

where a is positive constant and h_s is heat transfer parameter. The following suitable similarity transformation is introduced to the Eqs. (1)–(7)

$$u = axf'(\eta), v = -(av)^{1/2}f(\eta), \eta = (a/v)^{1/2}y, \theta(\eta) = (T - T_\infty)/T_\infty$$

$$u_p = axF'(\eta), v_p = (av)^{1/2}F(\eta), \theta_p(\eta) = (T_p - T_\infty)/T, \tag{8}$$

where ψ is the stream function defined as $u = \partial\psi/\partial y$ and $v = -\partial\psi/\partial x$. Equations (1)–(6) are now transformed to ordinary differential equations as follow

$$f''' + ff'' - f'^2 + \lambda_1 f'' f''' + \beta N(F' - f') - M \sin^2 \alpha_1 f' + \lambda \theta = 0, \tag{9}$$

$$\theta'' + \text{Pr} f \theta' + \frac{2}{3} \beta N(\theta_p - \theta) = 0, \tag{10}$$

$$F'^2 - FF'''' + \beta(F' - f') = 0, \tag{11}$$

$$\theta'_p F + \frac{2}{3} \frac{\beta}{\text{Pr} \gamma} (\theta - \theta_p) = 0 \tag{12}$$

and the boundary conditions (7) are reduced to

$$f(0) = 0, f'(0) = 1, \theta'(0) = -b(1 + \theta(0)) \text{ at } \eta = 0$$

$$f'(\eta) \rightarrow 0, F(\eta) \rightarrow 0, F'(\eta) \rightarrow f(\eta), \theta(\eta) \rightarrow 0, \theta_p(\eta) \rightarrow 0 \text{ at } \eta \rightarrow \infty \tag{13}$$

where a prime ($'$) denotes the differentiation with respect to η . N is the mass concentration of particle phase, M is the magnetic field parameter, β is the fluid-particle interaction parameter, Pr is the Prandtl number, γ is the specific heat ratio of mixture, b is the conjugate parameter for NH, λ_1 is the Williamson parameter and λ is the buoyancy parameter with Gr_x is the Grashof number and Re_x is the Reynolds number. The parameters can be denoted as

$$N = \rho_p/\rho, M = \sigma B_0^2/\rho a, \beta = 1/a\tau_v, \text{Pr} = \mu c_p/k, \gamma = c_s/c_p, b = -h_s(v/a)^{1/2},$$

$$\lambda_1 = \sqrt{2a^3/\nu} \Gamma x, \lambda = Gr_x/\text{Re}_x^2, Gr_x = g\beta * T_\infty x^3/\nu^2, \text{Re}_x = u_w(x)x/\nu$$

3 Results and Discussion

The Eqs. (9)–(13) are solved using RKF45 method in Maple software as the method is stable, easy to implement and having self-starting condition. The findings for the two phase flow of dusty Williamson fluid are visualized graphically in which

Table 1 Comparison of $\theta(0)$ when $M = \lambda_1 = \lambda = \beta = N = 0, \gamma \rightarrow \infty$ and $b = 1$

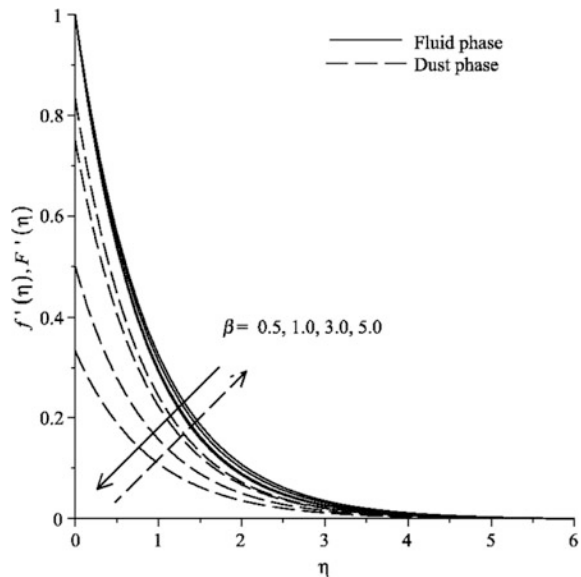
Pr	Salleh et al. [12]	Turkyilmazoglu [13]	Present
3	6.02577	6.05158546	6.05158577
5	1.76594	1.76039543	1.76039543
7	1.13511	1.11681524	1.11681523
10	0.76531	0.76452369	0.76452369
100	0.16115	0.14780542	0.14780574

the maximum boundary layer thickness for velocity and temperature profiles are $\eta_\infty = 6$ as its sufficient to obtain the boundary conditions asymptotically.

The pertinent values of $\alpha_1 = \pi/6, M = 0.2, \beta = N = \lambda = 0.5, b = \lambda_1 = \gamma = 0.1$ and $Pr = 10$ are fixed throughout the study for the computational purpose. The values of $\theta(0)$ between the previous works and present result are tabulated in Table 1. It is observed that a close agreement is attained and for that the present numerical results presented here is validated.

Figures 2 and 3 show the influences of β and λ_1 on the velocity profile. For dust (fluid) phase, the velocity profile decreases with an increase (decrease) in β as revealed in Fig. 2. It could be that, a large β attributed to the development of drag force which acts opposite to the fluid flow. The value of fluid velocity is observed to decrease about 2.56% while this value increase around 8.14% for dust phase when $\beta = 5$ at $\eta = 1.5$. The increasing of λ_1 is to boost the relaxation time of fluid and

Fig. 2 Effect of β on velocity profile



causes the enhancement of viscosity which observed decrease in fluid velocity as revealed in Fig. 3. Also, the dust particle velocity decreases with an increase in λ_1 .

Figure 4 displays the influences of α_1 with M on the magnitude of skin friction coefficient. Physically, the increment of both parameters generates the Lorentz force

Fig. 3 Effect of λ_1 on velocity profile

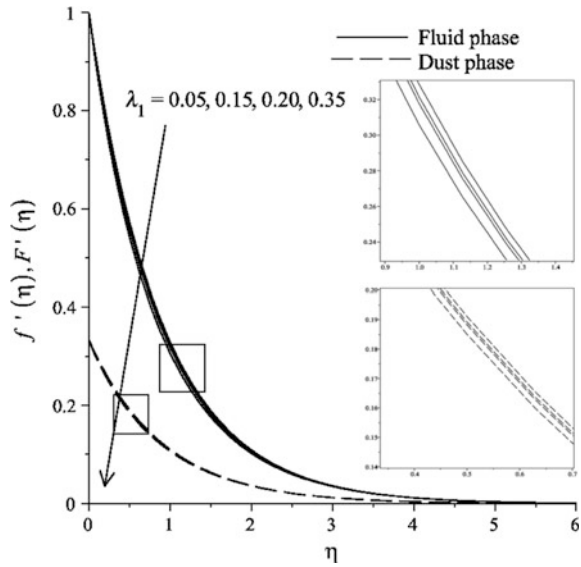


Fig. 4 Variation of α_1 on skin friction coefficient with M

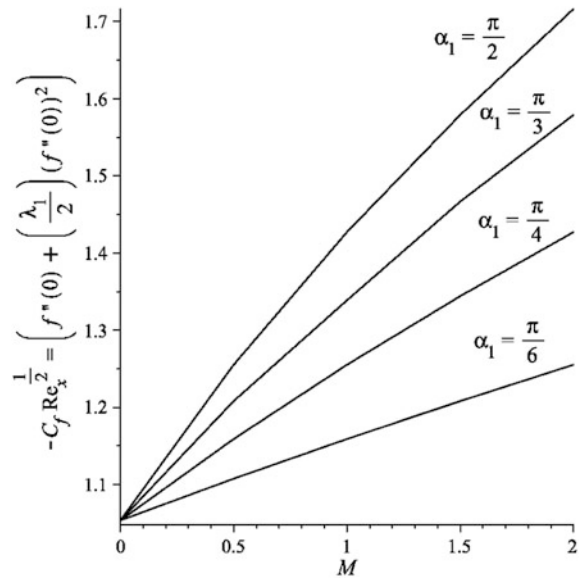
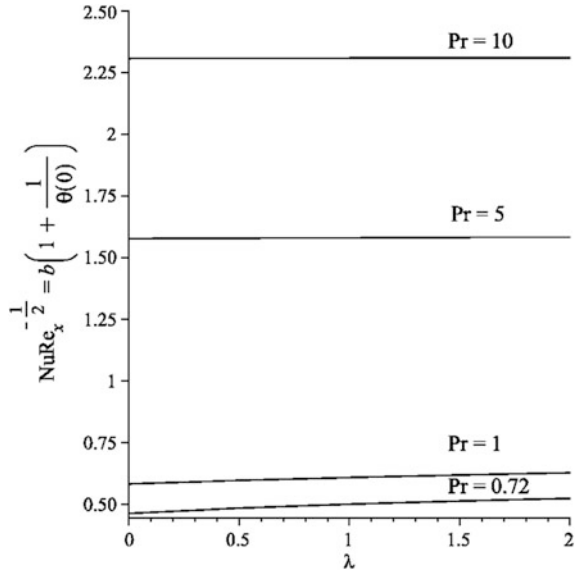


Fig. 5 Variation of b on Nusselt number with λ



which suppresses the fluid velocity and thus increases the magnitude value. Figure 5 shows the influence of Pr with λ on Nusselt number. The increasing of λ magnified the effect of buoyancy forces that implied to accelerate the fluid flow. At the same time, the viscous diffusivity is more pronounced for higher Pr and fluid absorbs more heat. Based on these facts, the surface temperature of fluid is reduced and thus the increasing trend of Nusselt number is expected.

4 Conclusion

In the present study, a dusty Williamson fluid on mixed convection flow is formulated as two-phase model and solved using RKF45 method. The study suggests that the dust particle influencing the Williamson fluid flow which resulted to decrease the velocity of fluid. Other than that, the two-phase flow model could be used to predict the contaminated fluid as in industrial flow process.

Acknowledgements The authors gratefully acknowledge the financial support received from Universiti Malaysia Pahang for (PGRS170397, RDU 160330, & RDU 170328).

References

1. Williamson, R.V.: The flow of pseudoplastic materials. *Ind. Eng. Chem.* **21**(11), 1108–1111 (1929). <https://doi.org/10.1021/ie50239a035>
2. Nadeem, S., Hussain, S.T., Lee, C.: Flow of a Williamson fluid over a stretching sheet. *Braz J. Chem. Eng.* **30**(3), 619–625 (2013). <https://doi.org/10.1590/S0104-66322013000300019>
3. Hayat, T., Shafiq, A., Farooq, M.A., Alsulami, H.H., Shehzad, S.A.: Newtonian and Joule Heating effects in two-dimensional flow of Williamson fluid. *J. Appl. Fluid Mech.* **9**(4) (2016). <https://doi.org/10.18869/acadpub.jafm.68.235.24646>
4. Nadeem, S., Hussain, S.T.: Flow and heat transfer analysis of Williamson nanofluid. *Appl. Nanosci.* **4**(8), 1005–1012 (2014). <https://doi.org/10.1007/s13204-013-0282-1>
5. Aurangzaib, Kasim, A.R.M., Mohammad, N.F., Shafie, S.: Effect of thermal stratification on MHD free convection with heat and mass transfer over an unsteady stretching surface with heat source, Hall current and chemical reaction. *Int. J. Adv. Eng. Sci. Appl. Math.* **4**(3), 217–225 (2012). <https://doi.org/10.1007/s12572-012-0066-y>
6. Aurangzaib, Kasim, A.R.M., Mohammad, N.F., Sharidan, S.: Unsteady MHD mixed convection flow with heat and mass transfer over a vertical plate in a micropolar fluid-saturated porous medium. *J. Appl. Sci. Eng.* **16**(2), 141–150 (2013). <https://doi.org/10.6180/jase.2013.16.2.05>
7. Naramgari, S., Sulochana, C.: MHD flow of dusty nanofluid over a stretching surface with volume fraction of dust particles. *Ain. Shams Eng. J.* **7**(2), 709–716 (2016). <https://doi.org/10.1016/j.asej.2015.05.015>
8. Bhatti, M.M., Zeeshan, A.: Analytic study of heat transfer with variable viscosity on solid particle motion in dusty Jeffery fluid. *Mod. Phys. Lett. B* **30**(16), 1650196 (2016). <https://doi.org/10.1142/S0217984916501967>
9. Kumar, K.G., Gireesha, B.J., Rudraswamy, N.G., Manjunatha, S.: Non linear thermal radiation effect on williamson fluid with particle-Liquid suspension past a stretching surface. *Results Phys.* (2017). <https://doi.org/10.1016/j.rinp.2017.08.027>
10. Siddiqua, S., Hossain, M.A., Saha, S.C.: Two-phase natural convection flow of a dusty fluid. *Int. J. Numer. Methods Heat Fluid Flow* **25**(7), 1542–1556 (2015). <https://doi.org/10.1108/HFF-09-2014-0278>
11. Arifin, N.S., Zokri, S.M., Kasim, A.R.M., Salleh, M.Z., Mohammad, N.F., Yusoff, W.N.S. W.: Aligned magnetic field on dusty Casson fluid over a stretching sheet with Newtonian heating. *Mal. J. Fund. Appl.* **13**(3), 244–247 (2017). <https://doi.org/10.11113/mjfas.v13n3.592>
12. Salleh, M.Z., Nazar, R., Pop, I.: Boundary layer flow and heat transfer over a stretching sheet with Newtonian heating. *J. Taiwan Inst. Chem. Eng.* **41**(6), 651–655 (2010). <https://doi.org/10.1016/j.jtice.2010.01.013>
13. Turkyilmazoglu, M.: Flow of a micropolar fluid due to a porous stretching sheet and heat transfer. *Int. J. Non-Linear Mech.* **83**, 59–64 (2016). <https://doi.org/10.1016/j.ijnonlinmec.2016.04.004>

Chapter 27

Unveiling the Asymmetric Adjustments of Policy Reaction Function in Indonesia



Lavaneesvari Manogaran and Siok Kun Sek

Abstract The standard Taylor rule assumes that monetary policy can be represented by a linear reaction function. This study extended previous research to investigate if there is any asymmetric adjustment in the policy reaction function of Indonesia as this country experienced a drastic shift in the monetary policy regime and exchange rate system from the impact of Asian financial crisis. To be more specific, we tend to reveal the behaviour of policy function towards variation in inflation and output gap over time changes. In the same way, we also seek to capture if there is any influence of exchange rate changes on policy function in two contrasting eras. Therefore, a nonlinear regression model (threshold approach) is applied to estimate the policy reaction function of Indonesia in two different sub-periods; pre- (1980Q1–1996Q4) and post- (2000Q1–2015Q4) inflation targeting (IT) regime. Our analysis with two threshold factors: exchange rate (*LREER*) for pre-IT and inflation (*LCPI*) for post IT, addressed asymmetric adjustments of policy rate towards inflation variation and output gap in both sub-periods. In the pre-IT, the policy rate is only responding to output gap. Turning to post-IT, the policy rate is reacting more actively to inflation variation than output gap but never to changes in the exchange rate. This validating a pure floating regime accompanied by inflation targeting policy framework aftermath crisis in Indonesia.

Keywords Taylor rule · Threshold approach · Asymmetric adjustments

L. Manogaran (✉) · S. K. Sek
School of Mathematical Sciences, Universiti Sains Malaysia, Gelugor, Penang, Malaysia
e-mail: resshma06@yahoo.com

S. K. Sek
e-mail: sksek@usm.my

© Springer Nature Singapore Pte Ltd. 2019
L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_27

217

1 Introduction

Two decades ago, studies on monetary policy framework are merely represented in a linear algebraic form of policy reaction function on the famous work of Taylor [1] called simple Taylor rule. This policy equation setting is illustrated with an interest rate (policy rate) as an instrument which mechanically adjusted to cater inflation variation and output gap to achieve price and output stability [2]. However, in recent years, there are increasing numbers of studies claiming the necessity to model the policy rule using a nonlinear policy function. Such nonlinearity setting captures different policy reactions (if any) due to sudden shocks or changes in the economy. This is especially for emerging economies which prone to external influences coexist with weak and financially unstable system [3]. In such most of them experienced large fluctuations in the financial market performance due to sudden shocks in the economy, where the monetary policy is adjusted to accommodate the fluctuations accordingly [4].

In this paper, we focus the study in Indonesia by considering this country as one of the emerging economies that experienced a downturn in the economy due to financial crisis (1997–98). Correspondingly, the monetary policy regime was also transformed to a flexible/floating exchange rate from a fixed/rigid system. Additionally, the implementation of a floating regime has been accompanied with an inflation targeting policy as the old policy regime failed to prevent over fluctuations in the economy from the impact of the crisis. Thus, the main purpose of this study is to inquire if there exist any asymmetric form of adjustment in the monetary policy of Indonesia under two different policy regimes with the presence of acute policy change and economic downturn from the effect of financial turmoil. At the same time, we also seek to unveil the behaviour of monetary policy in response to inflation and output gaps as well as exchange rate changes.

The threshold regression reveals some interesting findings. Firstly, the policy function of Indonesia is determined by two threshold factors: *LREER* for the pre-IT period and *LCPI* for the post-IT period which resulting the presence of nonlinear adjustments in the policy function. Before inflation targeting, the policy function is only responsive to changes in output gap which is limited at one threshold value. The policy rate is lessened to accommodate higher output gap. Aftermath inflation targeting, the policy function is responding in fluctuating magnitudes to inflation variation and output gap. In addition, there is no influence of exchange rate changes to policy rate. This rectifies, Indonesia is really practising a freely floating regime which is a fundamental requirement for a pure inflation targeting policy.

2 Literature Review

Once Taylor rule was a great benchmark for policymakers and economists to evaluate and predict current and future monetary policy [5]. Conversely, the rule was also criticised with several shortcomings and remains as an argument in the monetary policy practice in both advanced and emerging economies [6]. Among them is the nonlinear or asymmetric form of Taylor rule as most researchers diverting their studies from the original to a nonlinear form [7, 8].

In fact, this is proven with some empirical studies, where asymmetric form surpasses the symmetrical form in terms of its capability to capture the actual monetary policy function [2, 7, 9, 10].

As far as developing and emerging countries are concerned only limited studies are conducted. To address a few, Fatima and Malik [11] identified the formation of nonlinearity in the policy reaction function of Pakistan. Ncube and Tshuma [8] validated the nonlinearity behaviour in the policy reaction function of South Africa. Similarly, Baaziz [12] concluded that the Brazilian policy setting is better to be described with an asymmetric form of Taylor rule. To be more specific, Miles and Schreyer [13] detected nonlinearity style in the conduct of monetary policy in four emerging Asian economies (includes Indonesia). Likewise, Caporale et al. [14] examined five emerging economies and asserted that the policy reaction function of all these countries (also covers Indonesia) are relatively working in the form of nonlinear equation.

Therefore, now, the aim of our study is to close the research gap by exclusively focusing on Indonesia but in a different angle which excludes the financial crisis term and breaks the series of data into pre- and post-IT regime. This is to examine the presence of asymmetric structure in the policy reaction function of Indonesia in two contrasting eras.

3 Data and Methodology

In view of the changes in the policy regime due to financial crisis in Indonesia, the sample period of the study is divided into two sub-periods, the pre-IT (1980Q1 to 1996Q4) and post-IT (2000Q1 to 2015Q4). The time series data is obtained from *Thomson Reuters Datastream* and the basic variables are the central bank interest rate (*INT*) in percentage, consumer price index (*CPI*) in index form, actual gross domestic product (*GDP*) in US dollar and real exchange rate (*REER*) per US dollar. To streamline the data, all these variables are converted to natural logarithm (*LCPI*, *LGDP*, *LREER*) except *INT*. To proceed further, the CPI inflation (*CINF*) is obtained using *LCPI* deviates from its lagged 4 (proxy for annual rate). While the output gap (*GAP*) is developed using the Hodrick-Prescott filter by subtracting the *GDP* trend from *LGDP* and exchange rate changes (*DLREER*) is formulated from the first differencing of *LREER*.

In this analysis, the augmented Taylor rule (includes the exchange rate term) equation is adopted from Mehrotra and Sanchez-Fung [15] to estimate the policy reaction function of Indonesia:

$$i_t = a_0 + a_1(\pi_t - \pi^*) + a_2(y_t - \tilde{y}) + a_3\Delta e_t + a_4i_{t-1} + \varepsilon_t \quad (1)$$

$$i_t^* = c_0 + \alpha(\pi_t - \pi^*) + \beta(y_t - \tilde{y}) + \delta\Delta e_t \quad (2)$$

$$i_t = (1 - \rho)i_t^* + \rho i_{t-1} + \varepsilon_t \quad (3)$$

where i_t^* is the central bank nominal rate, i_t is the central bank actual or real rate, $(\pi_t - \pi^*)$ is the deviation between inflation and targeted rate, $(y_t - \tilde{y})$ is the output gap indicating the actual growth deviates from its potential rate, Δe_t is the changes in nominal exchange rate. In such $a_0 = (1 - \rho)c_0$, $a_1 = (1 - \rho)\alpha$, $a_2 = (1 - \rho)\beta$, $a_3 = (1 - \rho)\delta$ and $a_4 = \rho$. Where, ρ is the smoothing coefficient (takes the value between 0 and 1) shows the policy rate is adjusted slowly, c_0 is the constant term, α is the coefficient of inflation variation, β is the coefficient of output gap and δ is the coefficient of exchange rate changes. Then, Taylor rule augmented with smoothing term, i_{t-1}^* and exchange rate changes term, Δe_t is the reduced form of a policy rule in Eq. (2).

Whereby, to capture the possible nonlinear response in the policy rule the threshold regression or a simple form of nonlinear regression model from Bai and Perron [16] is executed. As such the current policy rule proceeds the following form:

$$i_t = I[x_t \geq x_1^*][a_0 + a_1(\pi_t - \pi^*) + a_2(y_t - \tilde{y}) + a_3\Delta e_t + a_4i_{t-1} + \varepsilon_{1t}] \\ + I[x_t < x_1^*][b_0 + b_1(\pi_t - \pi^*) + b_2(y_t - \tilde{y}) + b_3\Delta e_t + b_4i_{t-1} + \varepsilon_{2t}] \quad (4)$$

$$i_t = I[x_t \geq x_1^*][a_0 + a_1(\pi_t - \pi^*) + a_2(y_t - \tilde{y}) + a_3\Delta e_t + a_4i_{t-1} + \varepsilon_{1t}] \\ + I[x_2^* < x_t < x_1^*][b_0 + b_1(\pi_t - \pi^*) + b_2(y_t - \tilde{y}) + b_3\Delta e_t + b_4i_{t-1} + \varepsilon_{2t}] \\ + I[x_t \geq x_2^*][c_0 + c_1(\pi_t - \pi^*) + c_2(y_t - \tilde{y}) + c_3\Delta e_t + c_4i_{t-1} + \varepsilon_{3t}] \quad (5)$$

Equations (4) and (5) are the threshold equations with one and two threshold values, respectively. Further, i describes the threshold effect with threshold values x_1^* and x_2^* by searching at most 2 thresholds. Based on Akaike Information Criteria, $x_t = LREER_t$ is the threshold variable for pre-IT while $x_t = LCPI_t$ is the threshold variable for post-IT. Then, the test statistics exert the White heteroscedasticity consistent covariance and permit heterogeneous error distributions across thresholds.

4 Results and Discussion

We begin our analysis by examining the presence of unit root on variables with KPSS, Phillips-Perron and Breakpoint tests. All tests show very similar results with variables stationary at levels or $I(0)$. Moreover, the BDS test which is a test to examine the nonlinearity serial dependence in time-series data is also included in our analysis. In such the result shows rejection of independence in most cases with the existence of nonlinear structure in the data of both sub-periods. Tables of unit root tests and BDS test are omitted for limitation.

The results from the tests above granted the threshold regression estimation (refer Table 1). From the analysis, the policy function of Indonesia is determined by two threshold indicators, exchange rate (*LREER*) and inflation (*LCPI*) for pre- and post-IT, respectively. In general, the policy function is reacting asymmetrically to inflation and output gaps which can be noted at different threshold breaks in both sub-periods. In the pre-IT period, the policy rule is reacting only to output gap when the threshold value is $LREER \geq 4.6831$. Then, in the post-IT period, the policy rule is positively and strongly significant towards inflation variation in all threshold values but reacting adversely to output gap at limited threshold value (when $3.8945 \leq LCPI < 4.2542$). Basically, the results reveal the behaviour of policy reaction function in two different policy regimes. In pre-IT, Indonesia was practising the exchange-rate targeting (1970-July 1997). This coincides with the results, where no significant response of policy function to inflation variation. Hence, indicating inflation is not the main concern of the policy target. Although results highlighted responses of policy function towards output gap, somehow restricted only in one of the thresholds. In the post-IT period, we observe a very strong response in policy rate adjustment due to inflation variation. This is because under inflation targeting policy the foremost goal is to achieve a lower inflation rate for price stability. The results show that positive inflation gap leads to increase the policy rate, i.e. policy rate is set higher to bring down inflation. However, lower inflation may lead to lower output as this subject to inflation-output trade-off. This measure is observed as higher policy rate leads to negative output gap (potential output larger than the actual output).

Table 1 Threshold regression results

Variables	Pre-IT coefficients (Threshold: <i>LREER</i>)		Post-IT coefficients (Threshold: <i>LCPI</i>)		
	$LREER < 4.6831$	$LREER \geq 4.6831$	$LCPI < 3.8945$	$3.8945 \leq LCPI < 4.2542$	$LCPI \geq 4.2542$
<i>CINF</i>	30.5819	-8.9842	37.2439 ^a	27.4631 ^a	16.5806 ^a
<i>GAP</i>	3.9613	-5.7611 ^a	-2.4138	-5.5534 ^a	0.6782
<i>DLREER</i>	-26.0618 ^b	5.0450 ^b	1.1914	-0.2734	-0.9624
<i>INT(-1)</i>	0.8900 ^a	0.4650 ^a	0.2670 ^a	0.6870 ^a	0.6890 ^a
C	-1.3140	9.3931 ^a	8.2470 ^a	0.6543	1.2016 ^a
R ²	0.8235	0.8235	0.9938	0.9938	0.9938
Obs	31	33	8	17	39

Note a and b indicate significance at 1% and 5%, respectively

Nevertheless, looking at the influence of exchange rate changes towards policy function of Indonesia, there are contradicting results between the pre- and post-IT period. Where in the pre-IT, the policy rate is reacting inconsistently to exchange rate changes between the two threshold values which assure the exchange-rate targeting policy practiced in Indonesia before the financial crisis term. While in the post-IT, there is no relationship recorded between policy rate and exchange rate changes. This affirms the execution of pure floating regime after the adoption of inflation targeting in Indonesia as an act to overcome the crisis hit. Indeed, Indonesia is fulfilling one of the theoretical conditions for a strict inflation targeting policy framework by implementing a freely floating exchange rate system.

At last, we examine the residuals of estimations with two diagnostic tests (LM and ARCH tests). The outcomes failed to reject the null hypothesis of autocorrelation and heteroscedasticity, respectively. Hence, witnessing the reliability of our estimations' results.

5 Conclusion

Our findings indicating the asymmetric adjustments of Indonesia's policy function to inflation variation and output gap at different threshold values in both pre- and post-IT periods, respectively. Particularly, in pre-IT, the policy function is only responsive to output gap, yet restricted to one threshold value. This is different in post-IT, where the policy function is greatly influenced by inflation variation with more weights compared to output gap. The results justify that the intention of inflation targeting policy in Indonesia is to attain low inflation rate as to minimise the impact of currency crisis. Besides, the theory of 'impossible trinity' is also proven in the post-IT analysis as there is no relationship between policy rate and exchange rate changes which confirming a free-floating regime combined with an inflation targeting strategy for a small open economy like Indonesia.

Acknowledgements This study is supported by FRGS grant. 203/PMATHS/6711431.

References

1. Taylor, J.B.: Discretion versus policy rules in practice. *Carnegie-Rochester Conf. Ser. on Pub. Pol.* **39**, 195–214 (1993)
2. Petersen, K.: Does the federal reserve follow a non-linear Taylor rule? *Univ. Conn. Wor. Pap. Ser.* **37**, 1–19 (2007)
3. Mohanty, M.S., Berger, B.: Central bank views on foreign exchange intervention. *BIS Pap.* **73**, 55–74 (2013)
4. Calvo, G.A.: Capital markets and the exchange rate, with special reference to the dollarization debate in Latin America. *J. Mon. Cred. Ban.* **33**, 312–334 (2001)

5. Orphanides, A.: Monetary policy rules based on real-time data. *Amer. Econ. Rev.* **91**, 964–985 (2001)
6. Koenig, E.F., Leeson, R., Kahn, G.A.: *The Taylor Rule and the Transformation of Monetary Policy*. Hoover Press, Stanford, California (2012)
7. Castro, V.: Can central banks' monetary policy be described by a linear (augmented) Taylor rule or by a nonlinear rule? *J. Fin. Stab.* **7**, 228–246 (2011)
8. Ncube, M., Tshuma, M.M.: Monetary policy conduct based on nonlinear Taylor Rule: evidence from South Africa. *Afr. Dev. Ban. Gro. Wor. Pap. Ser.* **113**, 1–43 (2010)
9. Bruggemann, R., Riedel, J.: Nonlinear interest rate reaction functions for the UK. *Econ. Mod.* **28**, 1174–1185 (2011)
10. Chevapatrakul, T., Kim, T.H., Mizen, P.: The Taylor principle and monetary policy approaching a zero bound on nominal rates: quantile regression results for the United States and Japan. *J. Mon. Cred. Ban.* **41**, 1705–1723 (2009)
11. Fatima, M., Malik, W.S.: Choice of functional form in the nonlinear Taylor rule: The case of Pakistan. *Pak. Econ. Soc. Rev.* **53**, 225–250 (2015)
12. Baaziz, Y.: Estimating interest rate setting behavior in Brazil: A LSTR model approach. *Economies* **3**, 55–71 (2015)
13. Miles, W., Schreyer, S.: Is monetary policy non-linear in Indonesia, Korea, Malaysia, and Thailand? A quantile regression analysis. *Asian-Pac. Econ. Lit.* **26**, 155–166 (2012)
14. Caporale, G.M., Çatık, A.N., Helmi, M.H., Ali, F.M., Akdeniz, C.: Monetary policy rules in emerging countries: Is there an augmented nonlinear Taylor rule? *CESifo Wor. Pap.* **5965**, 1–39 (2016)
15. Mehrotra, A., Sanchez-Fung, J.R.: Assessing McCallum and Taylor rules in a cross-section of emerging market economies. *BOFIT Dis. Pap.* **23**, 1–22 (2009)
16. Bai, B.Y.J., Perron, P.: Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78 (1998)

Part III

Statistics

Chapter 28

A Comparative Study of Outlier Detection Methods in Poisson Regression



Faten Nabila Rustam Affandy and Sanizah Ahmad

Abstract Regression models using count data have a wide range of applications in engineering, econometrics, medicine and social sciences. Poisson regression models are widely used in the analysis and prediction of counts on potential independent variables. However, the presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates. In this study, three methods of identification of outlier are used which are DFFITS, DFBETAS and Cook's Square Distance (CD). The objective of the study is to investigate on the performance of the three detection methods (DFFITS, DFBETAS, and CD) in Poisson regression using simulation in R. A simulation study was performed with various regression conditions which include different number of predictors, sample sizes and percentage of outliers in the X-space, Y-space and both X-and Y-space. The best outlier detection method is the one that can detect the most number of outliers. Results show that for outliers in X-space and Y-space, DFFITS performs better in detecting outliers for all sample sizes with low percentage of outliers while DFBETAS performs better for most of samples sizes with high percentage of outliers. In both X-and Y-space, the best method in detecting outliers for small sample size with low percentage of outliers is DFFITS. However, for large sample size, CD and DFBETAS perform better in detecting low and high percentage of outliers, respectively. Similar results were obtained when these methods were applied to a real data set.

Keywords Poisson · Outlier · DFFITS · DFBETAS · Cook's square distance

F. N. R. Affandy (✉) · S. Ahmad
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Shah Alam, Shah Alam, Malaysia
e-mail: faten_affandy@yahoo.com

S. Ahmad
e-mail: sanizah@tmsk.uitm.edu.my

1 Introduction

Regression models using count data have a wide range of applications in engineering, econometrics, medicine and social sciences. Poisson regression models are useful in modeling count data where the response variable is the counted number of occurrences of the event. It also has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance [1]. In Poisson regression model, the number of events y has a Poisson distribution with a conditional mean that depends on individual characteristics according to the structural model [2]:

$$\lambda_i = E(y_i | x_i) = \text{Exp}(x_i' \beta) \quad (1)$$

Taking the exponential $x\beta$ forces the expected count λ to be positive, which is required for the Poisson distribution [3]. If the discrete random variable y follows the Poisson distribution, then:

$$p(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots \quad (2)$$

An outlier is a data point located far outside the norm for a variable or population [3]. Hawkins [4] described an outlier as an observation that “deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”. Presence of outliers should be taken care before doing any statistical analysis since extreme values of observed variables can distort estimates of regression coefficients.

2 Outlier Detection Methods

This study will focus on outlier detection methods in Poisson regression model, focusing on three methods, which are DFFITS, DFBETAS, and Cook’s Square Distance (CD). The DFFITS statistic is a scaled measure of the change in the predicted value for the i th observation and is calculated by deleting the i th observation with the formula given as

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)} \sqrt{h_{(ii)}}} \quad (3)$$

where \hat{y}_i and $\hat{y}_{(i)}$ are the prediction for point i with and without point i included in the regression, $s_{(i)}$ is the standard error estimated, and $h_{(ii)}$ is the leverage for the point. Value of DFFITS greater than 2 in absolute value indicates that the i th observation is a possible outlier. A size-adjusted cutoff is $2\sqrt{p/n}$, where n is the sample size and p is the number of parameters [5]. The DFBETAS statistic is the

scaled measure of the change in each parameter estimate and the calculation is by deleting the i th observation with the given formula:

$$DFBETAS = b - b_{(-i)} = \frac{(X^T X) x_i^T e_i}{1 - h_i} \quad (4)$$

where $b_{(-i)}$ denotes the coefficients estimated with the i th row x_i of X deleted, h_i denotes the i th row of H . Large positive and negative values of DFBETAS indicate observations that lead to large changes in the i th regression coefficient. The cutoff value 2 is to indicate influential observations and $2/\sqrt{p}$ as the size-adjusted cutoff [5]. Cook's square distance (CD) of unit i th is a measure base on the square of the maximum distance between ordinary least square (OLS) estimate based on all n points $\hat{\beta}$ and the estimate obtained when the i th point is not included, say $\hat{\beta}_i$. Examining cases with $CD_i^2 > 0.5$ and that case where $CD_i^2 > 1$ should always be studied. This distance measure can be expressed in a general form

$$CD_i^2 = \left(\frac{e_i^2}{p} \right) \left(\frac{h_{ii}}{1 - h_{ii}} \right) \quad (5)$$

where $i = 1, 2, \dots, n$, p is the number of parameters and h is the leverage for the point. Cook and Weisberg [6] suggest any i th observation with values $CD_i^2 > 1$ is considered as an outlier.

3 Simulation Study and Results

A simulation study with designs adopted from Nor [7] and Oyeyemi [8] was conducted to investigate the performance of the outlier detection methods in Poisson regression using R for 1000 replications. The first design matrix consists of two regressors which includes a constant, $X_1 \sim \text{Bin}(n, 0.5)$ and $X_2 \sim N(0, 1)$ with true values $\beta = (1, 0.2, 0.2)$. The second design matrix consists of three regressors with a constant, a simulated binomial variable, $X_1 \sim \text{Bin}(n, 0.5)$ and X_2 and $X_3 \sim N(0, 1)$ with true values $\beta = (1, 0.2, 0.2, 0.2)$. The response value Y_i was generated from a Poisson distribution with mean, λ . The sample sizes considered are small ($n = 20$), moderate ($n = 50$), and large ($n = 100$ and 200). The samples were then contaminated with 10, 20 and 30% levels of contamination.

3.1 X-Space Regression Outlier

For the first design matrix, the first percentage of observations in X_2 are replaced by $X_2 + 15$ and for the second design, X_2 and X_3 are replaced by $X_2 + 15$ and $X_3 + 15$, respectively. The simulation outputs give the probability of correctly identifying the

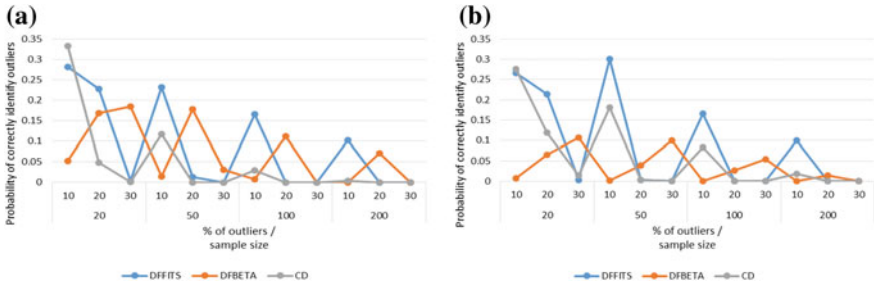


Fig. 1 Comparison of statistics for detection of outliers by sample sizes with outlier in X for **a** Two regressors with outliers in one X; **b** Three regressors with outliers in two Xs

actual number of outliers. The value 1.000 means that the method correctly identifies all the outliers in each of 1000 replicates while the value 0.000 implies the method fails to detect any outlier. Figure 1a shows that CD has the highest number of times outliers are detected (probability = 0.333) for small sample size with low % of outliers. However, for other sample sizes, DFFITS performs better. DFBETAS performs better as the % of outlier increases for larger n . From Fig. 1b, the best detection method for three regressors for small n with low % of outliers is CD. When n increases, DFFITS performs better for low % of outliers. DFBETAS shows better performance in detecting outliers in the X-space regression for moderate and large n with medium and large % of outliers.

3.2 Y-Space Regression Outlier

Outliers in the Y-space are introduced by replacing the first percentage of observations with Y^*25 . Figure 2a clearly shows that DFFITS performs better in detecting outliers for all sample sizes with low percentage of outliers where DFFITS has the highest performance in outlier detection in $n = 20$ with 20% of outliers

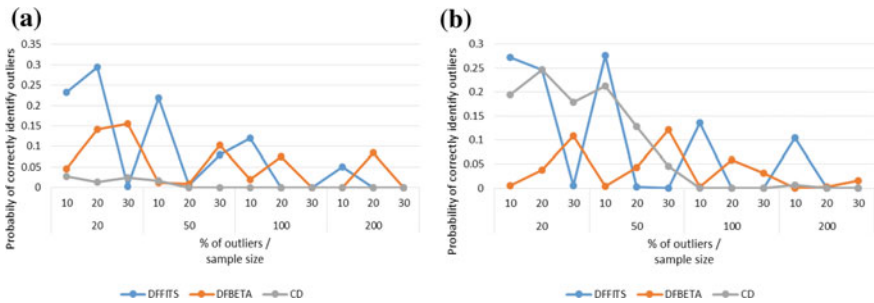


Fig. 2 Comparison of statistics for detection of outliers by sample sizes with outlier in Y for **a** Two regressors; **b** Three regressors

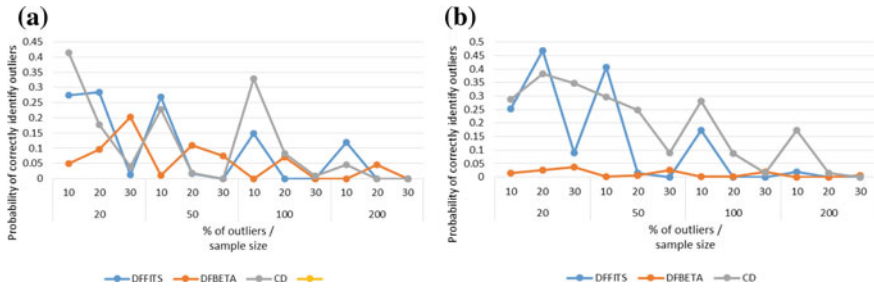


Fig. 3 Comparison of statistics for detection of outliers by sample sizes with outlier in both X and Y for **a** Two regressors with outlier in one X; **b** Three regressors with outlier in two Xs

(probability = 0.290). However, DFBETAS performs the best in every sample sizes with 30% of outliers. For two regressors with outliers in Y-space, none of the methods are able to detect 30% of outlier for $n = 200$. From Fig. 2b of three regressors, the best outlier detection for every sample sizes with low percentage of outlier in Y-space is DFFITS with probability of 0.272, 0.276, 0.136, and 0.105 for $n = 20, 50, 100$ and 200, respectively. For high % of outliers in small n , CD performs better. DFBETAS performs best when it comes to large n with high % of outliers.

3.3 X-and Y-Space Regression Outlier

For the first condition, the first (follows the percentage of the outlier injection) observations of X_2 are replaced by $X_2 + 15$ and Y are replaced by $Y*25$. For second condition, X_2 are replaced by $X_2 + 15$, X_3 by $X_3 + 15$ and Y are replaced by $Y*25$. Referring to Fig. 3a, for two regressors, CD performs best for $n = 100$ since it has the highest probability at all % of outliers. For small and moderate n , DFFITS and DFBETAS perform better for low and high % of outliers, respectively. Figure 3b shows that CD performs better for different n and % of outliers. Then, DFFITS performs better in detecting outlier for small and moderate n with low % of outlier. DFBETAS is worst among all methods since it only detects high % of outliers in large n .

4 Application on Real Data

The three methods are applied to a benchmark data set of nesting Horseshoe Crabs [9] with sample size of 173 female crabs and observation 165 as the outlier. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. The explanatory variables are female crab’s color

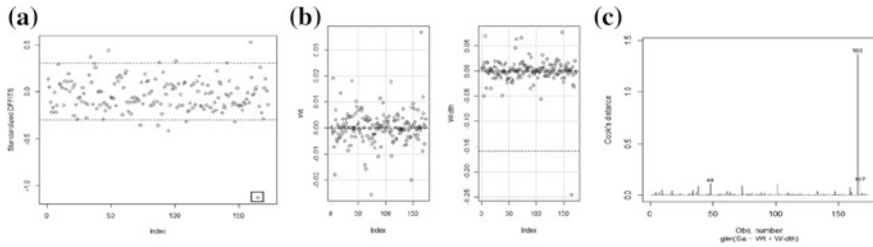


Fig. 4 a DFFITS, b DFBETAS, c CD

(*Color*), spine condition (*Spine*), weight (*Wt*), and carapace width (*Width*). The response outcome is the number of satellites (*Sa*) of female crab. The plots in Fig. 4 show that only CD is able to correctly identify observation 165 as the outlier. Here we can relate to the simulation result where CD performs best for large sample size with low percentage of outliers.

5 Conclusion

Results from the simulation studies proved that none of the three methods discussed can be identified as the best outlier identification method for Poisson models since all methods cannot detect the correct number of outliers. Further study is needed to investigate on the performance of other diagnostic methods available in the literature.

Acknowledgements The authors wish to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Shah Alam for the conference support fund.

References

1. Long, J.S.: Regression Models for Categorical and Limited Dependent Variables. SAGE Publication, New York (1997)
2. Algamal, Z.Y.: Diagnostic in poisson regression models. Electron. J. Appl. Stat. Anal. **2**, 178–186 (2012). <https://doi.org/10.1285/i20705948v5n2p176>
3. Jarrell, M.G.: A comparison of two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, for identifying multivariate outliers. Res. Sch. **1**, 49–58 (1994)
4. Hawkins, D.M.: Identification of Outliers. Chapman and Hall, London (1980). <https://doi.org/10.1002/bimj.4710290215>
5. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1980). <https://doi.org/10.1002/0471725153>
6. Cook, R.D., Weisberg, S.: Residuals and Influence in Regression. Chapman & Hall, New York (1982). <https://doi.org/10.1002/bimj.4710270110>

7. Nor, A.: Investigating the performance of mallows-type estimator in logistic regression model with the presence of outliers. Dissertation Master Sci. (Appl. Sci.) (2010)
8. Oyeyemi, G.M., Bukoye, A., Akayede, I.: Comparison of outlier detection procedures in multiple linear Regression. *Am. J. Math. Stat.* **5**(1), 37–41 (2015). <https://doi.org/10.5923/j.ajms.20150501.06>
9. Brockmann, H.J.: Satellite male groups in horseshoe crabs. *Limulus Polyphemus*. *Ethol.* **102** (1), 1–21 (1996). <https://doi.org/10.1111/j.1439-0310.1996.tb01099.x>

Chapter 29

A Modified Long Memory Model for Modeling Interminable Long Memory Process



Rosmanjawati Abdul Rahman and Sanusi A. Jibrin

Abstract In this paper, Autoregressive Fractionally Integrated Moving Average (ARFIMA) model was modified and was used for modeling the daily Malaysia Stock Price Index (MSPI). The long and slow decline in autocorrelation function of the data showed the presence of Long Memory (LM) structure. Therefore, the Mandelbrot and Lo rescaled-range tests were used to test the presence of LM. The ARFIMA model then is further extend to the Autoregressive Fractionally Unit Root Integrated Moving Average (ARFURIMA) model. The Geweke and Porter-Hudak (GPH), Local Whittle Estimator (LWE), and Hurst Exponent (HE) were used as the estimation methods to obtain the LM parameters d of both ARFIMA and ARFURIMA models. The best model was identified for each of ARFIMA and ARFURIMA models respectively based on the minimum Akaike Information Criteria (AIC) values. The best fitted model were specified as ARFIMA (2,0.989,0) and ARFURIMA (1,1.069,0). Having compared the residuals analysis of the two models, we conclude that the ARFURIMA model was better in estimating series that exhibit Interminable LM (ILM).

Keywords LWE · Long memory · Financial data · ARFIMA model · ARFURIMA model

R. A. Rahman · S. A. Jibrin (✉)
School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Gelugor,
Pulau Pinang, Malaysia
e-mail: sanusijibrin46@gmail.com

R. A. Rahman
e-mail: rosmanjawati@usm.my

S. A. Jibrin
Department of Statistics, Kano University of Science and Technology, P.M.B. 3244, Wudil,
Kano State, Nigeria

1 Introduction

Long memory (LM), also called long-range dependence is a phenomenon that may arise in the analysis of time series. It is traced either due to a gradual decline in the Autocorrelation Function (ACF) or as a marginal decrease in the Autocorrelation (AC) values of a time series. Box et al. [1] viewed LM as slow decays to zero at a hyperbolic rate and Charfeddine and Guegan [2] defined it as a slow varying function at infinity. According to Qu [3], if the spectral density of a scalar K is proportional to K^{-2d} as K tend to zero, then the process is said to have LM. Here, d is the LM parameter and also is called the degree of fractional differencing. The LM was first discovered in geophysical data by Hurst [4] and later in climatology, economic and financial data. Whittle [5], Mandelbrot [6] and McLeod and Hipel [7] were among early studies on the LM phenomenon.

The occurrence of LM can be determined by a rescaled range test as shown in Mandelbrot [6] and Lo [8]. Once the LM is confirmed by these tests, the LM parameter also can be estimated either by parametric or semiparametric procedures as discussed by [6, 9, 10]. In the last four decades, the degree of fractional differencing, d , had been estimated to be in the interval $0 < d < 1$, as shown by Nezhad [11], Jibrin et al. [12], Chaefeddine and Ajimi [13], Arouri et al. [14] and Jibrin [15]. Due to recent growth in financial markets, investors dwindling activities and other factors such as inflation, there are reasons that the degrees of fractional differencing is greater than unity (see Baillie et al. [16] and Dalla [17]).

This study considers series with linear and deterministic trend and in the same time, exhibits Interminable Long Memory (ILM). In order to overcome the problems of under and over differencing and to get the appropriate d values for the fractional differenced process, the fractional differencing operator or filter as suggested by Granger and Joyeux [19] will be modified. The modification lies in replacing d with $d + 1$. In this case, we let the fractional differencing estimator be $d^* = d + 1$, where d is such that $0 < d < 1$. The unity in the estimator indicates that the need to initially determine the first difference of any nonstationary, deterministic and interminable persistent series before estimating the d^* value.

The modified filter is then incorporated into the Box and Jenkins [20] Autoregressive Integrated Moving Average (ARIMA) such that to effectively accommodate the memory parameters in the range $1 < d < 2$. The Autoregressive Fractionally Integrated Moving Average (ARFIMA) model then was further extended to the Autoregressive Fractionally Unit Root Integrated Moving Average (ARFURIMA) model. Both models were compared and the best fitted model was determined for estimating series that exhibit ILM. The remaining of this paper is as follows. Section 2 describes the material and the methodology used. Section 3 provides results and discussion while Sect. 4 summarizes the main findings and offers some concluding remarks.

2 Material and Methodology

The data used for this study and discussion on the fractional differencing filter and model modification are outlined in this section.

2.1 The Dataset

Data for this study are obtained from the Morgan Stanley Capital International (MSCI) data stream and there are 5195 daily (5 days) Malaysia Stock Price Index (MSPI) between the 6th October 1997 and 1st September 2017.

2.2 Fractional Brownian Motion

The Fractional Gaussian Noise (FGN) process is obtained when an ordinary Brownian motion is differenced by certain degree of differencing value. There are two sensitive fractional Brownian motions (fBm); anti-persistent fBm, when $0 < H < 0.5$ and the other one is persistent fBm, when $0.5 < H < 1$. Here, H is called the Hurst index or Hurst parameters associated with the fBm where it is a real number ranged in $(0,1)$. These persistent and anti-persistent properties are both known as long-range dependence.

If Y_t is an interminable persistent fractal series with degree of fractional differencing $1 < d < 2$, then $Y_t^* = \Delta Y_t$, $t = 1, \dots, T$ so that $d = 1$. To define Y_t with respect to fBm, Y_t^* is said to be anti-persistent fBm, if $0 < d < 0.5$ and it is persistent fBm when $0.5 < d < 1$, while Y_t is called interminable persistent fBm if $1 < d < 2$.

2.3 White Noise Process

Consider $\{Y_t\}$, $t = 1, \dots, T$ is a nonstationary process with time-varying mean and variance. Hence, ARIMA (0,1,0) process is defined by:

$$(1 - L)Y_t = \varepsilon_t \tag{1}$$

where L is the backwards-shift element or operator. Here, if ε_t is independent and identically distributed random variable, then ε_t is a white noise process.

2.4 Fractionally Unit Root Integrated Process

{Y_t} is said to be Fractionally Integrated (FI) process if

$$(1 - L)^d(Y_t - \mu) = \varepsilon_t \tag{2}$$

where *L* is the backward shift operator, μ is the mean of the original series, ε_t is a white noise process, *d* is the LM parameter such that $0 < d < 0.5$. Now, by replacing *d* with *d* + 1 in (2), we get

$$(1 - L)^{d+1}(Y_t - \mu) = \varepsilon_t, \tag{3}$$

Hence, ε_t now is called a Fractionally Unit Root Integrated (FURI) process. The modified filter $(1 - L)^{d+1}$ can be simplified by using the standard binomial expansion as applied by Granger and Joyeux [19] and Hosking [21] as follows:

$$(1 - L)^{d+1} = 1 - \frac{(d+1)}{1!}L + \frac{d(d+1)}{2!}L^2 + \dots - \tag{4}$$

where *L* is the lag operator and *d* is a value such that $0 < d < 1$. If $f_y(\cdot)$ and $f_\varepsilon(\cdot)$ be the spectral density function of {Y_t} and { ε_t } respectively, then the spectrum of Y_t is given by:

$$f_y(k) = |1 - e^{-ik}|^{-2d} f_\varepsilon(k), \quad k \neq 0 \tag{5}$$

The modified filter is expected to produce series with properties such as a covariance stationary and ergodicity process, positive and bounded spectrum, the infinite spectral density at zero frequency, and with finite autocorrelations.

In addition, to effectively study time series that has memory parameters in the range, $1 < d < 2$, the modified filter is then incorporated into the Box and Jenkins [20] ARIMA model as follows;

$$\phi(L)(1 - L)^{d+1}(Y_t - \mu) = \theta(L)\varepsilon_t \tag{6}$$

where $\phi(L) = 1 - \phi_1L - \dots - \phi_pL^p$ and $1 + \theta_1L + \dots + \theta_qL^q$ and all the roots of $\phi(L)$ and $\theta(L)$ lies outside the unit circle. The series are considered as IML or interminable persistent when *d* = 1.25 and 1.75. Finally, (6) represents the modified model and is called the ARFURIMA model. The identification of this ARFURIMA (*p*, *d* + 1, *q*) model, is similar to Box and Jenkins approach.

The proposed ARFURIMA model is finally applied to the MSCI data and compared to the ARFIMA model. Accuracy measure analysis, serial correlation

tests and heteroscedasticity analysis are carried out based on the models residual to determine the best conditional mean model for the series that exhibited ILM or a FURI process.

2.5 Software

This study uses the Ox Professional version 7.00 [22] and gretl-2017c version 1.9.4 [23].

3 Analysis and Results

Figure 1 shows the time series plot of the MSPI data. The graph exhibits clear nonlinear and deterministic trends and overall, the series is not stationary.

Figure 2 is the Correlogram for the MSPI and it shows a slow decay in the autocorrelation which indicates the long memory processes.

The summary statistics and normality test for the MSPI fractionally differenced by using the fractional filters, $(1 - L)^d$ and $(1 - L)^{d+1}$ are displayed in Table 1. Results show that the modified filter produces a series with less variability and a reduction in both skewness and kurtosis.

The Mandelbrot [6] and Lo [8] tests confirm the incidence of LM at both level series and first difference of the MSPI as displayed in Table 2. As we can see the test statistic values: 30.91 and 21.86 at level series and 1.48 and 1.41 at first difference are greater than the critical value which is 0.809. The three LM estimators, Geweke and Porter-Hudak (GPH), Local Whittle Estimator (LWE), and Hurst Exponent (HE) produce inconsistent results at level series, with $d < 1$ and

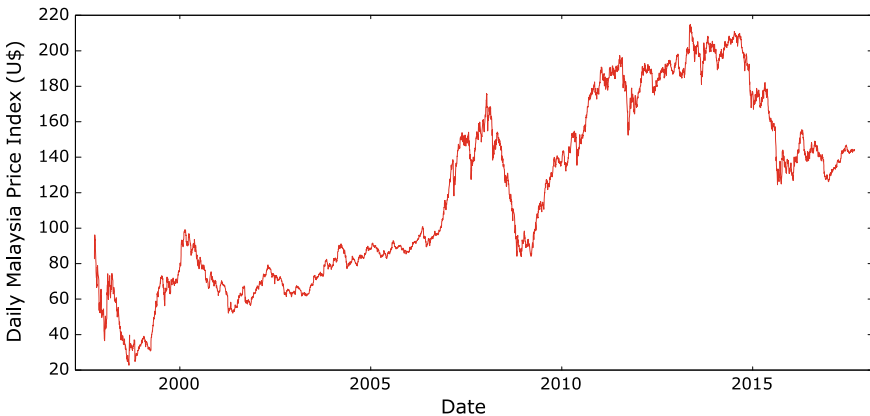


Fig. 1 The plot of daily prices for Malaysia MSCI stock index (1997–2017)

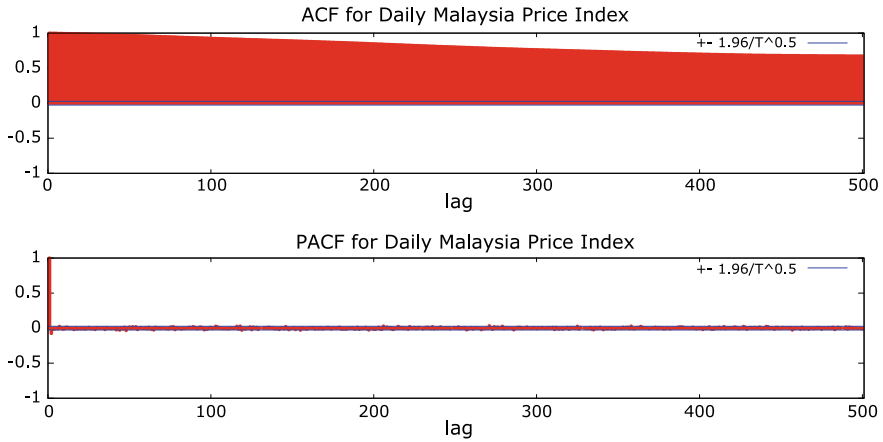


Fig. 2 Correlogram for daily prices for Malaysia MSCI stock index

Table 1 Summary statistics using the filter, the $(1 - L)^d$ and $(1 - L)^{d+1}$

Statistics	$X_t(1 - L)^{0.987}$	$\Delta X_t(1 - L)^{0.068+1}$
Mean	0.0312	0.0064
Stand. deviation	1.7933	1.3681
Skewness	18.792	-0.1257
Ex.Kurtosis	873.54	12.115
Shapiro-Wilk W	0.5809	0.9017

also show $d > 1$. However, after first differenced, all the three estimators indicate memory values are greater than unity. The mean of 1.069*, 1.0689* and 1.0665* is 1.068 which is the average degree of fractional differencing produced by the modified filter. Consequently, we conclude that it is convenient to use the ARFIMA and ARFURIMA model to study the daily MSPI data.

A comparison between all models at level and after first differencing between Table 3 and 4 shows that the ARFIMA (2,0)* and ARFURIMA (1,0)* provides the minimum AIC. This implies that the models are better fit to the data. Next, the residuals analyses are considered for both models.

The residuals analyses as shown in Table 5 are based on the corrected Ljung-Box test proposed by Francq and Zakoian [24], the Jarque-Bera [25]

Table 2 Long memory test and degree of fractional differencing estimation

Tests/estimators	MSPI (level series)	MSPI (first diff.)
Mandelbrot stat.	30.9083	1.4820
Lo statistic	21.8579	1.4073
HE	1.0026	1.0690*
LWE	0.9865	1.0689*
GPH	1.1325	1.0665*

The critical value for the test is 0.809

Table 3 ARFIMA (p,d,q) model identification

ARFIMA (p,q)	d	AIC
ARFIMA (1,1)	1.059(0.000)	20712.71
ARFIMA (0,1)	1.058(0.000)	20748.81
ARFIMA (1,0)	1.011(0.000)	20734.32
ARFIMA (2,0)*	0.989(0.000)	20706.02
ARFIMA (1,2)	0.654(0.010)	20707.51

Values in parenthesis are p-values

Table 4 ARFURIMA (p, d + 1, q) model identification

ARFURIMA (p,q)	d + 1	AIC
ARFURIMA (1,1)	1.176(0.000)	17992.82
ARFURIMA (0,1)	1.057(0.000)	17991.16
ARFURIMA (1,0)*	1.069(0.002)	17991.00
ARFURIMA (2,0)	1.215(0.059)	17992.59
ARFURIMA (1,2)	1.211(0.046)	17994.24

Values in parenthesis are p-values

Table 5 The models residuals analysis

Resid. analysis	ARFIMA (2,0)	ARFURIMA (1,0)
Ljung-Box test	127.418	93.4530
Jarque-Bera test	540562	10473.2
ARCH-LM test	344.835	341.369

residuals normality test and ARCH-LM or Engle [26] heteroscedastic test for the ARFIMA (2,0.989,0) and ARFURIMA (1,1.069,0). The results indicate that the ARFURIMA (1,1.069,0) is the most adequate model to study the daily MSPI data due to the minimum values of Portmanteau, Jarque-Bera and ARCH-LM test statistic compared to the ARFIMA (2,0.989,0) model. Besides, the ARFURIMA (1,1.069,0) model is also conform to principle of parsimony in modeling financial and economic data with few parameters compared to ARFIMA (2,0) model.

4 Conclusion

The results of this study show that the modified model which is the ARFURIMA is useful in a number of ways. First, our proposed filter provides the appropriate fractional degree of differencing, $d = 1.068$. It also produces better statistical properties such as stationary series with minimum values of certain statistics such as skewness, kurtosis and standard deviation, compared to the existing fractional differencing operator of Granger and Joyeux [19]. Second, the ARFURIMA long memory model as proposed for modeling ILM process indicates superiority over the ARFIMA model because of minimum measure of accuracy, the AIC. Also, the

ARFURIMA (1,1.069,0) model conform to principle of parsimony that explains financial and economic data with few parameters compared to ARFIMA (2,0) model. Third, the new model produces minimum test statistic for residuals analysis such as Portmanteau, Jarque-Bera, and ARCH-LM test and this shows that the ARFURIMA model is closed to a white noise process compared to the ARFIMA model. All these indicate that the identified ARFURIMA model is better in capturing some patterns in the original series which are not adequately explained by ARFIMA model which may lead to an erroneous statistical test and choice of unstable and inefficient model that may lead to bizarre forecast. Finally, since the study shows the presence of heteroscedasticity in the model's residuals, further research can be carried out by using the fractional integrated volatility model(s) in considering the effects of some stylized fact.

References

1. Box, G.C., Jenkins, G.E.P., Reinsel, G.M.: Time Series Analysis: Forecasting and Control, 4th edn. Wiley, Hoboken, NJ (2008)
2. Charfeddine, L., Guégan, D.: Breaks or long memory behaviour: an empirical investigation. *Phys. A* **391**(22), 5712–5726 (2012). <https://doi.org/10.1016/j.physa.2012.06.036>
3. Qu, Z.: A test against spurious long memory a test against spurious long memory. *J. Bus. Econ. Stat.* **29**(3), 423–438 (2011). <https://doi.org/10.11981/jbes.2010.09153>
4. Hurst, E.: Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* **116**, 770–799 (1951)
5. Whittle, B.Y.P.: On the variation of yield variance with plot size. *Biometrika* **43**(3), 337–343 (1956). 202.170.60.251
6. Mandelbrot, B.B.: Statistical methodology for non-periodic cycles: from the covariance to R/S analysis, vol. 1, pp. 259–290 (1972). <http://www.nber.org/chapters/c9433.pdf>
7. Mcleaoad, A.I., Hipel, K.W.: Preservation of the rescaled adjusted range I: a reassessment of the Hurst phenomenon. *Water Resour. Res.* **14**, 491–508 (1978). <https://doi.org/10.1029/wr014i003p00491>
8. Lo, A.W.: Long-term memory in stock market prices. *Econometrica* **59**, 1279–1313 (1991). <https://doi.org/10.2307/2938368>
9. Geweke, J., Porter-Hudak, S.: The estimation and application of long memory time series models. *J. Time Ser. Anal.* **4**(4), 221–238 (1983). <https://doi.org/10.18488/journal.8/2016.4.3/8.3.142.152>
10. Shimotsu, K., Phillips, P.C.B.: Local Whittle estimation of fractional integration. *Ann. Stat.* **33**, 1890–1933 (2005). <https://doi.org/10.1214/009053605000000309>
11. Nezhad, M.Z., Raoofi, A., Akbarzdeh, M.H.: The existence of long memory property in OPEC oil prices. *Asian J. Econ. Model.* **4**(3), 142–152 (2016). <https://doi.org/10.18488/journal.8/2016.4.3/8.3.142.152>
12. Jibrin, S.A., Musa, Y., Zubair, U.A., Saidu, A.S.: ARFIMA modelling and investigation of structural break(s) in West Texas Inter-mediate and Brent series. *CBN J. Appl. Stat.* **6**(2), 59–79 (2015). <http://www.econstor.eu/handle/10419/142106>
13. Charfeddine, L., Ajmi, A.N.: The Tunisian stock market index volatility: long memory vs. switching regime. *Emerg. Mark. Rev.* **16**, 170–182 (2013). <https://doi.org/10.1016/j.ememar.2013.05.003>

14. Arouri, M.E.H., Hammoudeh, S., Lahiani, A., Nguyen, D.K.: Long memory and structural breaks in modeling the return and volatility dynamics of precious metals. *The Qur. Rev. Econ. Financ.* **52**(2), 207–218 (2012). <http://dx.doi.org/10.1016/j.qref.2012.04.004>
15. Jibrin, S.A.: Data Mining and Modeling of Crude Oil Prices. M.Sc Diss. (2015). <http://scholar.google.com/scholar?cluster=117008-92651303445542&hl>
16. Baillie, R.T., Kapetanios, G., Papailias, F.: Bandwidth selection by cross-validation for forecasting long memory financial time series. *J. Empir. Financ.* **29**, 129–143 (2014). <https://doi.org/10.1016/j.jempfin.2014.04.002>
17. Dalla, V.: Power transformations of absolute returns and long memory estimation. *J. Empir. Financ.* **33**, 1–18 (2015). <https://doi.org/10.1016/j.jempfin.2015.05.002>
18. Granger, C., Joyeux, R.: An introduction to long memory time series models and fractional differencing. *J. Time Ser. Anal.* **1**(1), 15–29 (1980)
19. Box, G., Jenkins, G.: *Time series analysis: forecasting and control*, revised edn. Holden-Day (1976)
20. Hosking, J.R.M.: Fractional differencing. *Biometrika Trust.* **68**(1), 165–176 (1981). <https://doi.org/10.1093/biomet/68.1.165>
21. Laurent, S., Doornik, J.A.: *G@RCH 7.0: OxMetrics Package for testing LM, Estimating and Forecasting ARFIMA models* (2012)
22. Cottrell, A., Lucchetti, R.: *Gretl Function Package Guide, gretl documentation* (2016). <http://sourceforge.net/projects/gretl/files/manual/>
23. Francq, C., Zakoian, J.M.: Bartlett's formula for a general class of nonlinear processes. *J. Time Ser. Anal.* **30**, 449–465 (2009)
24. Jarque, C.M., Bera, A.K.: A test for normality of observations and regression residuals. *Int. Stat. Rev.* **55**(2), 163–172 (1987)
25. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007 (1982)

Chapter 30

Application of Functional Data Analysis in Streamflow Hydrograph



Jamaludin Suhaila

Abstract Streamflow data is often recorded at discrete time intervals, such as hourly, daily, monthly or annually and a hydrograph is normally used to represent the temporal variation of these flows in a graphical form. In practice, a river may have various shapes of flood hydrographs. The shape of a hydrograph varies in each river basin and each individual storm event. The aim of this study is to apply a functional data concept in hydrological applications using streamflow hydrographs as functional data. An entire hydrograph curve with respect to time can be considered as a single observation within the functional context. To analyse a streamflow hydrograph, functional descriptive statistics, functional principal components and functional outliers are among the functional data analysis tools introduced in this study. The functional principal component was adapted to find new functions that reveal the most important type of variation in the hydrograph curve, while the graphical methods (namely the rainbow plots) were used to visualise the functional data. The functional highest density region box-plot is employed to identify functional outliers. These methods were applied to the case study of flood analysis at the Sg. Kelantan River Basin, Malaysia. In conclusion, the functional framework is found to be more flexible in analysing the whole hydrograph and is able to make full use of information contained in the hydrograph.

Keywords Streamflow · Hydrograph · Functional data analysis · Rainbow plot · Outlier

J. Suhaila (✉)

Faculty of Science, Department of Mathematical Sciences,
Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: suhailasj@utm.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_30

245

1 Introduction

A flood event is commonly described by a multivariate characteristic of flood peak, volume and duration. In the past, hydrologist focused separately on these characteristic ignoring their dependence structure, but this resulted reduced accuracy in risk estimation. According to Yue et al. [1] a single variable frequency analysis could only provide limited assessment of extreme events in which it cannot provide a complete assessment of the probability occurrence if the hydrological event is characterised by a set of correlated variables. Therefore, multivariate approaches such as introducing the joint distribution of flood characteristics and employing copula models are proposed in previous studies [2, 3]. It is known that the severity of a flood not only depends on its flood peak, volume and duration, but also the shape of its hydrograph. In a hydrological framework, a hydrograph curve is often used to represent the temporal variation of flow over period of time. The information obtained from the hydrograph curve is required to determine the severity and the frequencies of extreme event.

Multivariate studies have improved the estimation accuracy in flood risk by providing the information on the dependency structure between flood characteristics; however, a flood cannot be characterised by finite number of characteristics but instead by its entire hydrograph as a curve [4]. In addition, using a limited number of characteristics could represent a loss of information in comparison to the overall available series since the total information that is available in a hydrograph is necessary for the effective planning of water resources and for the design and management of hydraulic structures [5].

The main objective of this study is to use streamflow hydrograph as functional data and employ statistical methods in the functional context. The concept of functional data may bring additional insight by looking the pattern and variation of hydrological variables in the form of smoothing curves. The developments of functional data make it possible to analyse particular data such as daily flow from a given station when describing the flow variation within a day, week, month or year. In addition, the information obtains from FDA approach could help the government in making decision that related to hydrological studies.

2 Data and Methods

Daily flow from Sg. Kelantan (Jam. Guillemard) with station code SF5721442 was used in the analysis. It is located at Kelantan River Basin with the northern latitude of $5^{\circ}45' N$ and eastern longitude of $102^{\circ}09' E$. Streamflow data in the period of 1980 to 2014 was used in the analysis.

Suppose we have a data set such as $Y_i = (y_1(t_1), y_2(t_2), \dots, y_i(t_T))'$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, T$, with $T = 365$ days, n is the number of years, and $y_i(t_j)$ is the flow measured at the day t_j of the i -th year. These discrete observed data are to be

converted into smoothing curves $x_i(t)$ as temporal functions with a base period of $T = 365$ days and k basis functions. Fourier bases are preferred since the flow data of the whole series present some seasonal variability and periodicity over the annual cycle. The choice of k can be determined to capture the flow variation. A linear combination of basis function is used for representing the functions, given as $x_i(t) = \sum_{k=1}^K d_k \phi_k(t)$ where d_k refers to the basis coefficient, ϕ_k is the known basis function while K is the size of the maximum basis required. The coefficients of the expansion d_k are determined by minimizing a least square criterion. It is essential to choose the number of basis functions that can reflect the characteristics of data. If a large number of basis functions are used, a penalty term can be added to ensure the regularity of the smooth function.

In classical statistics, measure of central tendency such as mean, median and mode is often used to describe the middle of the data set while dispersion measures such as variance and standard deviation is used to indicate the variability of the data set. Both measures are used to describe the shape of hydrograph curves. In hydrology, location curves can be used to profile and characterise a given river basin based on the behaviour or shape of flow data. Another functional method that can also be used to capture the variability of the function samples is through functional principal component analysis (FPCA).

The main idea of this extension is simply to replace vectors by functions, matrices by compact linear operators, covariance matrices by covariance operators, and scalar products in vector space by scalar products in L_2 space. After converting the data into functions, FPCA can be used to find new functions that reveal the most important type of variation in the curve data. The FPCA method maximizes the sample variance of the scores subject to orthonormal constraints. The method of FPCA is briefly described as follows:

- Let $x_i(t)$, $i = 1, 2, \dots, n$ be the functional observations obtained by smoothing the observed discrete observations $(x_i(t_1), \dots, x_i(t_T))$, $i = 1, 2, \dots, n$.
- Let $z_i(t) = x_i(t) - \bar{x}(t)$, $i = 1, 2, \dots, n$ be the centred functional observations where $\bar{x}(t)$ is the mean function. A FPCA is then applied to $z_i(t)$ to create a small set of functions, called harmonics that reveals the most important type of variation in the data.
- The first principal component $\xi_1(t)$ describes a weight function for the $z_i(t)$ that exists over the same range and accounts for the maximum variation. The first principal component yields the maximum variation in the functional principal component scores $f_{i1} = \int \xi_1(t) z_i(t) dt$ subject to the normalisation constraint $\int \xi_1(t)^2 dt = 1$.
- The next principal components $\xi_k(t)$ are obtained by maximising the variance of the corresponding scores $f_{ik} = \int \xi_k(t) z_i(t) dt$ under the constraints $\int \xi_k(s) \xi_j(s) ds = 0, k \geq 2, k \neq j$.

In order to explore, visualize and examine certain features of hydrograph curve such as outliers which cannot be captured via summary statistics, three graphical methods were introduced in this analysis. The methods were proposed by Hyndman

and Shang [6]. A rainbow plot is used for visualizing functional data while functional bagplot and functional high density region (HDR) boxplot is needed to identify outliers. Rainbow plot is a simple plot of all the data with the only added feature being a color palette based on an ordering of the data. In the functional context, their orderings are based on the functional depth or kernel density. The functional bagplot is a mapping of the bagplot of the first two robust principal component scores to the functional curves while the functional HDR boxplot is based on the bivariate HDR boxplot which is constructed using a bivariate kernel density estimate. A detail review of these three methods can be found in [6].

3 Results and Discussion

The smooth representation of flow data is done with a 365 day base period and k basis functions. It is recommended that to choose a number k which combines the quality of smoothing and a high percentage of explained variance by PCA analysis. Hence, the number $k = 183$ is chosen based on the quality of smoothing and a high percentage of explained variance. Figure 1a displays a functional hydrograph for year 2014. As shown in Fig. 1a, the highest flow was detected at the end of year at day 355th in December. Information on first derivative which represents the rate of flow is shown in Fig. 2b. A slight change in flow is observed from March to October in 2014 (from day 90th to 300th), but suddenly it resulted in high rate of change during November and December with nearly $6000 \text{ m}^3/\text{s}$.

The smooth location curves showing the mean, median and modal curves are presented in Fig. 2. The maximum flow is normally observed in the middle of November up to early January which can be considered as the Northeast Monsoon flow. Due to the contribution of heavy rainfall during the Northeast Monsoon, it is often found that heavy floods occurred during this period. This kind of flood is exhibited by the median curve which is higher than the mean and the mode. On the

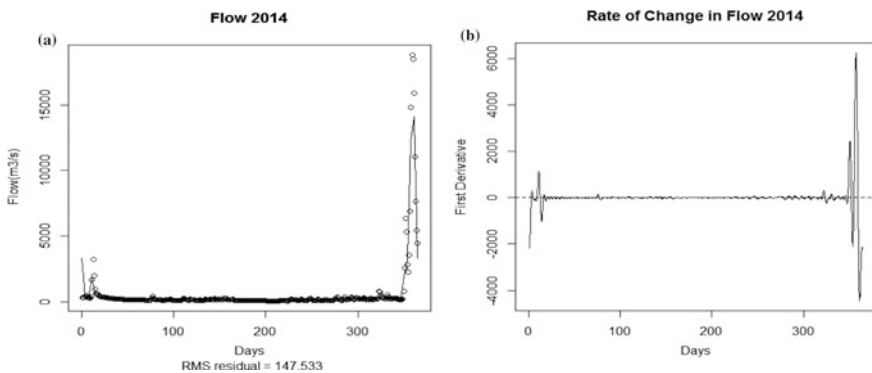
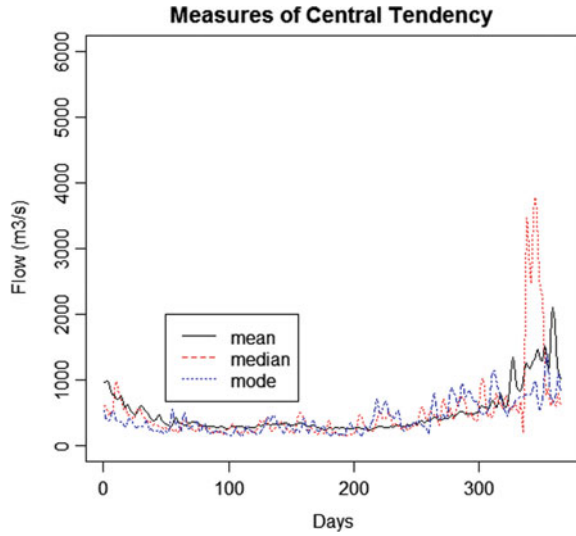


Fig. 1 a Functional hydrograph and b Rate of change in flow for year 2014

Fig. 2 Fourier smoothing location curves



other hand, low flow values are observed at the rest of the months. These location curves can be used in basin characterization rather than the usual descriptive summaries.

Five principal components are derived which are accounted for 73.8% of the total variance of the flow. The scores of the first two principal components were mapped onto Fig. 3. Several clusters can be classified according to these scores. The curves for 2014, 2009, 1988 and 1993 may be considered having a unique cluster of their own while for certain curves, they are high possibilities that they can be grouped together.

In order to justify the above unusual years, the outliers' detection methods were employed. Figure 4 displays the bivariate HDR and the associated functional HDR

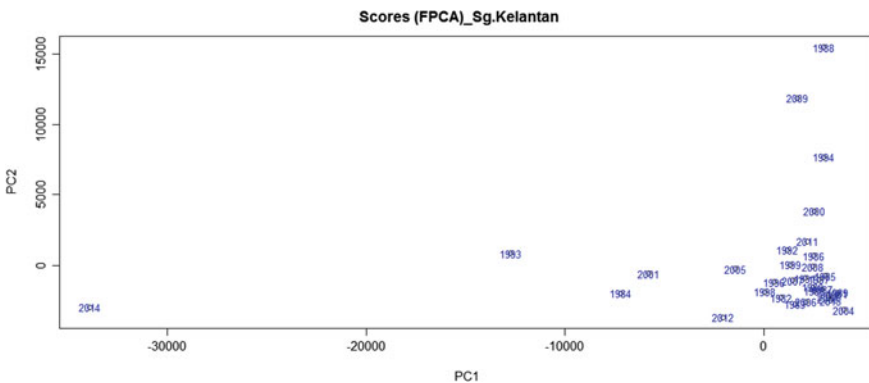


Fig. 3 Mapping of the scores of the first two principal components

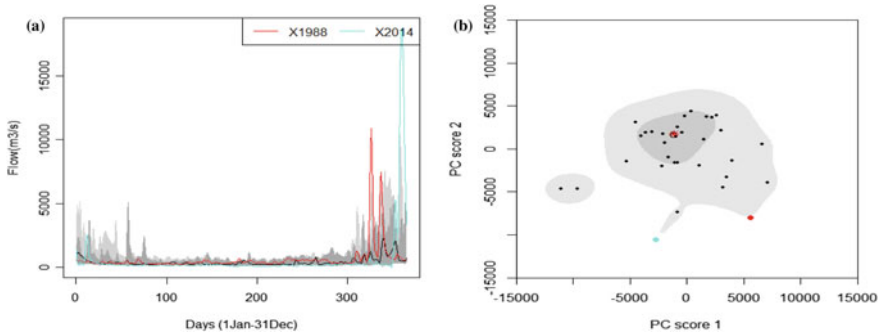


Fig. 4 a Bivariate score HDR box-plot and the corresponding b functional HDR box-plots with 99% of probability coverage

boxplots of the smooth flow curves for 99% of probability coverage. With the 99% coverage probability, the outliers detected in flow series are those curves represent 1988 and 2014. This indicates that these outliers have different magnitudes and shapes compared to the rest of data. The presence of bimodality in Fig. 4b indicates that the sample may come from two populations which need to reanalyse in future studies.

4 Conclusion

In the functional context, outlier curves are considered having different magnitudes and shapes compared to the rest of the observed curves. In the hydrological framework using the univariate and multivariate settings, the shape of the hydrograph curves is not normally considered in the analysis and cannot be captured even by using several variables. Therefore, the functional framework is more general and more flexible and can represent a large variety of hydrographs and able to make use the full information contained of the hydrograph. However, our findings are limited to exploratory analysis. Future work can be carried out in estimating the flood risk and considering inferential aspects such as modeling for prediction purposes.

References

1. Yue, S., Ouarda, T.B.M.J., Bobee, B., Legendre, P., Bruneau, P.: The Gumbel mixed model for flood frequency analysis. *J. Hydrol.* **226**, 88–100 (1999). [https://doi.org/10.1016/s00221694\(99\)00168-7](https://doi.org/10.1016/s00221694(99)00168-7)
2. Yue, S.: The bivariate lognormal distribution to model a multivariate flood episode. *Hydrol. Process.* **14**, 2575–2588 (2000)

3. Mitková, V.B., Halmová, D.: Joint modeling of flood peak discharges, volume and duration: a case study of the Danube River in Bratislava. *J. Hydrol. Hydromech.* **62**(3), 186–196 (2014). <https://doi.org/10.2478/johh-2014-0026>
4. Ternynck, C., Alaya, M., Chebana, F., Dabo-Niang, S., Ouarda, T.B.M.J.: Streamflow hydrograph classification using functional data analysis. *J. Hydrometeorol.* (2015). <https://doi.org/10.1175/jhm-d-14-0200.1>
5. Chebana, F., Dabo-Niang, S., Ouarda, T.B.M.J.: Exploratory functional flood frequency analysis and outlier detection. *Water Resour. Res.* **48**, W04514 (2012). <https://doi.org/10.1029/2011WR011040>
6. Hyndman, R.J., Shang, H.L.: Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph Stat.* **19**(1), 29–45 (2010)

Chapter 31

Bayesian Random Forest for the Classification of High-Dimensional mRNA Cancer Samples



Oyebayo Ridwan Olaniran and Mohd Asrul Affendi Bin Abdullah

Abstract The goal of many machine learning algorithms is to adequately identify the informative biomarkers in the biological samples useful for predicting disease outcome. Several algorithms have been proposed to perform this task using high-dimensional genomic messenger Ribonucleic Acid (mRNA) data. High-dimensionality poses serious problem in statistical analysis in terms of parameter estimation and inference. To address this problem, a powerful method has been developed called Random forest. Random forest was able to tackle high-dimensionality problem but it fails because it's more of computer program than a statistical learning method thus uncertainty in prediction cannot be quantified. In this paper, we develop Bayesian Random Forest (BRF) model for the classification of high-dimensional mRNA data. Bayesian procedures are the emerging solution to most applications of statistics in the recent time and in fact it has the least error rate in theory. In addition, they give appealing results in terms of parameter uncertainty, model uncertainty and data uncertainty. BRF model fitting and inference were achieved via Metropolis-Hasting (MH) MCMC algorithm. The model strength was illustrated using bake-off of 10 different mRNA cancer datasets. Results from data calibration established appreciable supremacy over competing methods.

Keywords Bayesian · Random forest · mRNA · High-dimensional · Classification

O. R. Olaniran (✉) · M. A. A. B. Abdullah
Faculty of Applied Sciences and Technology,
Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia,
Pagoh Educational Hub, 84600 Pagoh, Johor, Malaysia
e-mail: rid4stat@yahoo.com

M. A. A. B. Abdullah
e-mail: afendi@uthm.edu.my

1 Introduction

The growth in computer applications have enhanced collection and analysis of big datasets. Big datasets are often referred to as high-dimensional data in statistical parlance. The difficulties faced while analyzing big datasets has led to development of many statistical or machine learning procedures in the recent time [1]. In most areas of research especially bioinformatics, it is usual to have relatively small sample sized datasets collected on large number of features.

Random forests (RF), a tree based non-parametric method originally proposed by [2] is one of the popular methods for handling high-dimensional data, mainly because of its computational speed and high accuracy. Bayesian procedures are the emerging solution to most applications of statistics in the recent time in fact it has the least error rate in theory [3–6]. Chipman et al. [7] proposed Bayesian Additive Regression Trees (BART) which is a probabilistic approach to sum of trees model. However, BART is more of Bayesian approach to sum of trees model than to call it a Bayesian random forest. Specifically, BART did not incorporate bootstrapping of trees as in RF but a posterior distribution of trees. In addition, BART controls tree depth by imposing restrictive priors on tree with large daughter nodes. BART uses prior distribution specification as a pruning tool to avoid large trees [8]. Taddy et al. [9] proposed Bayesian forest (BF) as a nonparametric Bayesian approach to RF. They used posterior of trees instead of bootstrap of trees based on a nonparametric Bayesian model using multinomial draws. BF tried to mimic RF by replacing the bootstrapping procedure by [10] by its Bayesian counterpart (Bayesian Bootstrap, [11]). This implies BF focuses on the data generating process of RF but not its impurity measures.

Based on the aforementioned features of the Bayesian variation of Random Forest (RF), we observed that none of the existing methods fully captures the complete framework of RF and this affects their eventual results. Thus the goal of this research is to develop a complete Bayesian approach to Random Forest (RF). The method updates every aspect of RF using Bayesian reasoning. By way of example, we considered the case of binary classification with high-dimensional cancer datasets.

2 Decision Trees and Random Forests

Decisions trees is a class of methods under the broad Classification and Regression Trees (CART). The response variable of interest determines the type of model, such as decision trees if the response is categorical and regression trees if the response is continuous. CART do not have any statistical model but a set of steps called algorithm. CART modelling involves partitioning the feature space into M regions.

Formally, given training dataset $[y_i, x_{i1}, x_{i2}, \dots, x_{ip}, i = 1, 2, \dots, n]$, where y_i is a categorical outcome that assumes $k = 1, 2, \dots, K$ values and x_i is the vector of

features. CART algorithm automatically decides on the splitting variables and splitting point. After, successful partitioning of the response to R_1, R_2, \dots, R_M regions, the closest form of model that CART assumes is;

$$y = \sum_{m=1}^M \beta_m I(x \in R_m) \quad (1)$$

where β_m is a constant in region m . Estimating β_m requires the computation of an impurity function. For classification case, the commonly used impurity functions are Misclassification Error Rate (MER), Gini Index, and deviance [12].

Random Forest (RF) update built CART trees in two steps; (i) bootstrapping the training dataset J times to obtain a total of J trees (ii) Subsampling $l < p$ features without replacement at each split step in each j tree. Thus given a CART model $\mathfrak{S}(\hat{\beta}_m : x \in R_m)$, RF model is;

$$\hat{y} = \sum_{j=1}^J \mathfrak{S}_j(\hat{\beta}_m : x \in R_m) \quad (2)$$

RF has two tuning parameters, the number of trees J and number of subsampled features l . Breiman [2] suggested using at least $J = 200$ and $l = \sqrt{p}$ for classification task.

3 Bayesian Random Forests

Section 2 established the weakness of RF as the necessary tuning parameters are not chosen by any probabilistic law. The approach is nothing but a trial and error hence often referred to as black box method. A quick solution to avert trial and error is to select the tuning parameters by cross validation but at the expense of computation time. Therefore, the focus of this research is to modify RF forest by updating the two steps in (2) via Bayesian approach. For the bootstrapping step, we propose the Bayesian Simple Random Sampling With Replacement (BSRSWR) described by the posterior distribution in (3);

$$P(\pi|a, b) = \frac{\Gamma(n+a+b)}{\Gamma(a+1)\Gamma(b+n-1)} \pi^a (1-\pi)^{b+n}, 0 \leq \pi \leq 1 \quad (3)$$

where π is the probability of selecting any $i \in n$ in each j step, $\Gamma(d)$ is the gamma function evaluated at d , a is the prior expected number of times any $i \in n$ could be selected and b is its complement. It's clear that the density function in (3) is a resemblance of $Beta(a+1, b+n-1)$. A weighted CART tree $\mathfrak{S}(\hat{\beta}_m : x \in R_m)$

can be obtained using $\omega = \text{Beta}(a + 1, b + n - 1) \forall i \in n$. Similarly, for the subsampling of $l < p$ steps, we propose Bayesian Simple Random Sampling Without Replacement (BSRSWOR) with posterior density given in (4);

$$P(V|h, l, p, S, T) = \frac{\binom{S+V}{S+1} \binom{T+p-V}{T+l-h}}{\binom{S+T+p+1}{S+T+l+1}}, h \leq v \leq p-l+h \quad (4)$$

where V is the number of relevant features whose posterior is sought, h is the sample realization of relevant features, p is the total number of features, l is the number of subsampled features as in RF, S is the prior number of relevant features and T is the prior number of irrelevant features. If we denote the posterior density in (4) as δ , we can use δ to obtain a weighted splitting procedure where each impurity used at every splitting stage would be weighed by δ . For a Gini index impurity, we propose a weighted Gini index ϑ ;

$$\vartheta = \sum_{k=1}^K (1 - \delta) \hat{p}_{mk} (1 - \hat{p}_{mk}) \quad (5)$$

where \hat{p}_{mk} is the estimated class probability at each node m . The variable with weight $\delta \rightarrow 1$, will correspond to variable with minimal unweighted Gini index and therefore useful for further splitting step. If on the other hand $\delta \rightarrow 0$, implies the variable is not useful and therefore expected to yield a maximal unweighted Gini index. In this case, the proposed weighted Gini index returns the unweighted Gini index so that the variable is dropped at the splitting stage.

4 Application to Cancer Datasets

In this section we illustrate the application of Bayesian Random Forest (BRF) on published real data. We use the ‘‘bake-off,’’ approach of [7] to study the predictive performance comparison of BRF with competing methods on 10 different real cancer data sets. Table 1 presents the data set which is a subset of 22 datasets from package ‘‘datamicroarray’’ in R [13]. For each of the 10 data sets, we created 10 independent train/test splits by randomly selecting 9/10 of the data as a training set and the remaining 1/10 as a test set. Thus, $10 \times 10 = 100$ test/train splits were created (Fig. 1).

Based on each training set, each method was then used to predict the corresponding test set and evaluated on the basis of its predictive misclassification error rate and accuracy. The competing methods used alongside with BRF include Random Forest (RF), Bayesian Forest (BF), Gradient Boosting Machine

Table 1 The 10 datasets used in the bake-off and their associated dimensions

Cancer type	n	p
Colon Cancer	62	2000
Breast Cancer 1	168	2905
Lung Cancer	181	12533
Prostate Cancer	102	12600
Breast Cancer 2	49	7129
Leukemia Cancer 1	111	12625
Lymphoma Cancer	58	6817
CNS Tumor	60	7128
Myeloma Cancer	173	12625
Leukemia Cancer 2	50	10100

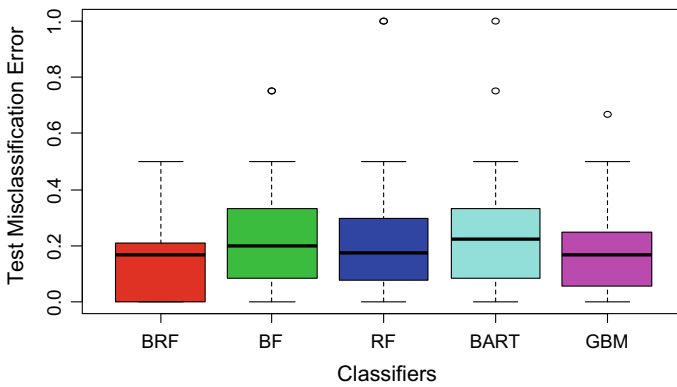


Fig. 1 Boxplot of test misclassification error rate (MER) for the five methods over 100 train/test partitions. BRF has the least MER with 25% of the MER equal zero. Also, the absence of outlying point(s) in BRF indicate that it is more stable than its competitors

(GBM) and Bayesian Additive Regression Trees (BART). Of all the five methods compared, GBM is the only frequentist method and also a major competitor of RF within the same classifier class [12].

5 Discussion and Conclusion

In this paper, we have established the weakness of RF and possible way to improve by formulating a probabilistic approach to tree sampling and split selection. We demonstrated the applicability of the method using 10 real cancer data sets. The individual and overall results in Table 2 show that in almost all the data sets used BRF accuracy is relative higher than its competitors. The result further shows that in any datasets used, BRF accuracy is bounded below at RF accuracy. This implies

Table 2 Accuracy of the methods in each and all of the 10 datasets

Cancer type	BRF	BF	RF	BART	GBM
Colon Cancer	87.14	74.76	82.38	72.62	80.71
Breast Cancer 1	79.19	73.27	75.63	76.25	76.80
Lung Cancer	99.44	98.89	99.44	99.44	97.78
Prostate Cancer	90.18	89.18	90.18	88.18	90.18
Breast Cancer 2	80.67	63.67	56.50	59.00	88.00
Leukemia Cancer 1	95.26	92.18	92.18	92.18	94.55
Lymphoma Cancer	90.36	82.86	86.61	77.86	88.21
CNS Tumor	74.33	67.05	64.81	66.81	61.48
Myeloma Cancer	80.99	79.26	79.26	79.26	78.01
Leukemia Cancer 2	79.67	66.67	70.33	68.33	69.00
Mean	85.72	78.78	79.73	77.99	82.47
SEM	11.98	15.55	17.39	16.47	14.37
25%	100.00	91.67	92.31	91.67	94.44
75%	79.44	66.67	70.44	66.67	75.00

that BRF accuracy will in most cases be higher than RF accuracy and at least RF accuracy. Therefore, it can be concluded that the Bayesian weighing scheme developed indeed correct the RF weakness.

Funding This work was supported by Universiti Tun Hussein Onn, Malaysia [grant numbers Vot, U607].

References

1. Lynch, C.: Big data: how do your data grow? *Nature* **455**(7209), 28–29 (2008)
2. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
3. Olaniran, O.R., Yahya, W.B.: Bayesian hypothesis testing of two normal samples using bootstrap prior technique. *J. Mod. Appl. Stat. Methods* **16**(2), 618–638 (2017). <https://doi.org/10.22237/jmasm/1509496440>
4. Olaniran, O.R., Olaniran, S.F., Yahya, W.B., Banjoko, A.W., Garba, M.K., Amusa, L.B., Gatta, N.F.: Improved Bayesian feature selection and classification methods using bootstrap prior techniques. *Anale. Seria Informatică* **14**(2), 46–52 (2016)
5. Olaniran, O.R., Affendi, M.A.: Bayesian analysis of extended cox model with time-varying covariates using bootstrap prior. *J. Mod. Appl. Stat. Methods* (2017) (in press)
6. Yahya, W.B., Olaniran, O.R., Ige, S.O.: On Bayesian conjugate normal linear regression and ordinary least square regression methods: a monte carlo study. *Ilorin J. Sci.* **1**(1), 216–227 (2014)
7. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010)
8. Pratola, M.T.: Efficient metropolis-hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Anal.* **11**(3), 885–911 (2016)

9. Taddy, M., Chen, C.S., Yu, J., Wyle, M.: Bayesian and empirical Bayesian forests (2015). [arXiv:1502.02312](https://arxiv.org/abs/1502.02312)
10. Efron, B.: Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics, pp. 569–593. Springer, New York (1992)
11. Rubin, D.: The Bayesian bootstrap. *Ann. Stat.* **9**, 130–134 (1981)
12. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Prediction, Inference and Data Mining*, 2nd edn. Springer, New York (2011)
13. Ramey, J.A.: Datamicroarray: collection of data sets for classification. <https://github.com/ramhiser/datamicroarray>, <http://ramhiser.com> (2016)

Chapter 32

Bayesian Statistical Modeling: Comparisons Between Poisson and Its Zero-Inflated Regression Model



Muhammad' Afif Amir Husin and Mohd Fadzli Mohd Fuzi

Abstract In this paper, we fit Poisson regression model and its Zero-Inflated version in Bayesian framework, to Malaysian motor vehicle claim count data, in order to study the differences between the models. The models are tested to Third Party Property Damage coverage data which contains sizeable amount of zero claims. The posterior distributions for both models are produced using Markov Chain Monte Carlo (MCMC) simulation to estimate their parameters. The results show that the Bayesian Zero-Inflated Poisson model has superiority over the standard Bayesian Poisson model based on the Deviance Information Criterion (DIC) values.

1 Introduction

In the general insurance business, the pricing of motor vehicle premium is based on risk factors or risk characteristics of drivers and vehicle and this method is called risk-based premium. Common risk characteristics or factors used for example are gender, age of driver, age of vehicle, vehicle model, vehicle cubic capacity (cc) and the geographical location of where the vehicle is registered. These characteristics can be linked to the claim frequency via regression models. In statistics, there are two main school of thought or framework in doing analysis, which are frequentist and Bayesian. In this paper, we will look at the modelling of claim frequency data using the Bayesian framework.

M. A. A. Husin (✉) · M. F. M. Fuzi
Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai, Malaysia
e-mail: muhammad.afif.amir.husin@gmail.com

M. F. M. Fuzi
e-mail: mohdfadzli@usim.edu.my

© Springer Nature Singapore Pte Ltd. 2019
L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference
on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_32

2 Bayesian Framework

In practice, regression models particularly Generalized Linear Model (GLM) are used to model count data. GLM is an extension of linear models and these include the exponential family of distributions among others Poisson, Binomial and Normal distributions.

Scollnik [1] used Bayesian method to fix the misrepresentation of Generalized Truncated Poisson which have been widely used by others. He applied the method to railway accidents data. Meanwhile, Denuit and Lang [2] modeled non-life insurance ratemaking by implementing Bayesian Generalized Additive Model to motor third party liability insurance data.

Bermudez and Karlis [3] implemented Bayesian multivariate Poisson models for insurance ratemaking. They proved that it is appropriate to use this model because it is possible that type of coverages for certain type of insurance have correlation between each other.

Donnelly and Wuthrich [4] used Bayesian model to predict the determinants of disability insurance. The study proposed to examine the difference between models with economic indicator and the one without those indicators. At the end of the study, authors found that credit spreads were the main indicator of claim development. Sanchez and Deniz [5] applied Bayesian methodology towards Australian automobile insurance dataset from 2004 to 2005. The results showed that Bayesian Power Series Distribution perform better than the classical model.

Fuzi et al. [6] applied Poisson, Negative Binomial, Generalized Poisson and quantile regression models in the Bayesian framework to Malaysian motor insurance claims frequency data and found that Generalized Poisson and median quantile regression fit the data well.

This study intended to focus on two models namely Poisson and Zero-Inflated Poisson. We aim to analyze the difference between the zero-inflated models with respect to its original model.

3 Methodology

Bayesian regression works by producing posterior distribution being simulated to estimate the parameters. The posterior distribution is generated by multiplying the likelihood function of a distribution with their prior distributions. In this case, the posterior distribution produced is the conditional probability of parameters given the frequency of claim. The data set used in this study is from the Third Party Property Damage coverage data, which consists of 1.2 million policyholders gathered from ten insurance companies provided by Insurance Services Malaysia (ISM).

3.1 Poisson Regression Model

Let Y_i be the frequency of accident for subject i and λ_i be the mean of outcomes with x_i is the vector of covariates, then the Poisson regression is as follows:

$$\Pr(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \quad (1)$$

$$\Pr(Y_i = y_i | \lambda_i) = \exp(y_i \ln \lambda_i - \lambda_i - \ln \Gamma(y_i + 1)) \quad (2)$$

The likelihood function is given by

$$l(y|\lambda) = \exp \left\{ \sum_{i=1}^n [y_i \ln \lambda_i - \lambda_i - \ln \Gamma(y_i + 1)] \right\} \quad (3)$$

$$= \exp \left\{ \sum_{i=1}^n [y_i (\ln e_i + x_i^T \beta) - e_i \exp(x_i^T \beta) - \ln \Gamma(y_i + 1) y_i] \right\} \quad (4)$$

In order to incorporate covariates and ensure positivity, the mean or the fitted value is assume to follow log-link, $E(Y_i|x_i) = \lambda_i = e_i \exp(x_i^T \beta)$ where e_i denotes exposure, x_i^T denotes vector of explanatory variables while $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)$

The prior distribution is normally distributed, given by

$$\pi(\beta|\beta_0, \sigma^2) = \exp \left[-\frac{1}{2\sigma^2} (\beta - \beta_0)^T (\beta - \beta_0) - \frac{p}{2} \ln \sigma^2 \right] \quad (5)$$

The posterior distribution is given by

This study will use Normal prior distribution for β and $\beta_0 = 0$ and $\tau = 0.001$ indicative of non-informative prior, where $\tau = \sigma^{-2}$ is known as accuracy parameter.

3.2 Zero-Inflated Poisson Regression Model

Let Y_i be variable for claim count with i th risk class, $i = 1, 2, \dots, n$, with n is the number of risk classes. Assuming ω is the probability that the response variable y_i for the i th risk class = 0, the distribution takes the form

$$\Pr(Y_i = y_i | \lambda_i, \omega) = \begin{cases} \omega + (1 - \omega) \exp(-\lambda_i), & y_i = 0, \\ (1 - \omega) \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, & y_i = 1, 2, \dots, \end{cases} \quad (6)$$

The likelihood function is given by

$$l(y|\lambda, \omega) = \prod_{i=1}^n \left[I_{y_i=0} \{ \omega + (1 - \omega) \exp(-\lambda_i) \} + I_{y_i > 0} (1 - \omega) \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right] \tag{7}$$

In Bayesian ZIP regression model, it is assumed that the mean parameter is linked to the covariates via log-linear predictor, $\lambda_i = e_i \exp x_i^T \beta$ and $\omega_i = e_i \cdot \frac{\exp Z_i^T \alpha}{1 + \exp Z_i^T \alpha}$

$$l(y|\lambda, \omega) = \prod_{j=1}^k \left[e_i \cdot \frac{\exp z_i^T \alpha}{1 + \exp z_i^T \alpha} + \left(1 - e_i \cdot \frac{\exp z_i^T \alpha}{1 + \exp z_i^T \alpha} \right) \exp -e_i \exp(x_i^T \beta) \right] \times \prod_{j=k+1}^n \left[\left(1 - e_i \cdot \frac{\exp z_i^T \alpha}{1 + \exp z_i^T \alpha} \right) \exp -e_i \exp(x_i^T \beta) \frac{(e_i \exp(x_i^T \beta))^{y_j}}{y_j!} \right] \tag{8}$$

This study will use Normal prior distribution for β and $\beta_0 = 0$ and $\tau = 0.0001$ where $\tau = \sigma^{-2}$ which known as accuracy parameter.

3.3 Markov Chain Monte Carlo (MCMC)

The process of estimating parameters of each covariate are done by using MCMC. This study will use Gibbs Sampler which introduced by Geman and Geman (1984) to optimize image sampling. The uniqueness of Gibbs Sampling is that the technique uses direct simulations from full conditionals. The validity of the algorithm is verified through the convergence to the stationary distribution.

3.4 Model Comparison

For comparison purpose, DIC will be computed for each model after MCMC is done. A regression model with a smaller value of DIC is favorable compared to the larger one. DIC can be defined as

$$DIC = p_D + \bar{D} \tag{9}$$

$$D(\theta) = -2 \log p(y|\theta) \tag{10}$$

\bar{D} represents posterior mean deviance while p_D represents the effective number of parameters of the model and calculated by subtracting deviance evaluated at the posterior from posterior mean of deviance while $p(y|\theta)$ is the data distribution. DIC is made up of two components: mean deviance which measure how well the data fit with the model and the other component which measure the complexity of the model.

4 Discussions

Table 1 shows descriptive statistics of the claim frequency. All claim frequency lies between 0 and 1077 while on average, it is equal to 31.76. For median of claim frequency, the value is 4.00 which indicate that half of the claim frequency concentrated below 4 while the other half scattered between 4 and 1077.

The posterior distribution results for each Bayesian regression model differs in terms of the number of iterations to achieve equilibrium distribution. Poisson distribution required 100,000 iterations while ZIP needed 300,000 iterations. ZIP needed more iterations mainly because it has more parameters than Poisson. Both models have discarded 25,000 samples as burn-in and applied one as thinning interval.

Figures 1 and 2 shows an example of traceplot of a covariate Local 2 for Poisson and ZIP's count part respectively. Its consistencies across iterations indicated that the posterior samples have mixed well and signaled that convergence is attained without problem. Table 1 gives the value parameter estimates, standard deviation, upper bound and lower bound for each covariate for both models.

Table 2 shows the posterior distribution results for both models. The parameter values are slightly different for all parameters for both models meanwhile the standard deviation value for all covariates are almost the same which results in only very small difference between Poisson and ZIP models.

Table 1 Descriptive statistics of claim frequency

Item	Value
Mean	31.76
Median	4.00
Max	1077.00
Min	0.00
3rd quartile	23.50
Standard deviation	81.53

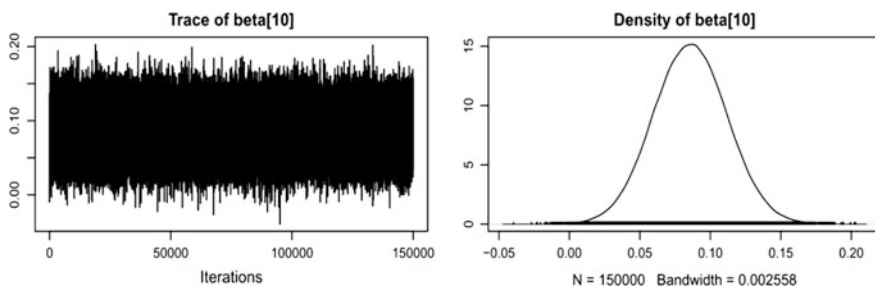


Fig. 1 Traceplot for Poisson model

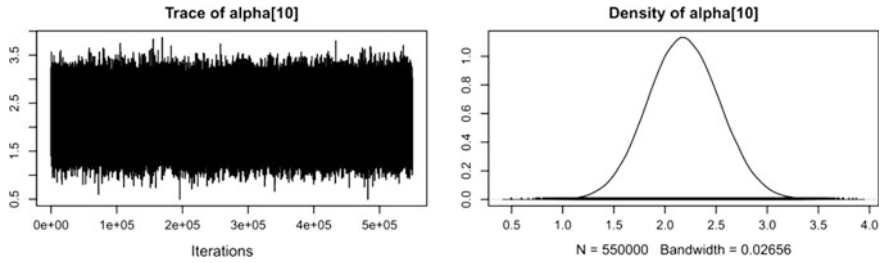


Fig. 2 Traceplot for ZIP model

Table 2 Posterior distribution results for Poisson and ZIP models

Risk class	Alpha/ Beta	Poisson				ZIP			
		Estimate	Standard deviation	Lower bound	Upper bound	Estimate	Standard deviation	Lower bound	Upper bound
<i>Zero-Inflated Part</i>									
Intercept	α_1	-	-	-	-	-4.549	0.431	-5.395	-3.705
Non-comprehensive	α_2	-	-	-	-	-0.045	0.216	-0.471	0.375
2-3 years	α_3	-	-	-	-	1.842	0.318	1.225	2.470
4-5 years	α_4	-	-	-	-	1.584	0.314	0.974	2.205
6-7 years	α_5	-	-	-	-	1.302	0.323	0.674	1.942
8+ years	α_6	-	-	-	-	1.495	0.372	0.772	2.229
1001-1300 cc	α_7	-	-	-	-	0.908	0.306	0.308	1.508
1301-1500 cc	α_8	-	-	-	-	2.100	0.354	1.409	2.796
1501-1800 cc	α_9	-	-	-	-	1.883	0.323	1.255	2.522
1801+ cc	α_{10}	-	-	-	-	2.187	0.352	1.506	2.884
Local 2	α_{11}	-	-	-	-	-1.670	0.372	-2.408	-0.949
Foreign 1	α_{12}	-	-	-	-	-0.648	0.381	-1.401	0.094
Foreign 2	α_{13}	-	-	-	-	-0.607	0.363	-1.323	0.100
Foreign 3	α_{14}	-	-	-	-	-1.876	0.438	-2.744	-1.031
East	α_{15}	-	-	-	-	-0.416	0.312	-1.033	0.192
Central	α_{16}	-	-	-	-	1.232	0.342	0.565	1.906
South	α_{17}	-	-	-	-	0.530	0.318	-0.091	1.154
East Malaysia	α_{18}	-	-	-	-	0.047	0.318	-0.577	0.671
<i>Count Part</i>									
Intercept	β_1	-4.879	0.040	-4.957	-4.800	-0.416	0.094	-0.599	-0.228
Non-comprehensive	β_2	0.955	0.015	0.928	0.983	0.458	0.082	0.301	0.629
2-3 years	β_3	0.903	0.023	0.857	0.948	0.426	0.095	0.251	0.627
4-5 years	β_4	1.100	0.024	1.054	1.146	0.021	0.055	-0.089	0.136
6-7 years	β_5	0.928	0.023	0.883	0.973	-0.008	0.092	-0.202	0.164
8+ years	β_6	0.648	0.023	0.604	0.692	0.040	0.054	-0.061	0.154
1001-1300 cc	β_7	0.024	0.033	-0.041	0.088	0.644	0.076	0.497	0.793
1301-1500 cc	β_8	0.246	0.034	0.180	0.312	0.382	0.077	0.227	0.530
1501-1800 cc	β_9	0.649	0.034	0.582	0.716	-0.054	0.074	-0.221	0.074

(continued)

Table 2 (continued)

Risk class	Alpha/ Beta	Poisson				ZIP			
		Estimate	Standard deviation	Lower bound	Upper bound	Estimate	Standard deviation	Lower bound	Upper bound
1801+ cc	β_{10}	0.507	0.035	0.438	0.576	-0.042	0.072	-0.191	0.097
Local 2	β_{11}	-0.289	0.033	-0.353	-0.225	0.068	0.091	-0.102	0.260
Foreign 1	β_{12}	-0.469	0.016	-0.500	-0.438	-1.034	0.106	-1.227	-0.805
Foreign 2	β_{13}	-0.431	0.024	-0.477	-0.384	-1.034	0.079	-1.167	-0.857
Foreign 3	β_{14}	-0.711	0.033	-0.775	-0.648	0.062	0.080	-0.086	0.233
East	β_{15}	-0.040	0.030	-0.098	0.018	0.404	0.081	0.250	0.576
Central	β_{16}	0.709	0.016	0.677	0.740	-0.061	0.076	-0.216	0.084
South	β_{17}	0.339	0.020	0.301	0.378	0.392	0.074	0.240	0.536
East Malaysia	β_{18}	0.174	0.021	0.133	0.216	0.020	0.059	-0.089	0.149
DIC		17121.8				-1201.48			

DIC values point out that ZIP model performs better than Poisson model. This is based on the lower value of DIC have been showed by ZIP model with -1201.48 in comparison to 17121.8 showed by Poisson model.

It is noticeable that the parameters for Poisson and ZIP’s zero-inflated part that contribute lowest risk to the claim frequency are Local 2, Foreign 1, Foreign 2, Foreign 3 and East. Whereas for ZIP’s count part, parameters that lower the claim frequency are Intercept, 1501–1800 cc, 1800+ cc, Foreign 1, Foreign 2 and Central. For Poisson model, the insignificant parameters are 1001–1300 cc and East. On the other hand, ZIP model’s count part produced 9 significant parameters which are Intercept, Non-comprehensive, 2–3 years, 1001–1300 cc, 1301–1500 cc, Foreign 1, Foreign 2, East and South.

5 Conclusions

This paper has applied two different models namely Bayesian Poisson and Zero-Inflated Poisson to TPPD data set of Malaysian motor insurance claims data. Both are done in the framework of the Bayesian regression model.

The posterior distribution is produced by the multiplication of prior distribution with the likelihood function for every model and simulated by using MCMC algorithm to estimate parameters for each model. At the end of MCMC simulation, DIC value is calculated. ZIP model has outperformed Poisson model according to DIC value. This is due to the characteristic of Poisson which assumes variance value is the same value with mean. Meanwhile, ZIP has the ability to capture excess zeros in comparison to Poisson model.

The DIC value of ZIP is -1201.48 while for Poisson is 17121.8 which shows that ZIP is a superior model than Poisson for this data.

References

1. Scollnik, D.P.M.: A Bayesian analysis of a simultaneous equations model for insurance rate-making. *Insur. Math. Econ.* **12**(3), 265–286 (1993). [https://doi.org/10.1016/0167-6687\(93\)90238-K](https://doi.org/10.1016/0167-6687(93)90238-K)
2. Denuit, M., Lang, S.: Non-life rate-making with Bayesian GAMs. *Insur. Math. Econ.* **35**(3), 627–647 (2004). <https://doi.org/10.1016/j.insmatheco.2004.08.001>
3. Bermúdez, L., Karlis, D.: Bayesian multivariate Poisson models for insurance ratemaking. *Insur. Math. Econ.* **48**(2), 226–236 (2011). <https://doi.org/10.1016/j.insmatheco.2010.11.001>
4. Donnelly, C., Wüthrich, M.V.: Bayesian prediction of disability insurance frequencies using economic indicators. *Ann. Actuar. Sci.* **6**(2), 381–400 (2012)
5. Pérez-Sánchez, J.M., Gómez-Déniz, E.: Simulating Posterior Distributions for Zero-Inflated Automobile Insurance Data, 47092 (2015)
6. Fuzi, M.F.M., Jemain, A.A., Ismail, N.: Bayesian quantile regression model for claim count data. *Insur. Math. Econ.* **66**, 124–137 (2016). <https://doi.org/10.1016/j.insmatheco.2015.11.004>

Chapter 33

BayesRandomForest: An R Implementation of Bayesian Random Forest for Regression Analysis of High-Dimensional Data



Oyebayo Ridwan Olaniran and Mohd Asrul Affendi Bin Abdullah

Abstract This paper presents methods of Bayesian inference for Random Forest (RF) procedure with high-dimensional data. The new methods termed Bayesian Random Forest (BRF) is developed to tackle sparsity in regression analysis of high-dimensional data. The bootstrap sampling and choosing of subsample variable size (*mtry*) procedures used by RF are replaced with the full Bayesian inference of binomial and hypergeometric random sampling from independent and dependent finite populations. Furthermore, the individual tree parameter estimates in the forest are obtained using a Metropolis-Hasting algorithm to achieve efficient posterior inference. We also introduced the application of the procedure in R via package *BayesRandomForest*. Monte-Carlo simulations of the Friedman five dimensional dataset of varying dimensions were used to demonstrate BRF relative efficiency with competing methods. Results from the simulations revealed that BRF is more efficient than the competing frequentist and Bayesian methods.

Keywords Random forest · Bayesian additive regression trees · High-dimensional · R

1 Introduction

The ensemble of tree-based methods has become popular choices for predicting qualitative response (classification) and quantitative response (regression) [1]. Ensemble methods train multiple learners to construct one learner from training

O. R. Olaniran (✉) · M. A. A. B. Abdullah

Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Johor, Malaysia

e-mail: rid4stat@yahoo.com

M. A. A. B. Abdullah

e-mail: afendi@uthm.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, https://doi.org/10.1007/978-981-13-7279-7_33

data. Random Forest (RF, [2]) and Gradient Boosting Machine (GBM, [3]) are the two well-established ensemble-based methods. The two methods focused on improving the unstable prediction problem in Classification and Regression Trees (CART, [2]). Apart from RF and GBM, other recently developed ensemble of tree-based methods are Bayesian Additive Regression Trees (BART, [4]), dynamic trees [5], Bayesian Forest and Empirical Bayesian Forest (BF; EBF, [6]) and Bayesian Additive Regression Trees using Bayesian Model Averaging (BART-BMA, [7]). [7] conducted a simulation study using [8] data initially motivated by Multivariate Adaptive Regression Splines (MARS) to compare the predictive performance of RF, BART and BART-BMA. The simulation results revealed that BART is only better than RF regarding computation time.

RF procedure requires selection of a bootstrapped sample of training data ($n \times p$) and the subsample of feature set p to create splits used in the split selection stage. The subsampled feature is often held fixed as $\approx \sqrt{p}$ for classification or $\approx p/3$ for regression. This subsample sizes do not take into account the number of relevant features in the entire feature set, thus the chance of selecting irrelevant features increases with increased p . Therefore, using the same data configuration, the predictive performance of RF reduces with increasing p . In this paper, we extend RF by introducing Bayesian weighted sampling and splitting and called it Bayesian Random Forest (BRF). The weighted splitting is achieved via Bayesian inference of the two sampling procedures involved in RF. We developed a fully Bayesian ensemble of tree procedure that is similar in spirit to RF. We also implemented the procedure in an ongoing R package “BayesRandomForest”.

2 Overview of Bayesian Random Forest

Bayesian Random Forest (BRF) is a Bayesian implementation of the nonparametric function estimates obtainable from regression trees. Given the training dataset $D = [Y_i, x_{i1}, x_{i2}, \dots, x_{ip}, i = 1, 2, \dots, n]$, where y_i assumes continuous values and x_i is the vector of features, BRF can be describe as;

$$Y = \sum_{j=1}^J \mathfrak{S}_j(\beta_m : x \in R_m) + \varepsilon \quad (1)$$

where β_m is an estimate of Y in region m [R_m], J is the number of trees in the forest, $\mathfrak{S}_j(\beta_m : x \in R_m)$ is a single regression tree, ε is the random noise that occurs in estimating β_m and its assumed to be independent and identically Gaussian distributed with mean zero and variance σ^2 over all trees. BRF model is very much similar to BART [4] but differs in terms of prior specification and posterior estimation approach.

3 Priors and Posterior Specification for Bayesian Random Forest

BRF has three major parameters which can be attributed to its ensemble nature. The first parameter is attributed to model uncertainty, which in this case is tree \mathfrak{S} uncertainty. Here we propose a uniform prior $\mathfrak{S}_0 \sim U(0, 1)$ such that $P(\mathfrak{S}_0) = 1$ for any candidate tree. We used this prior specification to retain the average weighing procedure of RF so that each tree \mathfrak{S} has equal right. The second form of prior is prior on terminal node parameter β_m , here we propose Gaussian prior $N(\mu, \sigma_\mu^2)$. We adapted the bootstrap prior technique [9–11] to obtain the prior hyperparameters μ and σ_μ^2 for each tree. For the parameter σ^2 , we propose the standard gamma default prior $G(\alpha, \theta)$ [12] with $\alpha = \theta$ such that $E[\sigma^2|G(\alpha, \theta)] = 1$. The complete prior specification for BRF is thus;

$$P(\mathfrak{S}_1, \mathfrak{S}_2, \dots, \mathfrak{S}_m) = \prod_{j=1}^J P(\mathfrak{S}_j, \beta_{jm}) P(\sigma^2) \quad (2)$$

$$P(\mathfrak{S}_1, \mathfrak{S}_2, \dots, \mathfrak{S}_m) = \prod_{j=1}^J P(\beta_{jm}) P(\sigma^2) \quad (3)$$

(3) follows from (2) since $P(\mathfrak{S}_0) = 1$. The posterior distribution using $L(\mathfrak{S}_1, \mathfrak{S}_2, \dots, \mathfrak{S}_m|y, x)$ and (3) is then obtain via Metropolis Hasting (MCMC) algorithm [13].

Now to mimic RF completely, we also specified some procedural priors similar in spirit to bootstrapping and features subsampling in RF. For the two procedures, we proposed Bayesian simple random sampling with replacement and Bayesian simple random sampling without replacement with posterior densities given in (3) and (5):

$$P(\pi|a, b) = \frac{\Gamma(n+a+b)}{\Gamma(a+1)\Gamma(b+n-1)} \pi^a (1-\pi)^{b+n}, \quad 0 \leq \pi \leq 1 \quad (4)$$

$$P(V|h, l, p, S, T) = \frac{\binom{S+V}{S+1} \binom{T+p-V}{T+l-h}}{\binom{S+T+p+1}{S+T+l+1}}, \quad h \leq v \leq p-l+h \quad (5)$$

where π is the probability of selecting any $i \in n$ in each j step, $\Gamma(d)$ is the gamma function evaluated at d , a is the prior expected number of times any $i \in n$ could be selected, b is its complement, V is the number of relevant features whose posterior is sought, h is the sample realization of relevant features, p is the total number of

features, l is the number of subsampled features as in RF, S is the prior number of relevant features and T is the prior number of irrelevant features. If we denote the posterior density in (4) and (5) as ω and δ , we then obtain a weighted Bayesian regression tree at each j step by weighing the data by ω and then weighing the impurity at each split by δ . For a Sum Squares Error (SSE) impurity, we propose a weighted impurity using;

$$SSE(\delta) = (1 - \delta) \left[\sum_{i=1}^{n_m} (y_i - \hat{\beta}_m)^2 \right] \quad (6)$$

where $\hat{\beta}_m$ is the posterior mean of β_m at each node m . The variable with weight $\delta \rightarrow 1$, will correspond to variable with minimal unweighted $SSE(\delta)$ and therefore useful for further splitting step.

4 Friedman Five Dimensional Data

Following [4, 6], [8] simulated data was used to compare the results of BRF, BF, RF, BART and GBM. The simulated datasets where x_1, \dots, x_p are $iid \sim U(0, 1)$ random variables and $\varepsilon \sim N(0, 1)$ were formulated as;

$$y = 10\sin(x_1x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon \quad (7)$$

The five methods were compared over various dataset sizes with five relevant features $[x_1, \dots, x_5]$ and complement of $p = [100, 500, 1000, 5000, 10000]$ as the irrelevant features. The predictive performance of the methods was assessed with 10 folds cross validation of Root Mean Squared Error (RMSE) at sample size $n = 50, 100$. All analyses were carried out using R with newly developed function ‘‘BayesRandomForest’’ accessible in [14] for BRF, ‘‘bartMachine’’ [1] for BART, ‘‘gbm’’ [15] for GBM, ‘‘randomForest’’ [16] for RF and ‘‘ranger’’ for [BF].

Table 1 shows the 95% credible interval for Bayesian-based methods (BRF, BF, BART) and confidence interval for frequentist-based methods (GBM and RF) of RMSE. The intervals are computed using the bake-off 10-folds cross-validation of $10 \times n$. The results show that 95% width for BRF is the lowest when compared to other methods. The least performing method is BART with the maximum width. The next in category in terms of stability of RMSE is RF. The poor performance of BART with increasing p in Figs. 1 and 2 is attributed to the nature of tree growth which involves using all features for each tree. This lead to overfitting and eventual poor performance in out of sample validation. Furthermore, the existing robust method to large p with low relevant feature is GBM because of its internally embedded feature selection procedure [14]. BRF results challenged this claim with stable and better results at sample size $n = 50, 100$.

Table 1 95% Credible and confidence interval of RMSE at sample size $n = 50, 100$

Method	$n = 50$			$n = 100$		
	2.5%	97.5%	Width	2.5%	97.5%	Width
BRF	15.32	25.26	9.94	9.82	12.17	2.36
BF	28.50	33.03	4.53	24.24	32.41	8.18
RF	24.77	31.44	6.67	14.46	21.63	7.17
BART	23.18	32.53	9.36	8.41	34.01	25.61
GBM	19.92	36.70	16.77	6.71	16.88	10.17

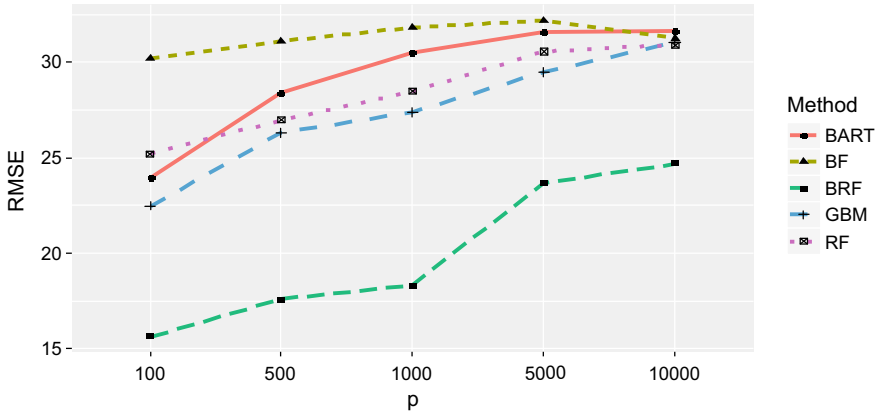


Fig. 1 RMSE with increasing p at sample size $n = 50$, as expected increasing p reduces the hypergeometric probability which increases RMSE of RF (purple dotted lines). Similarly, GBM, BRF and BART are affected with the increase but the effect is minimal on BRF. BF tends to be stable over p but the RMSE is on the high side

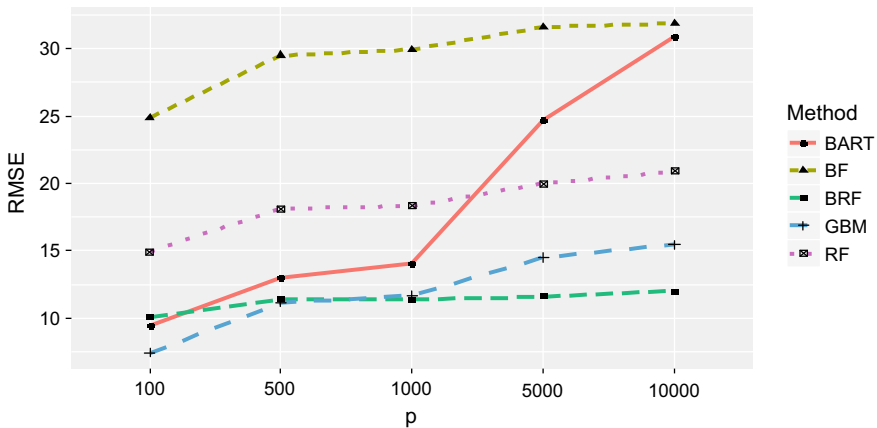


Fig. 2 RMSE with increasing p at sample size $n = 100$, as expected increasing p reduces the hypergeometric probability which increases RMSE of RF (purple dotted lines). The increase in sample size tends to balance the methods RMSE but BART, GBM, BF and RF are still affected with the increase in p

5 Conclusion

In this paper, we presented the theoretical framework for Bayesian Random Forest (BRF) for regression analysis of high-dimensional data. By way of example, We consider its application on simulated Friedman dataset with large p and fewer number of relevant features. We also compared the predictive performance of the method with some existing methods using RMSE via 10 folds cross-validation. The results observed from the simulation study shows that BRF is highly robust to large p small relevant feature issue at a reasonable sample size n when compared with its competitors.

6 Funding

This work was supported by Universiti Tun Hussein Onn, Malaysia [grant numbers Vot, U607].

References

1. Kapelner, A., Bleich, J.: bartmachine: machine learning with bayesian additive regression trees. (2014a)
2. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
3. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29** (5), 1189–1232 (2001)
4. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010)
5. Taddy, M.A., Gramacy, R.B., Polson, N.G.: Dynamic trees for learning and design. *J. Am. Stat. Assoc.* **106**(493), 109–123 (2011). <https://doi.org/10.1198/jasa.2011.ap09769>
6. Taddy, M., Chen, C.S., Yu, J., Wyle, M.: Bayesian and empirical Bayesian forests (2015). [arXiv:1502.02312](https://arxiv.org/abs/1502.02312)
7. Hernández, B., Raftery, A.E., Pennington, S.R., Parnell, A.C.: Bayesian Additive Regression Trees using Bayesian Model Averaging (2015). [arXiv:1507.00181](https://arxiv.org/abs/1507.00181)
8. Friedman, J.H.: Multivariate adaptive regression splines (with discussion and a rejoinder by the author). *Ann. Stat.* **19**(1), 67 (1991)
9. Olaniran, O.R., Yahya, W.B.: Bayesian hypothesis testing of two normal samples using bootstrap prior technique. *J. Mod. Appl. Stat. Methods* **16**(2), 618–638 (2017). <https://doi.org/10.22237/jmasm/1509496440>
10. Olaniran, O.R., Olaniran, S.F., Yahya, W.B., Banjoko, A.W., Garba, M.K., Amusa, L.B., Gatta, N.F.: Improved Bayesian feature selection and classification methods using bootstrap prior techniques. *Anale. Seria Informatică.* **14**(2), 46–52 (2016)
11. Olaniran, O.R., Affendi, M.A. Bayesian Analysis of Extended Cox Model with Time-varying Covariates using Bootstrap prior. *J. Mod. Appl. Stat. Methods.* In press (2017)

12. Yahya, W.B., Olaniran, O.R., Ige, S.O.: On Bayesian conjugate normal linear regression and ordinary least square regression methods: a monte carlo study. *Ilorin J. Sci.* **1**(1), 216–227 (2014)
13. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian Data Analysis*. CRC Press, Boca Raton, FL (2013)
14. Greg, R., et al.: *gbm: Generalized boosted regression models*. R package version 2.1.3 (2017). <https://CRAN.R-project.org/package=gbm>
15. Liaw, A., Wiener, M.: Classification and regression by random forest. *R News* **2**(3), 18–22 (2002)
16. Marvin N.W., Andreas Z.: *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C ++ and R*. *J. Stat. Softw.* **77**(1), 1–17 (2017) <https://doi.org/10.18637/jss.v077.i01>

Chapter 34

Bivariate Weibull Exponential Model Based on Gaussian Copula



Zakiah Ibrahim Kalantan and Mervet Khalifah Abd Elaal

Abstract The Weibull distribution is widely used as a life-time distribution in many fields such as reliability engineering and social science. The aim of this paper is to introduce a new bivariate model of Weibull and exponential distributions. The emerging based on Gaussian copula, which is a popular used in various applications like econometrics and finance. We discuss the goodness of fit test for copula and use both parametric and semi-parametric methods to estimate the model parameters. Finally, Simulation is studied to illustrate methods of inference and examine the satisfactory performance of the proposed distribution.

Keywords Weibull distribution · Exponential distribution · Gaussian copula

1 Introduction

Recently, there has been an increased interest in defining new generators for univariate continuous families of distributions by introducing one or more additional shape parameter(s) to the baseline distribution. For instance, Bourguignon et al. [3] proposed a generator of distributions called the Weibull-G class, and among others. The class of Weibull G distributions (WG) has received an increasing amount of attention in recent years. Many studies conducted based on the properties and inferences of Weibull G distributions with a consideration to their applications. In this paper, we introduce a bivariate Weibull exponential distribution in the dependence structure and illustrate its applicability. The (WG) probability density function (PDF) has the following.

Z. I. Kalantan (✉) · M. K. A. Elaal
Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: zkalanten@kau.edu.sa

M. K. A. Elaal
e-mail: khmervat123@yahoo.com

M. K. A. Elaal
Department of Statistics, Al-Azhar University, Cairo, Egypt

$$f(t, \alpha, \beta, \delta) = \frac{\alpha}{\beta^\alpha} \frac{g(t, \delta)}{1 - G(t, \delta)} \left\{ -\frac{\log[1 - G(t, \delta)]}{\beta} \right\}^{\alpha-1} \exp \left\{ -\left[-\frac{\log[1 - G(t, \delta)]}{\beta} \right]^\alpha \right\},$$

$$t \geq 0. \tag{1}$$

where $G(t, \delta)$ and $g(t, \delta)$, are Cdf and PDF of any baseline distribution depends on a parameter vector δ , t is in the range of $g(t, \delta)$, $\beta > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter. The (WG) distribution function (Cdf) is given by

$$F(t, \alpha, \beta, \delta) = 1 - \exp \left\{ \left[-\frac{\log[1 - G(t, \delta)]}{\beta} \right]^\alpha \right\}, \quad t \geq 0 \tag{2}$$

Various Class Weibull G distributions have been discussed such as Weibull Pareto distribution by Alzaatreh et al. [2]. Copulas are a general tool to construct multivariate distributions and measure the dependence structure between random variables. The paper of Abd Elaal [1] provided several methods of constructing bivariate distributions with copula functions. The main aim of this article is to introduce bivariate Weibull exponential (BWE) model based on the most used copula function named Gaussian copula with a suitable organization. The paper is organized as follows. Section 2 presents the bivariate Weibull exponential (BWE) model based on Gaussian copula function. The maximum likelihood estimates (MLEs) for the model parameters are demonstrated in Sect. 3. In Sect. 4, the flexibility of the model is explained. Finally, the performance of the suggested model using a simulation data is discussed Sect. 5.

2 Bivariate Weibull Exponential Distribution Based on Gaussian Copula

Suppose that $g(t)$ is exponential distribution where $g(t; r) = r \exp(-rt)$, $t > 0$, and $G(t; r) = 1 - \exp(-rt)$, then the Weibull-exponential (WE) distribution PDF and its Cdf are given by, respectively

$$f(t, \alpha, \beta, r) = \frac{\alpha r^\alpha}{\beta^{2\alpha-1}} t^{\alpha-1} \exp \left\{ -\left[\frac{rt}{\beta} \right]^\alpha \right\}, \quad t, \alpha, \beta, r > 0 \tag{3}$$

$$F(t, \alpha, \beta, r) = 1 - \exp \left\{ -\left[\frac{rt}{\beta} \right]^\alpha \right\}, \quad t, \alpha, \beta, r > 0 \tag{4}$$

The density of the WE distribution can be right-skewed, and has constant, increasing, decreasing, hazard rate. This fact implies that the BWE and WE

distributions can be very useful to fit different data sets with various shapes. Now, let T_1 and T_2 are following Weibull-exponential (WE) distribution then the bivariate Weibull-exponential (BWE) distribution which defined as the joint PDF of T_1 and T_2 based on Gaussian copula becomes

$$f(t_1, t_2, \alpha_j, \beta_j, r_j) = \prod_{j=1}^2 \frac{\alpha_j r_j^{z_j}}{\beta_j^{2z_j-1}} t_j^{z_j-1} \exp\left\{-\left[\frac{r_j t_j}{\beta_j}\right]^{z_j}\right\} \left\{\frac{1}{\sqrt{1-\rho^2}} (\exp[\frac{-\rho}{2(1-\rho^2)}\{\rho(z_1^2+z_2^2)-2z_1 z_2\}])\right\}, \tag{5}$$

where $t_j, \alpha_j, \beta_j, r_j > 0$, and $\rho \in [-1, 1]$ is a dependence parameter.

3 Parameter Estimation

In this section, we provide the estimation of the unknown parameters of BWE distribution. There are two approaches to fitting copula models; parametric and semi-parametric methods.

3.1 Parametric Methods of Estimation

There are two approaches to fitting BWE models. One approach is to estimate the marginal and copula parameters separately. The second approach is to obtain the estimation of the marginal and copula parameter from the pseudo-observations separately names modified ML (Fig. 1).

3.1.1 Maximum Likelihood Estimation (ML)

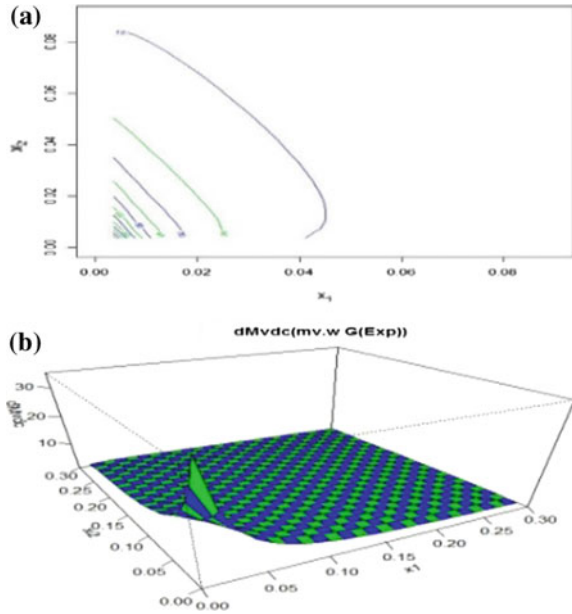
The log-likelihood function expressed as

$$\log L = \sum_{i=1}^n [\log f_1(t_{1i}) + \log f_2(t_{2i}) + \log c(F_1(t_{1i}), F_2(t_{2i}))], \tag{6}$$

The estimation of BWE distribution parameters obtained by ML in two-steps. The first step is estimating the parameters of marginal distribution F_1 and F_2 by MLE separately as

$$\log L_j = \sum_{i=1}^n \log f_j(t_{ji}), \quad j = 1, 2. \tag{7}$$

Fig. 1 BWE based on Gaussian copula: **a** Contour plot and **b** PDF curve



Then, estimating copula parameters by maximizing the copula density is;

$$\log L = \sum_{i=1}^n \log c(F_1(t_{1i}), F_2(t_{2i})) \tag{8}$$

By considering the first step with (WE) distribution, the parameters of each marginal distribution are estimated using MLE method. Now, if t_1, \dots, t_n is a random sample from $WE(\alpha_j, \beta_j, r_j)$, then the log-likelihood function $L(\alpha_j, \beta_j, r_j)$ becomes

$$\log L_j(t_j, \alpha_j, \beta_j, r_j) = n \log(\alpha_j) - n(2\alpha_j - 1) \log(\beta_j) + n \alpha_j \log(r_j) - \frac{r_j^{\alpha_j}}{\beta_j^{\alpha_j}} \sum_{i=1}^n t_{ji}^{\alpha_j}. \tag{9}$$

$$\begin{aligned} \frac{\partial \log L_j(t_j, \alpha_j, \beta_j, r_j)}{\partial \alpha_j} &= \frac{n}{\alpha_j} + n \log(r_j) - \frac{r_j^{\alpha_j} \sum_{i=1}^n t_{ji}^{\alpha_j}}{\beta_j^{\alpha_j}} \log \left\{ \frac{r_j \sum_{i=1}^n t_{ji}}{\beta_j} \right\} + -2n \log(\beta_j) \\ &= 0 \end{aligned} \tag{10}$$

$$\frac{\partial \log L_j(t_j, \alpha_j, \beta_j, r_j)}{\partial \beta_j} = \frac{-n(2\alpha_j - 1)}{\beta_j} + \frac{r_j^{\alpha_j} \sum_{i=1}^n t_{ji}^{\alpha_j}}{\beta_j^{\alpha_j + 1}} = 0. \tag{11}$$

$$\frac{\partial \log L_j(t_j, \alpha_j, \beta_j, r_j)}{\partial r_j} = \frac{n\alpha_j}{\alpha r_j} + \frac{\alpha_j r_j^{\alpha_j - 1} \sum_{i=1}^n t_{ji}^{\alpha_j}}{\beta_j^{\alpha_j}} = 0 \tag{12}$$

The solution of the system of nonlinear Eqs. (10), (11) and (12) gives the MLE of $\alpha_j, \beta_j,$ and r_j . Then copula density will be estimated as given,

$$\log L(\theta) = \sum_{i=1}^n \log c(\hat{F}_1(t_{1i}), \hat{F}_2(t_{2i})) \tag{13}$$

where $\hat{F}_1(t_1)$ and $\hat{F}_2(t_2)$ denote the ML estimates of the parameters from the first step. The solution of the nonlinear Eq. (13) gives the MLE of θ .

3.1.2 Modified Maximum Likelihood Estimation (MML)

We will propose a modified ML method to obtain the model parameters as follows. Firstly, the parameter estimation of marginal distribution F_1 and F_2 by MLE separately computed as $\log L_j = \sum_{i=1}^n \log f_j(t_{ji}), \quad j = 1, 2$.

The solution of the system of nonlinear Eqs. (10), (11) and (12) gives the MLE of $\alpha_j, \beta_j,$ and r_j . Secondly, estimate copula parameters by maximizing the copula density as

$$\log L(\theta) = \sum_{i=1}^n \log [c_\theta(\hat{U}_i, \hat{V}_i)] \tag{14}$$

where \hat{U}_i, \hat{V}_i are pseudo-observations computed from $\hat{U}_i = \frac{R_{1i}}{n+1} = \frac{n}{n+1} \hat{F}_1(t_{1i}), \hat{V}_i = \frac{R_{2i}}{n+1} = \frac{n}{n+1} \hat{F}_2(t_{2i}), R_{1i}, R_{2i}$ are respectively the ranks of t_{1i}, t_{2i} . It is important to respect that the margins Cdf.s are estimated parametrically from the first step.

3.2 Semi-parametric Methods of Estimation

This section presents the semi-parametric methods to estimate the copula model parameter.

3.2.1 Methods-of-Moments

Following Kojadinovic and Yan [6], let c be a bivariate random sample from Cdf $C_\theta[F_1(t_1), F_2(t_2)]$, where F_1 and F_2 are continuous Cdf.s and C_θ is an absolutely

continuous copula such that $\theta \in \mathcal{O}$, where \mathcal{O} is an open subset of R^2 . Furthermore, let R_1, \dots, R_n are the vectors of ranks associated with t_1, \dots, t_n unless otherwise stated. In what follows, all vectors are row vectors. Method-of-moments approaches are based on the inversion of a consistent estimator of a moment of the copula C_θ . The two best-known moments, Spearman's rho and Kendall's tau, are respectively given by

$$\rho(\theta) = 12 \int_{[0,1]^2} u v dC_\theta(u, v) - 3 \tag{15}$$

$$\text{and } \tau(\theta) = 4 \int_{[0,1]^2} C_\theta(u, v) dC_\theta(u, v) - 1 \tag{16}$$

$$\rho_n = \frac{12}{n(n+1)(n-1)} \sum_{i=1}^n R_{i,1} R_{i,2} - 3 \frac{n+1}{n-1} \tag{17}$$

$$\tau_n = \frac{4}{n(n-1)} \sum_{i=1}^n 1[t_{i,1} \leq t_{j,1}] 1[t_{i,2} \leq t_{j,2}] - 1 \tag{18}$$

The consistent estimators of these two moments can be expressed as
 When the functions ρ and τ are one-to-one, consistent estimators of θ is given by

$$\theta_{n,\rho} = \rho^{-1}(\rho_n), \theta_{n,\tau} = \tau^{-1}(\tau_n).$$

It called inversion of Kendall's (itau) and inversion of Spearman's rho (irho) respectively. For more information, see Kojadinovic and Yan [6].

4 Goodness of Fit Tests for Copula

The idea of this test is to compare the empirical copula with the parametric estimator derived under the null hypothesis see Dobrić and Schmid [4]. That is, test if C is well-represented by a specific copula C_θ

$$H_0 : C = C_\theta Vs. H_1 : C \neq C_\theta$$

Two approaches are commonly used in the literature to test the goodness of fit of a copula see Genest et al. [5]. The goodness of fit tests based on the empirical process

$$\mathbb{C}_n(u, v) = \sqrt{n} \{ C_n(u, v) - C_{\theta_n}(u, v) \} \tag{19}$$

where $C_n(u, v)$ is the empirical copula of the data of T_1 and T_2

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^n 1(U_{i,n} \leq u, V_{i,n} \leq v), u, v \in [0, 1], \tag{20}$$

$U_{i,n}, V_{i,n}$ are pseudo observations from C calculated from data as follows. $U_{i,n} = \frac{R_{1i}}{n+1}, V_{i,n} = \frac{R_{2i}}{n+1}, R_{1i}, R_{2i}$ are respectively the ranks of t_{1i}, t_{2i} . Here $C_n(u, v)$ is a consistent estimator and θ_n is an estimator of θ obtained using the pseudo observations. According to Genest et al. [5], and Kojadinovic et al. [6], the test statistics is the Cramer-von Miss and is defined as $S_n = \sum_{i=1}^n \{C_n(U_{i,n}, V_{i,n}) - C_{\theta_n}(U_{i,n}, V_{i,n})\}^2$.

5 Simulation Data

In this section, a new bivariate proposed model based on Gaussian copula is presented. The correlation measures Kendall’s tau and Spearman’s rho of two variables with BWE distribution are obtained and used to provide the values of copula parameter. Considering the following values of marginal and copula parameters of BWE distribution based on Gaussian copula with different sizes of sample (n = 30 and 150), where Gaussian copula parameter $\theta_G = 0.8$. The parameters estimations of the model by Gaussian copula and the corresponding bias, mean squared errors and relative mean squared errors based on 1000 replications are reported in Tables 1, 2, and 3.

To sum up, we observe the follows;

1. As expected, most results improve with increasing in sample size.
2. The selected values of $\alpha_1, b_1, r_1, \alpha_2, b_2, r_2$ and θ_G the bias, MSE and RMSE of the estimates $\hat{\alpha}_1, \hat{b}_1, \hat{r}_1, \hat{\alpha}_2, \hat{b}_2, \hat{r}_2$ and $\hat{\theta}_G$ become smaller as the sample size increased.
3. The efficient estimators of marginal parameters of the model differ according to the parameters. It seems that ML estimates $\hat{\alpha}_1, \hat{b}_1, \hat{r}_1, \hat{\alpha}_2, \hat{b}_2, \hat{r}_2$ and of the model are the same corresponding MML estimates.
4. For copula parameter, the MML provided efficient most estimates for the model with the marginal and Gaussian, copula parameters compared to ML, Itau, and Irho.

Now, goodness of fit test statistics using selected copula function for the marginals is preformed. The results in Table 3 show a non significant p-value obtained using parametric bootstrap for Gaussian copula function, which indicate that selected parametric copula function provides appropriate fit to the marginals. In

Table 1 The estimates, the bias, the mean squared errors and the relative mean squared errors of parameters by simulation study for BWE distribution based on Gaussian copula

Sample size		Estimates, bias, mean square errors and relative mean square errors of parameters						
		$\alpha_1 = 0.8$	$b_1 = 0.7$	$r_1 = 0.9$	$\alpha_2 = 0.7$	$b_2 = 0.8$	$r_2 = 0.75$	$\theta_G = 0.8$
n = 30	ML	0.734	0.837	0.932	0.8390	0.704	0.764	0.597
		0.066	0.137	0.032	0.139	0.096	0.062	0.203
		0.016	0.108	0.044	0.035	0.053	0.062	0.041
		0.021	0.155	0.048	0.050	0.067	0.083	0.052
	MML	0.734	0.837	0.932	0.839	0.704	0.764	0.813
		0.066	0.137	0.032	0.139	0.096	0.062	0.013
		0.016	0.108	0.044	0.035	0.053	0.062	0.000
		0.021	0.155	0.048	0.050	0.067	0.083	0.000
n = 150	ML	0.706	0.864	0.970	0.806	0.620	0.660	0.628
		0.094	0.164	0.070	0.106	0.180	0.025	0.172
		0.148	0.750	0.441	0.189	0.707	0.334	0.398
		0.185	1.071	0.490	0.270	0.886	0.445	0.498
	MML	0.706	0.864	0.970	0.806	0.620	0.660	0.804
		0.094	0.164	0.070	0.106	0.180	0.025	0.004
		0.148	0.750	0.441	0.189	0.707	0.334	0.008
		0.185	1.071	0.490	0.270	0.886	0.445	0.011

Table 2 The estimates, the bias, the MSE and the RMSE of parameters of correlation parameter by simulation study for BWE distribution based on Gaussian copula

Sample size	$\theta_G = 0.8$				
	Estimates	Bias	MSE	RMSE	Method estimation
n = 30	0.597	0.203	0.041	0.052	ML
	0.813	0.013	0.000	0.000	MML
	0.817	0.017	0.000	0.000	Itau
	0.821	0.021	0.000	0.001	IRho
n = 150	0.628	0.172	0.398	0.498	ML
	0.804	0.004	0.008	0.011	MML
	0.806	0.006	0.021	0.026	Itau
	0.804	0.004	0.012	0.015	IRho

Table 3 Goodness of fit test statistics with p-values and estimate of the copula parameter for Gaussian copula

Model	Statistic	p-value	$\hat{\theta}$	Method estimation
BWE	0.0235	0.3272	0.7949	MI
	0.0235	0.3422	0.79485	MML
	0.0270	0.2792	0.7548	Itau
	0.0261	0.3651	0.7625	Irho

addition, estimate of the copula parameter based on ML, MML, Itau, and Irho methods for the Gaussian copula. This estimates are used as initial value when fitting this copula model using BE marginals.

References

1. Abd Elaal, M.K.: Bivariate beta exponential distributions based on copulas. *Int. Organ. Sci. Res. J. Math.* **13**(3), 7–19 (2017). <https://doi.org/10.9790/5728-1303010719>
2. Alzaatreh, A., Famoye, F., Lee, C.: Weibull-Pareto distribution and its applications. *Commun. Stat. Theor. Methods* **42**(9), 1673–1691 (2013). <https://doi.org/10.1080/03610926.2011.599002>
3. Bourguignon, M., Silva, R.B., Cordeiro, G.M.: The Weibull–G family of probability distributions. *J. Data Sci.* **12**, 53–68 (2014)
4. Dobrić, J., Schmid, F.: A goodness of fit test for copulas based on Rosenblatt’s transformation. *Comput. Stat. Data Anal.* **51**(9), 4633–4642 (2007). <https://doi.org/10.1016/j.csda.2006.08.012>
5. Genest, C., Rémillard, B., Beaudoin, D.: Goodness-of-fit tests for copulas: a review and a power study. *Insur. Math. Econ.* **44**(2), 199–213 (2009)
6. Kojadinovic, I., Yan, J.: Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insur. Math. Econ.* **47**(1), 52–63 (2010)

Chapter 35

Claim Assessment of a Rainfall Runoff Model with Bootstrap



Wen Jia Tan, Lloyd Ling, Zulkifli Yusop and Yuk Feng Huang

Abstract Since the inception in 1954, researchers started to scrutinise the United States Department of Agriculture (USDA) Soil Conservation Services (SCS) rainfall runoff model with different field data after the model produced inconsistent runoff prediction results throughout the world. This paper re-assessed two key hypotheses used by SCS where $I_a = 0.2S$ and $\lambda = 0.2$ as a constant. The 112 original SCS data points were used to re-determine the correlation between I_a and S with Bootstrapping, BCa procedure. Both key hypotheses of SCS were proven to be statistical in-significant. Inferential statistics deduced that $I_a \neq 0.2S$ while λ is neither equal to 0.2 nor a constant at alpha = 0.01 level. Both hypotheses are not even applicable to the original dataset used by then SCS to formulate the rainfall runoff model. $I_a = 0.112S$ fitted SCS original data points better at alpha = 0.01 level. The 1954 SCS proposal of $I_a = 0.2S$ and $\lambda = 0.2$ committed type II error as pertain to its own dataset. Therefore, SCS rainfall runoff model cannot be blindly adopted. Practitioners of this model are encouraged to validate and derive regional specific relationship between I_a and S .

W. J. Tan · L. Ling (✉) · Z. Yusop · Y. F. Huang
Centre for Disaster Risk Reduction, Department of Civil Engineering, Lee Kong Chian
Faculty of Engineering & Science, Universiti Tunku Abdul Rahman. Jalan Sungai Long,
Bandar Sungai Long, 43000 Kajang, Malaysia
e-mail: linglloyd@utar.edu.my

W. J. Tan
e-mail: tanwenjia0104@utar.my

Z. Yusop
e-mail: zulyusop@utm.edu.my

Y. F. Huang
e-mail: huangyf@utar.edu.my

Z. Yusop
Centre for Environmental Sustainability and Water Security, Research Institute
for Sustainable Environment, Faculty of Civil Engineering Department,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Keywords Bootstrapping · Non-parametric inferential statistics ·
Runoff prediction · SCS

1 Introduction

In 1954, the United States Department of Agriculture (USDA), Soil Conservation Services (SCS), Natural Resources Conservation Service (NRCS) agency proposed a rainfall-runoff prediction model under Watershed Protection and Flood Prevention Act (PL-566) to address issues in flood management and to comply with the federal flood control programme [1]. The base rainfall-runoff model was proposed as:

$$Q = \frac{(P - I_a)^2}{P - I_a + S} \quad (1)$$

Q = Runoff amount (mm)

P = Rainfall depth (mm)

I_a = the initial abstraction (mm)

S = maximum potential water retention of a watershed (mm)

The initial abstraction (I_a) is also known as the event rainfall depth required for the initiation of runoff. SCS also hypothesized that $I_a = \lambda S = 0.20S$. The value of 0.20 was referred to as the initial abstraction coefficient ratio (λ), a correlation parameter between I_a and S . The value of 0.20 was also proposed as a constant while the substitution of $I_a = 0.20S$ simplified Eq. (1) into a common SCS runoff prediction model which was widely adopted by hydrology textbooks, official hydro design manuals and being incorporated into many design software and programmes after its inception in 1954. The simplified SCS runoff prediction model is:

$$Q = \frac{(P - 0.2S)^2}{P + 0.8S} \quad (2)$$

Equation (2) is subjected to a constraint that $P > 0.2S$, else $Q = 0$. However, there are increasing evidential study results leaning against the prediction accuracy of Eq. (2) and the hypothesis that $I_a = 0.20S$. Many researchers urged to perform regional hydrological conditions calibration instead of blindly adopting it as proposed by SCS [2, 3]. Although SCS proposed the model in 1954, some of the original field data and the origin of key assumptions were undocumented and untraced [2]. The only surviving document can be traced through the 1972 National Engineering Handbook, Sect. 4 (NEH4), Hydrology section (NEH4, 1972). NEH4 Chap. 10 Fig. 10.2 showed 112 data points in used by then SCS to illustrate the correlation between I_a and S of a watershed [4]. The same figure appears again in Fig. 10.1 of USDA, NRCS 2004, Part 630 Hydrology, National Engineering Handbook, Chap. 10 (both figures are shown in Fig. 1 below for comparison) [5].

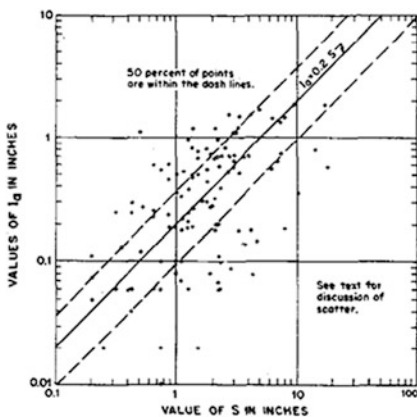
The determination of $\lambda = 0.20$ was based on simple correlation between I_a and S as given in those Figures. Despite that there was considerable scatter in the data, SCS adopted $\lambda = 0.20$ because 50% of the data points were within the range $0.095 < \lambda < 0.38$ [2, 4, 5]. This relationship of I_a to S was based on field data obtained from various un-identified small watershed locations in US [6]. I_a was the dependent while S was on the independent (logged) x-axis. Raw data points were plotted on the log-log scale graph where SCS hypothesized that $I_a = 0.2S$ as the result of so-called “linear fitting” to correlate I_a and S from those figures.

2 Inferential Statistics of 1954 SCS Claim Reassessment

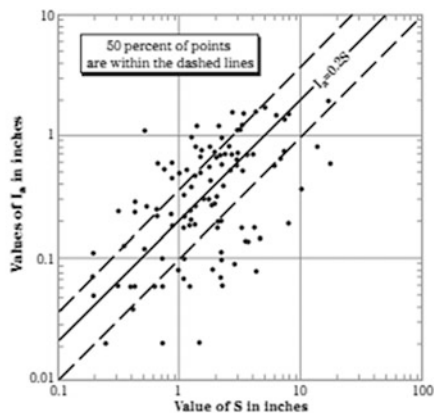
2.1 Assessment Methodology and Null Hypotheses

Since those 112 data points were plotted on the log-log scale graph the $I_a = 0.2S$ correlation proposed by SCS in 1954 should actually be a non-linear power function fitting form on a normal graph scale but its linear form was used by SCS to simplify Eq. (1) and produce Eq. (2) hence it is vital to assess the validity of $I_a = 0.2S$. To assess the 1954 SCS claim, all original data points used by SCS were scaled from Fig. 1 by this research to redetermine the correlation between I_a and S with linear regression through the origin fitting function ($y = mx$) again. The SCS claim reassessment was set forth by two Null hypotheses with first hypothesis stated as below:

H_{01} : $I_a = \lambda S$ where $\lambda = 0.2$ is valid to its original data points.



USDA 1972, NEH4 figure 10.2



USDA, NRCS 2004, Part 630 Hydrology

Fig. 1 Original I_a to S graphs (on log-log plotting scale)

Table 1 Inferential statistics of the regression model for the 1954 SCS dataset

Best regression model: $I_a=0.112S$			λ confident interval			
Gradient (λ)	Std. error	p value		Lower limit	Upper limit	Variation %
0.112	0.011	0.0001	95%	0.091	0.134	47.25
			99%	0.084	0.140	66.67
	0.022	0.0005	BCa 95%	0.077	0.172	123.38
			BCa 99%	0.069	0.193	179.71

One of the main hypotheses proposed by SCS was H_{01} from the correlation between I_a and S of its original data points. This study reassessed the validity of 1954 SCS claims through non-parametric inferential statistics. The assessment was conducted with inferential statistics using modern high computing capability computer and the IBM, PASW software. Besides fitting the scaled dataset with the best linear regression fitting model, Bootstrapping, BCa procedure (2,000 random sampling with replacement) was used to generate the fitting model’s gradient (λ) confidence intervals at alpha 0.05 and 0.01 levels [7] as shown in Table 1. Should the value of 0.20 falls within any of those confidence intervals, 1954 SCS’s claim was valid. The rejection of H_{01} implies that $I_a = 0.2S$ is invalid and not applicable even for the 1954 SCS dataset. The second Null hypothesis was stated as:

H_{02} : In $I_a = \lambda S$, λ is a constant and $\lambda = 0.2$.

H_{01} and H_{02} were used to simplify the proposed SCS runoff Eq. (1) where both hypotheses became the cornerstone of the simplified SCS runoff model, Eq. (2) which is commonly in use until today. It requires the acceptance of both hypotheses in order to justify the initiative to propose Eq. (2).

2.2 Inferential Statistics Assessment

Linear regression through the origin fitting function was conducted to correlate the 112 SCS data points. Bootstrapping, BCa procedure was used to produce confidence intervals of the gradient parameter (λ) at 95 and 99% levels for the best-fitted linear regression equation through original data points on the log-log scale. The inferential statistics of the 1954 SCS claim assessment are given in Table 1 for Null hypotheses assessment.

2.3 Fitting Models Comparison and Residual Analyses Assessment

The best regressed equation $I_a=0.112S$ was also compared to SCS initial proposal where $I_a = 0.2S$ in residual analysis of prediction accuracy against all data points for

Table 2 Model residual analyses results comparison

	$I_a=0.2S$	$I_a=0.112S$
<i>p value</i>	Not significant	0.01
<i>RSS</i>	38.245	23.868
Standard error of estimate	0.587	0.464
Residual: standard deviation	0.585	0.426
BCa 95% confidence interval span	[0.424, 0.740]	[0.348, 0.499]
BCa 99% confidence interval span	[0.377, 0.809]	[0.328, 0.530]
Residual: range	4.403	2.854

further assessment in Table 2. The comparison was based on residual sum of squares (*RSS*), and inferential statistics results generated from PASW such as: regression standard error of estimate, predictive model residual's standard deviation and respective residual range. Bootstrapping, BCa procedure (on 2,000 samples) was again used to generate residual's standard deviation confidence intervals at alpha = 0.05 and 0.01 levels, to assess the validity of $I_a=0.2S$. A better predictive model will have higher *E* index, lower *RSS*, standard error of estimate, less residual in standard deviation and smaller residual range.

3 Results and Discussion

$I_a=0.112S$ was identified by PASW as the best fitted linear equation (regressed through the origin) instead of $I_a=0.2S$. This result is also in line with previous research work that reported $I_a=0.111S$ as the best regression fitting outcome on scaled SCS data [8]. None of the gradient's confidence interval span includes $\lambda =0.2$ (either at 95% or 99%) thus H_{01} can be rejected at alpha = 0.01 level. Large λ variation (from 48 to 180%) between the lower and upper limit of the fitting gradient's (λ) confidence interval also rejected H_{02} at alpha = 0.01 level and proved that λ cannot be a constant but a variable. The best linear model on all the scaled data points from NEH-4 should be $I_a=0.112S$ which gave lower values of *RSS*, standard error of estimate, standard deviation in its residual spread and smaller residual range either under alpha = 0.05 or 0.01 level (Table 2). Thus it is a significant and better fitted result compared to the initial SCS proposal where $I_a=0.2S$ if linear regression fitting function (through the origin) is the preferred choice for its original dataset.

It is also not true to state that " I_a is linearly correlated to S " [4, 5] because SCS used the log-log scale graph to derive the correlation between I_a and S . This has been a common belief since 1954 and the statement appears in many hydrology text books, research articles and literature. The key proposal of $I_a = 0.20S$ is not a linear

correlation between I_a and S on regular graph scale. It would be more appropriate to restate that I_a and S are non-linearly correlated to each other.

The initial SCS hypothesis of $\lambda = 0.2$ was used to simplify the SCS base runoff model but now rejected by the reassessment up to $\alpha = 0.01$ level because all confidence interval spans did not include the value of 0.2 and as such, Eq. 2 became invalid and not applicable even to the original SCS dataset. The justification for $\lambda = 0.2$ and the proposal to fix it as a constant becomes weaker after this assessment. Blind adoption of Eq. 2 commits type II error. The rejection of both Null hypotheses paves the way for SCS model calibration. The key parameter λ can be derived according to the rainfall-runoff dataset and to calibrate Eq. 1 in order to formulate a new runoff model [2, 9–11].

4 Conclusion

Two key hypotheses proposed by SCS in 1954 were neither statistically significant nor applicable even to its original raw data points. Therefore, it is imminent to develop and adopt a regional specific calibration methodology instead of blindly adopting SCS simplified runoff model (Eq. 2) for any runoff prediction use. In conclusion, $I_a \neq 0.2S$ and its substitution into Eq. (1) to simplify it into Eq. (2) lacks statistical justification. Equation (2) is invalid and not even significant at $\alpha = 0.05$ level hence the 1954 SCS proposal of $I_a = 0.2S$ committed type II error.

References

1. Jiang, R.Y.: Investigation of runoff curve number initial abstraction ratio. Master Thesis, University of Arizona, Tucson, AZ, USA (2001)
2. Hawkins, R.H., Ward, T.J., Woodward, D.E., VanMullem, J.A.: Curve Number Hydrology: State of Practice. Reston, ASCE (2009)
3. Ling, L., Yusop, Z.: A micro focus with macro impact: exploration of initial abstraction coefficient ratio (λ) in Soil Conservation Curve Number (CN) methodology. In: IOP Conference Series: Earth Environment Science, (1) (2013)
4. Natural Resources Conservation Service (NRCS), National engineering handbook Part 630: Hydrology. U.S. Department of Agriculture, Washington, DC (1972)
5. Natural Resources Conservation Service (NRCS), National engineering handbook: Hydrology, USDA, Washington, DC (2004)
6. Rallison, R.E.: Origin and evolution of the SCS Runoff Equation. In: Symposium on Watershed Management, 21–23 July, Boise, Idaho, pp. 912–924 (1980)
7. IBM, SPSS Bootstrapping 21 guide. IBM Press (2012)
8. Hawkins, R.H., Khojeini, A.V.: Initial abstraction and loss in the curve number method. In: Arizona State Hydrological Society Proceedings, 15–17 April. Las Vegas, Nevada, pp. 115–119. Hjelmfelt, A. T. (1980)
9. Hjelmfelt, A.T.: Empirical investigation of curve number technique. J. Hydr. Eng. Div. ASCE **106**(9), 1471–1476 (1980)

10. Hawkins, R.H., Ward, T.J., Woodward, D.E., Vanmullem, J.A.: Progress report: ASCE task committee on curve number hydrology. *Managing Watersheds for Human and Natural Impacts*, pp. 1–12. ASCE, New York (2005)
11. Soulis, K.X., Valiantzas, J.D.: Identification of the SCS-CN parameter spatial distribution using rainfall-runoff data in heterogeneous watersheds. *Water Resour. Manage.* **27**(6), 1737–1749 (2013)

Chapter 36

Comparing the Influences of Monetary Versus Fiscal Policy on the Economy: The Case of Malaysia



Siti Fatimah Ismail and Sek Siok Kun

Abstract Low inflation and stable growth are the two main policy objectives targeted by many policymakers. However, both targets may not easily achieve. There are debates and issues discussed in comparing the performance between fiscal and monetary policies. In this study, empirical analyses were conducted to examine how influential both policies in determining the two economic variables of inflation and growth in Malaysia. The autoregressive distributed lags (ARDL) model was applied using the annual data of 1967–2016. Our results revealed that monetary policy (proxy by broad money) is influential in determining the inflation but not for the fiscal policy (proxy by government expenditure). In contrary, fiscal policy (proxy by government expenditure) is the main determinant to GDP growth. The study suggested for the co-implementation of monetary and fiscal policies in order to achieve low inflation and sustainable growth.

Keywords Monetary and fiscal policies · Inflation · Gross domestic product · Autoregressive distributed lag model

1 Introduction

The issues of the effectiveness of monetary and fiscal policy have traditionally attracted great attention among economists and researchers. The relative effectiveness of monetary and fiscal policy is not new. There are debates and issues discussed on comparing the performance between fiscal and monetary policies. A very high inflation level does not derive from monetary policy tools (money supply, interest rate etc.) alone but may also due to fiscal expenditure. Fischer et al. [6] showed that fiscal deficit is one of the main drivers of high inflation.

S. F. Ismail (✉) · S. S. Kun
School of Mathematical Sciences, USM, Gelugor, Malaysia
e-mail: citypalong83@gmail.com

S. S. Kun
e-mail: sksek@usm.my

In Malaysia, inflation showed an increasing rate at average 2.7% in 2016 [1]. In the past 2 years, two important steps were taken to strengthen the fiscal position. GST was introduced to increase the tax and subsidies were cut to free up budget resources. Fiscal policy intends to further narrow the budget deficit in 2016 to 3.1% of GDP. Falling oil prices prompted the government of Malaysia to revise down in January 2016 its revenue expectations and reduce budgeted operating and development expenditures. As for monetary policy, the central bank lowered its reserve requirement in January 2016 for banks to support economic growth.

In this study, empirical analyses were conducted to examine how influential both policies in determining the two economic variables of inflation and growth in Malaysia. The study revealed the effectiveness of monetary policy to accommodate price stability but the fiscal policy to stimulate GDP growth. The study suggested to the implementation of both policies in order to achieve low inflation and sustainable growth.

2 Literature Review

The monetary and fiscal policy is generally believed to be associated with growth and inflation. However, there is no agreement among the world's economists on which policy is more effective in determining the economic performance relative to the macro factors. Denbel et al. [4] investigated the relationship between inflation, money supply and economic growth in Ethiopia for the period of 1970–2010. The results revealed the presence of bi-directional causality between inflation and money supply and uni-directional causality from economic growth to inflation. The findings reported that inflation is a monetary phenomenon and negatively affected by economic growth. Kesavarajah and Amirthalingam [8] examined the nexus between money supply and inflation in Sri Lanka over the period of 1978–2010. They found the existence of a long-run relationship and there was significant causality from money supply to inflation.

Chaitip et al. [3] studied the influences of money supply on economic growth in AEC and ASEAN countries for the period from 1995 to 2013. They revealed that money supply was associated with economic growth in the long run while demand deposit was a negative correction on GDP growth. Tabar et al. [14] employed distributed lags (ARDL) method to investigate the relationship between money supply, prices, government expenditure and economic growth in Iran during the years 1981–2011. They found that all variables have a significant effect on economic growth. Besides that, Khosravi and Karimi [9] investigated the relationship between monetary, fiscal policy and economic growth in Iran and found that exchange rate and inflation have the negative impact on growth whereas government expenditure as fiscal policy has a positive impact on GDP growth. In other study conducted by Attari and Javed [2], the results indicated that there is a long-term relationship between the rate of inflation, economic growth and government expenditure in Pakistan.

Nguyen [11] focused the study in India, Indonesia and Vietnam for the period of 1970–2010 using Vector Error Correction Model. They reported that government spending has a statistically significant and positive effect on inflation in the long-run in all three countries. Mehrara et al. [10] investigated the nonlinear relationship between inflation and government expenditure using quarterly data span from 1990–2013. Smooth Transition Regression Model was employed and the results showed that under the regime of tight money or low growth of liquidity, government expenditure is not inflationary. However, in a regime of low growth of liquidity, government expenditure has a low inflationary impact but stimulates economic growth.

Obi and Ajana [12] also revealed evidence on the impact of money supply on inflation in ECOWAS member states, West African Monetary Zone (WAMZ) and West African Economic Monetary Union (WAEMU) for the period 1980–2012. However, Dupor [5] found almost no effect of government spending on inflation.

Based on the above literatures, there exist various models backing different views about the effectiveness of monetary and fiscal policy as economic policy instruments. Many studies showed that money supply and government spending have significant effect on growth and inflation.

3 Data and Methodology

3.1 Types and Sources of Data

This study used secondary data covering the year ranges from 1967–2016. Data were collected from the World Development Indicators of the World Bank. All data were transformed into natural log form. The details of the data were presented in Table 1.

3.2 Modeling the Influences of Monetary and Fiscal Policy on Economy

The monetary theory of inflation asserts that money supply growth is the cause of inflation. If money supply increases in line with inflation then there will be no inflation. In contrary, government spending may or may not be inflationary. Friedman [7] stated that it clearly will be inflationary if it is financed by creating money. Fiscal policy is extremely important in determining what fraction of total national income is spent by government and who bears the burden of that expenditure. Therefore, we constructed two equations which can explain the effect of macroeconomic variables (money supply, government expenditure, broad money and consumer price index) on inflation and growth. ARDL (Autoregressive

Table 1 Variables, symbols and measurements

Variables	Symbols	Measurements
Inflation	LCPI	Consumer price index
Economic growth	LGDP	Current U.S. dollars
Government expenditure	LGE	Current U.S. dollars
Broad money	LBM	Current U.S. dollars

Distributed Lags) model developed by Pesaran et al. [13] was applied for this purpose.

$$\Delta CPI_t = C + \phi \left(CPI_{t-1} - \theta'_1 LGDP_t - \theta'_2 LGE_t - \theta'_3 LBM \right) + \sum_{j=1}^{p-1} \lambda_j \Delta CPI_{t-j} + \sum_{j=0}^{q-1} \delta_j^* \Delta LGDP_{t-j} + \sum_{j=0}^{q-1} \delta_{2j}^* \Delta LGE_{t-j} + \sum_{j=0}^{q-1} \delta_{3j}^* \Delta LBM_{t-j} + \varepsilon_t \tag{1}$$

$$\Delta GDP_t = C + \phi \left(GDP_{t-1} - \theta'_1 LCPI_t - \theta'_2 LGE_t - \theta'_3 LBM \right) + \sum_{j=1}^{p-1} \lambda_j \Delta GDP_{t-j} + \sum_{j=0}^{q-1} \delta_j^* \Delta LCPI_{t-j} + \sum_{j=0}^{q-1} \delta_{2j}^* \Delta LGE_{t-j} + \sum_{j=0}^{q-1} \delta_{3j}^* \Delta LBM_{t-j} + \varepsilon_t \tag{2}$$

where CPI is the proxy of inflation, GDP is proxy of growth, BM is proxy of money supply and GE is a proxy of government spending; ϕ is the error correction coefficient, θ 's are the long-run coefficients, λ_j^* and δ^* 's are the short-run coefficient and Δ as the first-differenced operator.

4 Empirical Findings

Prior to the estimation, all variables were checked for stationarity using ADF unit-root test (see Table 2). All variables are stationary at first difference or I(1).

We further tested for bound testing to confirm the validity to apply ARDL model. The results of ARDL bound testing in Table 3 have detected long-run relationship in model 1 but inconclusive result in model 2. However, the significance of the error correction term in both models reconfirms the significance and existence of the long-run relationship.

The results of ARDL estimation as summarized in Table 3 show that all variables (CPI, GDP, BM and GE) in both models have the positive significant effect on inflation and economic growth in the short-run. However, only three variables

Table 2 ADF unit root test

Variable	Level	First difference	Level of integration
	t-stat	t-stat	
GDP	-1.9751	-5.3304***	I(1)
CPI	-1.9976	-3.7191**	I(1)
GE	-1.3835	-6.0045***	I(1)
BM	-1.7637	-6.1851***	I(1)

Note The asterisks (***), (**), (*) denote the statistically significant at 1%, 5% levels respectively

(GDP, BM and GE) are significant in the long-run. Higher GDP and broad money lead to higher inflation in the long-run. This is because higher GDP or faster economic growth feeds back to inflation. Also, the excess of money supply has been an important contributor to the rise in inflation. This supports the Monetarists proposition that inflation is a monetary phenomenon. Nevertheless, it is possible that increased spending could lead to an increase in GDP.

Comparing of the long-run results of the two models, we found that GDP and broad money are the main determinants to inflation while government spending is the main determinant to economic growth in the long-run. Therefore, fiscal policy is more influential on growth while monetary policy is influential on inflation.

Table 3 Estimation of ARDL and bound test

Model 1		Model 2	
Variables	ARDL (3,0,0,0)	Variables	ARDL (2,3,0,0)
<i>Long-run</i>		<i>Long-run</i>	
LGDP	0.1814*	LCPI	0.0821
LGE	-0.0001	LGE	0.4795**
LBM	0.1602***	LBM	0.1661
C	-1.3823*	C	-8.1417*
Error correction (ϕ_t)	-0.2761***	Error correction (ϕ_t)	-0.3801***
<i>Short-run</i>		<i>Short-run</i>	
Δ LGDP	0.0501**	Δ LGDP	-
Δ LGDP(-1)	-	Δ LGDP(-1)	0.2601
Δ LCPI	-	Δ LCPI	1.5225*
Δ LCPI(-1)	0.3966***	Δ LCPI(-1)	-2.4445*
Δ LCPI(-2)	-0.1616	Δ LCPI(-2)	1.1939*
Δ LGE	-0.0003	Δ LGE	-
Δ LGE(-1)	-	Δ LGE(-1)	0.1822*
Δ LBM	0.0442*	Δ LBM	0.0631
Bound test (F stat)	3.9459*	Bound test (F stat)	2.6482#

Note The asterisks (***), (**), (*) denote the statistically significant at 1%, 5% and 10% levels respectively. # indicates inconclusive result in bound testing

The speeds of adjustments are in negative values which due to the convergence to the equilibrium level. The results showed that the convergence rate of model 2 is higher than model 1 implying shorter time spent for economic growth to converge to its long-run equilibrium relative to inflation.

5 Conclusions

In this study, we focused our analysis on comparing the influences of monetary and fiscal policy on the economy. In addition, we also investigated the relative effect of GDP, money supply, government spending and CPI on inflation and growth. Applying the ARDL approach, we detected long-run effects of monetary variable on inflation while fiscal variable GDP respectively. Therefore, monetary policy is more influential on inflation while fiscal policy is more influential on GDP in the long-run. Broad money and GDP are positively correlated with inflation while government expenditures may stimulate economic growth. As a conclusion, monetary policy should co-implement with fiscal policy as so that the policy objectives of low inflation and stable growth are achievable in the long-run.

References

1. ADB: ASIAN Development Outlook, (June). ASIA'S Potential Growth, Manila, Philippines (2016)
2. Attari, M.I.J., Javed, A.Y.: Inflation, economic growth and government Pakistan: 1980–2010. Paper presented at international conference on applied economics, Istanbul, 27–29 June 2013
3. Chaitip, P., Chokethaworn, K., Chaiboonsri, C., Khounkalax, M.: Money supply influencing on economic growth-wide phenomena of AEC open region. Paper presented at international conference on applied economics, Kazan, Russia, 2–4 July 2015
4. Denbel, F.S., Ayen, Y.W., Regasa, T.A.: The relationship between inflation, money supply and economic growth in Ethiopia: co integration and causality analysis. **6**, 556–565 (2016)
5. Dupor, W.: How does government spending affect inflation. <https://www.stlouisfed.org/on-the-economy/how-does-government-spending-affect-inflation> (2016). Accessed 20 May 2017
6. Fischer, S., Sahay, R., Végh, C.A.: Modern hyper and high inflations. **3**, 837–880 (2002)
7. Friedman, M.: Quantity theory of money. In: Eatwell, J., Milgate, M., Newman, P. (eds.) *A Dictionary of Economics*, pp. 3–20. Macmillan (1987)
8. Kesavarajah, M., Amirthaligam, S.: The nexus between money supply and inflation in Sri Lanka. In: Abstract of the Jaffna University International Research Conference, Sri Lanka, 20–21 July 2012
9. Khosravi, A., Karimi, M.S.: To Investigation the relationship between monetary, fiscal policy and economic growth in Iran: autoregressive distributed lag approach to cointegration. **7**(3), 415–419 (2010)
10. Mehrara, M., Soufiani, M.B., Razaee, S.: The impact of government spending on inflation through the inflationary environment, STR approach. *World Sci. News* **137**, 153–167 (2016)
11. Nguyen, T.D.: Impact of government spending on inflation in Asian emerging economies: evidence from India, Vietnam, and Indonesia. *Dig. J. Mol. Med.* (2014). <https://doi.org/10.1142/s0217590816500338>

12. Obi, K.O., Ajana, U.Z.: Dynamic impact of money supply on inflation: evidence from ECOWAS member states. **6**, 10–17 (2015)
13. Pesaran, M.H., Shin, Y., Smith, R.P.: Bounds testing approaches to the analysis of level relationships. 289–326 (2001)
14. Tabar, F.J., Najafi, Z., Badooei, Y.S.: The relationship between money supply, prices, government expenditures and economic growth in Iran economy. 483–495 (2016)

Chapter 37

Comparison Between k -Means and k -Medoids for Mixed Variables Clustering



Norin Rahayu Shamsuddin and Nor Idayu Mahat

Abstract This paper compares the performance of k -means and k -medoids in clustering objects with mixed variables. The k -means initially means for clustering objects with continuous variables as it uses Euclidean distance to compute distance between objects. While, k -medoids has been designed suitable for mixed type variables especially with PAM (partition around medoids). By using a mixed variables data set on a modified cancer data, we compared k -means and k -medoids on internal validity set up in R package. The result indicates that k -medoids is a good clustering option when the measured variables are mixed with different types.

Keywords Mixed variables • k -means • k -medoids • Silhouette • Dunn index

1 Introduction

Cluster analysis (CA) deals with processes of recognizing similar objects and group them together while separating non-similar objects into other groups. The clustering strategies are many [1], but generally can be split into two major ideas; data point grouping and cluster point grouping. Data point grouping starts with individual object and subsequently group them together based on similarity until final clusters are met. Common cluster method of this type is hierarchical clustering. Whilst, cluster point grouping starts from some initial number of cluster and subsequently tries to reassign the objects to k clusters such that the error of misallocation is minimized. An example method of this type is k -means clustering. The execution of

N. R. Shamsuddin (✉)

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Malaysia

e-mail: norinrahayu@kedah.uitm.edu.my

N. I. Mahat

School of Quantitative Sciences, College of Arts and Sciences,
Universiti Utara Malaysia, Changlun, Malaysia

e-mail: noridayu@uum.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_37

303

clustering can be viewed as hierarchical or nonhierarchical. The former performs groups of objects through a tree type construction where similar objects are placed in the same node and the construction stop when there is no more object to group. In contrary, the nonhierarchical forms a grouping of objects into a pre-determined number of groups using iterative algorithm so that the goal of clustering is optimized. Between these two, discussion on nonhierarchical has attracted more interest among researchers due to its algorithmic challenge.

The principle of nonhierarchical clustering is to ensure the objects in dataset were assign to a set of disjoint clusters. The ability of nonhierarchical CA in partitioning the objects into homogeneous group based on similarity measure, has become an important multivariate analysis in diverse fields, from information science to social sciences and now significant in biological fields. The optimization procedure in nonhierarchical CA allows for reallocating objects other than the initial seed (centroid) to seek for minimized difference within clusters. As such, [2] has introduced a k -means method which give good judgment in term of within-class variance. However, this method limits for numerical values and is easily affected by outliers as it uses mean value for its seeds. Since k -means is remain popular in clustering for its simplicity, there were few proposal carried out to obtain its best performance [3, 4]. Huang [5] has extended the k -means algorithm, k -prototypes for both numerical and categorical variables and preserve its efficiency.

Alternative to k -means, k -medoids, determines center of clusters (termed as medoid) based on data points and gives minimal value in sum of dissimilarity to all objects. k -medoids model has gained much popularity as it is robust towards noise or outliers due to the minimization of total dissimilarity to other points [6, 7] compared to k -means model, and its appropriateness for mixed variables case. Partitioning around medoids (PAM) is the most well-known algorithm exemplify k -medoids which was discussed by Kaufman and Rousseeuw [8], but PAM only capable in handling small dataset (400–600 data) as it requires excessively computational time and memory storage. To handle this problem, a clustering large application (CLARA) was developed as an extension to k -medoids approached by Kaufman and Rousseeuw [8]. CLARA is a sampling based approach which randomly choose small portion of sample from dataset.

A decent quality of CA should produce a high homogeneity within cluster and high heterogeneity between clusters. However, the crucial part in clustering process which might deteriorate the quality of CA is in determining the exact number of clusters. There are abundant of algorithms proposed to confirm the number, yet particular algorithms show disagreement on the total number of cluster that should be generated [9]. Therefore, the purpose of this paper is to compare the performance of k -means and k -medoids in clustering objects into groups when the variables are of mixed types. This is achieved by determining the number of clusters that most representing the objects and validate the clusters groups using common clustering indicators namely Dunn Index and Silhouette. Principle of cluster validation are to seek for quality of CA as well as to justify the number of clusters.

In real cases, a simultaneous presence of mixed type of variables with different scale measures in a same data matrix such as binary, categorical (nominal or ordinal

scales) and continuous variables (interval or ratio scales), have created challenges in performing multivariate analysis [10]. If analyses are carried out without proper method in handling the scales, the outcome will effect on: (i) using incorrect analysis method, (ii) insignificant interpretation of the results and inaccurate conclusion and/or, (iii) illogical process of grouping the objects or attributes [11].

In cluster analysis, the fundamental information used is similarity/dissimilarity matrix of n objects. Common measurements to compute similarity/dissimilarity between objects are Euclidean, Manhattan, Mahalanobis, Minkowski, Canberra, Jacquard's, simple matching, Dice index, Russel and Roa, and many more. However, few of these measurements are mean for mixed variables. Euclidean is the most popular measure and has been applied in most algorithms. Nevertheless, it is only suitable for numerical variables [12].

The development of distance metric using mixed variables was devoted by Gower in 1970s. However, the study did not incorporate the information from ordinal data. However, this flaw was tackled by Podani [10] who has extended the study to allow for ordinal variables. Gower's similarity index has been widely used in support vector machine, machine learning, and pattern recognition, bioinformatics, molecular biology, epidemiology and other discipline areas. This paper aiming at investigating two clustering algorithms namely k -means and k -medoids when the measured variables are mixed. Section 2 of this paper briefs about the methodology used. The next Sect. 3 provides results of the investigation with discussion.

2 Methodology

2.1 Clustering Algorithm

2.1.1 k -Means

The goal of k -means is to partition the n objects into k predetermined number of clusters such that the within-cluster sum of squares is minimized. The squared error function, J , is defined as

$$J = \sum_{j=1}^k \sum_{i=1}^n \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where $x_i^{(j)}$ is object i if allocate in cluster j , c_j is the centroid of cluster j , and k is the number of desired clusters. k -means is fit for numerical variables, hence this study treated categorical variables as numeric to allow the computation. The R function `kmeans()` was used to derive the k -means clustering.

2.1.2 *k*-Medoids

k-means works based on centroids could be harm by the occurrence of outliers in a data. Alternatively, *k*-medoids chooses randomly *k* clustering centroids from the *n* objects and performs initial partition based on the closeness (high similarity) of each object to these *k* centroids. The iteration of partitioning processes is employed continuously until the goal of partitioning has achieved. In R-package, pam() and clara() functions are designed to search for *k* representative objects. In partitioning approaches, the number of cluster must be determined and one of the suitable method is silhouette approach. The mechanism in PAM is being carried out in two phases—(i) **BUILD** phase and, (ii) **SWAP** phase [6]. In BUILD phase, the initial medoids is being determine while SWAP phase tries to improve the quality of clusters by swapping objects to the nearest cluster.

2.2 *Internal Validation*

The internal validation was computed to measure the performance of clustering algorithm. The function of internal validation is obtained from package fpc() [13] for *k*-medoids, while clValid() determined the information for other clustering approaches [14]. We focus on compactness. Compactness assesses the degree of cluster separation between cluster as well as the cluster homogeneity. The best method for compactness is silhouette width and Dunn index. The silhouette value lies between $[-1, 1]$ where -1 determined poor clustered observation and vice versa if it approached 1. The Dunn index range between zeros to ∞ . Both methods should be maximized.

2.3 *Data Sets*

This study used cervical cancer (risk factor) [15] from UCI machine learning repository to compare both *k-means* and *k-medoids* in mixed variables case. The cancer cervical data consists of 36 variables and 858 number of objects. After a cleaning process, a total of 657 objects were used in this study. As for discussion purposes, only eight variables were used which comprise of Boolean and numerical information. The variables include age, number of sexual partner, age at first sexual intercourse, number of pregnancies, smoking and hormonal contraceptive status (yes; no), IUD information, and sexual transmitted disease (STD) status.

Table 1 Comparison between clustering methods

	<i>k-means</i>	<i>k-medoids</i>		Agglomerative	Divisive
		PAM	CLARA		
Number of cluster	2	4	2	2	2
Silhouette	0.2928	0.4419	0.3002	0.4006	0.2977
Dunn index	0.0852	0.0449	0.0292	0.1299	0.0146

3 Results and Discussion

The primary purpose of gathering the information on cervical cancer is to predict the individual patient's risk factors towards cervical cancer. Based on previous studies, the selected variables in our study are most factors contributed to the disease [16, 17].

Table 1 summarizes the clustering results for *k-means*, *k-medoids*, and two hierarchical clustering methods namely agglomerative and divisive. Despite of having mixed types of variables, most of the algorithms end with the same number of clusters, 2, except for *k-medoids*. The pam *k-medoids* with four clusters performs best in term of silhouette width, except for Dunn index. As mention in Sect. 2.2, the best selection methods for clustering must have large value for Dunn index and silhouette.

Interestingly, *k-medoids* performs the best based on the internal validation input. Moreover, it is able to retain the types of variables used in a dataset. Agglomerative comes as the second best, but *k-means* turns the worst. Such results could be contributed by the fact that *k-means* depends on the mean centroids which could be easily influenced by the distribution of the data.

The next effort of this findings is to further investigate both *k-means* and *k-medoids* in depth such as by adding more variables related to the risk of developing cervical cancer, which may help in improvised the clustering results. Our result had reveal some rare cases in which we are positive that it may create another group of clusters that had significant contribution in labelling the clusters.

References

1. Hennig, C.: Clustering strategy and method selection. In: Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (eds.) Handbook of Cluster Analysis, pp. 1–34. Chapman & Hall/CRC Press (2015)
2. Macqueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
3. Cui, X., Zhu, P., Yang, X., Li, K., Ji, C.: Optimized big data K-means clustering using MapReduce. J. Supercomput. **70**, 1249–1259 (2014). <https://doi.org/10.1007/s11227-014-1225-7>

4. Tzortzis, G., Likas, A.: The MinMax k-means clustering algorithm. *Pattern Recognit.* **47**, 2505–2516 (2014)
5. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: *In the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 21–34 (1997)
6. Kaufman, L., Rousseeuw, P.: Clustering by means of medoids. In: Dodge, Y. (ed.) *Statistical data analysis based on the L1 norm and related methods*, pp. 405–416. Faculty of Mathematics and Informatics, North-Holland (1987)
7. Jin, X., Han, J.: K-medoids clustering (2010). https://doi.org/10.1007/978-0-387-30164-8_426
8. Kaufman, L., Rousseeuw, P.J.: Partitioning around medoids (program PAM). In: *Finding Groups in Data: An Introduction to Clustering Analysis*, pp. 68–125 (1990)
9. Hu, X., Xu, L.: Investigation on several model selection criteria for determining the number of cluster. *Neural Inf. Process. Rev.* **4**, 1–10 (2004)
10. Podani, J.: Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* **48**, 331–340 (1999)
11. Sindik, J.: Two aspects of bias in multivariate studies: mixing specific with general concepts and “comparing apples and oranges”. *J. Sport. Sci. Med.* **3**, 23–29 (2014)
12. Lourenço, F., Lobo, V., Bação, F.: Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. *Measurement*, 1–18 (2004)
13. Hennig, C.: Package “fpc,” (2018)
14. Brock, G., Pihur, V., Datta, S.S., Datta, S.S.: *clValid*: an R package for cluster validation. *J. Stat. Softw.* **25**, 1–28 (2008). doi:citeulike-article-id:2574494
15. Fernandes, K., Cardoso, J.S., Fernandes, J.: Transfer learning with partial observability applied to cervical cancer screening. In: *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 243–250. © Springer International Publishing (2017)
16. Ezat, S., Puteh, W., Norin-Rahayu, S., Noor, S., Syed, A., Azhar, S., Aljunid, S.M., Science, Q., Lumpur, K., Latiff, J.Y.: HPV positivity and its' influencing factors among invasive cervical cancer women in Malaysia. *Int. J. Public Health* **1**, 13–22 (2011)
17. Roura, E., Castellsagué, X., Pawlita, M., Travier, N., Waterboer, T., Margall, N., Bosch, F.X., De Sanjosé, S., Dillner, J., Gram, I.T., Tjønneland, A., Munk, C., Pala, V., Palli, D., Khaw, K. T., Barnabas, R. V., Overvad, K., Clavel-Chapelon, F., Boutron-Ruault, M.C., Fagherazzi, G., Kaaks, R., Lukanova, A., Steffen, A., Trichopoulou, A., Trichopoulos, D., Klinaki, E., Tumino, R., Sacerdote, C., Panico, S., Bueno-De-Mesquita, H.B., Peeters, P.H., Lund, E., Weiderpass, E., Redondo, M.L., Sánchez, M.J., Tormo, M.J., Barricarte, A., Larrañaga, N., Ekström, J., Hortlund, M., Lindquist, D., Wareham, N., Travis, R.C., Rinaldi, S., Tommasino, M., Franceschi, S., Riboli, E.: Smoking as a major risk factor for cervical cancer and pre-cancer: results from the EPIC cohort. *Int. J. Cancer* **135**, 453–466 (2014). <https://doi.org/10.1002/ijc.28666>

Chapter 38

Data Analysis Comparison

Logit and Probit Regression

Using Gibbs-Sampler



Subanar

Abstract Binary regression using probit and logit regression is widely used in applied statistics. In the classical approach, the parameters of the models to be viewed as unknown constants and the maximum likelihood is the most popular inference method. If we know the prior distribution of the parameters then the Bayesian approach will be the suitable methods for data analysis as its ability to incorporate prior information which can increase the precision of parameter estimates. If certain prior distributions are particularly convenient for samples from certain other distributions then the explicit posterior distributions are easily been derived. For example, a random sample is taken from Poisson distribution and the prior distribution of θ is a Gamma distribution, then the posterior distribution of θ will be a Gamma distribution. The straight forward calculation of the posterior distribution is generally impossible if the parameter space is high dimensional form or have no explicit functional form. In that case, the role of computer intensive method in summarizing posterior distribution is conducted through Markov Chain Monte Carlo. In general, the Gibbs sampler is a technique to develop Markov Chain such that it can generate sample from the posterior distribution without calculating the density instead of simulating individual parameters from a set of p conditional distribution. In this paper, we study the Gibbs sampling applied to low birth weight study related to issues of implementation of the Millenium Development goal in special region of Yogyakarta. In addition, we shows the comparison between logit and probit estimates.

Keywords Binary probit and logit • Bayesian analysis • Markov chain monte carlo • Gibbs sampling

Subanar (✉)

Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
e-mail: subanar@ugm.ac.id; subanarseno7@gmail.com

1 Introduction

The development and application of Bayesian inference have an important role in the advancement of statistics. A lot of Bayesian application articles in science and engineering as well as activities related to Bayesian theoretical framework have been nicely published by Berger [1].

The basic elements of the Bayesian inferential approach consist of three elements, i.e. prior distribution, sampling distribution and posterior distribution, in which all inferences are based on. In case the posterior distribution parameters have no explicit functional form or the parameter space is high dimensional form, the role of computer intensive method in summarizing posterior distribution is conducted through Markov Chain Monte Carlo (MCMC). According to Albert [2], Albert and Chip [3] the MCMC algorithms are very attractive in sense that they are easy to set up and program and require relatively little prior input from the user. The MCMC is an algorithm to set up an irreducible, aperiodic Markov Chain where its limiting distribution converges to the posterior distribution. The Metropolis Hastings algorithm is a generic way to set up the Markov chain desired. There are three variants of the Metropolis Hastings algorithms. These first two variants are the independent chain and the random walk chain. Both variants need a proposal density to simulate the candidate of the parameter. The third variant of the Metropolis Hastings algorithm is called the Gibbs sampler which is used when we know the full conditional distribution of each parameter.

The Gibbs sampler, an algorithm in which we can set up a Markov chain simulation from the joint posterior distribution by simulating individual parameter from the set of conditional distribution. The book from Gelman [4] outlined the application of Gibbs sampler in many statistical methodologies both in classical (likelihood) and Bayesian. Section 2 presents the Gibbs sampling applied to low birth weight study of Niken Retnowati [5] related to the issues of implementation of the Millennium Development goal in special region of Yogyakarta. In addition we show the comparison between logit and probit estimates. Finally, Sect. 3 presents some concluding remarks.

2 Gibbs Sampling in Binary Regression Models

2.1 Gibbs Sampling

Suppose we are concerned with parameter vector $\theta = (\theta_1, \theta_2, \dots, \theta_p)$. The joint distribution between θ and data denoted by $[\theta|data]$ may be in a high dimension and it wants to be calculated. Suppose we have a set of conditional distribution

$$\begin{aligned} & [\theta_2 | \theta_1, \theta_3, \dots, \theta_p, data] \\ & \quad \vdots \\ & [\theta_p | \theta_1, \theta_2, \dots, \theta_{p-1}, data] \end{aligned}$$

where $[X|Y; Z]$ denotes the distribution of X conditional on the random variables Y and Z . Basically Gibbs sampling is a technique to develop Markov Chain such that it can generate sample from the posterior distribution without calculating the density instead of simulating individual parameters from a set of p conditional distribution. Simulating one value of each individual parameter from these distributions is called one cycle of Gibbs sampling. Under some mild conditions, the simulation will converge to the posterior distribution. There are two advantages of Gibbs sampling. First Gibbs-sampling is easy to implement if conditional posterior distribution is available or easy to be simulated using conditional distribution standard. Second, it does not need to know the proposal density.

2.2 Binary Regression Models

Suppose we observed independent binary random variables y_1, y_2, \dots, y_n with $y_i \in \{0,1\}$, $i = 1, 2, \dots, n$ and each y_i is Bernoulli distributed with probability of success is p_i . Related to the i -th response, we observe k covariate $x_{i1}, x_{i2}, \dots, x_{ik}$. Binary regression model is defined as $p_i = F(x_i^T \beta)$ with β is a $k \times 1$ vector of unknown parameters, $F(\cdot)$ CDF function which link the probability p_i with the linear function $x_i^T \beta$. If F is a logistic CDF then we have the logistic regression model, whereas the probit model is obtained if F is the standard normal CDF.

In general probit regression model using auxiliary variables has a representation

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$$

$$z_i = x_i^T \beta + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

$$P(y_i = 1) = P(z_i > 0) = \Phi(x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ik}\beta_k)$$

Let β has a prior distribution $\pi(\beta)$, then the posterior distribution is given by

$$\begin{aligned} \pi(\beta|y) & \propto \pi(\beta) \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \\ & \propto \pi(\beta) \prod_{i=1}^n \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i} \end{aligned}$$

If the sampling from the joint posterior distribution of z dan β and the vector parameter β is added with latent data $z = (z_1, z_2, \dots, z_n)$ then the Gibbs sampling automatically will work. Both conditional posterior distribution, $[z|\beta]$ and $[\beta|z]$ have convenient functional forms.

According to Albert [2], the conditional posterior distribution of β is

$$[\beta|z, data] \sim N_k\left((X^T X)^{-1} X^T Z, (X^T X)^{-1}\right)$$

where X is the design matrix for the problem. Let the value of the regression parameters of β are given, then $z_1, z_2 \dots z_n$ are independent, with

$$[z_i|\beta, data] \sim N_k(x_i\beta, 1)I(z_i > 0), \text{ if } y_i = 1$$

$$[z_i|\beta, data] \sim N_k(x_i\beta, 1)I(z_i < 0), \text{ if } y_i = 0$$

These results show us that given the value of β , we simulate the latent data z from truncated normal distributions where the truncation point is 0 and the side of the truncation depends on the values of the binary response. The Gibbs sampling algorithm in R probit regression is implemented using the function *bayes.probit*.

2.3 Data Analysis

We illustrate the Gibbs sampling on the low birth weight study first analyzed by Niken Retnowati [5]. In this study data were collected on 124 women who gave birth on Muhammadiyah Kotagede hospital in Yogyakarta. Among these 124 women $n_1 = 38$ of which had low birth weight babies and $n_0 = 86$ of which had normal birth weight babies. This investigation is part of giving input to local government program in reducing low birth rate in accordance to the Millenium Development goals. Six variables which were thought to be of importance were X_1 (previous premature birth history), X_2 (hipertension), X_3 (Anemia), X_4 (twin birth history), X_5 (birth-bleeding), and X_6 (Age).

After the test on the acceptance that low birth rate has Bernoulli distribution, and there are non Multi colliniarity among predictor variables (Retnowati [5]), we fitted the data to the probit regression model. After assesing the significance of the variables in the model we have X_2, X_3, X_5 and X_6 left. Continuing fitting the probit model on X_2, X_3, X_5 and X_6 we have the results of the fitting on Table 1.

Table 1 The results of fitting the probit regression model

Variable	Estimated coefficient	Standard error
Constant	2.73497	0.86126
X_2	0.55512	0.35126
X_3	0.98512	0.32494
X_5	1.40634	0.51353
X_6	-0.02610	0.02157

Table 2 The results of fitting the logit regression model

Variable	Estimated coefficient	Standard error
Constant	4.54856	1.52852
X_2	0.91237	0.58430
X_3	1.63327	0.54020
X_5	2.28439	0.88700
X_6	-0.04613	0.03730

As we have mentioned, besides the probit model we have the logit model to explain the behavior of a dichotomous dependent variable. Now we wish to fit the logit model to the low birth weight as an alternative to the probit model.

Based on the significant test we have variables X_2, X_3, X_5, X_6 left, consistent with the result of the fitted probit model. Again, by fitting the logit model on X_2, X_3, X_5 and X_6 we have the following results on Table 2.

As all variables in this case are significant, then we have logit model.

$$\hat{g}(x) = 4.54856 + 0.91237X_1 + 1.63327X_2 + 2.28439X_5 - 0.04613X_6$$

Comparing Logit and Probit Estimates

According to Gujarati [6], the estimates of the parameters, although logit and probit models qualitatively gives the mirror difference, the estimates of the parameters of the two models are not directly comparable because the basis variance of the two distribution is difference (the normal basis is one while the logistic distribution is $\pi^2/\sqrt{3}$).

But as Amemiya [7] suggest, a logit estimated of a parameter multiplied by 0.625 gives a fairly good approximation of the probit estimate of the some parameter. In the low birth weight case

$$\begin{aligned} 0.625 \times 4.54856 &= 2.84285 \\ 0.625 \times 0.91237 &= 0.57023 \\ 0.625 \times 1.63327 &= 1.0207 \\ 0.625 \times 2.28439 &= 1.42774 \\ 0.625 \times -0.04613 &= 0.02883 \end{aligned}$$

which are roughly equal to the corresponding probit estimate.

Using *R*-function *bayes.probit* with $m = 10000$ cycles we will get posterior distribution of β and its standard error by Gibbs sampling of the regression coefficients are

[1]	2.65257099	0.55630968	1.01151778	1.49689888
	-0.02774934	and		
[2]	0.89688646	0.35046230	0.32800254	0.53976585
	0.02173762			

These results are similar in value to the maximum likelihood estimates and their associated standard errors. It is expected since the posterior analysis was based on a non-informative prior on the regression vector β .

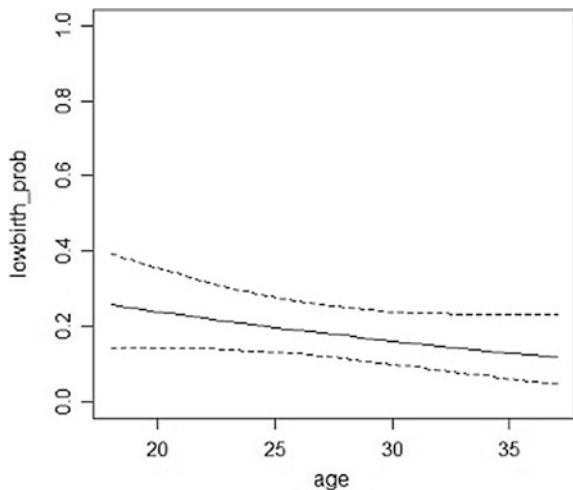
Since X_2 (hypertension), X_3 (Anemia), X_5 (birth-bleeding) and X_6 (age) all significant in this study, we are interesting to explore the probability of low birth rate

$$p = P(y = 1) = \Phi(\beta_0 + \beta_2X_2 + \beta_3X_3 + \beta_5X_5 + \beta_6X_6)$$

a function of these for variables. The *R*-function *bprobit.probs* is useful for computing a simulated posterior sample of probabilities for covariate sets of interest.

We compute the 5th, 50th, and 95th percentiles of the simulated sample of p for each the X_6 or ages values between 18 and 37 years women. Figure 1 shows the graph of these percentiles as a function of age. For each age, the solid line is the location of the median of the low birth weight and the interval between the dashed lines corresponds to a 90% interval estimate for this probability. The graph shows that the younger the women then the higher the probability of the low birth weight.

Fig. 1 The percentiles of the simulated sample of p for each ages



3 Concluding Remark

- Using Gibbs sampling the study on low birth weight shows the similarity of the posterior mean and standard deviation in value to the maximum likelihood estimates and their associated standard error.
- A logit estimated of a parameter multiplied by 0.625 gives a fairly good approximation of the probit estimate of the same parameter as Amemiya [7] suggest, is also true in the study on low birth weight.
- The younger the women then the higher the probability of the lower birth weight gives input to the local government in Yogyakarta to set programs in accordance to the Millennium Development goals program.

References

1. Berger, J.O.: Bayesian analysis: a look at today and thoughts of tomorrow. *JASA* **95**, 1269–1276 (2000). <https://doi.org/10.2307/2669768>
2. Albert, J.: *Bayesian Computational with R*. Springer (2009)
3. Albert, J., Chip, S.: Bayesian analysis of binary and polychotomous response data. *JASA* **88** (442), 669–679 (1993). <https://doi.org/10.2307/2290350>
4. Gelman, R.: *Bayesian Data Analysis*. Chapman and Hall (2003)
5. Retnowati, N.: Estimating the parameter of logistic regression using metropolis-gibbs algorithm. Unpublished thesis, Gadjah Mada University (2015)
6. Gujarati, D.N.: *Basic Econometric*. Mc Graw Hill (1995)
7. Amemiya, T.: Qualitative response models: a survey. *J. Econ. Lit.* **19**, 1483–1536 (1981)

Chapter 39

Dirichlet-Multinomial Estimation of Small Area Proportions of Socio-Economic Classes



Shirlee R. Ocampo, Harley Garcia and Mariel Uy

Abstract The hierarchical Bayesian Dirichlet-multinomial model is explored in this study to estimate small area proportions of socio-economic levels in regional and provincial levels using the Family and Income Expenditure Survey (FIES). The data include 38,484 households which are divided into three socio-economic classes such as low income, middle income and high income based on per capita income (PCI). Dirichlet-multinomial model was used to generate Bayesian small area estimates of households belonging to the low income, middle income, and high income socio-economic levels. Bayesian statistics were generated using Gibb's sampling and simulation techniques to provide estimates with small Markov Chain Monte Carlo (MCMC) standard errors. Results include direct and Bayesian estimates of households belonging to the three socio-economic groups in regional and provincial levels. Wide disparities in the distributions of low income, middle income, and high income are noted in the results. A notable advantage of Dirichlet-multinomial estimates is the absence of zero proportion in high income level in comparison with direct estimates. Regions and provinces were ranked based on the estimates obtained from the hierarchical Bayesian Dirichlet-multinomial model with comparison on rankings using the direct estimates.

Keywords Dirichlet-multinomial model · Hierarchical bayesian · Small area statistics · Socio-economic classes

S. R. Ocampo (✉) · H. Garcia · M. Uy
De La Salle University, Manila, Philippines
e-mail: shirlee.ocampo@dlsu.edu.ph

H. Garcia
e-mail: hha.garcia@yahoo.com

M. Uy
e-mail: marielgenevivuy@yahoo.com

1 Introduction

Official statistics and poverty literature in the Philippines generally focus on indicators concerning poor or food poor households. Despite some economic growth in the past, poverty persists and income distribution becomes inequitable. Not much research has been done about income distribution and very few studies dealt with the middle-income class vis-à-vis other income classes [1].

Systematic formation of data concerning the middle class has not yet been institutionalized in the Philippine Statistical System (PSS). Virola et al. [2] highlighted the socio-economic and demographic characteristics of the Filipino middle class. Their paper focused on the clustering of households based on per capita income (PCI) defining the income brackets of the low-income, middle-income, and high-income classes. Results emphasized at least three clusters for these socio-economic classes, rather than the usual poor or non-poor classification.

The demand for reliable statistics has been increasing exponentially. With direct estimates produced at provincial level having large standard errors, the researchers propose a hierarchical Dirichlet-multinomial Bayes approach in estimating proportions of low, middle and high-income classes in the Philippines.

2 Data and Methodology

This study used the Family Income and Expenditure Survey (FIES). This is a survey conducted by the government every three years where official poverty indexes are generated. FIES 2006 consists of 38,484 observations that summarize the data collected which were about the family income and expenditure that include levels of utilization by item in cash or in kind for the expenditure as well as for the sources of income. FIES in STATA was converted in SAS and Microsoft Excel where initial statistical analyses were performed. Variables used in the study include total family income, family size, weighing factor, socio-economic level, province and region. Per capita income (PCI) of each household was computed by dividing total family income by household size.

To estimate the proportion of socio-economic levels in the Philippines, the range of each income group from the study of Virola et al. [2, 3] as shown in Table 1 was used to categorize each household as either low income, middle income, or high income.

Table 1 Annual PCI brackets for the socio-economic levels

Low	Middle		High
	Minimum	Maximum	
<49,436	49,436	605,359	>605,359

The classifications are then weighted, and the proportions of low-income, middle-income and high-income classes were computed in regional and provincial levels.

Bayesian estimation utilizes the Baye's Theorem which states that

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} \quad (1)$$

where B_i 's are disjoint events partitioning the sample space and A is an event in the sample space. The joint posterior distribution can be derived as proportional to the product of the prior distribution $P(\theta)$ and the likelihood distribution $P(y|\theta)$. If θ is discrete, the posterior distribution can be computed by

$$P(\theta|y) = \frac{P(\theta) * P(y|\theta)}{\sum_{\theta} P(\theta) * P(y|\theta)}. \quad (2)$$

In estimating the proportions of socio-economic levels, Dirichlet-multinomial model is explored. The hierarchical Bayesian is a posterior Dirichlet-multinomial distribution composed of a multinomial likelihood distribution and a Dirichlet prior distribution. Let y_i be the number of times i will occur in category k , n_i be the total number of occurrences, k be the number of categories in which y_i 's can belong to, and π_i be the probability of observing y_i in category $1, 2, \dots, k$. Setting the order of inference for π_i ,

- (1) $\pi_i | \alpha_1, \alpha_2, \dots, \alpha_k \sim \text{Dirichlet}(k, \alpha_i)$, where i goes from 1 to k
- (2) $y_i | \pi_i \sim \text{Multinomial}(\pi_i, n_i)$.

The joint posterior distribution for π_i can be written as

$$P(\pi_i | y_i, \alpha_i) \propto \prod_{i=1}^n \pi^{y_i + \alpha_i - 1}. \quad (3)$$

Hence,

$$\pi_i | y_i, \alpha_i \sim \text{Dirichlet}(k, \alpha_1 + y_1, \alpha_2 + y_2, \dots, \alpha_k + y_i) \quad (4)$$

where the parameters α_j are assumed to be uniformly distributed.

Estimation of proportions of socio-economic levels under Dirichlet-multinomial models employed Gibb's sampling. Gibb's sampling is a Markov Chain Monte Carlo (MCMC) method for generating a sample from a joint posterior density. Consider random variables $[y_1, y_2, \dots, y_k]$. The conditional densities denoted as $[y_n | y_1, y_2, \dots, y_{k-1}]$ where n goes from 1 to k are all samples from the joint density $[y_1, y_2, y_3, \dots, y_k]$. Gibb's sampling generates a random sample from the joint density as follows: Let $[y_1^{(0)}, y_2^{(0)}, y_3^{(0)}, \dots, y_k^{(0)}]$ be the initial values. By sampling one observation from each conditional density, the first iteration is completed.

Hence, the first iteration is denoted as $[y_1^{(1)}, y_2^{(1)}, y_3^{(1)}, \dots, y_k^{(1)}]$. The iteration continues until the number of iterations specified is met. Estimation in hierarchical Bayesian models used the software WinBUGS or R.

3 Results and Discussion

Estimates of the proportions for low, middle and high-income classes were computed using direct and Dirichlet-multinomial estimation in the regional level. Direct and Bayesian Dirichlet-multinomial (DM) regional level estimates of the proportions of households belonging to socio-economic classes are shown in Table 2.

Table 2 Regional direct and bayesian DM estimates of proportions of socioeconomic classes

Regions	Low income		Middle income		High income	
	Direct (%)	DM (%)	Direct (%)	DM (%)	Direct (%)	DM (%)
I—Ilocos	82.22	82.22	17.73	17.73	0.04	0.04425
II—Cagayan Valley	81.47	81.47	18.42	18.42	0.11	0.10810
III—Central Luzon	71.57	71.57	28.29	28.29	0.14	0.14070
IV A—CALABARZON	67.02	67.02	32.88	32.88	0.10	0.10120
IV B—MIMAROPA	88.81	88.81	11.12	11.12	0.07	0.06672
IX—Zamboanga Peninsula	85.52	85.52	14.48	14.48	0.00	0.00016
V—Bicol	87.53	87.53	12.20	12.20	0.27	0.26830
VI—Western Visayas	84.65	84.65	15.28	15.28	0.07	0.06685
VII—Central Visayas	81.88	81.88	18.12	18.12	0.00	0.00008
VII—Eastern Visayas	86.31	86.31	13.50	13.50	0.19	0.19380
X—Northern Mindanao	82.55	82.55	17.39	17.39	0.06	0.05592
XI—Davao	83.77	83.77	16.23	16.23	0.00	0.00012
XII—SOCCSKSARGEN	88.41	88.41	11.49	11.49	0.10	0.10360
NCR	46.88	46.88	52.38	52.38	0.74	0.73890
CAR	71.78	71.78	27.93	27.93	0.29	0.29080
ARMM	96.18	96.18	3.82	3.82	0.00	0.00019
CARAGA	88.80	88.80	11.13	11.13	0.07	0.06976

The Bayesian Dirichlet-multinomial estimation yielded more realistic nonzero estimates with smaller standard errors in comparison to the direct estimates which included zero proportions with high standard errors. The regions with the highest estimates of low-income group are ARMM (96.18%), IV-B MIMAROPA (88.81%) and CARAGA (88.80%). The middle-income group has the highest estimate in NCR (52.38%) where many cities/urban areas are situated followed by nearby sub-urban areas like IV-A CALABARZON (32.88%). The proportions of the high-income group are too small in comparison to the other classes with the highest estimate in NCR (0.74%) followed by CAR (0.29%).

Small area statistics such as provincial level estimates are obtained. The provincial level estimates of low and middle-income groups from the direct and Bayesian Dirichlet-multinomial (DM) procedures are almost the same. Provinces with the highest proportions of low and middle-income households are shown in Table 3.

Table 3 Provincial estimates of proportions of low- and middle-income class

Provinces with highest proportions of low-income households		Provinces with highest proportions of middle-income households	
Provinces	Bayesian DM (%)	Provinces	Bayesian DM
Tawi-tawi	99.16	NCR fourth district	58.39
Sulu	97.47	NCR first district	54.42
Maguindanao	95.81	NCR second district	52.32
Basilan	95.17	Benguet	47.38
Lanao del Sur	94.35	Cavite	46.13
Saranggani	93.61	NCR third district	44.70
Romblon	92.32	Rizal	44.45
Masbate	91.85	Batanes	42.86
Davao Oriental	91.40	Pampanga	37.22
Compostela Valley	91.36	Laguna	37.12
Oriental Mindoro	91.17	Bulacan	36.76
Guimaras	90.85	Bataan	34.77
Sultan Kudarat	90.79	Batangas	26.19
Quezon	90.27	Aurora	26.16
North Cotabato	90.19	Nueva Vizcaya	25.92
Zamboanga del Norte	89.83	Biliran	25.32
Agusan del Sur	89.33	Ilocos Norte	24.04
Antique	89.20	La Union	23.17
Misamis Occidental	89.15	Davao del Sur	22.67
Negros Occidental	88.87	Cebu	22.47

The smaller domain provincial estimates are more helpful in policy making and program implementation. The provinces with the highest estimated proportions of low-income households are concentrated in Mindanao, namely, Tawi-tawi (99.16%), Sulu (99.47%), and Maguindanao (95.81%). These are also the provinces with the least estimated proportions of middle-income households. The provinces with the highest estimated proportions of middle-income households are concentrated in NCR, namely fourth district (58.39%), first district (44.98%) and second district (52.38%). These are usually the provinces with lower proportions of low-income households.

Table 4 shows the provinces with the highest and lowest proportions of high-income class. Notice that the proportions are significantly much lower compared to the proportions of low and middle-income groups. The advantage of using Bayesian Dirichlet-multinomial model over the direct method in estimating the proportions is shown in the result that the direct estimates included numerous zero proportions while the Bayesian Dirichlet-multinomial estimates included realistic nonzero estimates with smaller standard errors.

Table 4 Provincial direct and bayesian estimates of proportions of high income classes

Provinces	Direct (%)	Bayesian DM (%)	Provinces	Direct (%)	Bayesian DM (%)
Siquijor	1.70	1.70600	Misamis Occidental	0	0.00091
NCR fourth district	1.65	1.65100	Agusan del Norte	0	0.00087
Catanduanes	1.63	1.63500	Ilocos Norte	0	0.00085
Albay	0.63	0.62660	Agusan del Sur	0	0.00084
Mt. province	0.61	0.61450	Bataan	0	0.00078
Masbate	0.61	0.60800	Sultan Kudarat	0	0.00076
NCR first district	0.59	0.59200	Lanao del Sur	0	0.00075
NCR second district	0.57	0.56990	Compostela Valley	0	0.00074
Zambales	0.54	0.54060	Sorsogon	0	0.00071
Rizal	0.53	0.53340	Western Samar	0	0.00070
Benguet	0.51	0.51290	La Union	0	0.00068
Southern Leyte	0.51	0.51010	Oriental Mindoro	0	0.00064
Eastern Samar	0.49	0.49490	Lanao del Norte	0	0.00061
Zamboanga Sibugay	0.49	0.48920	Davao del Norte	0	0.00061
Nueva Vizcaya	0.43	0.42880	Maguindanao	0	0.00058
Occidental Mindoro	0.40	0.39730	Zamboanga del Norte	0	0.00055
Ilocos Sur	0.32	0.31810	North Cotabato	0	0.00047
Palawan	0.32	0.31530	Bukidnon	0	0.00043

4 Conclusions and Recommendations

In summary, the direct estimation provided high standard errors for the estimates, primarily because of the limited information given by the data at hand. By applying Bayesian estimation, the available information at hand was used in order to simulate more samples, hence standard errors for the estimates were significantly lower. The advantage of using hierarchical Bayesian models over the direct method in estimating the proportions are: (1) the estimates now follow a posterior distribution which can be computed by approximating the values for the parameters of the prior distribution and (2) the estimates are more precise compared to the direct method when the sample size is small.

This study will be further extended by using sequential Bayesian estimation by inclusion of available recent FIES data and other possible prior distributions. Thematic maps using statistical software like ArcGIS to visualize possible clustering of low-income and middle-income groups will be added. For future studies, the researchers suggest the following: (1) Consider the use of other Bayesian models and priors. (2) Estimate smaller area domains like municipalities which will further help in policy-making and program implementations. (3) Apply Bayesian small-area estimation methods to other official statistics such as poverty incidence, food poverty incidence and multi-poverty indices.

References

1. Virola, R.A., Addawe, M.B., Querubin, I.T.: Trends and characteristics of the middle-income class in the philippines: is it expanding or shrinking? In: Proceedings of 10th National Statistics Convention (2007)
2. Virola, R.A., Encarnacion, J.O., Balamban, B.B., Addawe, M.B., Viernes, M.M.: Will the recent robust economic growth create a burgeoning middle income class in the Philippines? In: Proceedings of 12th National Statistics Conference, Mandaluyong City, Philippines, Oct 1–2, pp. 2–13 (2013)
3. Lamberte, E., Reyes, M., Cabantog, R.: Going Beyond Income Measures in Mapping Poverty. National Statistics Office, Manila (2003)

Chapter 40

Efficiency of Fishery Production in Malaysia Using Data Envelopment Analysis



Anis Atiqah Abdul Rais, Siti Shaliza Mohd Khairi, Zalina Zahid
and Noor Asiah Ramli

Abstract Fish is among a low-fat and high-protein food sources that provides health benefits for human body. Meanwhile, fishery industries contributes to a wide range of employment and economic benefits for individuals and the country. As time past, the global fish stock has been fully exploited and the sign of recovery from it is tremendously low. Therefore, this research is conducted to find the efficiency level of fishery production in Malaysia using Data Envelopment Analysis (DEA). The Constant Return to Scale (CRS) with output orientation DEA model is applied. The result shows that Perak was the most efficient state throughout year 2004–2014 with mean score of 0.9440, followed by Perlis and Pahang with mean scores of 0.9036 and 0.8001 respectively. The least efficient state for the 11-year period was Negeri Sembilan with mean score of 0.1470. Overall, this research has successfully demonstrated the different efficiency score of fishery production for each state in Malaysia.

Keywords Fishery · Efficiency level · Data envelopment analysis · DMU

A. A. A. Rais (✉) · S. S. M. Khairi · Z. Zahid · N. A. Ramli
Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: anisabdrais@gmail.com

S. S. M. Khairi
e-mail: shalizakhairi@tmsk.uitm.edu.my

Z. Zahid
e-mail: zalina@tmsk.uitm.edu.my

N. A. Ramli
e-mail: asiahramli@tmsk.uitm.edu.my

1 Introduction

Fish is one of the most important protein sources for human to develop and build body tissues, muscles and bones. As for fishery industries, they provide a wide range of employment and economic benefits not only for individual but also for the country. Past studies found that about 20% of the global fish stock were fully exploited, over exploited or even depleted and it has been observed that the sign for fishery stock recovery is tremendously low (FFO [1]). By year 2030, Fishing Future Organization has predicted, the direct human consumption will exceed around 232 million metric tons, however, only 170 million metric tons able to be produced based on the current trajectory of the fishery production system. Three alternatives are suggested by Fishing Future Organization to prevent fish shortage includes reducing waste, improving fisheries as well as increasing numbers of aquaculture product (FFO [1]).

Many researchers conducted the investigation of fishery production globally and as in Malaysia it has been done much earlier by Aisyah et al. [2], Mustapha et al. [3] and liyasu et al. [4]. Efficiency can be defined as the degree of vessel performance at maximum level using the similar level of input as stated by Pascoe and Tingley [5]. Technical efficiency can be defined as the difference in the actual and potential output measured when the fixed and variable input is constant under the same level of observation [6]. Efficiency measurement has been widely used the application of Data Envelopment Analysis (DEA) in multiple sectors worldwide such as banking sectors, hospitality sectors as well as the agriculture sectors. According to Aisyah et al. [2], DEA is a useful tool to make comparison models in finding the efficiency of a production with multiple number of input and output. Narayankumar [7] has studied the level of fishery production in India by calculating the economic performance of fishery methods. The tabular analysis was conducted with the combination of value in fixed costs, operating cost per trip, gross revenue, net operating income per trip as well as annual revenue. The result shows the highest capital-productivity ratio for a day is Goa with score of 0.65 using purse seines method of fishing. The highest capital productivity ratio of multi-day operations is Satpati with score of 0.69 by using gillnet method.

Similar approach done by Geetha et al. [8], also use the economic efficiency of mechanized fishing in Chennai, Tamil Nadu. The operating ratio and labor productivity were calculated and the result indicates that the fishery production surpasses the breakeven point for all types of fishing units and the profit. In Malaysia, liyasu et al. [4], proposed the bootstrapped DEA method to estimate the bias-corrected technical efficiency (BCTE) of different culture systems and species of freshwater aquaculture for three states which are Perak, Selangor and Pahang.

Results show that technical efficiency score for all culture systems and species are below the optimal level. Therefore, the objective of this study is to investigate the efficiency of fishery productions in Malaysia towards the controlled variables imposed on the fishery activities.

2 Methodology

The empirical studies of DEA have been used widely for both finding the rate of efficiency as well as the technical efficiency by imputing multiple numbers of input and outputs of variables [9]. In fishery, DEA is used to measure the efficiency level of fishery productions based on the controlled variables used for fishery production. High level of efficiency score indicates the state has fully utilized the controlled input variables on conducting fishery activities.

In this study, data of 13 states in Malaysia for 11 years between 2004 and 2014 are used as Decision Making Units (DMUs) in DEA. The scope of the study covers all the Malaysian Fishery Water Zone and all fishery districts registered under the Department of Fishery Malaysia exclude Labuan. All data sources are taken solely from Department of Fishery Malaysia.

The calculation for level of efficiency score of fishery production for each state in Malaysia is conducted by using Efficiency Measurement System (EMS) programming. Efficiency measurement can be expressed by

$$\text{Efficiency} = \frac{\sum_{r=1}^n u_r \times y_r}{\sum_{i=1}^m v_i \times x_i} \tag{1}$$

Based on Table 1, u_r indicate weight of output of fishermen gain r for quantity of output y_r such that y_1 : fishery revenue and y_2 : total landing of marine fish. As for v_i indicate weight of input of fishermen gain r for quantity of input x_i such that x_1 :

Table 1 Input and output variables

Variables		Description
Input	Number of fishermen	Number of workers used for fishing activities based on licensed fishing vessel for each state in Malaysia
	Number of fish vessels	Number of vessels based on engine power of the boat for each state in Malaysia
Output	Landing of marine fish	Number of metric tons produce per year for each state in Malaysia
	Fishery revenue	Value of capture fisheries per year (RM Millions) for each state in Malaysia

number of fishermen and x_2 : number of fish vessels. The DEA model follows the Constant Returns to Scale (CRS) with output orientation. The model can be expressed as follows

$$\begin{aligned} \text{Maximize} &= \frac{\sum_{i=1} u_i y_i}{\sum_{i=1} v_i x_i} = \phi \\ \text{subject to} & \frac{\sum_{i=1} u_i y_i}{\sum_{i=1} v_i x_i} \leq 1 \end{aligned} \tag{2}$$

Where the value of ϕ indicate the efficiency score of fishery production for each state in Malaysia $0 < \phi \leq 1$. While x and y are the input and output variables respectively, i and j are the number of DMUs for each states ($i = 1, 2, \dots, 13; j = 1, 2, \dots, 13$), u_j and v_j are the weight attached for input and output variables. When the value of ϕ gradually increase and reach to the score of 1, this indicates that the state has become the most efficient state in Malaysia and fully utilized its input variables. On the other hand, when the value of ϕ is closer towards zero, it indicates the state is underutilized its input variables and show the state is inefficient on conducting fishery activities.

3 Analysis of Efficiency Score for Fishery Production in Malaysia

The analysis has been done using DEA to fulfil the objective of the study in determining the efficiency score for fishery production for each state in Malaysia. The period for data taken is 11 years based on the year of 2004 till 2014. Before the DEA analyses are conducted, the isotonicity and multicollinearity assumption for input and output variables must be fulfilled. This is to ensure the DEA model is valid to examine the efficiency score of fishery production. Therefore, Pearson’s correlation coefficient test is conducted and shown in Table 2.

Table 2 Correlation test of overall input and outputs for 2004–2014

	Input		Output	
	Number of fishermen	Number of fish vessel	Total fish landing	Total fish revenue
Number of fishermen	1			
Number of fish vessel	0.9393	1		
Total fish landing	0.676	0.5872	1	
Total fish revenue	0.6565	0.5414	0.9002	1

Table 3 Correlation test of DEA models with size of operations

Size of operations	Model 1	Model 2	Model 3	Model 4
No of fishermen	0.153101	-0.01395	0.004283	-0.06964
No of fish vessel	0.041036	-0.19105	-0.09024	-0.24847

Result in Table 2 shows that there are strong relationships between each input and each output. As the multicollinearity assumption must be fulfilled, the overall inputs and outputs must be separated into one to one DEA model. For instance, for each model creates involved only one input variable and one output variable such that number of fishermen against total fish landing as Model 1, number of fish vessels against total fish landing as Model 2, number of fishermen against total fish revenue as Model 3 and number of fish vessels against total fish revenue as Model 4.

Theoretically the concept of Constant Return to Scale (CRS) must be insensitive with the change in the size of operation [10]. Therefore, the correlation tests among the efficiency score for each model with the size of operation are conducted to validate assumption.

Table 3 shows that all DEA models are insensitive with the changes occur in the size of operations. Thus, it satisfies the characteristic of CRS whereby the efficiency score not influence with the changes occur in number of fishermen and number of fish vessels. Therefore, only one DEA model can be used as the main model to be used further for representing the efficiency score of fishery production for each state in Malaysia. Model 3 has the least or weakest relationship with the size of operation compared to other models. Therefore, Model 3 is used as the main model to evaluate the efficiency score. Model 3 is the DEA model involves the input variable of number of fishermen against total fish revenue.

In Table 4, the efficiency score of Model 3 for each year from 2004 until 2014 are shown. The mean efficiency score for fishery production show Perak has the highest score with mean score of 0.9440 throughout the 11 years. Then, followed by Perlis with mean score of 0.9036 and Pahang with mean score of 0.8001. The least efficient state is Negeri Sembilan with means score of 0.1470 which is the lowest score among all states in Malaysia.

Table 4 The efficiency score of Model 3, 2004–2014

States	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Mean
Johor	0.3221	0.5320	0.5320	0.5026	0.4983	0.2875	0.3740	0.4569	0.5085	0.6452	0.6055	0.4786
Kedah	0.5518	0.6149	0.6149	0.6644	0.5663	0.4711	0.3237	0.4080	0.4894	0.5733	0.9652	0.5675
Kelantan	0.2386	0.5246	0.5246	0.4220	0.3729	0.2126	0.1763	0.2934	0.3414	0.2579	0.3080	0.3338
Malacca	0.1992	0.1986	0.1986	0.1779	0.1803	0.0839	0.0782	0.0886	0.0980	0.1543	0.1848	0.1493
Pahang	1.0000	0.8787	0.8787	0.8503	1.0000	0.7360	0.6355	0.7333	0.7844	0.6940	0.6099	0.8001
Perak	0.9234	1.0000	1.0000	1.0000	0.9411	0.8147	0.8692	0.8493	0.9869	1.0000	1.0000	0.9440
Perlis	0.4426	0.9126	0.9126	0.9640	0.8859	1.0000	1.0000	1.0000	1.0000	0.8455	0.9768	0.9036
Penang	0.5127	0.7826	0.7826	0.8150	0.7171	0.3930	0.4548	0.7325	0.6737	0.6341	0.5607	0.6417
Sabah	0.2850	0.3126	0.3126	0.3316	0.3011	0.2257	0.2347	0.2405	0.2520	0.3033	0.3417	0.2855
Sarawak	0.4987	0.4790	0.4790	0.5382	0.6035	0.2150	0.2305	0.2722	0.3110	0.4061	0.4511	0.4077
Selangor	0.7114	0.7434	0.7434	0.5860	0.5075	0.5713	0.5443	0.4591	0.6079	0.4736	0.5242	0.5884
Terengganu	0.4772	0.6838	0.6838	0.5276	0.6556	0.3666	0.3094	0.3702	0.4516	0.4173	0.3402	0.4803
N.Sembilan	0.1098	0.1538	0.1538	0.1593	0.1418	0.1440	0.1579	0.1738	0.1160	0.1183	0.1881	0.1470

4 Conclusion and Recommendation

This study is conducted to determine the efficiency score of fishery production for all states in Malaysia. The calculation of efficiency score is calculated using DEA approach which accomplice all the inputs and outputs gathered into four different models. Then, after the DEA model validation is conducted, Model 3 which involved number of fishermen as input and the total fish revenue as output has become the main model to determine the technical efficiency score of fishery production. The result obtained show that Perak is the most efficient states in Malaysia throughout the year 2004 until 2014 with the mean score of 0.9440 while Negeri Sembilan is the most inefficient states in the fishery production because only able to reach the mean technical efficiency score of 0.1470. For future research, researchers can include different type of controlled variable for inputs and outputs such as in economic overview that incorporate with the operating cost and financial support by Government in term of grand, loan and incentive for better impact and broaden perspective on evaluating the efficiency level of fishery production.

Acknowledgements The authors would like to express their gratitude to Ministry of Education (MOE), Malaysia and Universiti Teknologi MARA (UiTM) for funding this research project. (600-IRMI/RAGS 5/3 (71/2015))

References

1. Fishing Future Organization: Getting to Eden-Building the ideal future for the global fish food system through the collective actions. <http://www.fishingfuture.org/> (2015). Accessed 4 Mar 2016
2. Aisyah, N., Arumugam, N., Hussein, M.A., Latiff, I.: Factors affecting the technical efficiency level of inshore fisheries in Kuala Terengganu, Malaysia. *Int. J. Agriculture Manag. Dev.* **2**, 49–56 (2012)
3. Mustapha, N.H.N., Aziz, A.A., & Hashim, N.M.H.: Technical efficiency in aquaculture industry using Data Envelopment Analysis (DEA) window: Evidences from Malaysia. *J. Sustain. Sci. Manag.* **8**(2), 137–149 (2013)
4. Iliyasa, A., Mohamed, Z.A.: Evaluating contextual factors affecting the technical efficiency of freshwater pond culture systems in Peninsular Malaysia: a two-stage DEA approach. *Aquac. Rep.* **3**, 12–17 (2016)
5. Pascoe, S., Tingley, D.: Capacity and technical efficiency estimation in fisheries: parametric and non-parametric techniques. In: *Handbook of Operations Research in Natural Resources*, pp. 273–294. Springer US (2007)
6. Tsitsika, E.V., Maravelias, C.D., Wattage, P., Haralabous, J.: Fishing capacity and capacity utilization of purse seiners using data envelopment analysis. *Fish. Sci.* **74**(4), 730–735 (2008)
7. Narayanakumar, R.: Economic efficiency in fishing operations-Technology, exploitation and sustainability issues. *Manual on World Trade Agreements and Indian Fisheries Paradigms: A Policy Outlook* (2012)
8. Geetha, R., Narayanakumar, R., Shyam, S.S., Aswathy, N., Chandrasekar, S., Srinivasa Raghavan, V., Divipala, I.: Economic efficiency of mechanised fishing in Tamil Nadu—a case study in Chennai. *Indian J. Fish.* **61**(1), 31–35 (2014)

9. Chen, Y., Li, Y., Liang, L., Salo, A., Wu, H.: Frontier projection and efficiency decomposition in two-stage processes with slacks-based measures. *Eur. J. Oper. Res.* **250** (2), 543–554 (2016)
10. Avkiran, N.K.: *Productivity analysis in the service sector with data envelopment analysis* 3rd edn., QLD, Australia (2006)

Chapter 41

Examining the PPP Theory for ASEAN-5 with Panel Data Analysis



Niri Martha Choji and Siok Kun Sek

Abstract This research paper examines the PPP theory for ASEAN-5 by utilizing panel data. PPP theory states that the exchange rate between two nations should be the same as the ratio of the aggregate price levels between the two nations. PPP is a tool used to determine the general economic welfare of countries. Because of its importance, a lot of effort has been put into testing the validity of the long-run purchasing power parity. This paper tries to examine if PPP exists in ASEAN-5 using very recent data set and Robust methods. Applying the Mean Group (MG) and the Pooled Mean Group (PMG) estimators, this paper finds sufficient evidence to support the PPP theory both in the short- and long-run. In long-run, domestic prices (LCPI) cause the nominal exchange rates to depreciate while the foreign price (LCPIUS) makes the nominal exchange rates to appreciate in ASEAN-5. However, in the short-run, only the foreign price is significant and causes the nominal exchange rates to appreciate. Consequently, depreciation makes exports cheaper, imports very expensive and cause inflation to increase. Nonetheless, appreciation of the nominal exchange rate will cause export to be more expensive, imports cheaper and thereby reducing inflation in ASEAN-5. In conclusion, the foreign price contributes more to the adjustments in the nominal exchange rates, thereby making the effect of nominal exchange rate appreciation more pronounced than depreciation in ASEAN-5.

Keywords Purchasing power parity · Mean group · Pooled mean group · ASEAN-5

N. M. Choji (✉)

Department of Mathematics, Plateau State University, Bokkos, Nigeria
e-mail: marthaniri@yahoo.com

N. M. Choji · S. K. Sek

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia
e-mail: sksek@usm.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_41

333

1 Introduction

The purchasing power parity theory is a very important theory in the area of international finance that is why so many researches have been carried out and are still been done by researchers. The theory says that the nominal exchange rate between two currencies should be the same as the ratio of aggregate price levels between the two currencies in order for the unit of currency of one nation to have the same power to purchase goods and services in a foreign nation [1]. The importance of the purchasing power parity theory is not limited to the fact that it is used to compare and measure national income/price levels among countries.

Due to its importance, many studies have utilized several methods in testing for the purchasing power parity empirically. A lot of previous studies utilized different tests of unit root, and tests of cointegration to find out if the theory of the purchasing power parity holds true. Unit root tests are mostly applied to the real exchange rates to see if the real exchange rates are stationary. If the real exchange rates are stationary, then purchasing power parity theory is said to hold, otherwise, purchasing power parity fails to hold. The cointegration tests, on the other hand, are applied to a combination of the nominal exchange rates and the aggregate price levels (the domestic and foreign prices). Cointegration occurs if a blend of at least two nonstationary series yield a long-run stationary relationship when the series have the same order of integration. If the null hypothesis of no cointegration is rejected implying a long-run relationship, then we say that the theory of purchasing power parity holds. Previous studies applied mostly the unit root tests and/or cointegration tests to validate the purchasing power parity.

Among empirical works that utilized the unit root tests to test for the purchasing power parity theory include the work of [2], who examined the long-run purchasing power parity (PPP) by utilizing panel method to test for unit roots in 84 countries using the Dollar as base currency. They found strong proof of PPP in countries closer to the U.S that are open to trade. They also showed that PPP is not valid for panels of Asian and African countries but is valid for panels of Latin American and European countries. Their discoveries revealed that country characteristics explain deviations from and support to long-run PPP. Furthermore, [3], re-examined the PPP hypothesis in 15 EU countries using a heterogeneous panel nonlinear unit root test in which symmetric or asymmetric exponential smooth transition autoregressive method is allowed in the alternative hypothesis. While the outcomes of the heterogeneous panel linear and symmetric nonlinear unit root tests are not in agreement with the PPP hypothesis, the heterogeneous panel asymmetric nonlinear unit root test they proposed gave support for the PPP hypothesis. They concluded

that the panel linear unit root test or the panel nonlinear unit root tests that do not consider asymmetry may be deceptive.

Recently, [4] found evidence in agreement with the PPP hypothesis for 20 African nations by utilizing nonlinear asymmetric unit root test by [3] by means of the sequential panel selection method (SPSM) of [5] from 1971I–2012IV which are accessible for a substantial number of African nations by [6]. Moreover, [7], found support for the PPP theory in 34 OECD nations from 1994 to 2013 using a new panel unit root test with sharp breaks and smooth shifts, a unique approach to panel unit root testing by [8]. Most recently, [9], investigated the validity of PPP for Turkey between January 2003 and June 2014 using the nonlinear threshold unit root testing procedure proposed by [10], in contrast to many previous studies in which linear unit root test methods have been employed. The results of the empirical analysis showed that PPP is not valid in Turkey.

However, studies that applied mainly the method of cointegration include [11] who found some evidence for the weak form of PPP when he applied the [12] cointegration method on data for 17 African countries from the period 1981–1994. Furthermore, [13] used the test proposed by [14] test which allows for heterogeneous coefficients of slope and for differences in the short-run dynamics of individuals of a panel to test the weak form of the purchasing power parity hypothesis on monthly data of 11 developed countries from 1918:01 to 1995:06. They found evidence in support of PPP when they used tradable goods in their price indices which indicate that the failure of PPP in previous studies could be the inclusion of non-traded goods in the aggregate price index. Moreover, [15] found proof to agree with the weak type of PPP against the strong type of PPP in G7 (France, Germany, Italy, Japan, the United Kingdom, the United States and Canada) Countries from January, 1973–December, 1998 using the maximum likelihood panel cointegration method of [16] which is robust in several dimensions in testing the strong form of PPP.

In addition, [17] applied the approach of Im et al. (2008 working paper) cointegration test using stationary instrumental variables to examine if PPP is valid on updated version of Taylor data set (Rev Econ Stat 84(1): 139–150, 2002) for 21 developed countries. They found evidence to support the validity of PPP using Taylor's data set. Finally, [18] tests the purchasing power parity hypothesis for a set of ASEAN-5 nations with data from 1968:I–2009:II. They utilised the Langrage Multiplier (LM) cointegration test of [19] that accommodates a cross-sectional dependence, substantial level of country specific heterogeneity as well as multiple structural breaks. The results of the panel cointegration test provide concrete evidence in support of purchasing power parity.

Most of the previous studies considered in the literature found support for the theory of the purchasing power parity except the work of [9] who used a time series method to test for the purchasing power parity. This is not surprising because the

time series unit root tests are known to have low power in rejecting the null of a unit root. All the other works who used panel methods found support for the purchasing power parity because panel data methods (panel unit root tests and cointegration) are known to have more power.

As mentioned earlier, most previous studies on purchasing power parity utilized the unit root tests and/or normal cointegration methods in testing for the validity of PPP in the long-run. This study uses the autoregressive distributed lag (ARDL) model approach to examine the validity of purchasing power parity in both the long-run and the short-run. The ARDL cointegration approach has more desirable qualities over the normal cointegration. It can be done when the variables are a mixture of $I(0)$ and $I(1)$ unlike the normal cointegration where the variables have to be $I(1)$. Moreover, the ARDL approach to cointegration can also be done when the variables are $I(0)$ or $I(1)$. One of the assumptions surrounding the ARDL model is that none of the variables should be $I(2)$. If any of the variables is $I(2)$, then the model is not valid. Because only a handful of studies utilized the ARDL approach, we, therefore, contribute to the literature in testing the purchasing power parity using the Mean Group (MG) and the Pooled Mean Group (PMG) ARDL approaches. However, results of the Hausman test revealed that the PMG estimator is preferred. The ARDL models in this study are valid since have $I(0)$ and $I(1)$ variables, and there is evidence of cointegration. This study found support for the purchasing power parity theory both in the long-run and short-run.

The remaining paper is organised thus; Sect. 2 presents the data and methodology for this study, Sect. 3 presents results and discussion on the results and finally, we conclude all investigations carried out in Sect. 4.

2 Data and Methodology

2.1 Data

In this paper, we used data collected from *Datastream*, Thomson Reuters. It is a set of monthly data for a group of ASEAN-5 countries starting from January 1996 to August 2016. The countries contained in our sample are; Malaysia, Indonesia, Thailand, Philippines, and Singapore. The data consist of the nominal exchange rate (local currency per 1USD), Consumer price index (CPI) for each country, and CPI for the US. The US was used as the base currency.

2.2 The Mean Group (MG) and the Pooled Mean Group (PMG) Estimators

Let an autoregressive distributive lag (ARDL (p, q_1, q_2)) of dynamic panel specification be of the form

$$\begin{aligned} LEXRATE_{it} = & \sum_{j=1}^p \lambda_{ij} LEXRATE_{i,t-j} + \sum_{j=0}^{q_1} a'_{ij} LCPI_{i,t-j} \\ & + \sum_{j=0}^{q_2} b'_{ij} LCPIUS_{i,t-j} + \varepsilon_{it} \end{aligned} \tag{1}$$

where $i = 1, 2, \dots, N$ represents the number of countries, $t = 1, 2, \dots, T$ is number of periods, $LEXRATE$ is the dependent variable; $LCPI$ and $LCPIUS$ are $K \times 1$ vector of explanatory variables; a'_{ij} and b'_{ij} are the $k \times 1$ coefficient vectors; λ_{ij} are scalars, and ε_{it} is the disturbance term p, q_1, q_2 are the lags of the dependent variable and the two independent variables. Suppose the variables in Eq. (1) are integrated of order one and are cointegrated, then the error term is a stationary process for all i . One of the main features of cointegrated variables is their sensitivity to any movement away from long-run equilibrium. This characteristic reveals an error correction where the short-run dynamics of the variables are affected by the movement away from equilibrium [20]. However, (1) can be written into an error correction form as;

$$\begin{aligned} \Delta LEXRATE_{it} = & \phi_i (LEXRATE_{it-1} + \beta'_{ai} LCPI_{it} - \beta'_{bi} LCPIUS_{it}) \\ & + \sum_{j=1}^{p-1} \lambda_{ij} \Delta LEXRATE_{it-j} + \sum_{j=0}^{q_1} a'_{ij} LCPI_{it-j} + \sum_{j=0}^{q_2} b'_{ij} LCPIUS_{it-j} + \varepsilon_{it} \end{aligned} \tag{2}$$

where $\phi_i = -1 \left(1 - \sum_{j=1}^p \lambda_{ij} \right)$; $\beta_{ni} = \sum_{j=0}^{q_l} n_{ij}$; $n = a, b$; $l = 1, 2$; $\tilde{\lambda}_{ij} = \sum_{m=j+1}^p \lambda_{im}$;

$$\tilde{n}_{ij} = \sum_{m=j+1}^{q_l} n_{im}.$$

By regrouping, we can write the error correction Eq. (2) as;

$$\begin{aligned} \Delta LEXRATE_{it} = & \phi_i(LEXRATE_{i,t-1} - \theta'_{ai}LCPI_{it} - \theta'_{bi}LCPIUS_{it}) \\ & + \sum_{j=1}^{p-1} \lambda_{ij}\Delta LEXRATE_{i,t-1} + \sum_{j=0}^{q_1-1} a'_{ij}\Delta LCPI_{i,t-j} \\ & + \sum_{j=0}^{q_2-1} b'_{ij}\Delta LCPIUS_{i,t-j} + \varepsilon_{it} \end{aligned} \tag{3}$$

$\theta_i = -\left(\frac{\beta_{mi}}{\phi_i}\right)$; $n = a, b$ shows the long-run/equilibrium relationship between the dependent and independent variables in Eq. (3). λ_{ij} is the short-run coefficient of the independent variable while $n_{ij} = (a, b)_{ij}$ are the short-run coefficients which show the short-run effects of independent variables on *LEXRATE* in Eq. (3). The quantity ϕ_i is the error correction speed of adjustment term. If $\phi_i = 0$, there is no proof of long-run relationship. This quantity is supposed to be negative and significant from a prior assumption that the variables return to a long-run equilibrium. The vector ϕ_i , is very important since it contains the long-run relationships between the variables.

In this paper, we estimate Eq. (3) using the Mean Group (MG) estimator by [21] and the Pooled Mean Group (PMG) estimator by [22]. With the MG, the model is fixed independently for every country and a simple mean of the coefficients is computed. With the MG estimator, the slope coefficients, intercepts, and error variances across groups differ. However, the PMG estimator uses a combination of both pooling and averaging. In addition to allowing the intercept, short-run coefficients, and error variances to be different across groups (same as the MG estimator), the PMG estimator also makes the long-run coefficients to be the same across groups. [22] developed a maximum likelihood procedure to estimate the parameters since (3) is nonlinear in the parameters.

3 Results and Discussion

Prior to the Mean Group and Pool Mean Group analyses, several panel unit root tests were conducted to check the stationarity properties of the data. We chose to carry out several panel unit root tests because most of the unit root tests have their flaws, combining all these tests allows us to take advantage of the average

Table 1 Panel unit root tests

Variables/tests	ADF-Fisher	Breitung	IPS	LLC	Pesaran	Remark
LEXRATE	23.5608***	0.0341	-2.3736***	-2.6177***	0.6870	I(0)
ΔLEXRATE	561.283***	-26.9227***	-33.2478***	-38.7375***	-10.9320***	
LCPI	12.5433	1.8825	0.1587	-4.3423***	0.6930	I(1)
ΔLCPI	346.168***	-16.9445***	-21.2124***	-22.9109***	-8.1910***	
LCPIUS	6.7003	2.6756	-0.0673	-4.0472***	0.4170	I(1)
ΔLCPIUS	394.101***	-23.3421***	-23.2905***	-25.8758***	-10.8380***	

where *** is significance at 1% level

Table 2 MG and PMG ARDL(1,1,1) for ASEAN-5 using the exchange rates

Variable	MG	PMG
LR		
LCPI	0.8305(1.0735)	0.7066(0.2701)***
LCPIUS	0.0911(2.8514)	-1.5087(0.4062)***
SR		
ΔLCPI	0.1208(0.1728)	0.1941(0.1981)
ΔLCPIUS	-0.4487(0.1455)***	-0.5505(0.1927)***
ec	-0.0338(0.0058)***	-0.3877(0.0071)***

The table shows coefficients and standard errors in parentheses. *** indicates significance at 1% level

performance of all these tests. Results of panel unit root tests are displayed in Table 1. The table shows results of the ADF-fisher, Breitung, IPS, LLC and Pesaran panel unit root tests for all the variables at levels and at first difference (LEXRATE and ΔLEXRATE, CPI and ΔCPI, CPIUS and ΔCPIUS). The panel unit root tests results generally show that the data is a mixture of I(0) and I(1) variables. Because there is a mixture of I(0) and I(1), we, therefore, run the ARDL models to estimate both long-run and short-run relationships.

Table 2 presents result of the ARDL models of the MG and PMG estimators. For the MG, in the long-run, both coefficients of LCPI and LCPIUS are not significant, meaning that both domestic and foreign prices do not impact the nominal exchange rate significantly. In the short-run, the coefficient of ΔLCPI not significant indicating that the domestic price does not impact the nominal exchange rate significantly. However, the coefficient of ΔLCPIUS is significant meaning that the foreign price has a significant impact on the nominal exchange rate. More specifically, in the short-run, one percent increase in the foreign price (ΔLCPIUS) results in 44.8% appreciation in the nominal exchange rate. The convergence coefficient is negative and significant showing the presence of cointegration. For the PMG however, in the long-run, both the coefficients of LCPI and LCPIUS are significant indicating that the domestic and foreign prices both impact the nominal exchange rate significantly. Here, the coefficient of the domestic price (LCPI) is significant and positive indicating that one percent increase in the domestic price leads to about

71% depreciation of the nominal exchange rate. However, the coefficient of the foreign price (LCPIUS) is negative and significant implying that one percent increase in the foreign price leads to about 151% appreciation of the exchange rate. In short-run, the coefficient of Δ LCPI is not significant indicating that the domestic price does not impact the nominal exchange rate significantly. But the coefficient of Δ LCPIUS is significant implying that the foreign price has a significant effect on the nominal exchange rate. Specifically, in the short-run, one percent increase in the foreign price (Δ LCPIUS) results in about 55% appreciation of the nominal exchange rate. The convergence coefficient is negative and significant indicating the presence of cointegration. The presence of cointegration in the estimates confirms that the ARDL models are valid. However, in choosing the best model between the mean group and the pooled mean group, the Hausman test revealed that the pooled mean group is the preferred estimator with $\text{Chi}2(2) = 1.95$ and $\text{prob} > \text{Chi}2 = 0.3777$.

4 Conclusion

This paper examined the purchasing power parity theory for ASEAN-5 from 1996 to 2016 using the MG and the PMG ARDL models. First and foremost, the stationarity properties of the variables were examined by means of several panel unit root tests. The outcomes of the tests indicate the variables are a mixture of $I(0)$ and $I(1)$, meaning that some are stationary (do not have unit root) and some are not stationary (have unit root). Furthermore, since the variables are a mixture of $I(0)$ and $I(1)$, we went ahead to estimate the MG and the PMG estimators. Results of the MG estimator indicate that only the foreign (Δ LCPIUS) has an effect on the nominal exchange rate in the short-run. The foreign price makes the nominal exchange rate to appreciate. On the other hand, results of the PMG estimator show that the domestic price (LCPI) has an impact in the long-run but not in the short-run on the nominal exchange rate, indicating that the domestic price makes the nominal exchange rate to depreciate. However, the foreign price (LCPIUS) has a significant impact on the nominal exchange rate both in the short- and the long-run. In the long- and short run, the domestic price makes the nominal exchange rate in ASEAN-5 to appreciate. The convergence coefficients for both the MG and the PMG are negative and significant indicating the presence of cointegration and hence giving evidence that the purchasing power parity theory holds for ASEAN-5. However, the Hausman test revealed that the PMG is preferred.

Moreover, both the domestic price and the foreign price have significant impacts on the exchange rate. The foreign price has more effect on the exchange rate than the domestic price. We observe that the domestic and foreign price cause the nominal exchange rate in ASEAN-5 to depreciate and appreciate respectively. Consequently, depreciation makes exports to be cheaper, imports more expensive and thereby cause inflation to increase. On the other hand, appreciation of the nominal exchange rate will cause export to be more expensive, imports cheaper and

thereby reducing inflation in ASEAN-5. In conclusion, the foreign price contributes more to the adjustments in the nominal exchange rates, thereby making the effect of nominal exchange rate appreciation more pronounced than depreciation in ASEAN-5.

References

1. Choji, N.M., Sek, S.K.: Testing for purchasing power parity in 21 African countries using several unit root tests. *AIP Conference Proceedings*, vol. 1830, pp. (080017)1–7 (2017). <https://doi.org/10.1063/1.4980999>
2. Alba, J.D., Papell, D.H.: Purchasing power parity and country characteristics: evidence from panel data tests. *J. Dev. Econ.* **83**(1), 240–251 (2007)
3. Emirmahmutoglu, F., Omay, T.: Reexamining the PPP hypothesis: a nonlinear asymmetric heterogeneous panel unit root test. *Econ. Model.* **40**, 184–190 (2014)
4. Bahmani-Oskooee, M., Chang, T., Lee, K.-C.: Panel asymmetric nonlinear unit root test and PPP in Africa. *Appl. Econ. Lett.* **23**(8), 554–558 (2015). <https://doi.org/10.1080/13504851.2015.1088132>
5. Chortareas, G., Kapetanios, G.: Getting PPP right: identifying mean-reverting real exchange rates in panels. *J. Bank. Financ.* **33**(2), 390–404 (2009). <https://doi.org/10.1016/j.jbankfin.2008.08.010>
6. Bahmani-Oskooee, M., Kones, A.: Real and nominal effective exchange rates of African countries during 1971Q1–2012Q4. *Appl. Econ.* **46**(17), 1961–1984 (2014). <https://doi.org/10.1080/00036846.2014.889805>
7. Jiang, C., Bahmani-Oskooee, M., Chang, T.: Revisiting purchasing power parity in OECD. *Appl. Econ.* **47**(40), 4323–4334 (2015). <https://doi.org/10.1080/00036846.2015.1026592>
8. Bahmani-Oskooee, M., Chang, T., Lee, K.-C.: Purchasing power parity in the BRICS and the MIST countries: sequential panel selection method. *Rev. Econ. Financ.* **4**, 1–12 (2014)
9. Karagöz, K., Saraç, T.B.: Testing the validity of PPP theory for Turkey: nonlinear unit root testing. *Procedia Econ. Financ.* **38**, 458–467 (2016)
10. Caner, M., Hansen, B.E.: Threshold autoregression with a near unit root. *Econometrica* **69**(6), 1555–1596 (2001)
11. Nagayasu, J.: Does the long-run ppp hypothesis hold for Africa? Evidence from a panel cointegration study. *Bull. Econ. Res.* **54**(2), 181–187 (2002)
12. Pedroni, P.: Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econom. Theory* **20**(3), 597–625 (1995). <https://doi.org/10.1017/S0266466604203073>
13. Jenkins, M.A., Snaith, S.M.: Tests of purchasing power parity via cointegration analysis of heterogeneous panels with consumer price indices. *J. Macroecon.* **27**(2), 345–362 (2005)
14. Pedroni, P.: Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econom. Theory* **20**(3), 597–625 (2004)
15. Carlsson, M., Lyhagen, J., Osterholm, P.: Testing for purchasing power parity in cointegrated panels. *IMF Working Papers*, vol. 7, pp. 1–21 (2007)
16. Larsson, R., Lyhagen, J.: Inference in panel cointegration models with long panels. *J. Bus. Econ. Stat.* **25**(4), 473–483 (2007)

17. Wallace, F.H.: Cointegration tests of purchasing power parity. *Rev. World Econ.* **149**(4), 779–802 (2013)
18. Munir, Q., Kok, S.C.: Purchasing power parity of ASEAN-5 countries revisited: heterogeneity, structural breaks and cross-sectional dependence. *Glob. Econ. Rev.* **44**(1), 116–149 (2015)
19. Westerlund, J.: Testing for panel cointegration with multiple structural breaks. *Oxf. Bull. Econ. Stat.* **68**(1), 101–132 (2006)
20. Blackburne III, E.F., Frank, M.W.: Estimation of nonstationary heterogeneous panels. *Stata J.* **7**(2), 197–208 (2007)
21. Pesaran, M.H., Smith, R.: Estimating long-run relationships from dynamic heterogeneous panels. *J. Econom.* **68**(1), 79–113 (1995)
22. Pesaran, M.H., Shin, Y., Smith, R.P.: Pooled mean group estimation of dynamic heterogeneous panels. *J. Am. Stat. Assoc.* **94**(446), 621–634 (1999)

Chapter 42

Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach



M. Fariz Fadillah Mardianto, Sri Haryatmi Kartiko and Herni Utami

Abstract Recently, forecasting time series data with trend and seasonal or trend-seasonal combinations with time series forecasting methods that are often used, are bound by assumptions that must be reached. If it does not reach the assumptions that exist, the forecasting process becomes longer. This study provides an alternative approach used for time series data forecasting that has trend-seasonal combination pattern using nonparametric regression. Some nonparametric regression approaches such as the kernel and the Fourier series can be done by considering the predictor as the time scale for a regular period. For the same data, using Nadaraya–Watson kernel approach and Fourier series in nonparametric regression gives different results. The result of prediction using nonparametric regression with the Fourier series approach is closer to the original data, when compared to the kernel approach. For each oscillation parameters inputted, nonparametric regression with the Fourier series approach always provides smaller MSE results than MSE in every bandwidth for Nadaraya–Watson kernel approach.

Keywords Nonparametric regression · Kernel approach · Fourier series approach · Trend-seasonal data

1 Introduction

Time series forecasting is a way to make prediction from a set data, each one being recorded at specific time. In time series analysis, often be found some data pattern, such as trend, seasonal, and combination between trend and seasonal or

M. F. F. Mardianto (✉)

Department of Mathematics, University of Airlangga, Surabaya, Indonesia
e-mail: m.fariz.fadillah.m@fst.unair.ac.id

M. F. F. Mardianto · S. H. Kartiko · H. Utami

Department of Mathematics, University of Gadjah Mada, Yogyakarta, Indonesia
e-mail: s-kartiko@yahoo.com

H. Utami

e-mail: herni.utami@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_42

trend-seasonal pattern. Those patterns will pass some procedures when time series analysis be used, because there are some assumptions must be satisfied. Time series—regression approach is an alternative to forecast time series data [1]. One of regression approach that be developed is nonparametric regression. Nonparametric regression accommodate relationship between predictor and response with unspecific pattern. Here, the predictor is period, and the response is data that will be used to forecast. Some approaches which often be used in nonparametric regression is Fourier series [2], kernel and local polynomial [3], spline [4, 5], histogram [6], and wavelet [7]. Nonparametric regression purpose is estimating response based on predictor based on an estimator function smoothly [3]. The advantage of using nonparametric regression approach is having high flexibility, because data is expected to approximate a regression curve form without being influenced with subjectivity, and nonparametric regression is not too bound with assumptions [8].

Nonparametric regression can be applied for time series data, where time represents a predictor, and value data for certain time as response. Estimation of nonparametric regression curve that be used based on kernel and Fourier series. Kernel estimator has flexible form and its mathematical calculations are easily adjusted [3]. Kernel density estimation is based on two parameters, bandwidth and kernel parameters. The most popular estimator in nonparametric regression based on kernel approach that be used is Nadaraya–Watson kernel estimator. It used because has good flexibility in nonparametric regression with kernel approach. Not only using kernel approach for nonparametric regression, but also, we use Fourier series estimator to estimate nonparametric regression curve. To obtain a more consistent and efficient estimator for large observations is better done with the Fourier series approach, instead the kernel approach is used [9]. Based on [9], simulation data with large observational, have a slower rate of convergence in cosine kernel estimator than Fourier series estimator. One of the advantages of nonparametric regression approach using Fourier series is capable to overcome data with periodic pattern, in this case represents with trigonometric function [10].

Relationship between Fourier series and kernel estimator explored by [11]. In that research, written that for the periodic case there is a relationship between kernel and Fourier series estimator. The conclusion obtained from that research is kernel estimator is Fourier series estimator by using all possible exponential functions so that it can be matched on the data. Nonparametric regression with kernel estimator for time series data proposed by [12]. Nonparametric regression with Fourier series estimator for time series data examined by [1]. However, data pattern based on time series plot more precise when approximated by Fourier series, especially seasonal pattern and trend-seasonal pattern. This idea based on concept that be derived from [9, 11]. This study determines comparison result using kernel and Fourier series approach to estimate nonparametric regression curve for time series data, which

have trend-seasonal pattern. The cases raised in this study are based on data [13], on forecasting the number of aircraft landing at an airport that be observed for 12 years.

2 Literature Review

2.1 Kernel Estimator

The idea of kernel estimator was introduced by Rosenblatt in 1956. Rosenblatt wanted a flexible approximation in order to make smooth estimator and get mathematical tractability. The tractability here means a smooth function can be controlled according to the characteristics of the data.

Consider a kernel function $K(t)$ with bandwidth $h > 0$ defined as follows:

$$K_h(t) = \frac{1}{h} K\left(\frac{t}{h}\right) \quad (1)$$

that satisfies some properties that be seen in [3]. Some kernel functions can be studied in [3]. A random variable $\{(T_i)\}_{i=1}^n$ can be approached with kernel function (1) with assumption independent and identically distributed with f as density function. Kernel density estimation depends on bandwidth h and kernel function K . Bandwidth h represents smoothing parameter that be used to determine smoothness of an estimated curve. Kernel density estimator for density function f can be defined as follows:

$$\begin{aligned} \hat{f}_h(t) &= \frac{1}{n} \sum_{i=1}^n K_h(t - T_i) \\ &= \frac{1}{nh} \sum_{i=1}^n K_h\left(\frac{t - T_i}{h}\right) \end{aligned} \quad (2)$$

Generally, (2) used to estimate a curve regression.

A popular approach in nonparametric regression is Nadaraya–Watson estimator. If there are n independent observation with $\{(Y_i)\}_{i=1}^n$ is a time series with observation time $\{(T_i)\}_{i=1}^n$, generally a nonparametric regression equation given as follows:

$$y_i = m_h(T_i) + \varepsilon_i, \varepsilon_i \sim IIDN(0, \sigma_i^2) \quad (3)$$

$m_h(T_i)$ is a kernel regression function, and ε_i is an error for each $i = 1, 2, \dots, n$ that independent and identically distributed with mean 0, and variance σ_i^2 . Estimator (3) can be gotten as follows:

$$\hat{y}_i = \hat{m}_h(T_i) \tag{4}$$

Function in (4) approached with Nadaraya–Watson estimator as follows:

$$\hat{m}_h(t) = \frac{\sum_{i=1}^n K_h(t - T_i) Y_i}{\sum_{j=1}^n K_h(t - T_j)} \tag{5}$$

2.2 Fourier Series Estimator

Fourier series is a periodic function with trigonometric basis that has high flexibility. Visually, Fourier series plot can represent a seasonal pattern. One of Fourier series part that often be used as estimator in nonparametric regression, developed from Cosine Fourier series that proposed by [2] as follows:

$$f(t_i) = \gamma t_i + \frac{\alpha_0}{2} + \sum_{k=1}^K \alpha_k \cos kt_i \tag{6}$$

Equation (6) is substituted to $f(t_i)$ in nonparametric regression equation $y_i = f(t_i) + \varepsilon_i$, here $\varepsilon_i \sim IIDN(0, \sigma_i^2)$. The result is nonparametric regression model with Fourier series approach as follows:

$$y_i = \gamma t_i + \frac{\alpha_0}{2} + \sum_{k=1}^K \alpha_k \cos kt_i + \varepsilon_i \tag{7}$$

with $\{(Y_i)\}_{i=1}^n$ is a time series, every Y_i are observed at t_i , $\gamma, \alpha_0, \dots, \alpha_k$ are parameter that the values are estimated. Oscillation parameter which represents the number of oscillation is symbolled with $k = 1, 2, \dots, K$. Estimator for (7) given as follows:

$$\hat{y}_i = \hat{\gamma} t_i + \frac{\hat{\alpha}_0}{2} + \sum_{k=1}^K \hat{\alpha}_k \cos kt_i \tag{8}$$

2.3 Comparison Measure

The comparison measure as a goodness indicator based on kernel and Fourier series nonparametric regression approach for trend-seasonal data is Mean Square Error (MSE). MSE also used to determine optimal bandwidth based on Generalized Cross Validation (GCV) criterion for kernel estimator in nonparametric regression. In Fourier series approach, MSE also used to determine oscillation parameter based on GCV. MSE needed in calculation GCV values because MSE is a numerator component in GCV formula. The formula of MSE and GCV for kernel estimator given in [3, 12], and for Fourier series estimator given in [2, 10], general formula either MSE or GCV can be found in [4, 5, 8]. Kernel and Fourier series estimator in nonparametric regression for forecasting trend-seasonal data compared based on smallest MSE value that be reached in optimal quantity of smoothing measure, like bandwidth for kernel estimator and oscillation parameter in Fourier series estimator.

3 Data and Procedure

The data taken from [13] about forecasting the number of aircraft landing at an airport each month observed for 12 years. Figure 1 presents the data pattern to be forecasted. Based on Fig. 1 it appears that there is trend-seasonal pattern on the

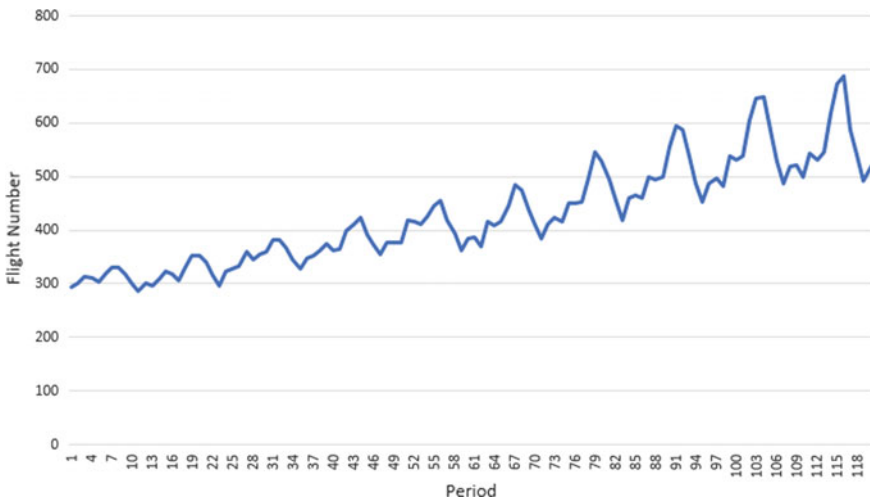


Fig. 1 Plot data with trend-seasonal pattern

flight number data. Furthermore, in sample data used to get estimator values, and out sample data used to forecast using both classical Nadaraya–Watson kernel and Fourier series estimator in nonparametric regression. The last step is comparing MSE value that be reached in optimal bandwidth for kernel estimator and oscillation parameter for Fourier series estimator.

4 Result and Discussion

First, trend-seasonal data forecasted using Nadaraya–Watson kernel estimator. Nadaraya–Watson kernel estimator in (5) approached with gaussian kernel, so that (5) becomes

$$\hat{m}_h(t) = \frac{\sum_{i=1}^n \exp\left(-\frac{1}{2} \left(\frac{t-T_i}{h}\right)^2\right) Y_i}{\sum_{j=1}^n \exp\left(-\frac{1}{2} \left(\frac{t-T_j}{h}\right)^2\right)} \tag{9}$$

where (9) used to forecast the number of flight for t period. The result for some bandwidth given in Table 1 as follows.

Second, trend-seasonal data forecasted using Fourier series estimator. The result for some oscillation parameter given in Table 2 as follows.

Nadaraya–Watson kernel estimator for optimal bandwidth 0,38 and Fourier series estimator for one oscillation parameter is used to forecast. The reason using one oscillation parameter for Fourier series, because MSE value for all oscillation parameter always smaller than kernel estimator, and other reason is holding parsimony model. The forecasting result, besides MSE value always smaller, forecasting trend-seasonal data using Fourier series more approximate to real data, compared with Nadaraya–Watson kernel estimator.

Table 1 The result of bandwidth determination

Bandwidth	0,10	0,20	0,30	0,32	0,35	0,38	0,40	0,50	1,00
MSE value	0,987	0,987	0,937	0,924	0,920	0,905	0,913	0,933	0,957

Table 2 The result of oscillation parameter determination

K	1	2	3	4	5	6	7	8	9
MSE value	0,395	0,386	0,386	0,374	0,352	0,355	0,369	0,374	0,375

5 Conclusions

Based on result and discussion, the Fourier series estimator is better than classical Nadaraya–Watson kernel to forecast trend-seasonal data. In this case, Fourier series estimator gives smaller MSE for all oscillation parameters, than all of bandwidths in Nadaraya–Watson kernel estimator. This conclusion supported by some studies about Fourier series properties that accommodate seasonal pattern. Here, trend pattern accommodated by linear function that be combined as additive equation with Fourier series. It means that Fourier series function that be used accommodate trend-seasonal data.

References

1. Bloomfield, P.: *An Introduction Fourier Analysis for Time Series*. Wiley, New York (2000)
2. Bilodeau, M.: Fourier Smoother and Additive Models. *Can. J. Stat.* **3**, 257–259 (1992). <https://doi.org/10.2307/3315313>
3. Hardle, W.: *Smoothing Techniques with Implementation in S*. Springer, New York (1990)
4. Eubank, R.L.: *Spline Smoothing and Nonparametric Regression*, 2nd edn. Marcel Dekker, New York (1999)
5. Wahba, G.: *Spline Model for Observational Data*. SIAM XII, Philadelphia (1990)
6. Green, P.J., Silverman, B.W.: *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London (1994)
7. Antoniadis, A., Bigot, J., Spatinas, T.: Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Stat. Softw.* **6**, 1–83 (2001). <https://doi.org/10.18637/jss.v006.i06>
8. Takezawa, K.: *Introduction to Nonparametric Regression*. Wiley, New Jersey (2006)
9. Li, F., Ionescu, C., Sminchisescu, C.: Random fourier approximations for skewed multiplicative histogram-kernels. *Ann. Stat.* **38**(6), 3321–3351 (2013). <https://doi.org/10.1007/978-3-642-15986-227>
10. Pane, R., Budiantara, I.N., Zain, I., Otok, B.W.: Parametric and nonparametric estimators in fourier series semiparametric regression and their characteristics. *Appl. Math. Sci.* **8**(102), 5053–5064 (2013). <https://doi.org/10.12988/ams.2014.46472>
11. Pujiastuti, C.E.: *Fourier Series and Kernel in Nonparametric Regression*. Gadjah Mada University, Yogyakarta (1996)
12. Harvey, A., Oryshchenko, V.: Kernel density estimation for time series data. *Int. J. Forecast.* **28**(1), 3–14 (2012). <https://doi.org/10.1016/j.ijforecast.2011.02.016>. Elsevier
13. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: *Time Series Analysis, Forecasting and Control*, 3rd edn. Wiley, New York (1976)

Chapter 43

Hydrological Trend Analysis in Johor



Norazlina Ismail, Werda Yelling and Nurfarhana Hassan

Abstract Rainfall, temperature and streamflow are climatic elements which play major roles in influencing the hydrological activities and human's life. This paper aims to investigate the trend of temperature and two hydrological data involving rainfall and streamflow in Johor. The rainfall data used in this paper was obtained from 10 stations for the period of 30 years (1986–2015), while the temperature and streamflow data used were obtained from 10 and nine stations respectively for the period of 10 years (2006–2015) for temperature and 30 years (1986–2015) for streamflow. In this paper, the Mann-Kendall test and Theil-Sen estimator were used to analyse the trend of the data. The results from Mann-Kendall test indicate that there are significant decreasing trends for annual rainfall at three stations which is Empangan Sg. Sembrong, Pintu Kawalan Parit Jawa and Felda Bukit Batu stations and a significant increasing trend at Pusat Kemajuan Per. Pekan Nanas station. While for temperature, the results show significant increasing trend at Batu Pahat, Mersing, Senai and Mardi Alor Bukit Pontian stations and a significant decreasing trend at Majlis Daerah Labis station. For the streamflow, the results show significant increasing trend at Sg. Endau station and significant decreasing trend at Sg. Kahang, Sg. Bekok, Sg. Lenggong, Sg. Johor and Sg. Segamat stations. The results for Theil-Sen estimator correspond with the results from the Mann-Kendall test. This trend analysis is important in predicting future pattern and preparing proper plans to face any upcoming disasters.

Keywords Trend analysis · Mann-Kendall · Theil-Sen estimator

N. Ismail (✉) · W. Yelling · N. Hassan
Johor Bahru, Malaysia
e-mail: i-norazlina@utm.my

W. Yelling
e-mail: Werda_Zichi94@yahoo.com

N. Hassan
e-mail: nurfarhanahassan@yahoo.com

1 Introduction

Climate is an important factor towards the changes on the environment that may cause several disasters. According to Malaysian Meteorological Department, Malaysia has a tropical climate with uniform temperature, high humidity and experiencing rainfall throughout the year. Malaysia experiences two monsoon seasons namely Northeast monsoon and Southwest monsoon seasons which occasionally lead to heavy rainfall and natural disasters such as flood, drought and haze. The heavy rainfall that causes flood may also lead to the rising of the water level and streamflow. These disasters have affected people's life and destabilize the economy. According to Shaari et al. [1], the 2006–2007 floods in Johor are considered as the most costly flood incident in the Malaysian history. Moreover, Shafie [2] stated that the devastating flood events were mainly caused by heavy rainfall from the monsoon and had affected thousands of life. Therefore, there are needs to investigate the behavior of the main factor that led to such disasters in order to predict the future pattern and for preventive measure.

Trend analysis is an approach to investigate the trend and changes in pattern of collected data over time. This statistical analysis is important for prediction of the future pattern so that a proper planning can be done to minimize the impact of the disasters. Few researchers have applied several statistical tests for trend analysis involving hydrological and meteorological data obtained at certain area in Malaysia. Jamaludin et al. [3] investigated rainfall trend at peninsular region in Malaysia from 1975–2004 using Mann-Kendall test. The result showed significant increasing trends during Northeast monsoon season and insignificant decreasing trends during Southwest monsoon season. Dindang et al. [4] investigated the trend of rainfall in Kuching, Sarawak for the period of 1968–2010 using Mann-Kendall test. The results showed an insignificant increasing trend for the annual rainfall data. Jamaluddin [5] investigated the trend of rainfall and streamflow in Bernam River and Slim River for the period of 1973–2012. The results from Mann-Kendall test indicated that there are increasing trends for both rainfall and streamflow. Amirabadizadeh [6] investigated the trend of rainfall and temperature at Langat River Basin using Mann-Kendall and Theil-Sen slope test. The results of the Mann-Kendall analysis showed that there is significant increasing trend at three to four stations and they are gradually increased based on Theil-Sen estimator. Sulaiman et al. [7] investigated the hydrological trend of streamflow at Pahang River using Mann-Kendall test. The results showed significant increasing trends at two stations. Jamaludin [8] analyzed the trend of temperature at 13 stations around peninsular area in Malaysia for the period of 32 years (1980–2011) using Mann-Kendall and Theil-Sen slope estimation test. The results showed significant increasing trend for annual mean temperature at almost all selected stations. These preceding studies on the trend analysis for hydrological and meteorological data were done and focused at only few selected stations in peninsular region in Malaysia, and selected districts and rivers in Sarawak, Selangor and Pahang.

In this study, we analyzed the trend of rainfall, temperature and streamflow specifically at the state of Johor. In recent decades, Johor has been developing enormously. The rapid changes and development are one of the factors towards climate change which lead to the growing of temperatures globally [6] and extreme rainfall events that causes flood. Thus, following the most devastating floods in 2006–2007, several flood and drought incidents in the past few years in Johor, the trend analysis for rainfall, temperature and streamflow in Johor is crucial in predicting the future pattern and preparing a proper plan to face any upcoming disaster. This present study intends to investigate and provide trend analysis for rainfall, temperature and streamflow in Johor by using Mann-Kendall test and Theil-Sen slope estimator.

2 Methodology

2.1 Study Area

Johor is one of the states in Malaysia that occasionally experiences natural disasters such as flood, drought and haze. Johor is located in the southern region of Malaysia. The weather in Johor has uniform temperature, high humidity and abundant rainfall throughout the year. The average monthly temperature recorded in Johor is around 25–28 °C and the average annual rainfall is around 2000 mm. Johor consists of 10 districts. Johor Bahru is the capital district and other districts include Kota Tinggi, Kluang, Tangkak, Muar, Mersing, Kulai, Pontian, Batu Pahat, and Segamat.

2.2 Data Set

The data for rainfall, streamflow and temperature in this study were obtained from Department of Irrigation and Drainage Malaysia (DID) and Malaysian Meteorological Department for the period of 30 years (1986–2015) for rainfall and streamflow and 10 years (2006–2015) for temperature. In this study, 10 stations were chosen based on the districts in Johor in which each station represents each district in Johor except for Tangkak district for temperature data, and Johor Bahru, Pontian and also Tangkak districts are not available for streamflow data.

2.3 Trend Analysis

In this study, Minitab software is used for non-parametric Mann-Kendall test and Theil-Sen slope estimation test to detect the trend in the hydrological and

meteorological data set. According to Yue et al. [9] non-parametric statistical test is commonly used in investigating the trend of rainfall and temperature compared with parametric tests because of its ability in dealing with non-normally distributed data. The studies on trend usually involve non-parametric Mann-Kendall test with Sen's slope estimation test [10]. In this study, Mann-Kendall test is used to detect the existence of the monotonic trend. The null hypothesis, H_0 for Mann-Kendall test states that there is no trend and the data is independent and identically distributed, while alternative hypothesis, H_1 states that there exist increasing or decreasing trend. Probability value (p-value) obtained from the test is used to test the significant trend. If the p-value is greater than α , the null hypothesis H_0 is failed to reject and it shows no existence of trend. In this study, the value of α is 0.05. The Theil-Sen estimator is then used to determine the magnitude of the trend detected by Mann-Kendall test. The Theil-Sen estimator is robust against non-normally distributed and missing data [11]. The magnitude slope of the trend can be estimated as follows [8]:

$$\beta_i = \text{median}\left(\frac{x_j - x_k}{j - k}\right) \text{ for } i = 1, 2, \dots, n \quad (1)$$

for all $j > k$, where x_j and x_k are the sequential data values at times j and k respectively.

3 Results and Discussion

The trend for the hydrological data is analysed by Mann-Kendall test and the trend magnitude is estimated by using the Theil-Sen estimation test. Both tests are carried out by using Minitab software. The results are presented in Table 1.

For Mann-Kendall test, positive value of Z indicates upward or increasing trend, while negative value indicates downward or decreasing trend. The significance of trend are based on the value of $|Z| > 1.6449$ and p-value less than $\alpha = 0.05$. Thus the null hypothesis, H_0 will be rejected and indicates that the trend is significant. Based on Table 1, the results for rainfall trend show significant downward trends at three stations which is Empangan Sg. Sembrong, Pintu Kawalan Parit Jawa and Felda Bukit Batu stations and a significant upward trend at Pusat Kemajuan Per. Pekan Nanas station. There is no evidence of the existence of the significant trends at the remaining six stations. For trend analysis on temperature, significant upward trend is detected at Batu Pahat, Mersing, Senai and Mardi Alor Bukit Pontian stations while significant downward trend is detected at Majlis Daerah Labis station. For trend analysis on streamflow, only one station shows significant upward trend which is Sg. Endau station. Significant downward trend is detected at five stations which is Sg. Bekok, Sg. Kahang, Sg. Lenggong, Sg. Johor and Sg. Segamat stations and there is no evidence of the existence of the significant trends at the remaining three stations.

Table 1 Results of Mann-Kendall trend test and Theil-Sen slope estimation test

Station		Mann-Kendall test				Sen's slope estimation (β)
		Z	p	Trend	Significant (p-value < α)	
<i>Rainfall</i>						
1931003	Empangan Sg. Sembrong	-2.2837	0.0112	Downward	Yes	-23.1875
2636170	Stor JPS Endau	-1.4987	0.0670	Downward	No	-19.7714
1437116	Stor JPS Johor Bahru	1.5700	0.0582	Upward	No	13.6600
1926001	Pintu Kawalan Parit Jawa	-4.1806	0.0000	Downward	Yes	-97.5238
1737001	Kota Tinggi	0.0357	0.4858	Upward	No	1.3909
2330009	Ldg. Sg. Labis	0.1249	0.4503	Upward	No	1.0364
1534002	Pusat Kemajuan Per. Pekan Nanas	1.7484	0.0402	Upward	Yes	12.5938
48672	Kluang	0.1071	0.4574	Upward	No	0.3071
47206	Hospital Tangkak	0.1071	0.4574	Upward	No	0.5458
47142	Felda Bukit Batu	-3.6396	0.0001	Downward	Yes	-47.4333
<i>Temperature</i>						
48670	Batu Pahat	2.6833	0.0036	Upward	Yes	0.07029
48672	Kluang	1.0733	0.1416	Upward	No	0.05094
48674	Mersing	1.7889	0.0368	Upward	Yes	0.09352
48679	Senai	2.5044	0.0061	Upward	Yes	0.11360
47117	Hospital Johor Bahru	0.1789	0.4290	Upward	No	0.01058
47142	Felda Bukit Batu	-0.3578	0.3603	Downward	No	-0.03773
47215	Felda Lenga	-0.7155	0.2371	Downward	No	-0.18186
47120	Hospital Kota Tinggi	0.0000	0.5000	Linear	No	0.01213
47214	Majlis Daerah Labis	-2.6833	0.0036	Downward	Yes	-0.04125
47116	Mardi Alor Bukit Pontian	1.9677	0.0245	Upward	Yes	0.08330
<i>Streamflow</i>						
2130422	Sg. Bekok	-1.6657	0.0479	Downward	Yes	-0.1666
2235401	Sg. Kahang	-1.9353	0.0265	Downward	Yes	-0.5938
2533474	Sg. Endau	2.9914	0.0014	Upward	Yes	1.0212
2237471	Sg. Lenggong	-3.9092	0.0000	Downward	Yes	-0.4053
1836402	Sg. Sayong	-1.5297	0.0630	Downward	No	-0.1924
2527411	Sg. Muar	-0.6119	0.2703	Downward	No	-0.2007
1737451	Sg. Johor	-2.3455	0.0095	Downward	Yes	-0.5038
1836403	Sg. Penggeli	-0.4079	0.3417	Downward	No	-0.0535
2528414	Sg. Segamat	-2.3295	0.0099	Downward	Yes	-0.4188

Theil-Sen estimator is then applied to determine the slope of the trend. Positive slope indicates upward trend while negative slope indicates downward trend. Based on the results from Mann-Kendall test, the slope is estimated based on the stations with significant trends only. For rainfall trend, the estimated slope revealed that Empangan Sg. Sembrong, Pintu Kawalan Parit Jawa and Felda Bukit Batu stations have a negative value of β , which indicates downward trend while Pusat Kemajuan Per. Pekan Nanas station has a positive value of β , which indicates upward trend. For temperature trend, the estimated slope revealed that Batu Pahat, Mersing, Senai and Mardi Alor Bukit Pontian stations have a positive value of β , which indicates upward trend while Majlis Daerah Labis station has a negative value of β , which indicates downward trend. For streamflow trend, the estimated negative value of slope at stations with significant trend involving Sg. Bekok, Sg. Kahang, Sg. Lenggor, Sg. Johor and Sg. Segamat stations indicates a downward trend. Based on these results, the rate of streamflow is expected to decrease at these stations in the next 30 years while only one station which is Sg Endau station has an increasing rate of trend. These results correspond with the results of Mann-Kendall test whereby positive slope indicates upward trend while negative slope indicates downward trend.

4 Conclusion

This study analyzes the trend of rainfall, temperature and streamflow in Johor. The rainfall and streamflow data used are from the period of 30 years (1986–2015), while the temperature data is from the period of 10 years (2006–2015). Mann-Kendall test and Theil-Sen estimator were used to analyse the trend of the data. Based on the result from Mann-Kendall test, significant decreasing trends were detected at three stations which is Empangan Sg. Sembrong, Pintu Kawalan Parit Jawa and Felda Bukit Batu stations and a significant increasing trend was detected at Pusat Kemajuan Per. Pekan Nanas station. For trend analysis of temperature, four stations which is Batu Pahat, Mersing, Senai and Mardi Alor Bukit Pontian stations each showed significant increasing trend while only one station at Majlis Daerah Labis showed a significant decreasing trend. For streamflow, the significant increasing trend was detected at Sg Endau station while Sg. Bekok, Sg. Kahang, Sg. Endau, Sg. Lenggor, Sg. Johor and Sg. Segamat stations shows a significant decreasing trend. The results of Mann-Kendall test were reinforced by the results obtained using Theil-Sen estimation test. Positive slope of Theil-Sen indicates upward trend while negative slope indicates downward trend. Thus the result of Theil-Sen estimator shows corresponding result with Mann-Kendall test whereby the negative slope was detected at station that has a significant decreasing trend from Mann-Kendall test while positive slope was detected at station that has a significant increasing trend. Therefore, it can be concluded that the trend analysis

results using Mann-Kendall Test are correspond with the estimated value of slope from Theil-Sen estimator. This trend analysis is important in predicting future pattern and preparing proper plans to face any upcoming disasters.

Acknowledgements The authors would like to acknowledge Ministry of Higher Education, Malaysia and Research Management Centre, UTM for the financial support through research grant vote number Q.J130000.2626.12J45 for this research.

References

1. Shaari, M.S.M., Karim, M.Z.A., Hasan-Basri, B.: Flood disaster and GDP growth in Malaysia. *Eur. J. Bus. Soc. Sci.* **4**(10), 27–40 (2016). <https://doi.org/10.21859/eulawrev-08023>
2. Shafie, A.: Extreme flood event: a case study on floods of 2006 and 2007 in Johor, Malaysia. M.Sc. thesis, Colorado State University (2009)
3. Jamaludin, S., Dheni, S.M., Zin, W.Z.W., Jemain, A.A.: Trends in Peninsular Malaysia rainfall data during the southwest monsoon and northeast monsoons seasons: 1975–2004. *Sains Malays.* **39**(4), 533–542 (2010)
4. Dindang, A., Taat, A., Phuah, E.B., Alwi, A.B.M., Mandai, A.A., Adam, S.F., Othman, F.S., Bima, D.A., Lah, D.: Statistical and trend analysis of rainfall data in Kuching, Sarawak from 1968–2010. Malaysian Meteorological Department (MOSTI) (2013)
5. Jamaluddin, I.B.: Analysis of trend in hydrologic system for Sungai Bernam basin. Project Report, Department of Civil Engineering Faculty of Engineering University Putra Malaysia Serdang, Selangor (2014)
6. Amirabadizadeh, M., Huang, Y.F., Lee, T.S.: Recent trends in temperature and precipitation in the Langat River Basin, Malaysia. *Adv. Meteorol.* (2015). <https://doi.org/10.1155/2015/579437>
7. Sulaiman, N.H., Kamarudin, M.K.A., Mustafa, A.D., Amran, M.A., Azaman, F., Abidin, I.Z., Hairoma, N.: Trend analysis of Pahang river using non-parametric analysis: Mann-Kendall's trend test. *Malays. J. Anal. Sci.* **19**(6), 1327–1334 (2015)
8. Jamaludin, S.: Temporal changes and variability in temperature series over Peninsular Malaysia. *AIP Conf. Proc. AIP* **1643**(1), 248–255 (2015). <https://doi.org/10.1063/1.4907452>
9. Yue, S., Pilon, P., Phinney, B., Cavadias, G.: The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrol. Process.* **16**, 1807–1829 (2002). <https://doi.org/10.1002/hyp.1095>
10. Machiwal, D., Jha, M.K.: *Hydrologic time Series Analysis: Theory and Practice*, p. 218. Springer, New Delhi (2012)
11. Zhang, S., Lu, X.X.: Hydrological responses to precipitation variation and diverse human activities in a mountainous tributary of the lower Xijiang. *China Catena* **77**(2), 130–142 (2009). <https://doi.org/10.1016/j.catena.2008.09.001>

Chapter 44

Implementing Correlation Dimension: K-Means Clustering via Correlation Dimension



Zakiah Ibrahim Kalantan

Abstract The estimation of intrinsic dimension is an essential step in the dimension reduction process. The intrinsic dimension can be estimated by fractal dimension estimation methods, which exploit the intrinsic geometry of a data set. The most popular concept from this family of methods is the correlation dimension. K-Means is most popular clustering algorithm for performing unsupervised learning in data mining. This paper we propose approaches to approximate the correlation integral. In addition, we propose a new selection for clusters via correlation dimension. The performance of the algorithm is discussed. Experimental results on an artificial and real-world data are used to demonstrate the algorithms and compare to other methodology.

Keywords Intrinsic dimension · Fractal dimension · Correlation dimension · Clustering · K-Means clustering

1 Introduction

Dimension reduction and clustering methods play important roles in data mining and pattern recognition. It is essential to capture the data information with minimum number of variables to overcome the curse of dimensionality. Most of dimension reduction methods require fixing the intrinsic dimension of the low-dimensional subspace in advance. The intrinsic dimension (ID) is defined as the minimum number of variables necessary (suffice) to describe the data without much loss of information [1, 2]. ID estimation methods can be classified as local methods such as K nearest neighbor and charting manifold which estimate the ID using the sub-regions of dataset, and global methods such as Fractal Dimension where the ID is estimated using the whole data set [3–5]. Fractal is a geometrical structure of an irregular object (a data set) where the object could be divided into substructures of

Z. I. Kalantan (✉)

Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: zkalanten@kau.edu.sa

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_44

359

same form, such as Sierpinski triangle and Koch snowflake. The fractal concept provides an ID estimation of a data set by non-integer value. Correlation dimension and box-counting dimension are commonly used as non-linear methods which are used to estimate the fractal dimension [5]. On other hand, clustering methods which determine the optimal grouping of observations within each cluster are similar. The similarity between pairs is based on some measured of distance. The identifying of clustering is made without detecting the dimension of clusters. This would be useful in order to estimate the intrinsic dimensionality of the data. Clustering methods could be classified to a hierarchical, such as single linkage and non-hierarchical clustering as K-Means [6].

Our approach aims to estimate the ID locally via correlation dimension. We try to identify subsets of points with low ID which are qualitatively different. The paper is organized as follows: Sect. 2 presents the cluster. In Sect. 3 correlation dimension is discussed. Section 4 illustrates there is our approach. In Sect. 5, there are the experimental results. Finally, conclusions are drawn in Sect. 6.

2 K-Means Clustering Method

K-Means clustering is an unsupervised learning method in which data points are partitioned into K clusters (Voronoi cells). Every data point is assigned to the cluster with nearest mean. As illustration, suppose $X = (x_1, x_2, \dots, x_n) \in R^D$ K-Means clustering tries to partition the n data points into $k (\leq n)$ clusters so as to minimize the within-cluster sum of squares (WCSS) [6].

$$\sum_{j=1}^k \sum_{i=1}^n \|x_i^j - c_j\|^2 \quad (1)$$

where c_i is the mean of points in k_i . This is equivalent to minimizing the pairwise squared deviations of points in the same cluster. Usually, the user needs to determine the number of clusters. One practical approach is to use Elbow method which examine the within-cluster dissimilarity as a function of the number of clusters.

3 Correlation Dimension

The correlation dimension estimates the dimension via a pairwise distances algorithm. Grassberger et al. [1, 3] proposed GP method that computes the correlation integral as

$$C(r) = \lim_{N \rightarrow \infty} \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N I(x_i - x_j \leq r) \tag{2}$$

where $I(\triangleright)_r$ is an indicator function, and $\|x_i - x_j\|$ is the Euclidean distance between data points. Hence, the correlation dimension is defined as

$$d_{cor} = \lim_{r \rightarrow 0} \frac{\log(C(r))}{\log(r)} \tag{3}$$

The correlation integral needs to be estimated for a ball tending to 0 with respect that $C(r) \propto r^d$. Additionally, the user needs to decide a suitable range of values of r to arrive the ID estimate [2]. Various techniques have been proposed to compute an estimate of the correlation dimension [4, 5] by using the Fisher’s maximum likelihood rule. Note that the correlation dimension is affected by data dimensionality, the variables correlation, the dimensionality of data set and the number of distance pairs. Furthermore, the number of data pairs is $\binom{N}{2} = \frac{1}{2}N(N-1)$, and it’s clear that $0 \leq C(r) \leq 1$. Now, given that $D(r) = \frac{\log(C(r))}{\log(r)}$ then $d_{cor} = \lim_{r \rightarrow 0} D(r)$.

4 Exploring Correlation Dimension via K-Means Method

The objective of this paper is to estimate local correlation dimension. We try to capture much data points using clusters that capture the useful information of the local growth curve. Firstly, we define a new approach to estimate the correlation dimension. Then, the ID is obtained locally using K-Means Clustering Method. We propose approach which aims to capture the distance pairs of $C(r)$ in a more effective way and consistent with the GP method. The correlation dimension is interpolated through the inferring of linear graph $(r, D(r))$. Practically, the correlation dimension is plotted versus r with radius close to 0. The range of radius contains a sufficient number of data points. Then, ID is determined as the intercept of the fitted linear equation. Specifically, consider a linear regression with least squares estimator a (intercept) and c (slope). Then the correlation dimension can be approximated as $D(r) = a + c r$, which at $r = 0$ gives $d_{cor} = D(0) = a$. We show this approach requires fewer data points and improves the calculation of the correlation dimension for various type of data [7, 8]. We determine the optimum number of clusters by applying Elbow method on reasonable sequence of K

between 2 and 10. Hence, the data points are grouped using k-mean clustering method.

For categorical data, the data is partitioned due to its type of category. After this step, the observations are sorted due to their cluster and split again into sub-regions. Finally, we estimate the change in the topology structure of outcome space using our approach of correlation dimension. Finally, the average of all ID estimates is computed.

5 Experimental Results

We deal with different types of datasets; artificial and real data. The data variables are scaled before the implementation. The computation results for each data example are presented in the following sections.

5.1 *Spiral Data*

The spiral data is artificial data with known ID = 1, it consists of two variables with 300 data points. The structure of the (scaled) data set is shown in Fig. 1a. Figure 1b illustrates the result of Elbow method. One finds that the suitable number of cluster is 8. The data set is partitioned to 8 clusters by K-Means Method, as displayed in Fig. 1c. Next, the data points are sorted and grouped to 8 different sub-regions. Then, the correlation dimension is estimated locally on each sub-region, and the estimate of IDs for each sub-region are 1.1212098, 0.6202766, 0.9886193, 1.0657393, 1.0372717, 0.7679395, 0.7633099, 1.6874269, respectively. Figure 1d displays the boxplot of local IDs. The average of IDs estimates is 1.006474 which is closest to the known ID. Note that the estimate of (global) ID over all data equals 1.232044. It is remarkable that the estimate of local correlation dimension provides a close value to known ID compared to the same method on all data.

5.2 *Koch Curve Data*

Koch snowflake is a fractal curve created by Nielsvon Koch 1904. Koch curve is a straight line that is divided up into three equal parts and creates the triangle. The base of triangle is removed and builds another triangle. Figure 3a shows the result of iterated process. The intrinsic dimension of the curve is $d = \frac{\log 4^n}{\log 3^n} \approx 1.2619$, and then the curve is expected to be more than a line and less than a plane. Practically, we sample 10000 data points. Hence, with 8 clusters the K-Means method is applied and the ID estimate for each cluster equals 1.232825, 1.255903, 1.156930,

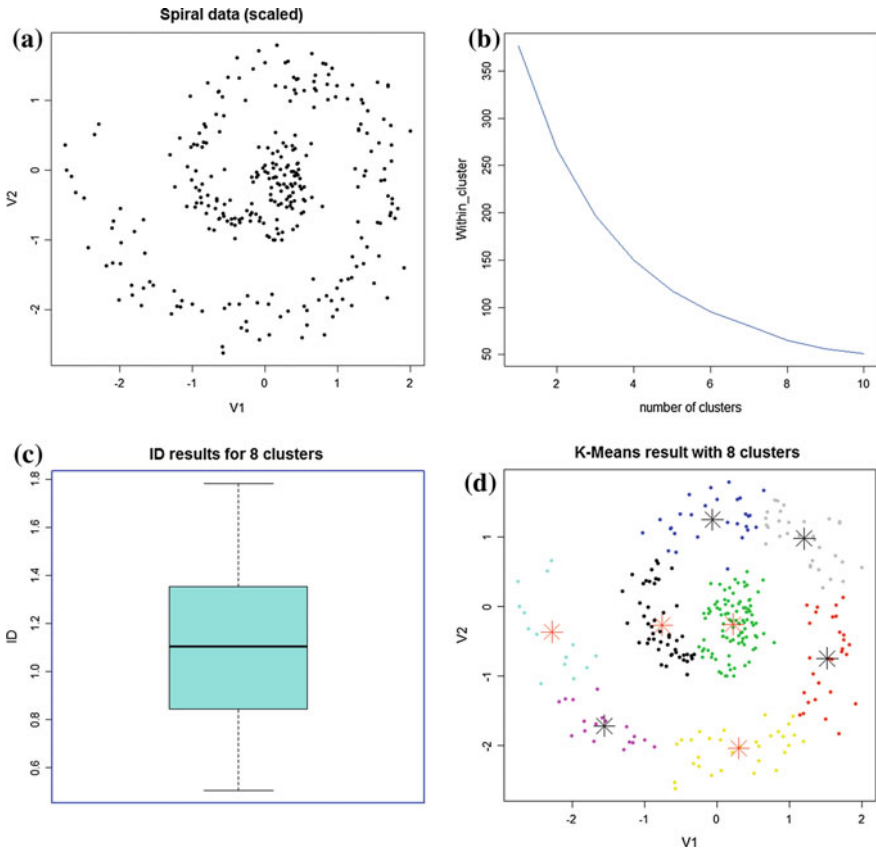


Fig. 1 Spiral data: **a** Scatter plot, **b** Scree plot of within cluster, **c** Boxplot of local correlation dimension with range of r from 0.4 to 0.7, **d** 8 clusters of data set

1.186678, 1.187084, 1.221890, and 1.231208, as displayed in Fig. 2c. The average of IDs is 1.21036 which is reasonable and close to the known ID.

5.3 Forensic Glass Fragments Data

Forensic Glass Fragments Data consists of 10 variables with 214 observations available in R package. For simplicity, we consider only five variables; Manganese (Mg), Aluminum (Al), Silicon (Si), Potassium(K) and types of fragments (WinF, WinNF, Veh, Con, Tabl, Head). Note that the variable type is used only to classify the data. Figure 3a shows the pairwise plot for the 4 measurements of the scaled dataset. The principal component analysis is used for this data to give an estimate of ID which is a linear ID. Figure 3b displays the screeplot of a principal component

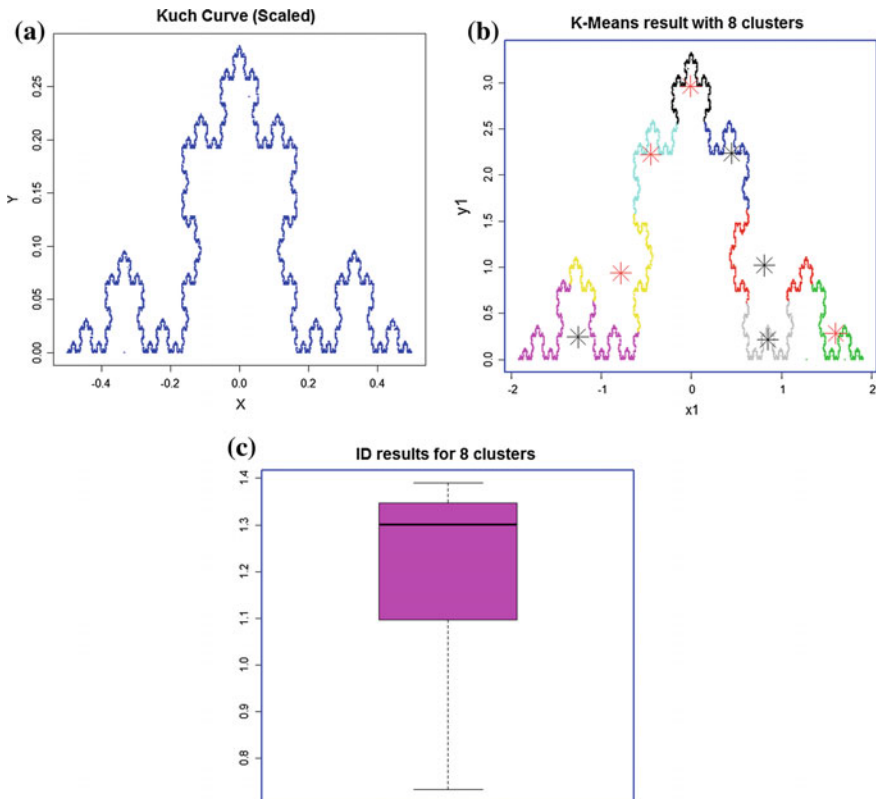


Fig. 2 Koch curve data: **a** The scatter plot, **b** The implementation of K-Means with 8 clusters, **c** Boxplot of local correlation dimension with range of r from 0.25 to 0.5

analysis on the (scaled) data set. One finds that the three components explain 89% while the two components explain 70% of the total variance. Obviously, when implementing the linear dimension reduction via PCA, users decide the amount of variance to be preserved. So depending on where one places the cut point, one would decide that the (linear) ID for this data is 2 or 3. This result is intuitive when considering the data, which do not possess a very pronounced inner structure. While the estimate of (global) correlation dimension, over all data, equals 2.886108.

Firstly, apply K-Means method with 6 clusters according to six types of the fragments involved, the result is displayed in Fig. 4a. Moreover, Table 1 illustrates the algorithm classification with respect to the types of data. Note that the cluster 5 and 6 contain 2 and 8 data points, respectively. Next, the data is ordered and grouped into 4 sub-regions where cluster 2, 5 and 6 are combined.

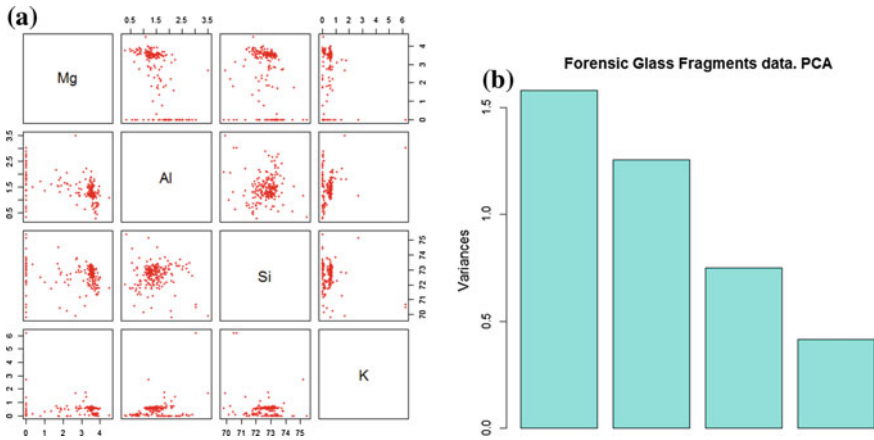


Fig. 3 Forensic Glass Fragments: **a** Pairwise plots, **b** Scree plot of four measurements

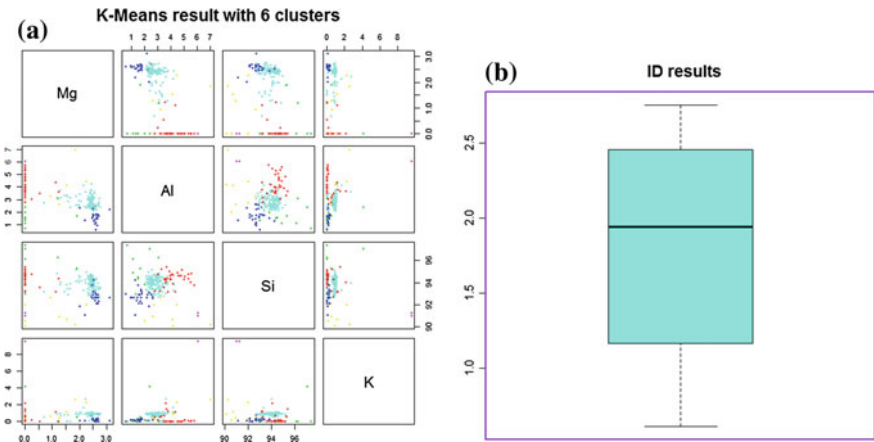


Fig. 4 Forensic Glass Fragments: **a** using K-Means method to find 4 clusters in the data set, **b** Boxplot of local correlation dimension with range of r from 0.25 to 0.5

Table 1 The distribution of data points in the clusters

Type of frag	1	2	3	4	5	6
WinF	0	0	20	5	0	0
WinNF	0	6	3	62	0	4
Veh	0	0	5	12	0	0
Con	7	0	0	3	2	1
Tabl	2	3	0	4	0	0
Head	23	1	0	2	0	3

Finally, the intercept method is implemented for each sub-region to estimate the intrinsic dimension. The ID estimate for each cluster equals 2.1592158, 0.6132457, 1.7230539 and 2.7525287. The average of the IDs obtained is 1.812011.

6 Conclusion and Discussion

The aim of this report was to provide a new approach to estimate the dimensionality locally via a global method (correlation dimension) from previous studies which indicate the ID local method always provides lower bound of ID estimates. While regardless to global method, the estimate of ID is greater than the local estimate. Our approach aims to overcome that issue and capture much data points using K-Means Clustering Method. The ID is estimated via fractal dimension since it provides a suitable indicator of spreading data. Additionally, it requires a few data points with quick computation that is easy to implement in different tasks of data mining. To sum up, the method has ability to implement data sets where we do not have enough information about the global structure and even with the categorical data.

References

1. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(10), 1404–1407 (2002)
2. Fukunaga, K., Olsen, D.R.: An Algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.* **20**(2), 176–183 (1971)
3. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. *Phys. D: Nonlinear Phenom.* **9**, 189–208 (1983)
4. Theiler, J.: Estimating fractal dimension. *J. Optical Soc. Am.* **7**, 1055–1073 (1990)
5. Mo, D., Huang, H.: Fractal-based intrinsic dimension estimation and its application in dimensionality reduction. *IEEE Trans. Knowl. Data Eng.* **24**(1), 59–71 (2012)
6. Rencher, A., Christensen, W.: *Methods of Multivariate Analysis*. 3rd edn. Willey (2012)
7. Kalantan, Z., Einbeck, J.: On the computation of the correlation integral for fractal dimension estimation. In: *International Conference on Statistics in Science. Business and Engineering (ICSSBE 2012)*, IEEE Conference Publications, pp. 80–85(2012). <https://doi.org/10.1109/icssbe.2012.6396531>
8. Einbeck, J., Kalantan, Z.: Intrinsic dimensionality estimation for high-dimensional data sets: new approaches for the computation of correlation dimension. *J. Emerg. Technol. Web Intell.* **5** (2), 91–97 (2013). <https://doi.org/10.4304/jetwi.5.2.91-97>

Chapter 45

Job Satisfaction Among Academic Staff: A Structural Equation Modeling Approach



Haslinda Ab Malek, Farah Farhana Mazli, Hafizah Sharif,
Noor Azira Mohammad and Isnewati Ab Malek

Abstract Job satisfaction is an affective reaction to an individual's work situation. It also can be defined as an overall feeling towards their job or career. For academic staff, satisfaction with their career may have strong implications for student learning. Specifically, an academician's satisfaction may influence the quality and stability of instruction given to students. Therefore, students also affected if the quality of education was poor. This research aims to study the factors that influence job satisfaction among academic staff in Universiti Teknologi Mara Negeri Sembilan, Kampus Seremban. A direct questionnaire was conducted in order to obtain a primary data in this research. A sample of 222 was obtained from different faculty such as Faculty of Computer and Mathematical Sciences, Faculty of Sport Science and Recreation and Faculty of Administrative Science and Policy Studies. This study involved three independent variables namely as working conditions, pay and promotion potential and relationships with others. Structural Equation Modeling approach was used in order to determine the factors that effect on job satisfaction. From the result, it found that, variable relationship with others (p -value = 0.044) gave a significant factor on job satisfaction. It showed that a good relationship with leader or supervisor affect on a job satisfaction among academic staff. Therefore, the management of the university may use the result in this study to set up proper implementation in order to increase level of satisfaction among academic staff.

H. Ab Malek (✉) · F. F. Mazli · H. Sharif · N. A. Mohammad · I. Ab Malek
Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara,
Cawangan Negeri Sembilan, Kampus Seremban, Seremban, Malaysia
e-mail: haslinda8311@ns.uitm.edu.my

F. F. Mazli
e-mail: farahfarhanamazli@gmail.com

H. Sharif
e-mail: hafizahsharif@gmail.com

N. A. Mohammad
e-mail: noorazira1992@gmail.com

I. Ab Malek
e-mail: isnewati@ns.uitm.edu.my

1 Introduction

Job satisfaction is an affective reaction to an individual's work situation. It also can be defined as an overall feeling towards their job or career. Furthermore, job satisfaction is said to be significant factor of productivity at work. For academic staff, satisfaction with their career may have strong implications for student learning. Specifically, an academician's satisfaction may influence the quality of instruction given to the students. Job satisfaction of academic staff in higher education was important because it influences their motivation and performance. Moreover, the job satisfaction also has an effect in delivering quality education services. Hence, students also affected if the quality of education was poor. Therefore, this study was concentrated on identify the factors that influence job satisfaction among academic staff.

According to Noordin [10], academic staff attitudes were affected by workplace conditions such as a positive and protected work environment, a loyal management, career succession, salary, work teams, peers, and the job itself. Other than that, academic staff's job satisfaction influences job performance, motivation, morale, attrition, and eventually students' performance [10]. Job satisfaction also can help the university to retain the potential academics and attract the new competent staff to the university.

In addition, working condition was very important factor in order to increase the job satisfaction. Most recently, Ghaffar [4] indicated that working and living conditions have greater impact on academic staff morale and motivation and their classroom performance. Moreover, working in an environment of cooperation and common respect was also significant factor to increase the academic staff's job satisfaction. The result from Mustapha [9] indicated that there was positive significant relationship between interpersonal relationships and job satisfaction.

Furthermore, salary was very primary factor of satisfaction for almost every type of employees. A study conducted by Oshagbemi [11] amongst UK academics, found a statistical important relationship between pay, rank of employees and their level of job satisfaction. In addition, the study of Grace and Khalsa [5], at Massachusetts higher education institution identified professional development and salary packages as the most important job satisfaction factors.

This study is important because it appraised the present situations of the academic staff in Universiti Teknologi Mara (UiTM) Negeri Seremban Kampus Seremban. The management in the university may use the result in this study to set up proper implementation in order to increase level of satisfaction among academic staff.

2 Methodology

2.1 Population and Sample Size

The population size of respondent is chosen from all academic staff in Universiti Teknologi Mara (UiTM) Negeri Sembilan Kampus Seremban. The faculties involved in this study was Faculty of Computer and Mathematical Sciences (FSKM), Faculty of Sport Science and Recreation (FSR), Faculty of Administrative Science and Policy Studies (FSPPP) and others faculty of academic staff. In this study, since the sampling frames were easy to obtain, the stratified sampling technique was used. This technique was used since the population was divided into several faculties.

According to Awang [2] suggested by Hair et al. [6], if the model have five or less latent constructs and each construct has more than three items, the minimum sample required was 100 samples. Since this research have four latent constructs so the condition of minimum 100 samples were used. The total of respondents from Faculty of Computer and Mathematical Sciences (FSKM) was 77, Faculty of Sport Science and Recreation (FSR) was 15, the Faculty of Administrative Science and Policy Studies (FSPPP) was 62 and others were 68.

2.2 Data Collection Method

This research used primary data since a direct questionnaire was used to acquire useful information from academic staff in Universiti Teknologi Mara (UiTM) Negeri Sembilan Kampus Seremban. The questionnaire used in this study was adapted from a questionnaire on job satisfaction by Al-Hinai [1]. This data collection method was chosen in this study because it permits respondent's time to complete without interference form.

2.3 Theoretical Framework

This study consists of one dependent variable which is the job satisfaction and 3 independent variables as shown in Fig. 1. The independent variables included working condition, pay and promotion potential and relationships with others.

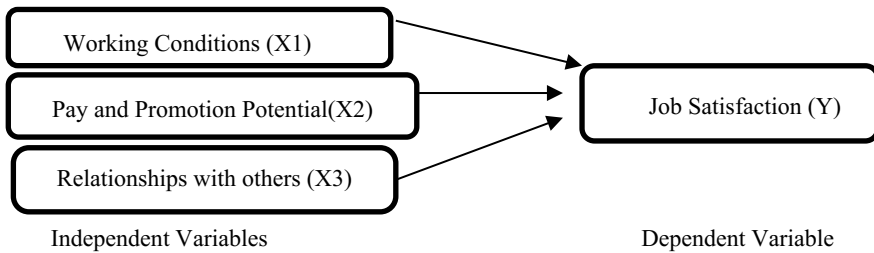


Fig. 1 Theoretical framework

2.4 Method of Analysis

Data obtained were quantitatively analyzed using the Structural Equation Modeling (SEM). Basically, structural equation modeling is similar with the multiple linear regressions. A two-step modeling approach was adopted. In the first step, the measurement model was tested to ensure that all latent constructs related to real variables.

In a second step, structural analysis was designed to examine the relationship between latent variables examined. The relationship between the variables will be tested only after ensuring adequate latent variables were measured. This procedure will reduce the risk of misinterpretation and bias. The data was analyzed by using statistical software namely as Analysis of Moment Structure (AMOS) version 23.0.

3 Analysis and Findings

3.1 Measurement Model

The assessment for unidimensionality, validity, and reliability for measurement models were required prior for modeling the structural model. The result shown in Table 1.

Based on Table 1, items C_1 , C_7 , D_6 , E_4 , and B_4 were deleted due to low factor loading. It shows that the unidimensionality was achieved since the measuring items for the inspective latent construct have acceptable factor loading (0.6 or higher). Furthermore, for the assessment of validity, the values of Average Variance Extracted (AVE) were measured for each variable. Table 1 revealed that the validity of measurement model was achieved since the value of AVE for each latent variable was more than 0.5. Moreover, the composite reliability (CR) also was calculated to measure how reliable the measurement model in measuring the intended latent constructs. Table 1 shows that the measurement model was reliable

Table 1 Assessing of unidimensionality, validity and reliability for measurement models

Construct	Item	Factor loading	Cronbach's alpha	Composite reliability (CR)	AVE
X ₁	C ₁	This item is deleted due to low factor loading			
	C ₂	0.60	0.766	0.868	0.574
	C ₃	0.81			
	C ₄	0.85			
	C ₅	0.85			
	C ₆	0.64			
	C ₇	This item is deleted due to low factor loading			
X ₂	D ₁	0.84	0.801	0.910	0.671
	D ₂	0.74			
	D ₃	0.81			
	D ₄	0.86			
	D ₅	0.84			
	D ₆	This item is deleted due to low factor loading			
X ₃	E ₁	0.78	0.786	0.760	0.444
	E ₂	0.61			
	E ₃	0.67			
	E ₄	This item is deleted due to low factor loading			
	E ₅	0.60			
Y	B ₁	0.83	0.722	0.917	0.651
	B ₂	0.92			
	B ₃	0.82			
	B ₄	This item is deleted due to low factor loading			
	B ₅	0.72			
	B ₆	0.65			
	B ₇	0.87			

because the value of CR for each variable was greater than 0.6. Thus, it can be concluded that the unidimensionality, validity and reliability of the measurement model was satisfied.

3.2 Structural Model

The structural model included 4 latent variables and 20 observed variables (Fig. 2). Working Conditions (X1), Pay and Promotion Potential (X2) and Relationships with others (X3) were the exogenous latent variables that directly influenced the endogenous latent variable (Job Satisfaction). Univariate analyses of all variables included in the model were examined for normality. Skewness indices (ranging from -1.755 to -0.467) attested that the distribution of data does not depart from

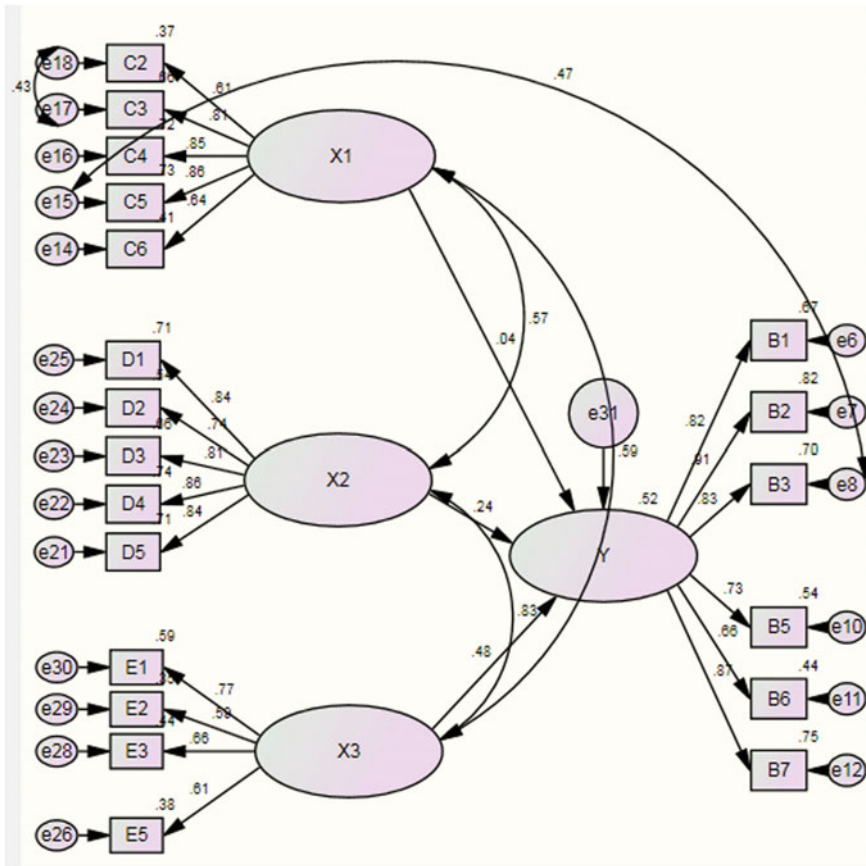


Fig. 2 The structural model

normality. It found that the value of skewness for every item was ± 2.0 . It indicated that no items exhibited significant skewness [3].

Based on Table 2, the Chi-square Goodness of Fit Test ($\chi^2 = 1.87$) showed that the model fits the data well [2]. Meanwhile, all fitness indices (NFI = 0.803, RFI = 0.769, TLI = 0.878, CFI = 0.896) that are close to 1 and (RMSEA = 0.094) which is value ranging between 0.05 and 1. It confirmed that this model is acceptably good [2, 7].

The value of coefficient of determination was 0.52. This figure indicated the contribution of exogenous constructs X_1 (Working Condition), X_2 (Pay and Promotion Potential) and X_3 (Relationships with Others) in estimating the endogenous construct Y was 52%.

Based on Table 3, it showed that only one variable was significant (p-value = 0.044) which is X_3 (Relationships with others). Meanwhile, variables X_1 (Working

Table 2 The assessment of fitness for the structural model

Name of category	Name of index	Index value	Comments
Absolute fit	RMSEA	0.094	The required level is accepted
Incremental fit	CFI	0.896	The required level is accepted
Parsimonious fit	Chisq/df	302.712/162=1.87	The required level is achieved

Table 3 The regression weight for X₁, X₂ and X₃ in predicting Y

Path		Path	The actual beta value	Standard error	Critical ratio	P-value
Y	<-	X ₁	0.020	0.052	0.380	0.704
Y	<-	X ₂	0.141	0.119	1.189	0.235
Y	<-	X ₃	0.451	0.224	2.014	0.044

Condition) and X₂ (Pay and Promotion Potential) seem not significant since the p-value of both variables was 0.704 and 0.235 respectively which is greater than significant value (0.05).

4 Conclusion

Job satisfaction was either a global feeling about the job or a related constellation of attitudes about various aspect of the job. The main objective in this paper was to investigate the factors that influence job satisfaction among academic staff. Preliminary findings in this study showed that variable ‘the relationships with others’ was significantly influence the job satisfaction among academic staff. This result was in line with previous studies [1, 6, 8].

Academicians should be provided with proper guidance and counseling by the institutions. Thus, academicians will be aware of their duties and working conditions in the university. Other than that, to reduce the conflicts with co-workers or leaders, the authorities should provide clear guidelines, so that academicians will be aware of their role and there will no ambiguity in understanding of their works. This research will provide a platform to researchers in digging other factors that influence job satisfaction by interviewing more academics staff in higher education.

References

1. AL-Hinai, Z.A.: A study on the factors affecting job satisfaction of academic staff in higher education institution. In: 13th International Academic Conference, Antibes. IISES. ISBN 978-80-87927-05-2 (2014)
2. Awang, Z.: Structural Equation Modelling Using AMOS Graphic, pp. 1–97. UiTM Shah Alam (2012)

3. Bolt, M.A.: Chapter 6 Statistical Analysis (1999). Retrieved from <http://scholar.lib.vt.edu/theses/available/etd-040999201108/unrestricted/CHP6.PDF>, 10 Oct 2015
4. Ghaffar, A.: Factors affecting job satisfaction level of academic staff in Pakistan. *J. Acad. Pract. Anal.* **1**(2), 45–59. ISSN 2222-1735 (2013)
5. Grace, D.H., Khalsa, S.A.: Re-recruiting faculty and staff: the antidote to today's high attrition. *Indep. Sch.* **62**(3), 20–27 (2003)
6. Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: *Multivariate Data Analysis*, 7th edn. Prentice Hall, New Jersey (2010)
7. Hu, L., Bentler, P.M.: Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* **6**, 1–55 (1999)
8. Khalid, S.: Job satisfaction among academic staff: a comparative analysis between public and private sector universities of Punjab, Pakistan. *Int. J. Bus. Manag.* **7**(1) (2012)
9. Mustapha, N.: Measuring job satisfaction from the perspective of interpersonal relationship and faculty workload among academic staff at public universities in Kelantan, Malaysia. *Int. J. Bus. Soc. Sci.* **4**(15) (2013)
10. Nordin, F.: Levels of job satisfaction amongst Malaysian academic staff. *Asian Soc. Sci.* **5**(5). ISSN 1911-2017 (2009)
11. Oshagbemi, T.: Correlates of pay satisfaction in higher education. *Int. J. Educ. Manag.* **14**(1), 31–39 (2000)

Chapter 46

Machine Learning Using H2O R Package: An Application in Bioinformatics



Azian Azamimi Abdullah and Shigehiko Kanaya

Abstract Bioinformatics is an interdisciplinary field that combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. In systems biology, many analytical methods such as mass spectrometry and DNA sequencing generate a large amount of data and advanced statistical and bioinformatics tools are urgently needed to analyze such data. In this study, machine-learning methods using the H2O package in R software were proposed to classify 341 volatile organic compounds (VOCs) based on their molecular structure. Using nine types of molecular fingerprints, including one newly proposed fingerprint (COMBINE) to represent the molecules, 72 classification models were generated to predict biological activities of VOCs by four machine-learning methods, which are deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM). The models were evaluated by an external validation set containing 120 VOCs from other sources. Based on computational results, the best classification model was developed by COMBINE fingerprint trained with GBM method with predictive accuracy at 94.4% and the obtained mean-squared error (MSE) value was 0.3952804. We found that the combination of molecular fingerprints and machine-learning methods can be used for predicting biological activities of VOCs. It is recommended to use COMBINE fingerprint trained with GBM method in the context of classifying VOCs. GBM method has advantage in term of computational speed and requires less parameter for optimization compared to other machine-learning methods.

A. A. Abdullah (✉)

Biomedical Electronic Engineering Programme, School of Mechatronic Engineering,
Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: azamimi@unimap.edu.my

S. Kanaya

Computational Systems Biology Laboratory, Graduate School of Information Science,
Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan
e-mail: skanaya@gtc.naist.jp

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_46

375

1 Introduction

Machine learning is the subfield of computer science, which an application of artificial intelligence (AI) that provides systems the ability to automatically learn from experience without being explicitly programmed. Machine learning can be categorized into two types; unsupervised and supervised learning. There are many machine-learning algorithms available out there, however, it depends on the data itself in order to choose the most suitable algorithms to solve a problem. Recently, machine-learning methods are popular in bioinformatics and quantitative structure–activity relationships (QSAR), which usually predicting the unknown property values of a test set of molecules based on the known values for a training set [1–3]. In this study, we applied machine-learning approaches in bioinformatics field, where four machine-learning algorithms were used to classify 341 volatile organic compounds based on their molecular structure. We developed 72 classification models trained with four types of supervised machine learning methods, which are deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM) using H2O package in R software [4].

2 Proposed Methods

2.1 Datasets

In this study, we have accumulated 341 volatiles emitted by various biological species associated with 11 types of biological activities. We prepared a database, KNApSAcK Metabolite Ecology Database for accumulation and search of volatile metabolite-species relationships, as shown in Fig. 1. This database is currently available and can be accessed freely at <http://kanaya.naist.jp/MetaboliteEcology/top.jsp> [5].

2.2 Molecular Fingerprints

The molecular fingerprint of a chemical compound is a binary vector indicating the substructures it contains. In this study, eight types of molecular fingerprints are used to represent the molecules, which are PubChem (PubChem, 881 bits), CDK (CDK, 1024 bits), Extended CDK (Extended, 1024bits), MACCS (MACCS, 166 bits), Klekota-Roth (KR, 4860 bits), Substructure (Sub, 307 bits), Estate (Estate, 79 bits), and atom pairs (AP, 780 bits). We also proposed a new type of fingerprint, by combining all features and substructures obtained by these fingerprints (COMBINE, 9121 bits). We converted the SDF files of all 341 VOCs into binary fingerprints using ChemDes software [6]. There are 11 classes of biological

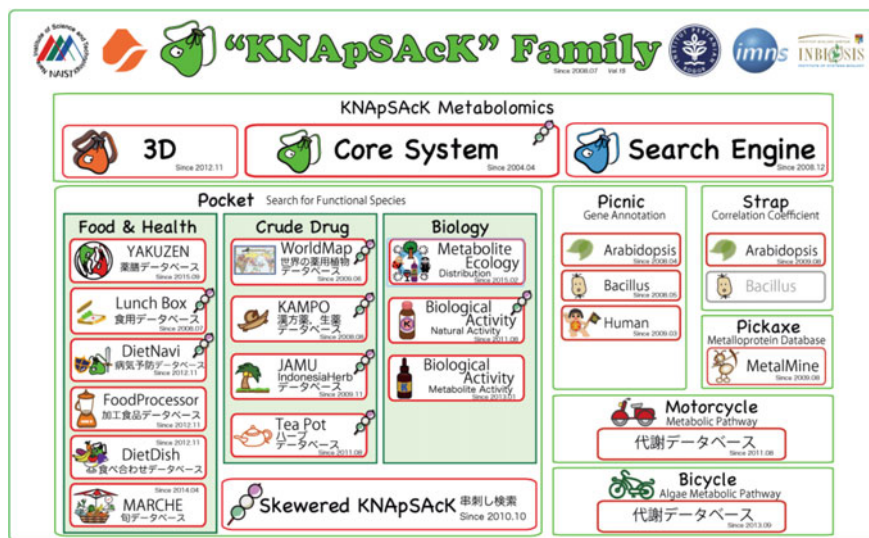


Fig. 1 The main window of KNApSACK metabolite ecology database

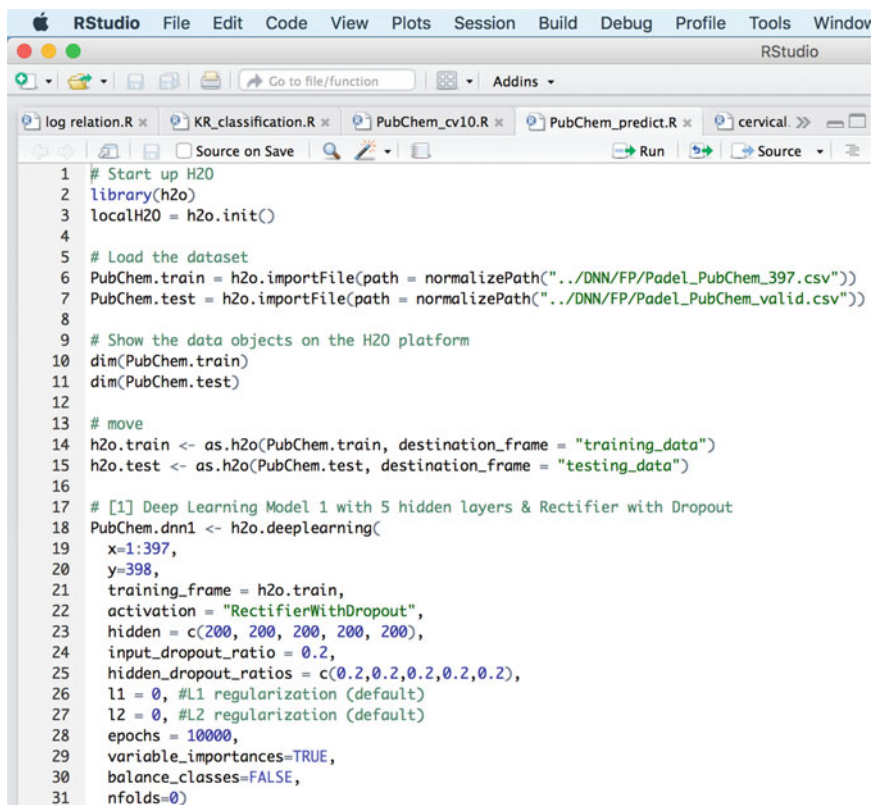
activities, which have been used as outputs for the classification model. The VOC-Substructure-Biological activities relations can be represented as a matrix, as shown in Table 1 where rows represent VOCs and columns represent substructures of molecular fingerprints. We added one additional column to represent biological activities for each of VOCs.

2.3 Machine Learning Using H2O R Package

We implemented four types of supervised machine learning methods for classification of VOCs, which are deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM) using H2O

Table 1 Representation of VOCs, substructures and biological activities as a two-dimensional matrix

VOCs	Substructures						Biological activities
	S ₁	S ₂	S ₃	S ₄	...	S _M	
VOC ₁	1	0	1	1	...	0	Antimicrobial
VOC ₂	1	1	0	0	...	0	Biomarker
VOC ₃	0	1	0	1	...	0	Defense
...	
VOC _N	1	0	0	0	...	1	Odor



```
1 # Start up H2O
2 library(h2o)
3 localH2O = h2o.init()
4
5 # Load the dataset
6 PubChem.train = h2o.importFile(path = normalizePath("../DNN/FP/Padel_PubChem_397.csv"))
7 PubChem.test = h2o.importFile(path = normalizePath("../DNN/FP/Padel_PubChem_valid.csv"))
8
9 # Show the data objects on the H2O platform
10 dim(PubChem.train)
11 dim(PubChem.test)
12
13 # move
14 h2o.train <- as.h2o(PubChem.train, destination_frame = "training_data")
15 h2o.test <- as.h2o(PubChem.test, destination_frame = "testing_data")
16
17 # [1] Deep Learning Model 1 with 5 hidden layers & Rectifier with Dropout
18 PubChem.dnn1 <- h2o.deeplearning(
19   x=1:397,
20   y=398,
21   training_frame = h2o.train,
22   activation = "RectifierWithDropout",
23   hidden = c(200, 200, 200, 200, 200),
24   input_dropout_ratio = 0.2,
25   hidden_dropout_ratios = c(0.2,0.2,0.2,0.2,0.2),
26   l1 = 0, #L1 regularization (default)
27   l2 = 0, #L2 regularization (default)
28   epochs = 10000,
29   variable_importances=TRUE,
30   balance_classes=FALSE,
31   nfolds=0)
```

Fig. 2 H2O package in R software

package in R software [7–10]. H2O is an open-source math and in-memory prediction engine for big data science developed by 0 × data. Figure 2 shows a part of machine learning coding using the H2O R package.

2.4 Performance Metrics

The performance of multi-classification models was measured by mean squared error (MSE) and accuracy (%) value. We conducted two experiments: (1) Using all datasets as training, (2) Using 10-fold cross-validation technique. In this technique, the compounds were randomly divided into ten parts, where nine parts were used for training and remaining part for testing. This process is carried out ten times in such a way that each part was used once for testing. We also validated the data by using 120 external new datasets, which were obtained from other sources.

3 Results and Discussion

Figure 3 shows the performance of 72 classification models (MSE value) by using all datasets as training and 10-fold cross validation technique. Based on this figure, we can observe that the classification model No 23 (PubChem fingerprint trained with GBM method) gives good results in both experiments. This model obtained MSE value = 0.1214795 when using all datasets as training and MSE value = 0.39318013 in case of 10-fold cross validation technique. The results show that GBM method is good at predicting biological activities of VOCs. The left side points, which consist of many fingerprint types and DNN methods suffered from over-fitting problems. The performance of DNN is good when using all datasets as training, however it becomes worst when we used 10-fold cross validation technique, such as model No 12 (Klekota-Roth fingerprint trained with DNN4 method) and model No 36 (Extended fingerprint trained with DNN4 method).

We also evaluated the performance of all 72 models in term of classification accuracy. Classification accuracy is the ratio of correct predictions to total predictions made and often presented as a percentage by multiplying the result by 100. Figure 4 shows the performance of 72 classification models in term of accuracy value (%) by using all datasets as training and 10-fold cross validation technique. Based on Fig. 4, we observed that the classification model No 7 (COMBINE fingerprint trained with GBM method) gives good results in both experiments. This model obtained accuracy value of 94.4% when using all datasets as training and 57.7% in case of 10-fold cross validation technique. This result somehow is aligned with our previous result shown in Fig. 3, where we proved that GBM appears to be a very effective and efficient algorithm, compared to other machine-learning methods.

It is important to show that the recommended fingerprint and machine learning method will also apply to other new datasets that have been not been part of the

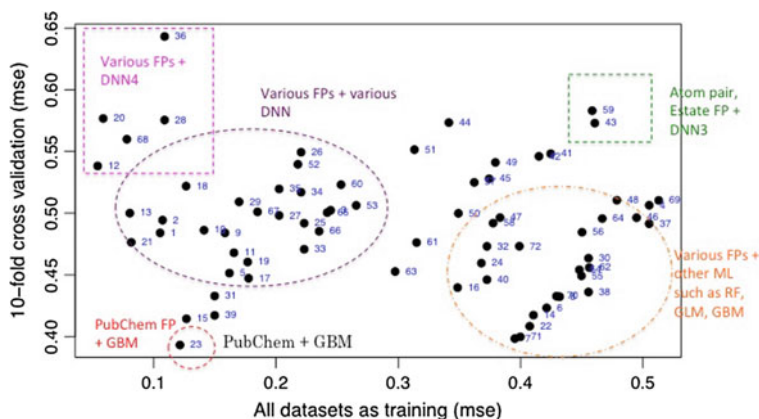


Fig. 3 Performance of 72 classification models in term of MSE value

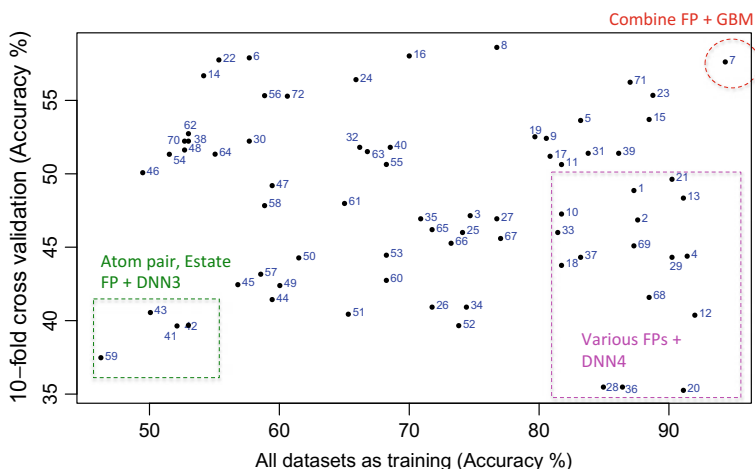


Fig. 4 Performance of 72 classification models in term of accuracy (%) value

model building. Thus, 120 additional new VOC datasets were selected from other sources. We found that 73 VOC were predicted correctly, which give about 60.8% accuracy. The low accuracy of prediction results is because of small sample of our training datasets, which is not sufficient enough for the model to learn and predict new data. Most of VOCs are predicted as biomarker, maybe because the training data was dominated by biomarkers. Also, 12 of the 14 odor VOCs were predicted as biomarkers. It is maybe because odor and biomarker VOCs are chemical structurally similar. It is recommended to increase the quantity of sample datasets so that the data for each of activity are well balanced and this can increase the prediction accuracy.

4 Conclusions

We have developed 72 classification models for the classification of VOCs by 9 types of fingerprints and trained by deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM). Based on the computational results, PubChem and COMBINE fingerprints were recommended as the input for the prediction model. Gradient boosting machine (GBM) method can outperform deep neural network (DNN) in term of classifying VOCs, in our case. GBM method has advantage in term of computational speed and requires less parameter for optimization, compared to other machine learning methods. Hence, we highly recommend using GBM for the classification of VOCs based on chemical structures.

In future, more VOCs can be accumulated, and comprehensive analysis can be performed in the context of human healthcare and chemical ecology. The prediction outcome may be useful for the discovery of novel agricultural tools and also for the non-invasive identification of biomarkers in the medical diagnostic field.

References

1. Mitchell, J.B.: Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**(5), 468–481 (2014)
2. Karthikeyan, M., Vyas, R.: Machine learning methods in chemoinformatics for drug discovery. In: *Practical Chemoinformatics*, pp. 133–194. Springer, India (2014)
3. Libbrecht, M.W., Noble, W.S.: Machine learning in genetics and genomics. *Nat. Rev. Genet.* **16**(6), 321–332 (2015)
4. Intelligence, M.: *High Performance Machine Learning in R with H2O* (2015)
5. Abdullah, A.A., et al.: Development and mining of a volatile organic compound database. *Biomed. Res. Int.* (2015)
6. Dong, J., et al.: ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* **7**(1), 1–10 (2015)
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2001)
9. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
10. Cook, R.J.: Generalized linear model. *Encycl. Biostat.* **6**(2), e16104 (1998)

Chapter 47

Modeling of Risk for Diabetes Mellitus and Hypertension Using Bi-response Probit Regression



Suliyanto and M. Rifada

Abstract One of statistical analysis used to find relation among categorical response variable either categorical or continue predictor variables is logistic regression analysis. Logistic regression has two link functions as logit and probit link. Assumption applied on bi-response probit regression model is that both of response variables are connected. Diabetes mellitus and hypertension are a related disease. They can occur at the same time and are known as comorbidity diseases, i.e. diseases that may exist on the same patients. Hypertension sufferers have chances to become diabetic patients. Moreover, hypertension may be possessed by diabetic patients. Therefore, the aim of this study was to design a model of risk for diabetes mellitus and hypertension occurrence simultaneously using bi-response probit regression approach and identify significant factors influencing diabetes mellitus and hypertension. Based on the result of chi-square test to find the relation between diabetes mellitus and hypertension, Pearson Chi-Square scored 15.009 while p-value 0.000. It means that there was a closed relation on both diabetes mellitus and hypertension so it can be concluded that both of response variables is dependent. Furthermore, based on the smallest AIC's score was obtained the best bi-response probit regression model with significant factors influencing diabetes mellitus and hypertension occurrences were Body Mass Index (BMI), systolic blood pressure, and diastolic blood pressure.

Keywords AIC · Bi-response probit regression · Diabetes mellitus · Hypertension

1 Introduction

The relation between the risk of the occurrence of a disease and the factors that influence it would be more useful if formulated in mathematical models to determine how much influence factors that significantly affect the probability of the

Suliyanto · M. Rifada (✉)
Universitas Airlangga, Surabaya, Indonesia
e-mail: marisa.rifada@fst.unair.ac.id

Suliyanto
e-mail: yanfit@yahoo.com

occurrence a person exposed to a disease. One of statistical analysis that be used to find relation among categorical response variable and categorical or continuous predictor variables is logistic regression analysis [1]. Acheampong [2] conducted a study to determine the relation between diabetes with hypertension and other cardiovascular risk factors among diabetic patients in Komfo Anokye Teaching Hospital, Kumasi. The main objective is to modeling the relation between diabetes and hypertension and significant risk factors. However, the study only modeled the risk of the occurrence of diabetes and hypertension independently.

Diabetes and hypertension is a disease closely linked. They often occur together so it is considered ‘comorbidity’, i.e. a disease that may exist on the same patients. Hypertension sufferers may have a risk of developing diabetes. And vice versa, the risk of hypertension can also be experienced by diabetics [3]. According to research of Rifada et al. [4], there is a close relation between diabetes and hypertension in patients at Surabaya Hajj General Hospital Indonesia, so it would be more realistic if it were modeled together (simultaneously). Rifada et al. [4] used bi-response logistic regression approach with a logit link function. In the logistic regression, there are link function logit and probit. Cakmakyapan and Goktas [5] have done a simulation to compare the accuracy of the classification between the link function logit and probit, and the results showed that the logit link function is better used for a number of large samples of more than or equal to 500 samples, whereas for the probit link function more well used to the small sample size that is less than or equal to 200 samples. Therefore, in this study will be developed other approach of bi-response logistic regression with probit link function (bi-response probit regression model). Bi-response probit regression model is a regression model that consists of two interconnected response variables, each of which consists of two categories, while the predictor variables are categories or continuous. The aim of this study is to design a model of occurrence risk of diabetes and hypertension simultaneously by using bi-response probit regression approach and identify factors that significantly influence the occurrence of diabetes and hypertension.

2 Theoretical Estimation Model

According to [6], In the bi-response probit regression model, we assume that both unobserved response variables, Y_1^* and Y_2^* , follow regression model:

$$Y_1^* = X\beta_1 + \varepsilon_1 \quad (1)$$

$$Y_2^* = X\beta_2 + \varepsilon_2 \quad (2)$$

$\beta_1 = [\beta_{10} \ \beta_{11} \ \dots \ \beta_{1p}]^T$, $\beta_2 = [\beta_{20} \ \beta_{21} \ \dots \ \beta_{2p}]^T$, $X = [1 \ x_1 \ \dots \ x_p]$, error ε_1 and ε_2 are assumed has standard Normal distribution and

$corr(\varepsilon_1, \varepsilon_2) = \rho$. So, $Y_1^* \sim N(X\beta_1, 1)$ and $Y_2^* \sim N(X\beta_2, 1)$. The observed categorical response variables, Y_1 and Y_2 , each can be obtained by the unobserved response variables, Y_1^* and Y_2^* , as follows:

$$Y_1 = \begin{cases} 0 & , \text{ if } Y_1^* \leq 0 \\ 1 & , \text{ if } Y_1^* > 0 \end{cases} \quad \text{and} \quad Y_2 = \begin{cases} 0 & , \text{ if } Y_2^* \leq 0 \\ 1 & , \text{ if } Y_2^* > 0 \end{cases}$$

Because each response variables Y_1^* and Y_2^* has Normal distribution, so (Y_1^*, Y_2^*) has bivariate Normal distribution. The pdf of standard bivariate Normal for $(\varepsilon_1, \varepsilon_2)$ can be defined by:

$$\phi(\varepsilon_1, \varepsilon_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(\varepsilon_1^2 - 2\rho\varepsilon_1\varepsilon_2 + \varepsilon_2^2)\right) \tag{3}$$

Let $z_1 = -X\beta_1$ and $z_2 = -X\beta_2$, then the joint probability is defined as follow:

$$\begin{aligned} p_{00}(X) &= P(Y_1 = 0, Y_2 = 0) = \Phi(z_1, z_2) \\ p_{11}(X) &= P(Y_1 = 1, Y_2 = 1) = 1 - \Phi(z_1) - \Phi(z_2) + \Phi(z_1, z_2) \\ p_{10}(X) &= P(Y_1 = 1, Y_2 = 0) = \Phi(z_2) - \Phi(z_1, z_2) \\ p_{01}(X) &= P(Y_1 = 0, Y_2 = 1) = \Phi(z_1) - \Phi(z_1, z_2) \end{aligned}$$

Where $\Phi(\cdot)$ is the cdf of standard bivariate Normal.

We obtain estimator of parameters on bi-response probit regression model use Maximum Likelihood Estimation (MLE) method. The likelihood function can be written as follows:

$$L(\beta) = \prod_{i=1}^n p_{11i}^{y_{11i}} p_{10i}^{y_{10i}} p_{01i}^{y_{01i}} [1 - p_{11i} - p_{10i} - p_{01i}]^{1-y_{11i}-y_{10i}-y_{01i}} \tag{4}$$

Based on Eq. (4), we obtain log likelihood function as follow:

$$\ell(\beta) = \sum_{i=1}^n [y_{11i} \ln p_{11i} + y_{10i} \ln p_{10i} + y_{01i} \ln p_{01i} + y_{00i} \ln p_{00i}] \tag{5}$$

The solution of β is gotten by minimizing of log likelihood function and we get implicit equation, so we use numerical method i.e. Newton Raphson iteration. Testing the parameter simultaneously using hypothesis as follows:

$$\begin{aligned} H_0 : \beta_{11} = \beta_{12} = \dots = \beta_{1p} = 0 \quad \text{and} \quad H_0 : \beta_{21} = \beta_{22} = \dots = \beta_{2p} = 0 \\ H_1 : \text{at least one } \beta_{rs} \neq 0, \text{ where } r = 1, 2 \text{ and } s = 1, 2, 3, \dots, p \end{aligned}$$

Test statistic which is used is

$$G^2 = 2 \sum_{i=1}^n \left[y_{11i} \ln \left(\frac{p_{2i} - p_{01i}}{p_{2i}^* - p_{01i}^*} \right) + y_{10i} \ln \left(\frac{p_{1i} - p_{2i} - p_{01i}}{p_{1i}^* - p_{2i}^* - p_{01i}^*} \right) + y_{01i} \ln \left(\frac{p_{01i}}{p_{01i}^*} \right) + y_{00i} \ln \left(\frac{1 - p_{1i} - p_{01i}}{1 - p_{1i}^* - p_{01i}^*} \right) \right] \tag{6}$$

For $n \rightarrow \infty$ and $G^2 \sim \chi_{df}^2$ with the critical areas of rejected H_0 if $G^2 > \chi_{df, \alpha}^2$, degrees of freedom (df) is the number of model parameters under population reduced by the number of model parameters below H_0 .

If the parameter test results are simultaneously significant, then proceed with partial test parameters with the following hypothesis:

$$\begin{aligned} H_0 : \beta_{rs} &= 0 \\ H_1 : \beta_{rs} &\neq 0, \text{ where } r = 1, 2 \text{ and } s = 0, 1, 2, \dots, p \end{aligned}$$

Test statistic which is used is

$$Z = \frac{\hat{\beta}_{rs}}{SE(\hat{\beta}_{rs})} \tag{7}$$

Statistics Z has standard Normal distribution, so that the decision is reject H_0 if $|Z| > Z_{\alpha/2}$ or P-value $< \alpha$.

3 Results and Discussion

In this study, response variables are the occurrence of Diabetes (Y_1) i.e. the condition of patient in the Poly Disease in Surabaya Hajj General Hospital whose the doctor’s diagnosis sufferer diabetes or not and the occurrence of Hypertension (Y_2) i.e. the condition of patient in the Poly Disease in Surabaya Hajj General Hospital whose the doctor’s diagnosis sufferer hypertension or not. Some predictor variables in this study are Age (X_1), BMI (Body Mass Index) (X_2), Waist circumference (X_3), Systolic blood pressure (X_4) and Diastolic Blood Pressure (X_5).

The assumption on bi-response probit regression model i.e. both response variables have a relation with each other. We use Chi-square test to determine the relation between two categorical response variables. By using SPSS software, we obtain the Pearson Chi-square value as shown in the following Table 1:

Table 1 The chi-square test of two response variables

	Value	P-value
Pearson chi-square	15.009	0.000
N	81	

Table 1 shows that the *Pearson Chi-Square* value is 15.009 and p-value is 0.000. Because p-value (0.000) < alpha (0.05), it means that there is a relation between two response variables i.e. the occurrence of Diabetes and Hypertension in Surabaya Hajj General Hospital and conclude that two response variables are mutually dependent.

Based on testing parameters of model simultaneously, we obtain Wald Chi-square value (G^2) of 36.56 and p-value of 0.0001 alpha (0.05). So, conclude that there are at least one of predictor variables significantly influence to occurrence of Diabetes and Hypertension. The next step, we test for model parameters partially to determine the effect of each predictor variable significantly to the occurrence of Diabetes and Hypertension, which minimum significant predictor variables against one occurrence of Diabetes or Hypertension. Based on the results of parameter estimation by using bi-response probit regression model with all predictor variables is obtained AIC value of 171.7551 and the results are shown as follows.

Based on Table 2, we obtain the most not significant predictor variable is waist circumference because the p-value is greatest, so removed from the model. Furthermore, we estimate bi-response probit regression model without waist circumference variable and obtain AIC value of 168.9957 and the result are given below.

Based on Table 3, the predictor variable age is not significant to occurrence of Diabetes and Hypertension thus excluded from the model. The estimation of Bi-response probit regression model without predictor variable age and waist circumference obtain AIC value of 168.6188 and the results are given in Table 4 as follows.

Table 2 The results of parameter estimation for occurrence of Diabetes and Hypertension

Predictor variables		Diabetes		Hypertension	
		Estimation	<i>P-value</i>	Estimation	<i>P-value</i>
X ₁	Age	0.020871	0.191	-0.0132521	0.398
X ₂	BMI	0.1469638	0.006	0.0918033	0.039
X ₃	Waist circumference	0.0063887	0.632	-0.0077461	0.488
X ₄	Systolic blood pressure	0.0321375	0.006	0.0337199	0.002
X ₅	Diastolic blood pressure	-0.0326805	0.095	0.0095976	0.565
	Constants	-7.007567	0.001	-6.505739	0.001

Table 3 The results of parameter estimation for occurrence of Diabetes and Hypertension without waist circumference variable

Predictor variables		Diabetes		Hypertension	
		Estimation	<i>P-value</i>	Estimation	<i>P-value</i>
X ₁	Age	0.0217813	0.177	-0.014456	0.351
X ₂	BMI	0.1538163	0.002	0.0825282	0.048
X ₄	Systolic blood pressure	0.0343205	0.003	0.0323811	0.002
X ₅	Diastolic blood pressure	-0.0330663	0.097	0.0085768	0.606
	Constants	-6.928065	0.001	-6.652449	0.000

Table 4 The results of parameter estimation for occurrence of Diabetes and Hypertension without age and waist circumference variables

Predictor variables		Diabetes		Hypertension	
		Estimation	<i>P-value</i>	Estimation	<i>P-value</i>
X ₂	BMI	0.1504701	0.002	0.0846183	0.044
X ₄	Systolic blood pressure	0.0410603	0.000	0.0276436	0.002
X ₅	Diastolic blood pressure	-0.0427574	0.025	0.0149215	0.329
	Constants	-5.71889	0.002	-7.389802	0.000

Table 4 shows that all predictor variables are significant to both occurrence of diabetes and hypertension. Based on the smallest AIC value criterion, then we obtain the best bi-response probit regression model with significant predictor variables are BMI, systolic blood pressure and diastolic blood pressure can be written as:

$$y_1^* = -5.71889 + 0.1504701 X_2 + 0.0410603 X_4 - 0.0427574 X_5 \tag{8}$$

$$y_2^* = -7.389802 + 0.0846183X_2 + 0.0276436X_4 + 0.0149215X_5 \tag{9}$$

So, based on Eqs. (8) and (9) we get:

$$z_1 = -y_1^* = 5.71889 - 0.1504701X_2 - 0.0410603X_4 + 0.0427574X_5 \tag{10}$$

$$z_2 = -y_2^* = 7.389802 - 0.0846183X_2 - 0.0276436X_4 - 0.0149215X_5 \tag{11}$$

Suppose that a patient has BMI of 21.33, systolic blood pressure of 125 and diastolic blood pressure of 74, then we obtain probability that patient affected by diabetes and hypertension are as follows:

$$\begin{aligned} p_{11} &= 1 - \Phi(z_1) - \Phi(z_2) + \Phi(z_1, z_2) \\ &= 1 - \Phi(0.54087288) - \Phi(1.02525267) + \Phi(0.54087288, 1.02525267) \\ &= 1 - 0.7057024 - 0.847378 + 0.633831 = 0.0807506 \end{aligned}$$

$$\begin{aligned} p_{10} &= \Phi(z_2) - \Phi(z_1, z_2) = \Phi(1.02525267) - \Phi(0.54087288, 1.02525267) \\ &= 0.847378 - 0.633831 = 0.213547 \end{aligned}$$

$$\begin{aligned} p_{01} &= \Phi(z_1) - \Phi(z_1, z_2) = \Phi(0.54087288) - \Phi(0.54087288, 1.02525267) \\ &= 0.7057024 - 0.633831 = 0.0718714 \end{aligned}$$

$$p_{00} = \Phi(z_1, z_2) = \Phi(0.54087288, 1.02525267) = 0.633831$$

So, a patient with BMI of 21.33, systolic blood pressure of 125 and diastolic blood pressure of 74 has the greatest probability for not affected by diabetes and

hypertension of 0.633831. The value of classification accuracy from the best bi-response probit regression model on occurrence of diabetes and hypertension in Surabaya Hajj General Hospital is 64.1975%.

References

1. Agresti, A.: *Categorical Data Analysis*, 2nd edn. Wiley, Hoboken, New Jersey (2003). <https://doi.org/10.1002/0471249688>
2. Acheampong, A.Y.: *The Relation between Diabetes with Hypertension and other Cardiovascular Risk Factors Using Logit and Probit Model*, A Thesis submitted to University of Science Technology, Kumasi (2011). 10.1.1.845.6674
3. Epstein, M., Sowers, J.R.: Diabetes mellitus and hypertension. *Hypertension* **19**, 403–418 (1992). <https://doi.org/10.1161/01.hyp.19.5.403>
4. Rifada, M., Chamidah, N., Norrachma, S.N.: Pemodelan Risiko Kejadian Diabetes Mellitus dan Hipertensi berdasarkan Regresi Logistik Birespon. *Jurnal Statistika* **5**(2) (2017). ISSN 2338–3216
5. Cakmakyapan, S., Goktas, A.A.: Comparison of binary logit and probit models with a simulation study. *J. Soc. Econ. Stat.* **2**(1), 1–17 (2013)
6. Bokosi, F.K.: Household poverty dynamics in malawi: a bivariate probit analysis. *J. Appl. Sci.* **7** (2), 258–262 (2007). <https://doi.org/10.3923/jas.2007.258.262>

Chapter 48

On Bootstrapping Using Smoothed Bootstrap



Sulafah Binhimd

Abstract The standard bootstrap method was introduced as a resampling method for statistical inference; it is a computer based method for assigning measures of accuracy to statistical estimates. The bootstrap sample is obtained by randomly sampling n times, with replacement, from the original sample. Thereafter, different versions of bootstrap were developed such as smoothed bootstrap. The smoothed bootstrap method uses n observations to create $n + 1$ intervals, and then sample the observations from these intervals. This paper will discuss the smoothed bootstrap method and compare it to the standard method via simulation studies.

Keywords Bootstrap · Smoothed bootstrap · Prediction interval · Resampling method

1 Introduction

The standard bootstrap method was introduced as a resampling method for statistical estimates by Efron [1]. It is a computer based method for assigning measures of accuracy to statistical estimates, and based on independent observations. Initially this method was used to estimate the distribution of statistics, and then developed to work with various statistical inferences. The basic idea of the bootstrap is estimating the characteristics of the probability distribution for a random variable of interest. The standard bootstrap sample is obtained by randomly sampling n times, with replacement, from the original sample. Banks [2] described the smoothed Efron's bootstrap (smoothed bootstrap method) as a new version of the bootstrap, he smooths Efron's bootstrap by linear interpolation histospline smoothing between the jump points of empirical distribution. It is using n observations to create $n + 1$ intervals, and then sample the observations from them. The observations in

S. Binhimd (✉)

Faculty of Science, Department of Statistics, King Abdulaziz University,
Jeddah, Saudi Arabia

e-mail: shamad@kau.edu.sa

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,

https://doi.org/10.1007/978-981-13-7279-7_48

smoothed bootstrap sample are sampled from the intervals in between the data observations and also outside the data range, not restricted to the values in the data. Banks [2] used the confidence intervals to study the performance of this version of bootstrap and for comparing basic bootstrap strategies, he showed that the smoothed bootstrap has a good results in some cases. This paper will discuss the performance of smoothed bootstrap method using the prediction intervals, and compare it with standard bootstrap method. The smoothed bootstrap is suggested here because it is not widely discussed although it gives good results in some cases.

In Sect. 2 of this paper we present an overview of the standard bootstrap and smoothed bootstrap methods. Section 3 will illustrate a small simulation study which is used to compare the standard bootstrap with smoothed bootstrap method. Finally, in Sect. 4, some conclusions from the study are shown.

2 Bootstrapping

In this section we present a description of the two bootstrap methods.

2.1 *Standard Bootstrap*

The standard bootstrap method is used for specifying the measures of accuracy of the sample estimate, especially the standard error. It is generating an empirical estimate of the sampling distribution of the statistic by using the Monte Carlo sampling, which builds on drawing a large number of samples from the original sample and obtaining the statistic for each sample. Utilizing the bootstrap method to estimate the standard error does not require complex calculations, and it is available for any estimator. It is a useful method when the sample size is not sufficient for statistical inference.

A standard bootstrap sample $x^* = (x_1^*, x_2^*, \dots, x_n^*)$ is obtained by randomly sampling n times with replacement from the original sample x_1, x_2, \dots, x_n , see [3]. The size of the original sample can be different to the size of a bootstrap sample. The basic steps of the standard bootstrap method are draw B random samples of size n with replacement from the original sample and calculate the statistic of interest to estimate the sampling distribution.

2.2 *Smoothed Bootstrap*

A smoothed version of bootstrap, which discuss in this paper, was introduced by Banks [2]. He smooths the standard bootstrap by linear interpolation histospline

smoothing between the jump points of empirical distribution to produce the smoothed empirical distribution. Histospline is a smooth density estimate based only on the information in a histogram. The procedure is as follows:

1. Take the data set of n observations which are real valued data, one dimensional on a finite interval.
2. These n observations partition the intervals between $x_0, x_1, x_2, \dots, x_n, x_{n+1}$ into $n + 1$ intervals, where x_0 and x_{n+1} are the end points of the possible finite range.
3. Put uniformly distributed probabilities $\frac{1}{n+1}$ over each interval.
4. Sample n observations from the distribution.
5. Find the statistic of interest.
6. Repeat Steps 4 and 5 B times to obtain B bootstrap samples.

A smoothed bootstrap sample does not consist just of the observations from the original sample but of points from the whole possible data range. Banks [2] used confidence regions to compare his method with other bootstrap methods, at different values of α .

3 Simulation Study

In this section we have performed simulation studies for the performance of the smoothed bootstrap method by creating prediction intervals, and we did the same for the standard bootstrap method for comparison. Various types of bootstrap prediction intervals can be used, see [4–8]. In [9], the bootstrap method was used to construct prediction intervals for observations from the Birnbaum-Saunders distribution. They applied the bootstrap percentile method and found good coverage results.

In this simulation study we applied the bootstrap percentile prediction interval with the mean value (and the variance and the 75-th percentile in a similar way), as follows:

1. Draw an actual sample of size n from a specific distribution, giving x_1, \dots, x_n . Then draw a second sample of size m from the same distribution, giving y_1, \dots, y_m with mean value \bar{y} , this will be used as the future sample to check the performance of the bootstrap prediction intervals.
2. Use the actual sample to draw B smoothed bootstrap samples of size m as described above. Calculate the mean of each sample, giving m_j for $j = 1, \dots, B$.
3. Construct a $100(1 - \alpha)\%$ prediction interval for the mean by defining the lower bound to be the $\frac{\alpha}{2} \times B$ -th value in the ordered list of the values m_j and the upper bound to be the $(1 - \frac{\alpha}{2}) \times B$ -th value in this list (use the largest integer if these values are not integer).
4. Check if the prediction interval from Step 3 contains the mean \bar{y} of the future sample from the underlying distribution as resulted from Step 1.

This procedure is applied repeatedly (say 100 times) to get a pointer of the coverage of these intervals, that is the proportion of intervals which contain the mean value of the future sample from the underlying distribution. For perfect coverage the probability of the $100(1 - \alpha)\%$ interval containing that mean value should of course be $100(1 - \alpha)\%$. We also considered the variance and 75-th percentile q_{75} . Also, the same procedure has been applied with the standard bootstrap method, just change Step 2 to draw the standard bootstrap samples. We draw $B = 1000$ bootstrap samples using each of the two methods. For the simulations reported in this paper, the underlying distribution is Beta distribution as an example of continuous distribution on a finite interval $[0,1]$, with different values of parameters, Beta (0.5,0.5), Beta (2,5) and Beta (5,1), these are symmetric, positive skewed and negative skewed shaped, respectively, and we took $m = n$. The tables below are present the proportions of correct coverage of the value of the statistic of interest from the future sample, for several values of n and $\alpha = 0.05$.

From Table 1, the smoothed bootstrap method, in most cases, has higher coverage probability than standard bootstrap method. This is logical because the smoothed bootstrap have more variation; it is not restricted to the values in the actual sample.

Tables 2 and 3 show the same results with different values of parameters. This study shows that both methods are under coverage but the smoothed bootstrap method is close to the target value in most cases. The results from this study, for both methods when the statistic of interest is the mean or variance or the 75-th percentile, are not different.

Table 1 Coverage of 95% prediction interval when the original sample was from Beta (0.5,0.5)

Mean					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.80	0.72	0.86	0.85	0.68
Standard bootstrap	0.65	0.64	0.84	0.86	0.70
Variance					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.81	0.79	0.79	0.77	0.76
Standard bootstrap	0.62	0.64	0.81	0.73	0.70
q_{75}					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.71	0.66	0.83	0.88	0.70
Standard bootstrap	0.71	0.65	0.82	0.88	0.71

Table 2 Coverage of 95% prediction interval when the original sample was from Beta (2,5)

Mean					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.86	0.77	0.80	0.85	0.75
Standard bootstrap	0.77	0.74	0.75	0.82	0.74
Variance					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.91	0.89	0.84	0.86	0.77
Standard bootstrap	0.70	0.73	0.74	0.76	0.67
q_{75}					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.77	0.80	0.77	0.75	0.81
Standard bootstrap	0.75	0.80	0.78	0.74	0.80

Table 3 Coverage of 95% prediction interval when the original sample was from Beta (5,1)

Mean					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.76	0.82	0.77	0.79	0.77
Standard bootstrap	0.74	0.83	0.74	0.78	0.77
Variance					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.83	0.75	0.75	0.77	0.76
Standard bootstrap	0.81	0.73	0.73	0.76	0.76
q_{75}					
$n = m$	20	50	100	200	500
Smoothed bootstrap	0.80	0.74	0.82	0.81	0.80
Standard bootstrap	0.77	0.74	0.81	0.81	0.81

4 Conclusions

This paper has discussed the standard bootstrap and smoothed bootstrap methods, and uses the prediction interval in the simulation study. Both methods have under coverage probability, but the smoothed bootstrap method has higher coverage probability and appears to be close to the target value in most cases. That is show that the smoothed bootstrap method is a useful alternative to the standard bootstrap for prediction.

References

1. Efron, B.: Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979). <https://doi.org/10.1214/aos/1176344552>
2. Banks, D.L.: Histospline smoothing the Bayesian bootstrap. *Biometrika* **75**, 673–684 (1988). <https://doi.org/10.1093/biomet/75.4.673>
3. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman and Hall (1993)
4. Mojirsheibani, M.: Iterated bootstrap prediction intervals. *Stat. Sin.* **8**, 489–504 (1998)
5. Adakeye, K.S., Lamidi, M.A., Osanaiye, P.A.: Prediction interval: a tool for monitoring outbreak of some prominent diseases. *Glob. J. Maths* **2**, 41–46 (2010)
6. Mojirsheibani, M., Tibshirani, R.: Some results on bootstrap prediction intervals. *Can. J. Stat.* **24**, 549–568 (1996). <https://doi.org/10.2307/3315333>
7. Goncalves, S., Perron, B., Djogbenou, A.: Bootstrap prediction intervals for factor models. *J. Bus. Econ. Stat.* **35**, 53–69 (2017). <https://doi.org/10.1080/07350015.2015.1054492>
8. Errouissi, R., Cardenas-Berrera, J., Meng, J., Castillo-Guerra, E., Gong, X., Chang, L.: Bootstrap prediction interval estimation for wind speed forecasting. In: *Energy Conversion Congress and Exposition (ECCE)*. IEEE (2015). <https://doi.org/10.1109/ecce.2015.7309931>
9. Lu, M.C., Chang, D.S.: Bootstrap prediction intervals for Birnbaum-Saunders distribution. *Microelectron. Reliab.* **37**, 1213–1216 (1997)

Chapter 49

On the Markov Chain Monte Carlo Convergence Diagnostic of Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification



Nur Iriawan, Kartika Fithriasari, Brodjol Sutijo Suprih Ulama,
Irwan Susanto, Wahyuni Suryaningtyas
and Anindya Apriliyanti Pravitasari

Abstract The Bidikmisi scholarship program is an education assistance program by the government of Indonesia which aims to achieve equitable access and learning opportunities at University. Bidikmisi acceptance status having a binary type (i.e. 0 and 1) produces a structure of Bernoulli mixture model with two components. The characteristics of each component can be identified through the Bernoulli mixture regression modeling by involving the covariates of Bidikmisi scholarship grantees. The estimating parameter of Bernoulli mixture regression model was performed using Bayesian-Markov Chain Monte Carlo (MCMC) approach. One of the challenges in using Bayesian-MCMC algorithm is determining the convergence of the sampler to the posterior distribution which is typically assessed using diagnostics tools. In this paper, we present that the diagnostics tools such as Geweke method, Gelman-Rubin method, Raftery-Lewis method and Heidelberger-Welch method can give different results to conclude MCMC convergence. The improvement of convergence indicators occurs on Gelman-Rubin method and Heidelberger-Welch method when the number of iterations is increased.

Keywords Markov chain monte carlo · Convergence diagnostic · Bernoulli mixture regression · Bayesian computation · Bidikmisi

N. Iriawan (✉) · K. Fithriasari · I. Susanto · W. Suryaningtyas · A. A. Pravitasari
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: nur_i@statistika.its.ac.id

I. Susanto
e-mail: irwansusanto@staff.uns.ac.id

B. S. S. Ulama
Department of Business Statistics, Institut Teknologi Sepuluh Nopember, Surabaya,
Indonesia
e-mail: brodjol_su@statistika.its.ac.id

1 Introduction

Bernoulli mixture model is developed based on mixture distribution which has an adaptive capability to represent data pattern in data-driven analysis perspective [1]. Bidikmisi acceptance status that has a binary type can be formed as Bernoulli mixture model with two components. The characteristics of each component can be identified through the Bernoulli mixture regression model by involving the covariates of Bidikmisi scholarship recipients.

The inference for Bernoulli mixture regression model with Bayesian-Markov Chain Monte Carlo (MCMC) can overcome a particular challenge on computational aspects. Nevertheless it encounters a weakness that relates with the convergence of estimation process. Therefore the parameter estimation process of Bayesian Bernoulli mixture regression model has to be assessed on its convergence achievement.

2 Methodology

2.1 Bayesian Bernoulli Mixture Regression Model

Nadif and Govaert [2] introduced Bernoulli mixture model which was further developed by Grun and Leisch [3] in the generalized linear model framework. Suppose $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ is a random sample of a binary vector which has a linear relationship with covariates X_1, X_2, \dots, X_p on each Y_i such that

$$\eta_i = g(\mu_i) = g(E(Y_i|X)) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \sum_{j=1}^p \beta_j X_{ij} \tag{1}$$

where η is linear predictor, $g(\cdot)$ is the link function which is defined as logit function for Bernoulli distribution, μ_i is expected value of random variable Y_i and β is regression parameter. Therefore Bernoulli mixture regression model can be defined as

$$f(\mathbf{Y}|L, \boldsymbol{\pi}, \mathbf{X}, \boldsymbol{\beta}) = \sum_{\ell=1}^L \pi_\ell p_\ell(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_\ell) \tag{2}$$

where L is the number of mixture components, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ is the mixture proportion which has property $\sum_{\ell=1}^L \pi_\ell = 1$ and $p_\ell(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_\ell) \sim Be(\text{logit}_\ell^{-1}(\boldsymbol{\mu}))$, i.e., $p_\ell(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_\ell)$ has a Bernoulli distributed with parameter $\text{logit}_\ell^{-1}(\boldsymbol{\mu})$ with $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)'$, $\mathbf{X}' = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ and $\boldsymbol{\beta}_\ell = (\beta_1, \beta_2, \dots, \beta_p)'$. Let $\Theta =$

$(\theta_1, \theta_2, \dots, \theta_d)' = (\beta_1, \dots, \beta_L, \pi)'$ denote all unknown parameters appearing in the Bernoulli mixture regression model. The posterior probability distribution $f(\Theta|Y, L, X)$ can be represented as $f(\Theta|Y, L, X) \propto f(Y|\Theta, L, X)p(\Theta)$ where $p(\Theta)$ is the prior distribution of Θ and $f(Y|\Theta, L, X)$ is the mixture likelihood of Eq. (2).

2.2 Markov Chain Monte Carlo Convergence Diagnostics

As referred in [4], in Bayesian inference perspective, if Markov chain is convergent imply that the chain reaches the true posterior distribution. Thus the convergence of estimated parameters should be checked in order to get the true posterior inference for parameters.

Based on Gelman and Rubin [5], if there are m Markov Chains which are mutually independent and it has been taken a number of T iterations, $t = 1, 2, \dots, T$, MCMC convergence can be monitored through estimation of potential scale reduction factor (PSRF). If the PSRF value is close to 1, then every m Markov Chains converge to the true posterior distribution.

As stated in [6], the diagnostic test of Geweke compute indicator Z that has means of subsamples A, $\bar{\theta}^A$, and means of subsamples B, $\bar{\theta}^B$ as the beginning and the end of samples respectively. Considering $\sigma_{(\bar{\theta}^B - \bar{\theta}^A)}$ is an estimated standard deviation of difference $\bar{\theta}^B - \bar{\theta}^A$ and Z asymptotically follows the standardized normal distribution, $Z \sim N(0, 1)$, so if $|Z| > 2$ then the chain is not convergent.

Raftery and Lewis [7] defined N_{\min} as the minimum number of iterations that would be needed to achieve the required estimation precision for some function of parameter. If the value of dependence factors, $I = N/N_{\min}$, is greater than 5, then it implies convergence failure of Markov chain.

Heidelberger and Welch [8] proposed the method consists of two tests for assessing convergence. Firstly, a stationary test which verifies whether the Markov chain occurs from a stationary stochastic process. Secondly, the half-width test which determines if there is sufficient sample size for a chain to estimate the mean values of the process with appropriate accuracy. Markov chain has not convergent when it fails to meet those two tests.

2.3 Data and Model

The Data in this research are Bidikmisi 2015 data of all districts in East Java Province Indonesia which are composed of 33,603 Bidikmisi registrants from 35 regencies. Research variables used in this study consisted of the response variable (Y) and the predictor variable (X) which is constructed by dummy variables as follows

Y : the acceptance status of Bidikmisi scholarship (1 = accepted, 0 = not accepted)

X_1 : father's job is formed by dummy variables d_{11} , d_{12} , d_{13} , and d_{14} .

d_{11} : farmer, fisherman or others job which relate with agriculture.

d_{12} : civil servants, police, and army.

d_{13} : entrepreneur.

d_{14} : private employees

X_2 : mother's Job is formed by dummy variables d_{21} , d_{22} , d_{23} , and d_{24} .

d_{21} : farmer, fisher and others job which relate with agriculture.

d_{22} : civil servants, police, and army.

d_{23} : entrepreneur.

d_{24} : private employees

X_3 : father's education is formed by dummy variables d_{31} , d_{32} , and d_{33} .

d_{31} : not continue to school.

d_{32} : elementary, junior high, or senior high school graduate level.

d_{33} : higher education level

X_4 : mother's education is formed by dummy variables d_{41} , d_{42} , and d_{43} .

d_{41} : not continue to school.

d_{42} : elementary, junior high, or senior high school graduate level.

d_{43} : higher education level

All of dummy variables defined above are valued by 1, and otherwise are 0.

The Bernoulli mixture regression model which has to be estimated is defined by

$$f(\mathbf{y}|\boldsymbol{\pi}, \mathbf{x}, \boldsymbol{\beta}) = \pi_1 Be\left(\frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})}}\right) + \pi_2 Be\left(\frac{e^{g_2(\mathbf{x})}}{1 + e^{g_2(\mathbf{x})}}\right) \quad (3)$$

with π_1 and π_2 are mixture proportions which have properties $0 \leq \pi_1 \leq 1$, $0 \leq \pi_2 \leq 1$ and $\pi_1 + \pi_2 = 1$. $f(\mathbf{y}|\boldsymbol{\pi}, \mathbf{x}, \boldsymbol{\beta})$ represents two mixture components namely a component of wrong acceptance condition and a component of right acceptance condition. While $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ are formed by

$$g_1(\mathbf{x}) = \beta_{(1)0} + \beta_{(1)11}d_{11} + \beta_{(1)12}d_{12} + \beta_{(1)13}d_{13} + \beta_{(1)14}d_{14} + \beta_{(1)21}d_{21} + \beta_{(1)23}d_{23} + \beta_{(1)24}d_{24} + \beta_{(1)31}d_{31} + \beta_{(1)32}d_{32} + \beta_{(1)33}d_{33} + \beta_{(1)41}d_{41} + \beta_{(1)42}d_{42} + \beta_{(1)43}d_{43}$$

and

$$g_2(\mathbf{x}) = \beta_{(2)0} + \beta_{(2)11}d_{11} + \beta_{(2)12}d_{12} + \beta_{(2)13}d_{13} + \beta_{(2)14}d_{14} + \beta_{(2)21}d_{21} + \beta_{(2)23}d_{23} + \beta_{(2)24}d_{24} + \beta_{(2)31}d_{31} + \beta_{(2)32}d_{32} + \beta_{(2)33}d_{33} + \beta_{(2)41}d_{41} + \beta_{(2)42}d_{42} + \beta_{(2)43}d_{43}.$$

In that model, there are two parameters $\pi_\ell, \beta_{(\ell)kj}$ which have to be estimated. The prior distributions which are implemented for the each of the parameters are $p(\pi_\ell) \sim \text{dirichlet}(\vartheta)$ and $p(\beta_{(\ell)kj}) \sim N(\mu, \sigma)$. Coefficient of $\beta_{(\ell)kj}$ indicates the number of units (as coded) of change in $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ between the category for which dummy variable $d_{kj} = 0$ and the category for which dummy variable $d_{kj} = 1$.

3 Results

Computation of estimated parameters was performed on OpenBUGS [9]. Whereas diagnostic processes of MCMC convergence were done through R software with convergence diagnosis and output analysis (CODA) package [10]. In the first process, we generated 10,000 iterations which produced $\pi_1 = 0.604$ and $\pi_2 = 0.396$ as a significant estimated mixture proportions. CODA diagnostic for some parameters which have convergent problems is described on Table 1.

Referring to Table 1, Gelman-Rubin method shows that MCMC is convergent for all estimated parameter. The Geweke method indicates MCMC for the estimated parameter $\hat{\beta}_{41}$ on $g_1(x)$ is not convergent. Whereas Raftery-Lewis method gives unconverted MCMC on estimated parameters $\hat{\beta}_0, \hat{\beta}_{11}$ and $\hat{\beta}_{21}$ in $g_1(x)$ and $g_2(x)$. Based on Heidelberger-Welch method [8], MCMC for estimated parameter $\hat{\beta}_{24}$ is failed to converge. It means that by discarding of 10% increment until 50% of the iterations, the stationary tests are still failed. Therefore, when the simulation is run 10,000 iterations, there is only Gelman-Rubin which has a convergent indicator for all parameters. If Markov chain does not converge to the posterior distribution of parameters, then valid inferences of parameters on the Bernoulli mixture regression model cannot be accomplished. Based on [9], we conducted further simulations, i.e., 100,000 iterations in order to know the effect of increased number iterations on MCMC convergence. The significant estimated mixture proportions are $\pi_1 = 0.6041$ and $\pi_2 = 0.3959$. The result of CODA diagnostic is presented on Table 2.

In regard to Table 2, it can be seen that the Gelman-Rubin method and the Heidelberger-Welch method present an improvement of indicator values. Those

Table 1 Indicator values of CODA diagnostics with 10,000 iterations

Param	Sig. est. value		Gelman-Rubin		Geweke		Raftery-Lewis		Heidel-Welch	
	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$
$\hat{\beta}_0$	1.201	-1.79	1.01	1.03	0.223	-0.11	18.9	23.7	passed	passed
$\hat{\beta}_{11}$	-1.288	0.87	1.00	1.01	0.306	0.15	9.03	12.8	passed	passed
$\hat{\beta}_{21}$	-2	1.14	1.02	1.01	-0.77	0.115	20.7	20.3	passed	passed
$\hat{\beta}_{24}$	-1.3	0.07	1.01	1.00	-0.69	-0.71	4.54	3.63	failed	failed
$\hat{\beta}_{41}$	-0.257	-0.07	1.00	1.00	2.084	0.185	3.71	4.01	passed	passed

Table 2 Indicator values of CODA diagnostics with 100,000 iterations

Param	Sig. est. value			Gelman-Rubin			Geweke			
	Raftery-Lewis						Heidel-Welch			
$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	$g_1(x)$	$g_2(x)$	
$\hat{\beta}_0$	1.19	-1.72	1.00	1.00	1.594	-1.77	53	25.2	passed	d passed
$\hat{\beta}_{11}$	-1.287	0.87	1.00	1.00	-0.50	1.87	13.8	13	passed	passed
$\hat{\beta}_{21}$	-1.99	1.13	1.00	1.00	-1.39	1.28	38.7	19.2	passed	passed
$\hat{\beta}_{24}$	-1.29	0.067	1.00	1.00	-2.09	1.09	4.91	3.92	passed	passed
$\hat{\beta}_{41}$	-0.256	-0.069	1.00	1.00	-0.74	-1.00	3.74	3.68	passed	passed

results can confirm significant outcome of estimated parameters. Four predictor variables, i.e., father’s job, mother’s job, father’s education and mother’s education significantly influence $g_1(x)$ and $g_2(x)$ which can affect on the parameter of Bernoulli distribution according to the Eq. (3). Nevertheless, the Geweke method shows inconsistent diagnostic results, whereas, the Raftery-Lewis method has not significant change on convergent indicator results and still gets unconvergent MCMC indicator values on $\hat{\beta}_0$, $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$. These results affirm that application of MCMC diagnostic tests to Bayesian Bernoulli mixture regression model can assure classification of two acceptance conditions in Bidikmisi, i.e., wrong acceptance condition and right acceptance condition.

4 Conclusion

The MCMC diagnostic methods give different results to conclude MCMC convergence. On the Gelman-Rubin method and the Heidelberger-Welch method, the increasing number of iterations improve convergence indicators for estimated parameters of Bernoulli mixture regression model.

Acknowledgements The Authors are grateful to DRPM-Kemenristekdikti Indonesia which supported this research under PUPT research grant no. 608/PKS/ITS/2017.

References

1. Iriawan, N.: *Pemodelan dan Analisis Data-Driven*. ITS Press, Surabaya (2012)
2. Nadif, M., Govaert, G.: Clustering for binary data and mixture models-choice of the model. *Appl. Stoch. Mod. Bus. Ind.* **13**, 269–278 (1997). [https://doi.org/10.1002/\(SICI\)1099-0747\(199709/12\)13:3/4%3c269:AID-ASM321%3e3.0.CO;2-7](https://doi.org/10.1002/(SICI)1099-0747(199709/12)13:3/4%3c269:AID-ASM321%3e3.0.CO;2-7)

3. Grun, B., Leisch, F.: Finite mixtures of generalized linear regression models. In: Shalabh, C. H. (ed.) *Recent Advances in Linear Models and Related Areas*, pp. 205–230. Springer, Heidelberg (2008)
4. Alkan, N.: Assessing convergence diagnostic tests for bayesian cox regression. *Comm. Stat. Sim. Comp.* **46**(4), 3201–3212 (2017). <https://doi.org/10.1080/03610918.2015.1080835>
5. Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**(4), 457–511 (1992). <https://doi.org/10.1214/ss/1177011136>
6. Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Smith, A.F.M., Dawid, A.P., Berger, J.O. (eds.) *Bayesian Statistics 4*, pp. 169–193. Oxford University Press, New York (1992)
7. Raftery, A.E., Lewis, S.M.: How many iterations in the gibbs sampler? In: Bernardo, J.M., Smith, A.F.M., Dawid, A.P., Berger, J.O. (eds.) *Bayesian Statistics 4*, pp. 762–773. Oxford University Press, New York (1992)
8. Heidelberger, P., Welch, P.D.: Simulation run length control in the presence of an initial transient. *Oper. Res.* **31**(6), 1109–1144 (1983). <https://doi.org/10.1287/opre.31.6.1109>
9. Lunn, D., Spiegelhalter, D., Thomas, A., Best, N.: The BUGS project: evolution, critique and future directions (with discussion). *Stat. Med.* **28**(25), 3049–3082 (2009). <https://doi.org/10.1002/sim.3680>
10. Plummer, M., Best, N., Cowles, K., Vines, K.: CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**(1), 7–11 (2006)

Chapter 50

Road Fatalities Using Logistic Regression



Isnewati Ab Malek, Nurul Najihah Mohd Salim, Siti Naffsikhah Alias, Nurul Akilah Mohd Zaki and Haslinda Ab Malek

Abstract Safety and accident issues are considered as important problems in the world. Road accident issues would have a more conspicuous countenance in Malaysia. Since 4% of fatal accidents increase annually, prevention of this accidents issues needs to be controlled. This study focused only on the Seremban area to identify the characteristics of road fatalities and to investigate the probable factors that contribute to the fatal accident. It was conducted on 381 road users where 224 were fatally injured in road accidents in Seremban during 2015. The risk of involvement in fatal rather than nonfatal accidents for gender was higher among males than among females. The middle age group (31–40 years old) is the most vulnerable to fatal accidents. In accidents that occurred during the night, the risk of death was higher than during the day. The risk of death for vehicle types was higher in other vehicles (bus, pedestrians, bulldozer and other uncategorized vehicles) compared to motorcycle, car and lorry/truck. Major accident causes in road user fatalities were over speed. The risk for fatal injury in a road traffic accident was estimated using logistic regression adjusting for gender, age, time of day of accident, vehicle types and accident cases. Among the variables obtained, two independent variables were found significantly associated with fatal accidents; namely vehicle types (lorry/truck and others) and accident cases (over speed and sudden change signal). The findings show meaningful interpretations that can be used for future safety improvement in Seremban.

I. A. Malek (✉) · N. N. M. Salim · S. N. Alias · N. A. M. Zaki · H. A. Malek
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Cawangan Negeri Sembilan, Kampus Seremban, Shah Alam, Malaysia
e-mail: isnewati@ns.uitm.edu.my

N. N. M. Salim
e-mail: annajihah93@gmail.com

S. N. Alias
e-mail: naffsikhah@gmail.com

N. A. M. Zaki
e-mail: nurulakilah5@gmail.com

H. A. Malek
e-mail: haslinda8311@ns.uitm.edu.my

1 Introduction

Transportation is essential to the success of both the economy and life quality in rural and urban areas Oyedepo [1]. Therefore, numerous vehicles are being introduced in this globalization era. This has resulted in the number of vehicles on the road to increase and this can lead to certain phenomena's such as road accidents. Accident is an unwanted event that occurs unexpectedly, resulting in damage or harm. Car collisions are one of the most serious collisions that commonly result in severe injuries or fatalities Dabbour [2].

Statistically, fatal accidents increase by 3.1% and the injury rate rose by 6.7% from 2011 to 2012, National Highway Traffic Safety Administration [3]. Malaysia also faces almost the same case right now. On the other hand, The Royal Malaysian Police [4] has issued a statement that a total of 303 accident cases arises in a total of 330 fatalities. Taking into consideration the alarming rise in road accidents observed around the world, firm action should be taken.

Accident severity is of special concern to researchers in traffic safety since this research is aimed not only at prevention of accidents but also at reduction of severity. One way to accomplish the latter is to identify the factors that influence road traffic accidents in Seremban. Therefore, the objective of conducting this research is to describe the characteristics of road fatalities and to investigate the probable factors that contributes to the fatal accidents in Seremban.

Road fatalities can be caused by many factors. A study by Wedagama [5] investigated the impact of the accident related factors on road fatalities using logistic regression. They included seven predictor variables in their model development and the outcome of the study shows that the most significant variables are age and gender.

According Yan [6], the accident time is measured by day and night. From the author's studies, the relative accident involvement ratio for night is clearly lower than daytime. Besides, there is a significant difference between accident time and age groups. For the youngest and oldest groups, rear-end accidents are more likely to happen during daytime than night; for middle age groups, the difference is not very apparent.

The types of vehicle that are always involved in road accident are motorcycle, car, van, and heavy vehicles such as buses and trucks. Accidents involving cars is significant. According to Yan [6] accidents involving cars was 69.38%, the most heavily involved in an accident compared to other vehicles.

There are many factors that cause fatal accidents. One of the factors is caused by the bad behavior of the road users which is over speeding. This fact has been proved by many researchers and shows a significant lead to the road accidents. In a recent study by Haadi [7], it was indicated that approximately 30% of road accidents are caused by speeding. Shankar et al. [8] also believed that the effect of driving over the speed limit impacts the severity of the accident.

The other factor that causes of an accident is because of alcohol and drugs. A person, who takes drug or alcohol, also does not have a healthy mind and can

commit errors while driving. Alcohol and drugs can affect the driving skills, it causes general impairment of the brain and the function of the nervous system [7].

A study on this fatality due to road accidents would provide clear motives on what factors or elements are needed to be highlighted to find solutions on the rise in road accidents. This study can be as evidence of what kind of conditions contributed to accidents. This, in turn, can protect a person’s life. Therefore, authorities and government can benefit from this research and improve any shortcomings on road conditions and create awareness to all drivers.

2 Methodology

2.1 Source of Data

The sources of data for this study were secondary data which are indispensable for most business research. How accidents happen and what factors contribute towards the accident is important to be investigated. Therefore, the way forward to reduce mortality rate among road users could be by identifying the factors and analyzing the data gathered from victims so that the pattern of fatality can be discovered in Seremban. Hence, this study obtained data from the Police District Headquarters of Seremban (IPD Seremban). This study was conducted on 381 road users in Seremban during 2015.

2.2 The Characteristics of Road Accidents in Seremban

Based on the first objective which describes the respondent profile of road accidents, descriptive analysis is suitable for describing the characteristics. The descriptions of variables in this study are as follows (Table 1).

Table 1 Descriptions of variables used in the study

No	Variables	Descriptions	
1	Status of road accidents	0 = Non-fatal	1 = Fatal
2	Gender	1 = Male	2 = Female
3	Age	1 = below and 20 3 = 31–40 5 = 51–60	2 = 21–30 4 = 41–50 6 = above 60
4	Time of day	1 = Day	2 = Night
5	Vehicle type	1 = Motorcycle 3 = Lorry/Truck	2 = Car 4 = Others
6	Accident causes	1 = Over speed 3 = Sudden change signal 5 = Using phone	2 = Drunk 4 = Sleepy

2.3 Factors that Contributes to the Road Fatalities in Seremban

This study used logistic regression to determine the factors that contribute to the fatal accidents in Seremban. Logistic regression is a regression model suitable for modelling binary outcomes taking on value 0 and 1. The target binary outcomes for severity of road accident fatality (Y) takes on two values; 1 = fatal; 0 = non-fatal.

The underlying principle of binomial logistic regression, and its statistical calculation, was quite different to ordinary linear regression. Ordinary regression uses ordinary least squares to find a best fitting line and comes up with coefficients that predict the change in the dependent variable for one-unit change in the independent variable, however, logistic regression estimates the probability of an event occurring. The dependent variable is the population proportion or probability (P) that resulting outcome is equal to 1. The odd ratios for each of the independent variables can be estimated by using parameters obtained for the independent variables.

The specific form of the logistic regression model is:

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i}} \tag{1}$$

Setting up the log odds as the target variable for the regression, the regression equation looks like:

$$\text{Logit}(P) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i \tag{2}$$

where:

- β_0 the model constant.
- β_1 the parameter estimates for the independent variables.
- $\beta_i X_i$ set of independent variables (I = 1, 2, ..., n).
- P probability ranges from 0 to 1.
- $\ln\left(\frac{p}{1-p}\right)$ the natural logarithms range from negative infinity to positive infinity.

Odds is the nearest to the natural. Probability and odds are the same thing, but the probabilities are restricted to the range 0–1, while odds go from 0 to infinity (Table 2).

Table 2 The association of odds ratio

Odds ratio	Association
Odds ratio = 1	Outcome of odds does not affect by exposure
Odds ratio > 1	Outcome of higher odds associated with exposure
Odds ratio < 1	Outcome of lower odds associated with exposure

3 Results and Analysis

The analysis revealed the characteristics of road fatalities in Seremban by using descriptive analysis. Data that has been received and reviewed, as well as taken into consideration for the deaths in road accidents are gender, age, time of day, vehicle types and accident cause.

In summarizing, it is concluded that the number of fatal accident rates is higher than non-fatal accidents. In terms of gender, percentage of men involved in fatal accidents are higher than women. Furthermore, those aged between 31 and 40 years, which were studied reached the highest record of age. In addition, in this study showed that the time of day is more dangerous at night than during the day. This may also be because of the weather and the darkness of night. For the last variable, the types of vehicle which are most at risk are other vehicles. The high risk of fatal accidents is 'others' followed by motorcycle, car and lorry/truck.

Logistic regression was performed to investigate the probable factors that contribute to the fatal accidents in Seremban. The full model containing all predictors was statistically significant since the p-value = 0.00 which is less than 0.05, χ^2 (10, 54.712) as shown in Table 3.

From Table 4, the lorry/truck showed a positive B-logit value (2.099). This indicates that the higher number of lorry/truck drivers, the more likely they involved with fatal accidents by 716.1% compared to motorcycles. The variable measuring others showed a positive B-logit value (1.187). This indicates that the greater number of others (bus/pedestrian), the more likely they are to be involved with fatal accidents by 227.7% compared to motorcycles.

In addition, over speed showed a positive B-logit value (0.692). This indicates that over speed is likely to cause fatal accidents by 99.8% compared to by using phone. The sudden change signal showed a negative B-logit value (-0.734), indicates that sudden change signal is less likely to cause fatal accidents by 52.0% compared to by using phone.

The lorry/truck is a significant predictor, according to the p-value ($p = 0.011$), in the Table 4. The odds ratio for this variable is 8.161, which is more than 1. This indicates that the lorry/truck drivers are more likely to cause fatal accidents than motorcycle. The others are also a significant predictor, according to the p-value ($p = 0.000$). The odds ratio for this variable is 3.277, which indicates that the 'others' are more likely to cause fatal accidents than motorcycle.

Moreover, over speed and sudden change signal are significant predictors with value 0.014 and 0.013 respectively. The odds ratio for over speed is 1.998, a value more than 1. This indicates that over speed is more likely to cause fatal accidents

Table 3 Omnibus test of model coefficient

Chi-square	Degree of freedom	P-value
54.712	10	0.000

Table 4 Summary of variables

	Coefficient	Wald	P-value	Odds ratio
Gender	-0.296	1.504	0.220	0.744
Time of accident	0.246	1.099	0.295	1.279
Age	-0.066	0.551	0.458	0.936
Vehicle type		20.578	0.000	
Car	0.605	2.228	0.136	1.832
Lorry/truck	2.099	6.437	0.011	8.161
Others	1.187	17.524	0.000	3.277
Causes		26.036	0.000	
Over speed	0.692	6.090	0.014	1.998
Drunk	-1.249	1.986	0.159	0.287
Sudden change signal	-0.734	6.143	0.013	0.480
Sleepy	-0.331	0.350	0.554	0.718

than using phone. However, the odd ratio for sudden change signal is 0.480, which is less than 1. This indicates that sudden change signal is less likely to cause fatal accidents than using phone while driving.

4 Conclusions

As shown in Table 4, only two of the independent variables made a unique statistically significant contribution to the model, type of vehicles (lorry/truck and others) and causes (over speed and suddenly changes signal). The strongest predictor of reporting fatal accidents is vehicles type which is a lorry/truck, an odds ratio of 8.161. This indicates that road user who were involved in fatal accidents were more likely than those who did not involve in fatal accidents, controlling all other factors in the model.

For further research, it is suggested to increase the sample size due to logistic regression acquiring large sample size. Factors that cause road fatalities have been outlined in this paper. From the results, this study can be used to develop strategies to prevent and reduce fatal accidents in Seremban. The finding of this study also can be considered as guideline for a future study.

References

1. Oyedepo, O.J., Makinde, O.O.: Accident prediction models for Akure—Ondo carriageway, Ondo state southwest Nigeria; using multiple linear regressions. *Afr. Res. Rev.* **4**(2), 30–49 (2010)
2. Dabbour, E.: Using logistic regression to identify risk factors causing rollover collisions. *Int. J. Traffic Transp. Eng.* **2**(4), 372–379 (2012). [https://doi.org/10.7708/ijte.2012.2\(4\).07](https://doi.org/10.7708/ijte.2012.2(4).07)

3. National Highway Traffic Safety Administration: Traffic Safety Facts. Motor Vehicle Crash Data from FARS and GES (2012)
4. Royal Malaysian Police: <http://www.rmp.gov.my/> (2015)
5. Wedagama, D.M.P., Dissanayake, D.: The influence of accident related factors on road fatalities considering Bali province in indonesia as a case study. *J. East. Asia Soc. Transp. Stud.* **8** (2009)
6. Yan, X., Radwan, E., Abdel-Aty, M.: Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accid. Anal. Prev.* **37**(6), 983–995 (2005)
7. Haadi, A.-R.: Identification of factors that cause severity of road accidents in ghana: a case study of the northern region. *Int. J. Appl. Sci. Technol.* **4**(3) (2014)
8. Shankar, V., Mannering, F., Barfield, W.: Statistical analysis of accident severity on rural freeways. *Accid. Anal. Prev.* **28**(3), 391–401 (1996)

Chapter 51

Robust Clustering on Spatial Torrential Rainfall Patterns



Shazlyn Milleana Shaharudin and Norhaiza Ahmad

Abstract Peninsular Malaysia has a tropical climate that is characterized by three monsoons. The aim of this study is to identify the spatial distribution patterns of the daily torrential monsoon rainfall in Peninsular Malaysia by clustering the most relevant principal directions between its day correlations. In such climate, the daily rainfall variability between monsoons typically differ and its daily rainfall patterns are influenced by the different wet days. Thus, the clustering results on such data would tend to generate too few clusters that are imbalanced when the different rainfall days are given equal weights in highlighting the spatial rainfall patterns. In this study, we use a robust correlation measure by applying a clustering method on the principal component loadings of the daily torrential rainfall based on a weighted correlation matrix of a Tukey's biweight M-estimate. The findings indicate ten distinct rainfall patterns that appear to display the dominant role extended by the complex topography and exchange monsoons of the peninsular.

Keywords Tukey's biweight correlation · Robust correlation · k-means cluster analysis · Principal component analysis · Pearson correlation

1 Introduction

Classification studies in large-scale dimensions of weather data are required to characterise synoptic and dynamic climatology patterns. Clustering techniques preceded with principal component analysis (PCA) are often combined to identify

S. M. Shaharudin (✉)

Faculty of Science and Mathematics, Department of Mathematics,
Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
e-mail: shazlyn@fsmt.upsi.edu.my

N. Ahmad

Faculty of Science, Department of Mathematics, Universiti Teknologi Malaysia,
81310 UiTM, Johor Bahru, Johor, Malaysia
e-mail: norhaiza@utm.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_51

key spatial patterns in the data by reducing the number of variables for clustering cases. A typical procedure is to first employ a PCA based T-mode correlation to reduce the dimension of the data set and k-means clustering technique to generate rainfall patterns for a particular region. These two step approach has proven to provide a clear distinction within the data, highlighting those days which presented more similarities in their spatial rainfall patterns [1].

The use of configuration points of entities between rows and columns of data in accordance with Pearson correlation matrix is a common classification method used for identifying rainfall patterns. Here, Pearson correlation matrix is more commonly used in the derivation of T-mode correlation to measure similarity between the daily rainfall [2]. Pearson correlation between two variables is defined as covariance of the two variables divided by the product of their standard deviations that perform well on normally distributed independent data. Here, each point is equally weighted.

In analysing characteristics of rainfall pattern in Malaysia, it is important to realize that the daily rainfall variability between monsoons and regions differ, the daily rainfall distribution is far from normally distributed and that daily rainfall pattern is influenced by the different wet days. Thus, applying Pearson correlation on such data might not be suitable as different days carry unequal weights and it will affect the clustering result in highlighting spatial rainfall patterns.

Tukey's biweight correlation matrix is introduced for the analysis of spatial distribution torrential rainfall patterns, as an alternative of the Pearson correlation matrix in the T mode. We have used a Tukey's biweight correlation matrix based on an M-estimate approach which is more flexible and performs well under a variety of data distributions [3].

2 Data

In order to illustrate our new correlation matrix, we employed PCA based Tukey's biweight correlation to daily rainfall data from 75 rain gauge stations over Peninsular Malaysia. We obtained the data from Jabatan Pengairan dan Saliran (JPS) for the period 1975–2007. In this study, we focus on the occurrence of episodes on extreme rainfall event described as torrential rainfall. We have selected days that exhibited torrential characterized based on criteria described in order to retain only those days that exhibited torrential character. It was therefore necessary to choose some criteria that would lead to the establishment of a threshold, in order to allow for a clear distinction between what constitutes a day of torrential rainfall in the Peninsular Malaysia region and what does not. Area with a tropical climate with 60 mm/day is the most common threshold applied for this purpose. By filtering days with rainfall more than 60 mm for at least 2% of overall stations, we managed to obtain 250 days and 15 rainfall stations which in turn are suffice enough to represent the main torrential centers.

3 Methods

3.1 Principal Component Analysis

PCA is concerned with explaining the variance covariance structure of a set of variables through a reduction of the large dimension matrix to some smaller data set. Shaharudin and Ahmad [4] recommended using more than 70% of cumulative percentage of total variation as a method to extract the variation of modes in large data set. The most satisfactory result is found by using unrotated PCs due to rotated PCs will cause much more complex results [1]. The results of PCA are usually discussed in terms of component scores and loadings. In the study of identifying patterns, usually researchers were using loadings to carry out further analysis.

In climate data, the approach that used to derive the typical torrential rainfall patterns consists in subjecting the T mode analysis to principal component analysis (PCA). T mode is applied in order to analyse spatial fields in different times and useful to extracting and reproducing the circulation types, quantifying their frequency and showing the dominant weather periods in them [5].

3.2 Pearson Correlation

In identifying spatial rainfall patterns, Pearson correlation is typically used in PCA for calculating its eigenvectors and eigenvalues. Pearson correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. Typically, Pearson correlation is used to measure the distance (or similarities) before implementing a clustering algorithms. The Pearson correlation coefficient between two vectors of observation is as follows:

$$r_{ij} = \frac{\sum_{i=1}^n (X_i \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_i)^2 \sum_{j=1}^n (X_j - \bar{X}_j)^2}} \quad (1)$$

where X_i and X_j refer to the vectors of rainfall day in matrix data X with n rainfall day, with \bar{X}_i and \bar{X}_j refer to the mean of the vectors for $i, j = 1, 2, \dots, n$.

3.3 Tukey's Biweight Correlation

Tukey's biweight has been well established as a resistant measure of location and scale for multivariate data. Tukey's biweight function is one of the family member in M-estimators used in robust correlation estimates. M-estimator has a function of

determines the weights assigned to the observations in the data set which call it as derivative function, ψ . The ability of M-estimator is to down weight point that are far from data center and because of that, M-estimator are more resistant to outlying values than Pearson correlation. The derivative function is derived as follows:

$$\psi(u) = \begin{cases} u(1 - u)^2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \tag{2}$$

It can be seen that if $|u|$ is large enough, then $\psi(u)$ diminish to zero. Another aspect that can be seen on M-estimators is its breakdown point. Intuitively, the breakdown point is the proportion of incorrect observations that can cause an incorrect result [6]. In this study, several trials were conducted in Tukey’s biweight with the breakdown set to 0.0, 0.2, 0.4 and 0.5. The biweight with a breakdown of 0.4 is much more accurate and efficient in most situations compared to other investigated breakdown points.

The first procedure when using biweight estimate of correlation is calculating the location estimate, \tilde{T} and followed by updating the shape estimate, \tilde{S} . \tilde{S}_{ij} is a covariance between two vectors, \tilde{X}'_i and \tilde{X}'_j , where acting as a resistance estimate in the (i,j)th element of \tilde{S} . The formula of biweight correlation of these two vectors is written as follows:

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \tag{3}$$

with

$$\mathbf{T}_n^{(k+1)} = \frac{\sum_{i=1}^n \mathbf{X}_i w(\mathbf{u}_i^{(k)})}{\sum_{i=1}^n w(\mathbf{u}_i^{(k)})} \quad k = 0, 1, 2, \dots \tag{4}$$

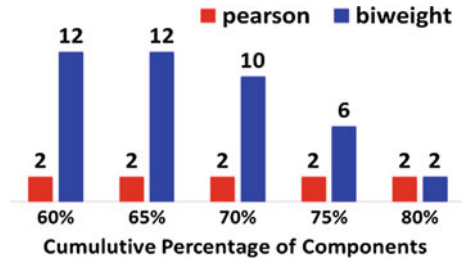
$$\mathbf{S}_n^{(k+1)} = \frac{\sum_{i=1}^n w(\mathbf{u}_i^{(k)}) (\mathbf{X}_i - \mathbf{T}^{(k+1)}) (\mathbf{X}_i - \mathbf{T}^{(k+1)})^t}{\sum_{i=1}^n w(\mathbf{u}_i^{(k)}) (\mathbf{u}_i^{(k)})} \tag{5}$$

where $\mathbf{T}_n^{(k+1)}$ is a location vector and $\mathbf{S}_n^{(k+1)}$ is a shape matrix for $k = 0, 1, 2, \dots$

4 Results and Discussion

Figure 1 shows the number of clusters obtained using two different approaches in PCA based correlation. PCA based Pearson correlation, produces only two clusters regardless of the cumulative percentage of variation used. PCA based Tukey’s biweight correlation shows differentiating patterns on the number of clusters produced at different cumulative percentage of variation used. In hydrological studies particularly in identifying rainfall patterns, it is more reasonable to obtain more than

Fig. 1 Bar chart shows the number of clusters obtained using two different approaches in PCA based correlation



two clusters to explain the various types of rainfall patterns. In identifying spatial rainfall pattern, the clustering output at 70% cumulative percentage of variation on PCA-based Tukey’s biweight correlation (10 clusters) is chosen respectively.

The main features of the clustering result are discussed to verify the distinction between the clusters with respect to their significant locations and period of monsoon occurrence for the torrential rainfall patterns based on the recommended settings in the previous methodology section. In defining the spatial characteristics of torrential rainfall pattern in Peninsular Malaysia, ten clusters are obtained.

In each torrential rainfall patterns, there are distinct locations that are especially affected by torrential rainfall, and this summary can also be drawn from the illustrated of Table 1. It is quite evident that, in general overview, Terengganu is the location most affected by torrential rainfall events. Six of the ten rainfall patterns indicate that the torrential rainfall patterns during the northeast monsoon experiencing the heaviest rain in the eastern region of the Peninsula. As shown in Table 1, rainfall pattern 7 (RP) where located in Kg. Jabi (Terengganu) is interesting to note that this rainfall pattern is the most frequent ones.

RP 1 and RP 2 exhibit moderate rainfall in the whole region in Peninsular Malaysia, with a general increase for most of the highland uplands, such as Bukit Bendera (Pulau Pinang). The maximum torrential rainfall for RP 1 occurs in Pintu Air Bagan Air Itam (Pulau Pinang) and RP 2 in Bukit Bendera (Pulau Pinang), but

Table 1 Summary of the ten rainfall pattern groups obtained for daily torrential rainfall

Rainfall pattern (RP)	Region	Location	Days included
RP 1	Northern	Pintu Air Bagan, Air Itam	17
RP 2		Bukit Bendera	17
RP 3	Eastern	Kota Bahru	18
RP 4		Dungun	19
RP 5		Kemasek	27
RP 6		Kemaman	29
RP 7		Kampung Jabi	41
RP 8		Kampung Menerong	32
RP 9		Endau	28
RP 10		Kuantan	22

the main feature of this particular pattern is the wide distribution of the torrential rainfall areas, which comprise the entire region.

RP 3 represents heavy rainfall over the east region with maximum torrential rainfall in Kota Bahru (Kelantan). In this pattern, the intensity of the torrential rainfall decreases from east to the southwest region. Torrential rainfall occurs mostly in Kelantan due to strong influenced by the northeast monsoon and occurrence of sea breeze. Furthermore, Kota Bahru (Kelantan) is located near the coast. Therefore, this exposes the region to more rainfall during that period.

RP 4, RP 5, RP 6, RP 7 and RP 8 are characterized by heavy torrential and intense rainfall occurring in the eastern region. All the patterns are located in Terengganu at different areas where RP 4 received higher rainfall in Dungun, RP 5 received heavy rainfall in Kemasek, RP 6 received maximum torrential rainfall in Kemaman, RP 7 received maximum torrential rainfall in Kampung Jabi and heavy torrential rainfall occurred in Kampung Menerong at RP 8. Distribution of rainfall patterns for each group is significantly different due to different altitudes where the rainfalls are observed. Northwestern region received less rainfall caused by blocked by the Titiwangsa Range which possibly affects most of the rainfall stations along the western part of Peninsular Malaysia. Meanwhile, the eastern part in Peninsular Malaysia is considered as wettest area due to the strong influenced by northeast monsoon that bring heavy rainfall to east region in the period of November until March.

RP 9 represents substantial rainfalls with maximums in Endau, Mersing where the topography is defined as lowland area. Naturally, water of the rainfall will flow from high to low area. Hence, when the northeast monsoon brings heavy rainfall to that area, this make the location receives heavy rainfall. Furthermore, without ranges or mountains, the region is more likely to encounter rainfall. Due to the location concentrated close to the coast, the occurrence of sea breeze is also one of the major factors that cause this region to receive maximum rainfall.

RP 10 shows that torrential rainfall of Kuantan (Pahang) is concentrated to urban area where it is characterized by higher population density and vast human features in comparison to areas surrounding it. Normally, plants especially largest trees in urban area are difficult to find as urban areas undergo continuous built up of urban development that is within a labor market. Therefore, when northeast monsoon bring heavy rainfall in that region, it will directly receive heavy rainfall without any restriction from the largest trees.

5 Conclusion

In this paper, PCA based Tukey's biweight correlation has been proposed to identify daily spatial torrential rainfall patterns. The purpose is to introduce an alternative correlation matrix due to the issues when dealing with rainfall data particularly when the variability of the data is differ, the data is far from normally distributed and the data is influenced by the different wet days. This study shows a

substantial improvement in the cluster partition with PCA based Tukey's biweight correlation than Pearson's to avoid inaccurate imbalanced clusters. Most of the RPs particularly affect the eastern region, exhibiting a strong rainfall maximum mostly in Terengganu. The majority of the location that received heavy rainfall is located in the coast and this is the one of the factor that heavy rainfall occurred in the region.

References

1. Fragoso, M., Gomes, P.T.: Classification of daily abundant rainfall patterns and associated large-scale atmospheric circulation types in southern Portugal. *Int. J. Climatol.* **28**, 537–544 (2008)
2. Wickramagamage P (2010) Seasonality and spatial pattern of rainfall of Sri Lanka: exploratory factor analysis. *Int. J. Climatol.* **30**, 1235–1245
3. Hardin, J., Mitani, A., Hicks, L., VanKoten, B.: A robust measure of correlation between two genes on a microarray. *BMC Bioinform.* **8**, 220 (2007). <https://doi.org/10.1186/1471-2105-8-220>
4. Shaharudin, S.M., Ahmad, N.: Modeling, design and simulation systems. In: CCIS, vol. 752, pp. 216–224. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-6502-6_19
5. Compagnucci, R.H., Araneo, D., Canziani, P.O.: Principal sequence pattern analysis: a new approach to classifying the evolution of atmospheric systems. *Int. J. Climatol.* **21**, 197–217 (2001)
6. Choulakian, V.: Robust q-mode principal component analysis in L1. *Comput. Stat. Data Anal.* **37**, 135–150 (2001)

Chapter 52

Robust Logistic Regression in Application to Divorce Data



Sanizah Ahmad and Rosa Shafiqah Azureen Mohamad Rosni

Abstract In logistic regression, the parameters are commonly estimated by using maximum likelihood estimator (MLE). However, MLE is easily affected when outliers appear in the data. The objective of this study is to compare the performance between the MLE and four robust methods namely the Mallows-type, Conditionally Unbiased Bounded Influence (CUBIF), Bianco and Yohai (BY), and Weighted Bianco and Yohai (WBY) which are available in R by applying to a sample of real dataset on potential divorce which contains outliers. The performance of the parameter estimators is determined by using the chi-square arcsine transformation with the best estimator having the smallest value. The result in this study found that the WBY method proved to be best robust method in estimating the parameters of potential divorce data.

Keywords Outliers · Divorce · Robust · Maximum likelihood · Logistic regression

1 Introduction

Logistic regression is a predictive analysis that is suitable to use when the dependent variable (Y) is categorical. The purpose of logistic regression is to predict the probability of an occurrence ($Y = 1$) and a non occurrence ($Y = 0$) of an event [1]. In practice, a common statistical method used to estimate the parameters of the model is the Maximum Likelihood Estimator (MLE). However, the MLE works poorly when the model assumptions are violated [2]. The reason is because the performance and validity of logistic regression model will be affected by the

S. Ahmad (✉) · R. S. A. M. Rosni
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: sanizah@tmsk.uitm.edu.my

R. S. A. M. Rosni
e-mail: rosashafiqah@yahoo.com

presence of outliers. Outliers can be defined as the value of observations deviating from the expected range [3].

However, to apply the MLE by removing the outliers is not a good decision to proceed. Hence, corresponding remedial is important task before modelling the binary outcomes. Robust estimators are introduced by researchers as a remedial technique that is not easily affected by outliers. In logistic regression, there are several familiar robust techniques available in the literature such as the Mallows type estimator (MALLOWS), the Conditionally Unbiased Bounded Influence Function estimator (CUBIF), the Bianco and Yohai estimator (BY), and Weighted Bianco and Yohai estimator (WBY). In short, this study is intended to study the performance and make comparison between MLE and four robust logistic regression estimators mentioned by applying to a sample of divorce data.

2 Methodology

Binary logistic regression is one of the regression applications when there are two possible outcomes in the response variable. Therefore, the response variable (Y) in this study is represented by taking on the values 1 (for success) and 0 (for failure). Since the data is concentrating on marriage dissolution, the response variable (Y_i) will be coded as 1 for divorce and 0 for not divorce. In this study, we let (Y_i) to be a Binomial random variable since its distribution having two possible outcomes which 1 for success and 0 for failure.

2.1 Maximum Likelihood Estimator

The main purpose of logistic regression is to estimate the parameters, β . In this study, we consider multiple logistic regression since there are more than one independent variables involved. Hence, a multiple logistic regression model is $\beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1}$ where is p the number of parameters. We consider Y as $n \times 1$ vector response with probabilities π for $Y = 1$, and $1 - \pi$ for $Y = 0$. The independent variables (X) and random errors (ε) is treated as $n \times k$ matrix with $k = p + 1$ and $n \times 1$ vector, respectively. Then, a logit transformation for a logistic regression model in terms of $\pi = (X)$ can be stated as follows:

$$g(X) = \log_e \left(\frac{\pi}{1 - \pi} \right) = X\beta \quad (1)$$

The method of Maximum Likelihood (ML) estimates the parameters by working with the logarithm of the joint probability function which can be stated as follows:

$$\begin{aligned} \log_e g(Y_1, \dots, Y_n) &= \log_e \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \\ &= \sum_{i=1}^n [Y_i \log_e \pi_i + (1 - Y_i) \log_e (1 - \pi_i)] \end{aligned} \tag{2}$$

The presence of outliers will lead to biasedness in MLE [4]. Hence, in this study, MLE will be used to compare its performance with four existing robust logistic regression models which are MALLOWS, CUBIF, BY and WBY.

2.2 The Mallows-Type Estimator (MALLOWS)

The Mallows-type estimator with minimized weighted log-likelihood function was introduced by [5] where the weights of Mallows-type are dependent on covariates. The Mallows-type estimator can be defined as follows:

$$\sum_{i=1}^n w_i [y_i \log p_i(\beta) + (1 - y_i) \log(1 - p_i(\beta))] \tag{3}$$

with $w_i = W(h_n(x_i))$ where W is a non-increasing function hence $W(u)u$ is bounded. W depends on a parameter $c > 0$ and $W(u)$ as follows [6]:

$$W(u) = \left(1 - \frac{u^2}{c^2}\right)^3 I(|u| \leq c) \tag{4}$$

2.3 The Conditionally Unbiased Bounded Influence Function Estimator (CUBIF)

Another familiar robust method for logistic regression is Conditionally Unbiased Bounded Influence Function (CUBIF) estimator. The bounded-influence estimator main point is to introduce a bound on the influence function [5]. The CUBIF estimator minimized the efficiency of covariance matrix to a bound on a measure of gross error sensitivity. It is designed from a consistent M-estimator in the form of $E(\Psi(y, x, \beta)|x_1 = 0) = 0$. Then, the optimal function of Ψ can be defined as:

$$\Psi(y, x, \beta, \mathbf{B}) = W(\beta, y, x, \mathbf{B}) \left\{ y - g(\beta^T x) - c(\beta^T x) - c \left(\beta^T x, \frac{b}{h(x, \mathbf{B})} \right) \right\} x \tag{5}$$

2.4 The Bianco and Yohai Estimator (BY)

The purpose of Bianco and Yohai (BY) estimator is to minimize the weight of total deviance and down weight the outlier. Hence, the outliers in the data can be down-weighted by BY estimator. The BY can be stated as follows:

$$M(\beta) = \sum_{i=1}^n [\rho(d^2(\pi_i(\beta), y_i)) + q(\pi_i(\beta))] \tag{6}$$

2.5 The Weighted Bianco and Yohai Estimator (WBY)

The Weighted Bianco and Yohai (WBY) estimator is an extension from Bianco and Yohai (BY) estimator. This is because it has extra weights to minimize the weight of deviance. Therefore, the WBY can be defined by:

$$\sum_{i=1}^n w(x_i) [\Psi(d_i^2(\beta))(y_i - \pi_i(\beta)) - E_\beta(d_i^2(\beta))(y_i - \pi_i(\beta)) | x_i] \tag{7}$$

2.6 Measure of Performance

In robust logistic regression, a performance measure for evaluating the goodness of fit test is considered by using the χ^2 arcsine transformation [7] which have been used by many researchers [1, 4, 8]. The χ^2 arcsine transformation formula is as given:

$$\chi_{arc}^2 = \sum_{i=1}^n 4 [\arcsin \sqrt{y_i} - \arcsin \sqrt{\pi_i}]^2; i = 1, 2, \dots, n \tag{8}$$

where $\sqrt{\pi_i}$ presents the fitted probabilities for $i = 1, 2, \dots, n$. The purpose of arcsine is to normalize the data in percentages or proportions when the distributions fit the Bernoulli distribution. Hence, the arcsine transformation will change a Bernoulli random variable to one which is nearly normal and the variance is dependent on the parameter π_i . The smaller the χ^2 arcsine value, the better the goodness of fit.

Table 1 Summary of descriptive statistics for independent variables

Variable	Unit	Min	Max	Mean	Median	Std dev
Duration of marriage (X_1)	month	12	492	103.1	90.0	96.3
Husband's age (X_2)	year	20	52	28.7	27.0	6.3
Wife's age (X_3)	year	20	50	25.8	24.5	5.5
Number of children (X_4)		0	7	1.8	2	1.5

3 Application on Real Data

In this study, a set of secondary data was obtained from one of the Syariah Lower Court in Malaysia. The observations of 150 applications of divorce cases were randomly selected in the year 2016. The variables involved in this study are duration of marriage (X_1), age of husband at marriage (X_2), age of wife at marriage (X_3), and number of children (X_4). The response variable is marital status (Y) where the occurrence of divorce is denoted as $Y = 1$ (82.67%) and the non-occurrence of divorce as $Y = 0$ (17.33%). Table 1 shows the summary of descriptive statistics for independent variables. Based on the table, the minimum duration of marriage is 12 months which indicates that after 12 months, some married couples decided to end their marriage. Meanwhile, the mean duration of marriage is 8.6 years. Moreover, the average ages at marriage for the husband and for the wife are 28.69 and 25.84 years, respectively.

The average number of children for divorced couples is 1.81 which is approximately two children per family. Figure 1 provides boxplots of each independent variable to check for any outliers. It can be observed that outliers do exist in all of the independent variables. Due to the presence of outliers in the divorce data set and the impact of these outlying observations on the MLE method, therefore robust logistic regression methods are more appropriate to be used. Table 2 provides the estimated parameters, standard error and values of goodness-of-fit measure for five estimators. The table shows the MLE produces the highest value of χ^2 arcsine. Hence, it is proven that the MLE failed to perform as a good parameter estimate when outliers are present in the dataset. However, WBY method produced the lowest χ^2 arcsine indicating that this is the most efficient robust method in the presence of outliers in the data.

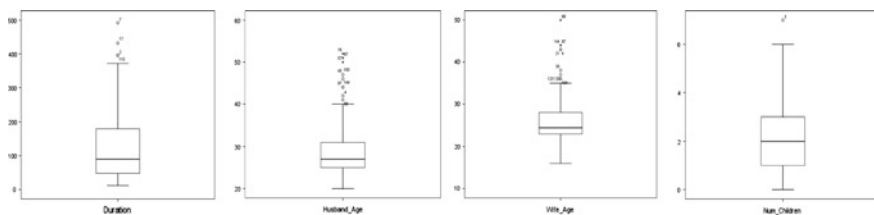


Fig. 1 Boxplots for the four independent variables

Table 2 Summary of estimated parameters (Est), standard error (s.e), and goodness-of-fit measure

Method		MLE	MALLOWS	CUBIF	BY	WBY
Intercept	Est.	4.6038	4.6649	4.9154	4.4122	2.7858
	s.e.	1.2933	1.3054	1.3199	1.3984	2.6644
X ₁	Est.	0.0031	0.0035	0.0032	0.0030	0.0070
	s.e.	0.0035	0.0035	0.0035	0.0032	0.0048
X ₂	Est.	-0.1198	-0.1337	-0.1572	-0.1149	-0.1398
	s.e.	0.0375	0.0395	0.0426	0.0562	0.0989
X ₃	Est.	0.0175	0.0288	0.0437	0.0168	0.0854
	s.e.	0.0434	0.0444	0.0459	0.0513	0.0855
X ₄	Est.	-0.1624	-0.1614	-0.1453	-0.1557	-0.1504
	s.e.	0.2006	0.2025	0.2025	0.1733	0.2038
χ^2_{arc}		525.9093	521.0463	523.304	517.742	392.751

4 Conclusion

The results in this study found that the classical MLE behaves poorly in the presence of outliers; hence alternatively, robust methods are preferred. Out of the four robust methods discussed, the WBY method performs better than the rest and therefore the WBY is highly recommended provided that outliers are detected in advance. Other than using boxplots, other recommended diagnostic methods are also available such as DFFITS, DFBETAS, Cook’s Distance and many more. However, the MLE performs best when the data set is clean or free from outliers. In terms of using divorce data, it is recommended to consider socioeconomic variables such as education level, income and employment status of husband and wife [9].

Acknowledgements The authors wish to thank Universiti Teknologi MARA (UiTM) Shah Alam for providing the internal grant (600-IRMI/MyRA 5/3/LESTARI(0129/2016) and Syariah Lower Court for providing the data.

References

1. Syaiba, B.A., Habshah, H.: The performance of classical and robust logistic regression estimators in the presence of outliers. *Pertanika J. Sci. Technol.* **20**(2), 313–325 (2012)
2. Bellio, R., Ventura, L.: An introduction to robust estimation with R functions. In: *Proceedings of 1st International Work*, pp. 1–57 (2005)
3. Sarkar, S.K., Midi, H., Rana, M.: Detection of outliers and influential observations in binary logistic regression: an empirical study. *J. Appl. Sci.* **11**(1), 26–35 (2011)
4. Sanizah, A., Habshah, M., Norazan, M.R.: Robust estimators in logistic regression: a comparative simulation study. *J. Mod. Appl. Stat. Methods* **9**(2), 502–511 (2010)

5. Kunsch, H.R., Stefanski, L.A., Carroll, R.J.: Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *J. Am. Stat. Assoc.* **84**, 460–466 (1989)
6. Carroll, R.J., Pederson, S.: On robust estimation in the logistic regression model. *J. Roy. Stat. Soc. B* **55**, 693–706 (1993)
7. Kordzakhia, N., Mishra, G.D., Reiersølmoen, L.: Robust estimation in the logistic regression model. *J. Stat. Plan. Inference* **98**(1), 211–223 (2001)
8. Croux, C., Haesbroeck, G.: Implementing the bianco and yohai estimator for logistic regression. *Comput. Stat. Data Anal. J.* **44**, 273–279 (2003)
9. Sanizah, A., Hasfariza, F., Norin Rahayu, S., Nur Niswah Naslina, A.: Determinants of marital dissolution: a survival analysis approach. *Int. J. Econ. Stat.* **2**, 348–354 (2014)

Chapter 53

SMOTE Approach to Imbalanced Dataset in Logistic Regression Analysis



Amirah Hazwani Abdul Rahim, Nurazlina Abdul Rashid,
Asmahani Nayan and Abd-Razak Ahmad

Abstract Logistic regression is a classification model that is commonly used in bankruptcy studies. The classifier works well when data is balanced. However, imbalanced data set is found in almost all bankruptcy studies. The most common approach to deal with imbalanced data set is by selecting and matching the samples from both bankrupt and non-bankrupt samples. The problem of imbalanced data and the approach taken to deal with it can affect a good predictive model. The objective of the study is to improve the classification accuracy of a logit model when data is heavily loaded to one side. The approach taken is by using SMOTE sampling. The study used SMEs categorized under the accommodation and food service activities, and the hotel sector. There are 14 explanatory variables involved. The result from this study confirmed that the AUC and sensitivity values from SMOTE Logistic Regression (SLR) model is higher than the AUC and sensitivity values of a logit model.

Keywords Imbalanced data · SMOTE sampling · Logistic regression

A. H. A. Rahim (✉) · N. A. Rashid · A. Nayan · A.-R. Ahmad
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kedah,
Malaysia

e-mail: amirah017@kedah.uitm.edu.my

N. A. Rashid

e-mail: azlina150@kedah.uitm.edu.my

A. Nayan

e-mail: asmahanin@kedah.uitm.edu.my

A.-R. Ahmad

e-mail: ara@uitm.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_53

1 Introduction

Most financial distress prediction (FDP) studies ignored the problem that arises when the data collected on distressed and healthy firms are imbalance. Past studies on financial distress prediction based their studies on a balanced data set, which regularly leads to an overestimation of a distress model's ability to recognize distressed firms [1].

In real life, there are a lot of paired imbalanced data sets [2] and when dealing with imbalanced class distributions one needs to address the typical two class classification problem, that is one data set (the majority) is bigger than the minority dataset. The imbalance causes the decrease in the prediction accuracy in the minority class datasets. Lakshmi [3], biases in classification and leads to poor generalization performance [4]. The classification algorithms of distressed models of previous studies are meant to be used when the datasets are balanced. These algorithms treat the two class samples as equal. Disregarding misclassification rate of minority class can prompt serious issues in numerous studies [3].

The classification accuracy of most studies involving imbalanced data set is questionable [5]. Most classifying models such as neural network, decision tree and logistic regression perform perfectly when the dataset is balanced. There are several methods to handle imbalanced problem at both the algorithm level and data level. At the data side, several resampling methods are used to balance class distribution, such as undersampling majority class and oversampling minority class [6]. SMOTE (Synthetic Minority Over-sampling Technique) [7] is a well-received approach used to generate new minority class data, from the nearest neighbors which could enlarge decision boundary closer to the majority class. The SMOTE approach introducing artificial examples or instances and these examples are produced based on the elements of the original dataset so that they turn out to be similar to the original examples of the minority class [8]. This study used SMOTE sampling method on the classification of logistic regression model. The aim of this research is to explore the effect of over sampling on the performance of quantitative bankruptcy prediction model on real highly imbalanced datasets.

2 Data

The dataset were data of Small and Medium Enterprises (SMEs) obtained from Suruhanjaya Syarikat Malaysia (SSM). There are many sectors of SMEs. This study focused on SMEs classified under the *transportation and storage, accommodation and food service activities*, and *hotel* sectors only. It contains 601 failed and 26284 non-failed Malaysian SMEs in the three sectors for the period of 1999–2013. Financial ratios were used in this study as the covariates. Table 1 shows the financial ratios used in this study.

Table 1 Financial ratios

Label	Financial ratio
LIQUIDITY1	CA/CL
LIQUIDITY2	WC/TA
LIQUIDITY3	EBIT/IE
PROFIT1	EBIT/TA
PROFIT2	NI/S
PROFIT3	EBIT/S
PROFIT4	NI/SE
PROFIT5	EBIT/SE
PROFIT6	NI/TA
LV1	TL/TA
LV2	CL/TA
LV3	FL/TA
LV4	TL/SE
LV5	LTL/(LTL+SE)

Note Current Assets (CA), Current Liability (CL), Working Capital (WC), Total Asset (TA), Profit/Loss Before Tax = (EBIT), Sales Revenue (S), Net Income (NI), Total Liability (TL), Long Time Deferred Liability (FL), Total Liability (TL), Shareholder Equity (SE), Long Term Liability (TLT)

3 Method

3.1 Sampling Strategies

To deal with imbalanced dataset, the approach taken in this study was to use a sampling technique known as Synthetic Minority over Sampling Technique (SMOTE), which is available in ‘Imbalanced package’ of R. SMOTE over-samples the minority class sample by generating new minority examples by interpolating between examples of the minority class. After sampling was done, logistic regression was used to determine the accuracy of the model.

3.2 Performance Measures

There are four performance measures to be considered, which are accuracy rate (Acc), Sensitivity (Sen), Specificity (Spec) and Precision rate (Pre). The accuracy rate refers to Table 2.

Table 2 Confusion Matrix

Actual class	Predicted	
	Positive (bankrupt)	Negative (non-bankrupt)
Positive (bankrupt)	True positive (TP)	False negative (FN)
Negative (non-bankrupt)	False positive (FP)	True negative (TN)

4 Results and Analysis

The main dataset was split into estimation and validation samples. Table 3 shows the number of cases in each sample. Using the original approach, the estimation sample consists of 15656 non-bankrupt firms and 382 bankrupt firms. The original data is highly imbalanced. After SMOTE sampling was done, the ratio of minority to majority cases is 2153: 2501. Similarly, the validation sample of the original dataset consists of firms leaning towards the non-bankrupt category with 10,353 firms versus 241 bankrupt firms. Again, after SMOTE sampling was done, class distribution of minority to majority cases becomes 1410:1671.

Table 4 shows a classification table for the original and SMOTE sampling dataset. The results generated showed a highly imbalanced cases if the original approach was taken compared to SMOTE sampling method.

Table 5 shows the results for the logistic and SMOTE logistic. The specificity is high and almost 100% for logistic while sensitivity is 0.36%. Estimation part of sensitivity SMOTE logistic is 57.23% and specificity is 58.83%. Sensitivity, precision rate and area under curve of SMOTE logistic are better than the logistic bankruptcy model.

Table 3 Sampling

Sampling methods	Estimation		Validation	
	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt
Original	15656	382	10353	241
SMOTE	2153	2501	1410	1671

Table 4 Classification table for sampling method

Sampling methods	Actual class	Estimation		Validation	
		(Non-bankrupt)	(Bankrupt)	(Non-bankrupt)	(Bankrupt)
Original	(Non-bankrupt)	13292	1	8819	0
	(Bankrupt)	274	1	170	0
SMOTE	(Non-bankrupt)	1086	760	707	488
	(Bankrupt)	796	1065	459	781

Table 5 Performance measures

Performance measure	Logistic		SMOTE logistic	
	Estimation	Validation	Estimation	Validation
Sensitivity	0.36	0.00	57.23	62.98
Specificity	99.99	100.00	58.83	59.16
Accuracy rate	98.00	98.11	58.03	61.11
Precision rate	50.00	0.00	58.36	61.54
Area under curve	61.00	61.00	66.00	66.66

5 Conclusion and Discussion

In this study, the results show that the SMOTE logistic regression approach is more accurate compare to logistic regression model. “No sampling” method in logistic gave the highest test accuracy, however the results have been biased toward the majority class as the classifiers tend to limit the misclassification by way of classifying all samples into the majority class. Ares under curve, sensitivity and precision rate generated by SMOTE logistic is better than logistic bankruptcy model. For future research, the application of different sampling method such as under-sampling and Borderline-SMOTE on the other machine learning model should be carried out to increase the quality of prediction.

References

1. Sun, J., Shang, Z., Li, H.: Imbalance-oriented SVM methods for financial distress prediction: a comparative study among the new SB-SVM- ensemble method and traditional method. *J. Oper. Res. Soc.* **65**, 1905–1919 (2014). <https://doi.org/10.1057/jors.2013.117>
2. Shi, B., Wang, J., Qi, J., Cheng, Y.: A novel imbalanced data classification approach based on logistic regression and fisher discriminant **2015** (2015)
3. Lakshmi, T.J.: A Study on Classifying Imbalanced Datasets, pp. 141–145 (2014)
4. H. Engineering, Cai, Y., Li, Y.: Oversampling method for imbalanced **34**, 1017–1037 (2015)
5. Hanifah, F.S.: SMOTE bagging algorithm for imbalanced dataset in logistic regression analysis (Case: Credit of Bank X) **9**(138), 6857–6865 (2015)
6. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-smote: a new over- sampling method in imbalanced data sets learning. In: *Advances in Intelligent Computing*, pp. 878–887 (2005)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
8. Alghamdi, M., Al-mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford Exercise Testing (FIT) project, pp. 1–15 (2017)

Chapter 54

Statistic for Outlier Detection in Circular Functional Relationship Model



Mohd Syazwan Mohamad Anuar, Abdul Ghapor Hussin
and Yong Zulina Zubairi

Abstract During the last few years, researchers have shown strong interest on the subject of outlier detection in both linear and circular for Error-in-Variables (EIV) Models. Recently, the studies of outlier detection on circular variables models using row deletion method are widely explored; in particular in regression and EIV models for circular variables. In this paper, we have proposed a new measure of mean circular error using cosine function for circular functional relationship model. We also used the row deletion method to detect observations that affect the measure the most, thus identifying them as outlier. The corresponding cut-off points are identified via simulation studies.

Keywords Circular functional relationship model · Outlier · Cosine function · Row deletion

1 Introduction

Outlier detection in any dataset has been widely discussed by many researchers. It is because their occurrence will give a significant difference towards the analysis and lead to phenomenon of under study. It is also important to identify the outlier first before further analysis is done in research. In circular functional relationship model,

M. S. M. Anuar (✉)

Centre of Defense Foundation Studies, National Defense University of Malaysia,
Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia
e-mail: m.syazwan@upnm.edu.my

A. G. Hussin

Faculty of Defense Science and Technology, National Defense University of Malaysia,
Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia
e-mail: ghapor@upnm.edu.my

Y. Z. Zubairi

Centre for Foundation Studies in Sciences, University of Malaya, 50603 Kuala Lumpur,
Malaysia
e-mail: yzulina@um.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference
on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_54

the outlier detection still new and more studies can be conducted to explore the problem of outlier.

This because numerous studies on outlier detection are mostly done in circular regression model, see, Rambli [1], Abuzaid et al. [2], Abuzaid et al. [3] and Ibrahim et al. [4]. In this paper we will discuss detail about the outlier detection in circular functional relationship model via row deletion method. The first outlier detection in functional relationship model was Hussin et al. [5] for circular variables via complex form.

Recently, procedure of outlier detection for functional relationship model using deletion row has been discovered but it is for linear variables, see Shamsudheen et al. [6]. Then, Mokhtar et al. [7] used clustering technique to detect an outlier in the same linear functional relationship model.

The model is called an EIV model if the model is observed with error in both response and explanatory variables. There are two types of EIV models and they are functional relationship model and structural relationship model. Sufficient theories and methods on parameter estimation of the EIV models can be found in Fuller [8], Cassela and Berger [9] and Chernov [10].

The first functional relationship model was introduced by Hussin [11] called as Pseudo-replication in Functional Relationship model. Then, Hussin [12] has developed an unreplicated complex linear functional relationship model for circular variables by extending the Laycock [13] complex linear regression model. Hussin et al. [14] has improved the model by finding the estimate error of the concentration parameter for any ratio of two concentration parameters, λ using properties of the Bessel function.

Latest, Satari et al. [15] proposed circular functional relationship model. The model is an extension of the regression model for circular variables proposed by Downs and Mardia [16]. We will utilize this model to propose a new statistics of the circular error for outlier detection.

Study on outlier detection for circular functional relationship model was first done by Satari [17] using clustering technique. Thus in this paper, we aim to develop a row deletion approach to detect outlier for this model using cosine statistic.

2 Functional Relationship Model for Circular Variables

Suppose we have circular random variables X and Y with the observe value $x_i = (-\pi, \pi]$ and $y_i = (-\pi, \pi]$ for $i = 1, \dots, n$, respectively. Assume that the pairs of observation (x_i, y_i) are measured with errors $\delta_i \sim VM(0, \kappa)$ and $\varepsilon_i \sim VM(0, \nu)$, respectively where δ_i and ε_i are both independently distributed with von Mises distribution. For any fixed independent angle X_i , the Satari et al. [15] functional relationship model given by

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i$$

where

$$Y_i = \beta + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2} (X_i - \alpha) \right\}. \quad (1)$$

Given that Y_i is dependent random angle, ω is a slope parameter in the close interval $[-1, 1]$, while α and β are angular parameters.

By assuming that the ratio of the error concentration parameters for above model is $v = \lambda\kappa$, $\lambda > 0$, is fixed and known, then the likelihood function for a set of observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \dots, (x_n, y_n))$ can be written as

$$L(\alpha, \beta, \omega, \kappa, X_1, \dots, X_n | \lambda, (x, y)) = \frac{1}{(2\pi)^{2n} [I_0(\kappa) I_0(\lambda\kappa)]^n} \exp \left[\sum_{i=1}^n \kappa \cos(x_i - X_i) \right] \exp[\lambda\kappa \cos(y_i - Y_i)]. \quad (2)$$

The log-likelihood function for the model may be given by

$$\log L(\alpha, \beta, \omega, \kappa, X_1, \dots, X_n | \lambda, (x, y)) = -2n \log[2\pi] - n \log[I_0(\kappa)] - n \log[I_0(\lambda\kappa)] + \sum_{i=1}^n \kappa \cos(x_i - X_i) + \sum_{i=1}^n \lambda\kappa \cos \left(y_i - \left(\beta + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2} (X_i - \alpha) \right\} \right) \right). \quad (3)$$

3 Definition of New Statistics for Outlier Detection

The statistics proposed by Shamsudheen [6] for linear functional relationship model and we will apply this statistics for the circular functional relationship model. The statistics may be defined by

$$\text{FMCEc} = 1 - \frac{1}{2n} \left[\sum_{k=1}^n \cos(x_k - \hat{x}_k) + \sum_{k=1}^n \cos(y_k - \hat{y}_k) \right] \quad (4)$$

Where n is the sample size, \hat{x}_k is the estimated values of x_k and \hat{y}_k is the estimated values of y_k for any given model and $\text{FMCEc} \in [0, 2]$.

As the result, the observation of x_k or \hat{y}_k is defined as an outlier in the data set if the circular distance between x_k or \hat{y}_k and the estimated values \hat{x}_k and \hat{y}_k is expected to be relatively large respectively. If the distance between x_k or \hat{y}_k and the estimated values \hat{x}_k and \hat{y}_k is large, thus the summation of all circular distances as well as the

value of the FMCEc statistics will increase. If with removal of the k th observation denoted by $FMCEc_{(-k)}$, it will decrease the value of the statistic. Then, the k th observation is assumed to be an outlier in the data set. We will find the maximum absolute difference between the value of the statistics for reduced (after removal of the k th observation) and full data sets by

$$FDMCEc = \max_k (|FMCEc - FMCEc_{(-k)}|). \tag{6}$$

Then, the k th observation will be assumed to be an outlier if FDMCEc exceeds a certain pre-specified cut-off point. In this study, the cut-off point will be derived from simulation study where percentile point of test statistic are used.

4 Results

Table 1 tabulated the cut-off points for the FDMCEc statistics at different levels of significance. The result from Table 1 shows that, for fixed sample size, the cut-off points for the FDMCEc statistics exhibit a decreasing trend as κ gets larger. The smallest value of the cut-off points will be at 0.00061 and the largest will be at 0.04960.

Table 1 The simulated cut-off points for FDMCEc statistic sorted by sample size

n		10%	5%	1%
20	5	0.02817	0.03996	0.0496
	10	0.02488	0.03405	0.04324
	15	0.02008	0.02281	0.03815
	20	0.01356	0.02217	0.02616
40	5	0.01749	0.02179	0.02284
	10	0.00343	0.00774	0.01167
	15	0.00328	0.00575	0.01144
	20	0.00193	0.00387	0.00857
60	5	0.01008	0.01333	0.01569
	10	0.00553	0.00696	0.00902
	15	0.00225	0.00447	0.00897
	20	0.00139	0.00165	0.00309
80	5	0.00472	0.00688	0.00837
	10	0.00144	0.00211	0.00378
	15	0.00104	0.00198	0.00412
	20	0.00061	0.00089	0.00201

5 Conclusion

In this paper, we consider the problem of outlier detection in the Satari et al. functional relationship model for circular variables based on the FDMCEc statistic. The sampling behavior is investigated using simulation study. The result shown that cut-off points will decrease as κ increases and as the sample size increases, the cut-off points will decrease. Further study can be conducted on the power of performance of FDMCEc and then these procedures also can be applied to real dataset.

References

1. Rambli, A.: Outlier detection in circular data and circular-circular regression model, MSc. Thesis, University of Malaya (2011)
2. Abuzaid A.H., Hussin A.G., Mohamed I.B.: Detection of outliers in simple regression model using mean circular error statistic. *J. Stat. Comput. Simul.* **83**(2), 269–277 (2013). <https://doi.org/10.1080/00949655.2011.602679>
3. Abuzaid, A.H., Mohamed, I.B., Hussin, A.G., Rambli, A.: COVRATIO statistic for simple circular regression model. *Chiang Mai J. Sci.* **38**(3), 321–330 (2011)
4. Ibrahim, S., Rambli, A., Hussin, A.G., Mohamed, I.: Outlier detection in a circular regression model using COVRATIO statistic. *Commun. Stat. Simul. Comput.* **42**(10), 2272–2280 (2013). <https://doi.org/10.1080/03610918.2012.697239>
5. Hussin, A.G., Chik, Z.: On estimating error concentration parameter for circular functional model. *Bull. Malays. Math. Sci. Soc.* **26**, 181–188 (2003)
6. Shamsudeen M.I.: Bootstrapping and outlier detection problem in linear functional relationship model for circular data, MSc. Thesis, National Defense University Malaysia (2014)
7. Mokhtar, N.A., Zubairi, Y.Z., Hussin, A.G.: Parameter estimation of simultaneous linear functional relationship model for circular variables assuming equal error variances. *Pak. J. Stat.* **31**(2), 251–265 (2015)
8. Fuller, W.A.: *Measurement Error Models*. Wiley (1987). ISBN 0-8218-5117-9
9. Casella, G., Berger, R.L.: *Statistical Inference*. Duxbury Thomson Learning (2001). <https://doi.org/10.1002/bimj.4710370219>
10. Chernov N., *Circular and Linear Regression: Fitting circles and lines by least squares*. Taylor and Francis Group (2011). ISBN: 978-1-4398-3591-3
11. Hussin, A.G.: Pseudo-replication in functional relationship with environmental application, PhD Thesis, University of Sheffield, England (1997)
12. Hussin, A.G.: The unreplicated complex linear functional relationship model and its application. *Bull. Malaysian. Math. Soc. (Second Series)* **21**, 79–86 (1998)
13. Laycock, P.J.: Optimal regression: regression models for directions. *Biometrika* **62**, 305–311 (1975)
14. Hussin, A.G., Abuzaid, A.H.: Detection of outliers in functional relationship model for circular variables via complex form. *Pak. J. Stat.* **28**(2), 205–216 (2012)
15. Satari, S.Z., Hussin, A.G., Zubairi, Y.Z., Hassan, S.F.: A new functional relationship model for circular variables. *Pak. J. Statist.* **30**(3), 397–410 (2014)
16. Down, T.D., Mardia, K.V.: Circular regression. *Biometrika* **86**(3), 683–697 (2002). <https://doi.org/10.1093/biomet/89.3.683>
17. Satari, S.Z.: Parameter estimation and outlier detection for some types of circular model, PhD. Thesis, Universiti Malaya (2015)

Chapter 55

Structural Breaks in Malaysian Shariah Compliant Indices



Ida Normaya Mohd Nasir and Mohd Tahir Ismail

Abstract This paper contributes to the literature on the detection of structural breaks in the mean and variance of Malaysian Shariah compliant stock indices (FTSE Emas Shariah and FTSE Hijrah Shariah) and their main index, FTSE Bursa Malaysia KLCI. The existence of structural changes in mean is tested using the Bai and Perron procedure. On the other hands, we use Inclan and Tio (1999) Iterated Cumulative Sums of Squares (IT-ICSS) algorithm and Sansó et al. (2004) modified ICSS algorithm procedures to identify any structural changes in series variance. Based on the Bai and Perron test results, there is no structural break reported in Shariah compliant indices. However, the structural breaks in Shariah compliant indices are only reported in the variance of the data series. This study also provides some possible explanations for the cause of these structural shifts in both mean and variance.

Keywords Structural change · ICSS · Modified ICSS · Shariah compliant indices · Malaysia

1 Introduction

The Shariah compliant stock index is one of the most important branches of the Islamic capital market. Governed by Shariah laws, there are five major principles of operation; preventing any practice of usury, the risks are shared between the entrepreneurs and investors, any speculation or *gharar* is forbidden, compliance of

I. N. M. Nasir (✉) · M. T. Ismail
School of Mathematical Sciences, Universiti Sains, George Town, Malaysia
e-mail: idanormaya@gmail.com

M. T. Ismail
e-mail: m.tahir@usm.my

I. N. M. Nasir
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM)
Kedah, Bedong, Malaysia

the aqad with the stated contract and any economic activities must be legal within the Shariah aspects aspect [1].

A growing body of literature has focused on the performance and efficiency of Shariah compliant indices with their conventional counterparts. However, far too little discussion about the presence of structural breaks in the return series of Shariah compliant indices. As reported in literature, failure to consider for structural breaks in modelling could lead to model misspecification [2] and increase the degree of volatility persistence which will lead to forecast failure [3].

Malaysia has earned its reputation as one of the leading global centers for Islamic finance. In 2009, Malaysia was recorded as the largest Shariah compliant equity market among the Islamic finance centers in the world [4]. Approximately 74.2% of the total securities traded in the Bursa Malaysia were Shariah compliant.¹ Due to the rapid progressive of Malaysian Islamic finance, the Malaysian Shariah compliant indices are selected for this study.

2 Methodology

2.1 Testing for Structural Breaks in Mean

The shift in the mean specification is identified by using the Bai and Perron [5, 6] (BP henceforth) test. The proposed linear model with l breaks (or $l + 1$ regimes) is given below:

$$y_t = \beta x_t' + \delta_j z_t' + \varepsilon_t \quad (1)$$

where y_t denotes the dependent variable at period t , $t = T_{j-1} + 1, \dots, T_j$ where $j = 1, \dots, l + 1$ while x_t and z_t are vectors of covariates with $(p \times 1)$ and $(q \times 1)$, respectively. β and δ_j are the corresponding beta coefficients for x_t and z_t , respectively. Here, ε_t represents the residuals at period t . In this study, the sequential procedure was chosen. According to Bai and Perron [6], the sequential procedure provides better information criteria to select break points.

¹<https://www.sc.com.my/data-statistics/islamic-capital-market-statistics/>.

2.2 Testing for Structural Breaks in Variance

2.2.1 Inclan and Tiao (1994) Iterative Cumulative Sum of Squares Statistic (IT-ICSS)

Inclan and Tiao [7] develop an iterative cumulative sum of squares statistic (IT-ICSS) to test the null hypothesis of a constant unconditional variance against the alternative hypothesis of a break in the unconditional variance. The Inclan-Tiao statistic is given by:

$$IT = \sup_k \left| (T/2)^{0.5} D_k \right| \tag{2}$$

where $D_k = (C_k/C_T) - (k/T)$, with $C_k = \sum_{t=1}^k \varepsilon_t^2$ for $k = 1, \dots, T$. and C_T is the sum of the squared residuals from the whole sample period. The null hypothesis of constant variance is rejected if the maximum $\left| (T/2)^{0.5} D_k \right| >$ the critical value. The 95th percentile critical value for the asymptotic distribution is 1.358 [7, 8]. However, the method is designed for independent and identically distributed (i.i.d) processes which are impractical for financial data which often evidence of conditional heteroscedasticity.

2.2.2 Sansó et al. (2004) Modified ICSS Algorithm

Sansó et al. [9] find certain drawbacks in the IT-ICSS where the algorithm neglects kurtosis properties and the conditional heteroskedasticity. To overcome of the problem, they extended the IT-ICSS algorithm and introduced the κ_1 test. Similar to the IT-ICSS, κ_1 assumes that the residual series is an i.i.d with zero mean and constant variance. However, the κ_1 test is free from nuisance parameters produced by the IT-ICSS test, where κ_1 is defined as:

$$\kappa_1 = \sup_k \sqrt{T} |B_k| \tag{3}$$

where $B_k = \frac{C_k - \frac{k}{T}C_T}{\sqrt{\left(\frac{1}{T} \sum_{t=1}^T \varepsilon_{t-1}^4\right) - \frac{1}{T}C_T}}$, $k = 1, 2, \dots, T$. If $\varepsilon^2 \sim iid(0, \sigma^2)$ and $E(\varepsilon_t^4) \equiv$

$$\frac{1}{T} \sum_{t=1}^T \varepsilon_{t-1}^4 < \infty, \text{ thus } \kappa_1 \Rightarrow \sup_r |W^*(r)|$$

Sanso et al. [9] introduced κ_2 test to address leptokurtic distribution and persistence in the conditional variance series. The adjusted statistic is given by:

$$\kappa_2 = \sup_k \sqrt{T} |G_k| \tag{4}$$

where $G_k = \frac{1}{\sqrt{\hat{\lambda}}} [C_k - (\frac{k}{T}) C_T], k = 1, 2, \dots, T, \hat{\lambda} = \hat{\gamma}_0 + 2 \sum_{i=1}^m [1 - \frac{i}{m+1}] \hat{\gamma}_i, \hat{\gamma}_i = \frac{1}{T} \sum_{t=i+1}^T (\varepsilon_t^2 - \hat{\sigma}^2)(\varepsilon_{t-1}^2 - \hat{\sigma}^2)$ and $\hat{\sigma}^2 = \frac{1}{T} C_T$.

The lag truncation parameter, m is estimated using the procedure in Newey and West [10] estimator. This study will implement the IT, κ_1 and κ_2 techniques to detect the presence of potential structural break in the variance of both Shariah and conventional indices series.

3 Results and Discussion

The historical data used in this study are the daily closing price indices of FTSE Bursa Malaysia Kuala Lumpur Composite Index (FBMKLCI), FTSE Bursa Malaysia Emas Shariah Index (FBMS) and FTSE Bursa Malaysia Hijrah Shariah Index (FBMHS) from 1st March 2007 to 28th April 2017. There are a total of 2652 observations for each financial market obtained from Thomson Reuters DataStream. Let p_t denotes daily closing stock price index and corresponding daily return at time t is calculated as follows:

$$r_t = 100 \times (\log p_t - \log p_{t-1}) \text{ for } t = 1, 2, \dots, 2652 \tag{5}$$

3.1 Descriptive Statistics

Table 1 provides descriptive statistics of the daily return series. All the indices show positive expected returns, having the same level of volatility and negatively skewed. The kurtosis coefficients suggesting that the series are leptokurtic. The Jarque–Bera (JB) statistics are statistically significant and thus, suggesting that the distribution of all the indices are not standard normal. The residual diagnostics suggest that there is an ARCH effect in the series. The Ljung-Box serial correlation (Q-stats) shows the presence of serial correlation up to 12 lags. All the indices are stationary as indicated by Augmented Dickey and Fuller (ADF) and the Phillips and Perron (P-P) unit root tests.

Table 1 Summary of descriptive statistics

	FBMKLCI	FBMS	FBMHS
Mean	0.0064	0.0079	0.0097
Std. Dev	0.3219	0.3370	0.3456
Skewness	-1.1591	-1.5445	-1.1694
Kurtosis	19.0319	24.4944	21.8283
Jarque-Bera	28994.63 (0.0000) ^a	52106.45 (0.0000) ^a	39777.09 (0.0000) ^a
Q (12)	37.890 (0.0000) ^a	47.8193 (0.0000) ^a	41.190 (0.0000) ^a
Q ² (12)	231.41 (0.0000) ^a	157.9984 (0.0000) ^a	204.17 (0.0000) ^a
ARCH test	27.0242 (0.0000) ^a	8.5328 (0.0000) ^a	10.7169 (0.0000) ^a
LM test	13.7872 (0.0000) ^a	3.9924 (0.0000) ^a	3.6339 (0.0000) ^a
ADF	-46.5125 (0.0001) ^a	-46.0556 (0.0000) ^a	-46.2678 (0.0001) ^a
PP	-46.5099 (0.0001) ^a	-44.2267 (0.0001) ^a	-44.2798 (0.0001) ^a

^aSignificant at 5% level

3.2 Testing for Structural Breaks in Mean

Table 2 summarizes the BP test results for the daily return series of FBMKLCI, FMBS and FBMHS. The structural breakpoint test shows that the null hypothesis that there is no structural break is failed to reject in all Malaysian Shariah compliant indices. It suggests that the Shariah compliant series is stable in the mean. However, only one structural break in detected on the mean of FBMKLCI return series.

3.3 Testing for Structural Breaks in Variance

Table 3 presents the number of sudden changes in the variance, the time periods identified as when such sudden changes have occurred and the number of observation where the break is identified for the daily return series of FBMKLCI, FBMS and FBMHS indices. Variance breaks are identifies in conventional (FBMKLCI) and Shariah-compliant (FBMS and FBMHS) in the range of one to eight breaks. There are few similar break dates located at 15th March 2007, 26th July 2007, 23rd August 2007, 3rd January 2008, 7th March 2008 suggesting that Shariah-compliant indices provide almost the same variance changes than conventional indices. This finding is consistent with the study reported by Charles et al. [11].

Table 2 Bai and Perron Test

	FBMKLCI	FBMS	FBMHS
F-stat	9.2296	8.6600	4.8237
Scaled F-test	9.2296	9.1491	5.7323
Break date	30.10.2008	-	-

Table 3 Sudden changes in volatility

Indices	IT-ICSS		κ_1		κ_2	
	No of breaks	Date	No of breaks	Date	No of breaks	Date
FBMKLCI	6	15.3.07, 26.7.07, 23.8.07, 3.1.08, 7.3.08	1	31.7.09	1	31.7.09
FBMS	8	15.3.07, 4.5.07, 26.7.07, 23.8.07, 3.1.08, 28.1.08, 7.3.08, 17.3.08	6	25.6.09, 4.8.11, 1.12.11, 17.3.14, 6.10.14, 4.2.16	1	25.6.09
FBMHS	3	3.2.08, 7.3.08, 17.3.08	6	25.6.09, 4.8.11, 18.10.1, 10.9.13, 20.11.14, 4.2.16	1	25.6.09

The structural changes in the variance of FBMKLCI, FBMS and FBMHS are related to various important announcement and enhancement made by Bursa Malaysia. Similar with other countries, Malaysian stock indices (both conventional and Shariah compliant) also suffer the impact of the 2008 global financial crisis and it might give some structural change in the variance of the data series.

Based on the findings, the KLCI adoption of FTSE global index standards on 6th July 2009 gives no impact on the structural of the FBMKLCI index series. However, the new board structure implemented on 3rd August 2009 have the structural impact and need to take into consideration in the FBMKLCI stock index study. Under the new structure, Main and Second Boards will be merged into single board and be called Main Markets. The MESDEX markets (for technology and high growth companies) will be called ACE Markets. The detail structure is provided in Bursa Malaysia website.

However, unlike conventional counterparts, the announcement of KLCI adoption made on 24th June 2009 give structural impact on the variance of Shariah compliant stock indices. There are also a few breaks recorded on the late of 2013 and 2014. This might be due to the revision on the Shariah screening methodology made by Bursa Malaysia effective on 29th November 2013.

4 Conclusion

There are various events that contribute to the structural instability in Malaysian stock indices. Unlike conventional indices (FBMKLCI), the Shariah compliant indices (FBMS and FBMHS) do not display structural breaks in the mean. However, the structural breaks are detected on the variance of the Shariah compliant data series. Therefore, the structural breaks in variance of Shariah compliant indices need to be considered in the study of volatility to avoid inaccurate

estimation. This study suggests the use of κ_2 test to detect structural breaks in variance of Shariah compliant indices due to characteristics of the data series which exhibits high kurtosis.

Acknowledgements The authors would like to extend their sincere gratitude to the Ministry of Higher Education Malaysia (MOHE) for the financial supports received for this work under FRGS grant (203/PMATHS/6711604).

References

1. Bacha, O.I.: New issues in islamic capital market development: risk management and islamic capital markets. In: Islamic Capital Markets Conference, Securities Commission, March 2002, Kuala Lumpur, Malaysia (2002)
2. Andreou, E., Ghysels, E.: Detecting multiple breaks in financial market volatility dynamics. *J. Appl. Econom.* **17**(5), 579–600 (2002)
3. Clements, M.P., Hendry, D.F.: *Forecasting Economic Time Series*. Cambridge University Press, Cambridge (1998)
4. PricewaterhouseCoopers.: *Gateway to Asia: Malaysia, International Islamic Finance Hub*, vol. 3. Malaysia International Islamic Financial Center: Shaping Islamic finance together (2010)
5. Bai, B.Y.J., Perron, P.: Estimating and testing linear models with multiple structural. *Econometrica* **66**(1), 47–78 (1998)
6. Bai, B.Y.J., Perron, P.: Computation and analysis of multiple structural change models. *J. Appl. Econom.* **18**(1), 1–22 (2003)
7. Inclan, C., Tiao, G.C.: Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Am. Stat. Assoc.* **89**(427), 913–923 (1994). <http://doi.org/10.2307/2290916>
8. Aggarwal, R., Inclan, C., Leal, R.: Volatility in emerging stock markets. *J. Financ. Quant. Anal.* **34**(1), 33–55 (1999). <http://doi.org/10.2307/2676245>
9. Sansó, A., Aragón, V., Carrion, J.: Testing for changes in the unconditional variance of financial time series. *Revista de Economía Financiera* **4**, 32–53 (2004)
10. Newey, W.K., West, K.D.: Automatic lag selection in covariance matrix estimation. *Rev. Econ. Stud.* **61**, 631–653 (1994)
11. Charles, A., Darné, O., Pop, A.: Risk and ethical investment: empirical evidence from Dow Jones Islamic indexes. *Res. Int. Bus. Financ.* **35**, 33–56 (2015)

Chapter 56

Teenage Driving Behavior Modeling Using Deep Learning for Driver Behavior Classification



Muhamad-Husaini Abu-Bakar, Rizal Razuwan and Syafiq Kamal

Abstract Speed transition is mainly contributed from road condition which is the change from straight to cornering path. This transition introduces a sharp non-linearity in driving speed, and this behavior is significantly changed with different driver ages. Because the non-linearity has occurred, the current linear model is inaccurate to modeling the driving speed. The aim of this paper is to model the driving speed which has a sharp transition in the road path. 100 samples of driving data from the designated track were used as an input to Deep Learning (DL) architecture. DL architecture that combines a linear model that is fitted using L1 regularization, a sequence of ReLU activation layers and develop with Keras framework associated with R studio. The designated track has a distance of 700 m with 4 corners and 4 straight paths, and the driver is among the teenager's age in the range of 20 to 25 years old. The input for the DL is GPS coordinates and output are driving speed for each coordinate. As a result, DL model successfully developed with 4% error as compared with validation data. Even though the error increases at the (selected location) transition point with maximum value was 6%, the value is considered small for the proposed model to be accepted. As a conclusion, the model has a capability in modeling the sharp non-linearity in the road path. The model further significantly improves the driver behavior for early crash prediction.

Keywords Deep learning · Driving behavior · Naturalistic driving data · Machine learning

M.-H. Abu-Bakar (✉) · R. Razuwan · S. Kamal
System Engineering and Energy Laboratory, Section of Manufacturing,
University Kuala Lumpur Malaysian Spanish Institute,
Kulim Hi-Tech Park, 09000 Kulim, Kedah, Malaysia
e-mail: muhamadhusaini@unikl.edu.my

R. Razuwan
e-mail: rizal.razuwan@s.unikl.edu.my

S. Kamal
e-mail: msyafiq.kamal@s.unikl.edu.my

1 Introduction

The driving behavior classification recently on the verge of a significant transformation with drastic research and developments around the world. For example, the Intelligent Transportation System (ITS) evolution toward a remarkable transition from ITS to Data-Driven Intelligent Transportation System (D2TS) in transportation system today [1]. Predicting the speed of a vehicle for a future point on the road ahead is a vital sub-task of many advanced safety systems [2]. Speed prediction of a car became a challenging topic, for the vehicle speed was influenced by several factors, such as road types, weather condition, traffic condition, road curvature, and driving behavior. In contrast with L1 regularization function vehicle speed factors, the vehicle must continuously adapt the driving behavior with optimal speed for safety and comfort.

Multiple approaches in a form of machine learning have been proposed for learning representation features of driving behavior from naturalistic driving data such as Artificial Neural Network (ANN), Convolutional Neural Network (CNN), State-Space Neural Network (SSNN), and Long Short Term Memory (LSTM) recurrent neural network [3]. Speed prediction approaches can be classified into three categories such as model-based, data-driven, and hybrid model. Typically, model-based approach is originated from theoretical model and the hybrid approach integrates both model-based and data-driven approach [4].

In this work, Deep Learning (DL) is introduced because of DL model was capable of capturing sharp non-linearity in driving speed by exploiting the correlation of the high dimensional set of explanatory variables [4]. L1 regularization penalty is used to prevent overfitting the input data. Next, preventing vanishing and explosion gradient of activation functions in DL by using ReLU activation functions. Next, DL model is validated by using a statistical approach.

2 Methodology

Figure 2 shows a developed data logging system and designated track for collecting driving speed data at particular GPS position. The system was built in standalone mode using a lithium-ion battery as a power source for 4 h duration. The driver's speed and position data were calculated with Ublox Neo 6 M GPS module which was selected due to low-cost and acceptable accuracy. The antenna was used to reinforce signal strength for preventing data loss. The system was placed in the vehicle during a data collection process.

A 100 samples were selected among University of Kuala Lumpur Malaysian Spanish Institute (UniKL MSI) students which students age ranging between 18–23 years old and compulsory to have a driving license. 100 sample was taken from the students for this work. The students were instructed to driving a car in the designation track. Figure 1 shows a designation track in UniKL MSI. Because of a

designation track was in campuses, the experiment was conducted at 9 to 10 p.m which is an out of peak hour. This free-flow condition is important to ensure that only driver's driving reflects driving behavior. Otherwise, the traffic flow might disturb a driver focus and potentially increase the complexity of the developed model.

GPS data frequency acquisition was set to 1 Hz to prevent biases effect in speed distribution and loss of information. This effect can affect the average, median, or other speed percentiles [5]. Although the frequency of acquisition was set up to 1 Hz, there still exists a problem where the speed and coordinates of data were different in each driving behavior. This was caused by speed transition in driving behavior which introduces a sharp non-linearity in speed distribution at the designated track. Therefore, the selected track was divided into 8 road paths consist of 4 straight paths and 4 cornering paths to overcome this non-linearity. Each of the road paths was given a label to identify each path which was shown in Fig. 1.

2.1 Deep Learning (DL) Architecture

DL architecture was constructed using Keras framework library in R studio with Tensorflow backend and a particular class of DL used is a feed-forward neural network in this work. The Deep Learning (DL) model was used to predict the speed for each path label in designated track by giving the 100 samples of GPS data for data training. The input is coordinates and output is the speed of the vehicle for driving behavior features learning by DL architecture. An optimal L1 regularization



Fig. 1 Designated track

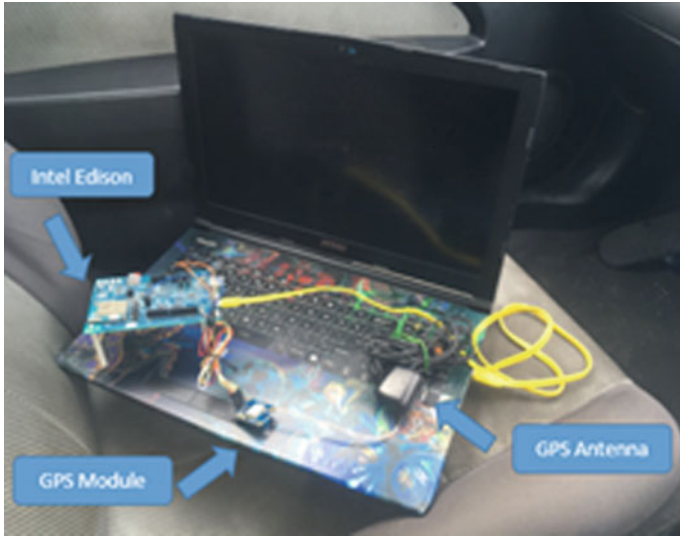


Fig. 2 Data logging system

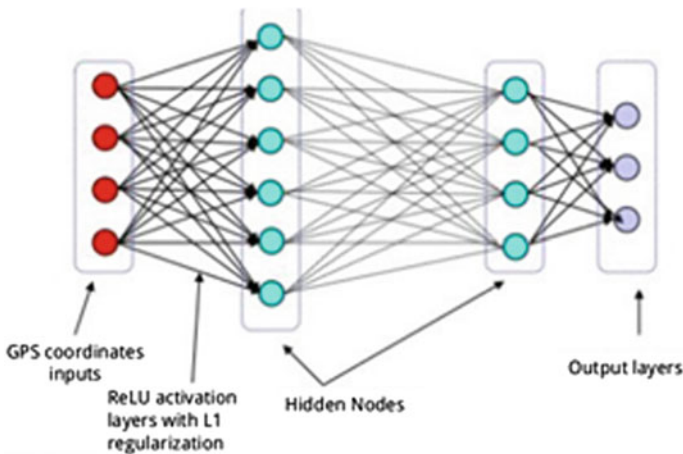


Fig. 3 Deep Learning model architecture

penalty was added in DL to penalized least squares calculation to increase predictive accuracy performance. A sequence of layers ReLu was selected for an activation layer. The speed of each driver was predicted by DL and was compared with actual speed as shown in Fig. 4. The predicted and validation speed mean was calculated and shows 4% of mean squared error (MSE) (Fig. 3).

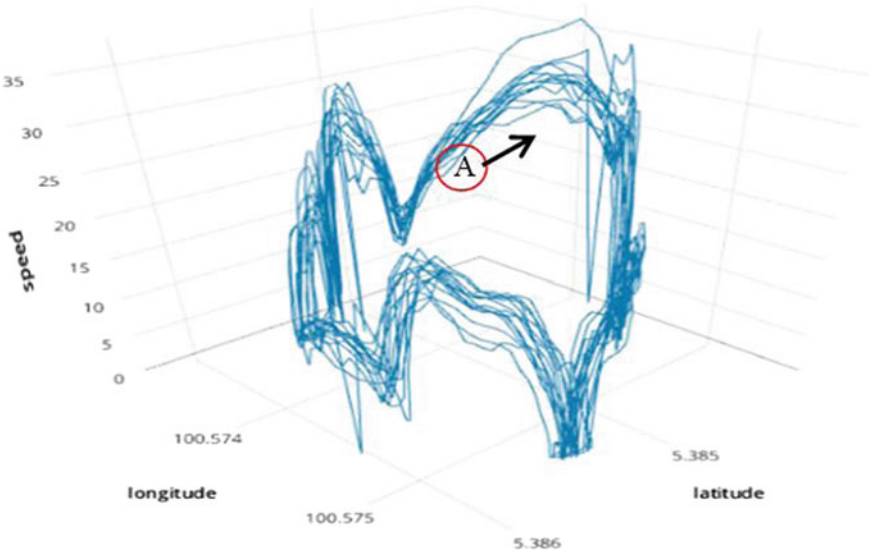


Fig. 4 3d plot of GPS data

3 Results and Discussion

Figure 4 shows a 3d plot of GPS data and validation method used to validate our DL model. The 3d plot consists of speed profile versus longitude and latitude of 30 samples from 100 samples. Moreover, speed profile demonstrates each driver has a different speed at same GPS location and indicates the speed is dependent on driver's behavior. The indicator was labeled A in a 3d plot to prove that the speed is decreasing while entering a cornering path from a straight path. Interestingly, the standard deviation on this area is 2.9 which explaining the deviation of driver speed is small. Particularly, ordinary drivers are slowing down before entering a cornering path. Be-sides that, the speed that has a larger standard deviation 4 significantly shows that each driver has a different style of driving behavior. Therefore, this inconsistency in driving behavior requires a development, evaluation, and validation of modeling model.

In order to validate DL model, the range of speed distribution in selected road path location at Fig. 4 was predicted. There are two paths were chosen for validation purpose which are a full path, straight path and Corner path as shown in Fig. 1. Table 1 shows a percentage error every path. Percentage error was calculated by the absolute difference between actual data and predicted data and divided with actual data.

Percentage error for a full path is 6% and reduce to 4% when an only straight path is considered. The error increase to 7% at the corner path. Even though the deviation of speed is low at corner path, the model error is highest. This occurs due

Table 1 Validation percentage errors of Deep Learning model

Model type	Percentage errors (%)
Deep learning model	6
Straight path (id: S3)	4
Cornering path (id: C4)	7

to a sharp transition of driving path from straight to corner. At this corner a driver required to give more focus on a braking system to maintain in the right track. Thus, the driving behavior drastically changes and increase the difficulties to the model for capture the behavior. As consequence the error increase as compared to a straight path. However, within 7% error, the model is still acceptable and can be used in modeling the driving behavior.

In the straight path, the error is smallest, because of in this area the driver in comfort zone where they can drive with no constraint. Thus, the driver free to employ their own driving style and less focus on the braking system. Because of low frequency in using a braking system, the speed profile become smooth and less sharp transition occurs. This explains why in a straight path the model error is small as compared to corner path.

4 Conclusion

The purpose of current study was modeling the teenage driving behavior which introduced sharp non-linearity in driving speed. In this paper driving speed successfully modeled with DL approaches when factor such as road path and teenagers was considered. The performance of DL model was evaluated and showing outstanding performance of predicting speed with a small deviation between predicted and actual values. The percentage errors are 6% within predicted and actual speed values. As a conclusion, the DL model has a capability in modeling the sharp non-linearity in the road path. This work contributes further improvements of driving behavior for early crash prediction. In future, the model will be expanded to a real track, and the performance of the model will be evaluated. The model further can be used in the decision-making process for an autonomous driving vehicle.

References

1. Zhang, J., Wang, F.Y., Wang, K., Lin, W.H., Xu X., Chen, C.: Data-driven intelligent transportation systems: a survey. *IEEE Trans. Intell. Transp. Syst.* **12**(4), 1624–1639 (2011). <https://doi.org/10.1109/tits.2011.2158001>
2. Schroedl, S., Zhang, W.: Predicting driving speed using neural networks. In: *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, vol. 1, pp. 402–407 (2003). <https://doi.org/10.1109/itsc.2003.1251985>

3. Zhu, X., Yuan, Y., Hu, X., Chiu, Y.C., Ma, Y.L.: A Bayesian network model for contextual versus non-contextual driving behavior assessment. *Transp. Res. Part C Emerg. Technol.* **81**, 172–187 (2017). <https://doi.org/10.1016/j.trc.2017.05.015>
4. Polson, N.G., Sokolov, V.O.: Deep learning for short-term traffic flow prediction. *Transp. Res. Part C: Emerg. Technol.* **79**, 1–17 (2017). <http://doi.org/10.1016/j.trc.2017.02.024>
5. Dyrud, L., Jovancevic, A., Brown, A., Wilson, D., Ganguly, S.: Ionospheric measurement with GPS: receiver techniques and methods. *Radio Sci.* **43**(6), 1–11 (2008). <http://doi.org/10.1029/2007RS003770>

Chapter 57

Temporal Patterns Analysis of Paddy Production in Sri Lanka



N. B. W. I. Udeshika and T. M. J. A. Cooray

Abstract Rice is the staple food of Sri Lanka and the most important agriculture sub-sector in the country. Within the last five years paddy production showed a different pattern when compared with the past 25 years. Hence the modelling recent patterns are more complex than in the past. Most of the studies on paddy production have been carried out based on descriptive statistics using short term data or based on regression analysis. But it would be more worth to do an analysis based on time series such as univariate and multivariate, because the production indicates diverse behaviours with time. Forecasting paddy production is important for planning purposes and making import policies of the government. Seasonal paddy production data from 1952 to 2015 were essentially used for this study. Time series plots, ACF and PACF graphs were inspected to identify the trend, seasonal effect and stationary of the series. Box-Cox Transformation method was used to modify the distributional shape of the data series so as to apply time series theories. Two diverse time series approaches were developed to understand the past trends in paddy production with the purpose of forecasting future production. Mean Absolute Percentage Error (MAPE) was used to measure the prediction accuracy of two models. SARIMA model was given better forecast to the actual values when compared with VEC model.

Keywords Seasonal autoregressive moving average (SARIMA) model • Vector error correction model (VEC) • Mean absolute percentage error (MAPE) • Co-integration test

N. B. W. I. Udeshika (✉) • T. M. J. A. Cooray
University of Moratuwa, Moratuwa, Sri Lanka
e-mail: imaliudeshika88@gmail.com

T. M. J. A. Cooray
e-mail: cooray@uom.lk

1 Introduction

1.1 Background of the Study

Rice is the main crop cultivated by the majority of farmers in rural areas and it is the staple food of approximately 20 million inhabitants in Sri Lanka. Hence, rice sector makes a significant contribution to the economy of Sri Lanka. About 30% of the crop sector contribution to the agricultural GDP (Gross Domestic Product) is from rice sector. Approximately 800,000 farm families, which are about 20% of the population, depend on paddy cultivation for their livelihood [1–3].

In Sri Lanka paddy cultivation is mainly divided into two seasons (time periods) known as “Maha” and “Yala” which are associated with the two monsoons. Maha season is the main season in paddy cultivation associated with the north-east monsoon during the period of September to March. Yala is the secondary season which is associated with south-west monsoon during the period of May to August [1]. However, the whole area devoted for paddy is not being cultivated due to number of reasons such as shortage of water during the seasons, prevailing unsettled conditions on the ground, etc.

1.2 Rice Cultivation in Sri Lanka

Rice is the single most vital crop occupying 34% (0.77/million hectares) of the total cultivated area in Sri Lanka. On average 560,000 hectares are cultivated during Maha and 310,000 hectares during Yala making the average annual extent sown with rice to about 870,000 hectares. About 1.8 million farm families are engaged in paddy cultivation island-wide. Sri Lanka currently produces 2.7 million tonnes of rough rice annually and satisfies around 95% of the domestic requirement [1–3].

It is projected that the demand for rice will increase at 1.1% per year and to meet this, the rice production should grow at the rate of 2.9% per year [3]. Increasing the cropping intensity and national average yield are the options available to achieve this production target [1].

2 Literature Review

The study on ‘Forecasting Paddy production in Sri Lanka’ [4] investigated the past, present and future trends of paddy production in Sri Lanka and a time series model is developed to detect the long term trend and prediction for future changes of paddy production for the three leading years. According to the study “Sense and non sense rice price controls in Sri Lanka “which was done based on the rice prices

from 2005–2008 univariate time series models have modeled and the relationship between the rice prices and the paddy prices have been measured by the correlation coefficient [5].

Relationship between climatic factors and rice production in Sri Lanka is analyzed by the study “Some Agroclimatological Aspects of Rice Production in Sri Lanka” [6]. As a result rainfall has been found as the foremost factor controlling the cultivation system. Further the relationships between sown and harvested acreage, yield and rainfall are measured based on correlation and regression analysis.

3 Methods

3.1 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

Log transformation is applied to the data series to remove the non constant variance of the series. Four SARIMA models are developed for log transformed series. Numerous statistics such as AIC, SBC, R2, DW...etc. are used to identify the most adequate model among formed SARIMA models. Validity of the assumptions of the fitted model are checked by considering results of the hypothesis tests specifically; Box–Pierce test, Serial Correlation LM Test and Histogram Normality Test.

3.2 Vector Error Correction (VEC) Model

Integration order of the secondary data series is investigated to build up a multivariate approach for the paddy production. Consequently Vector Error correction model (VEC) is fitted including disequilibrium term. VEC Lag exclusion wald test, Portmanteau Test for Autocorrelations and Lagranch Multiplier test are used to examine the goodness of fit of the formed VEC model.

4 Results and Discussion

4.1 Descriptive Analysis

According to the Fig. 1 in each year, average yield is greater than the final production this can be due to wastage or natural disasters. However the difference is getting reduced compared with the past. Both average yield and production shows an upward trend and a seasonal pattern.

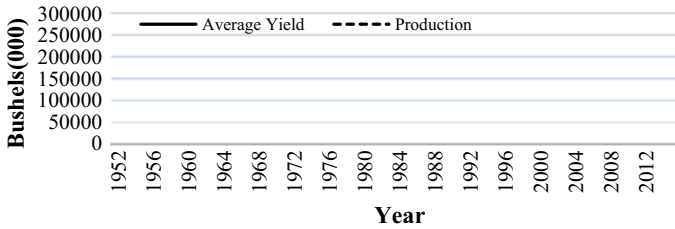


Fig. 1 Paddy production versus average yield

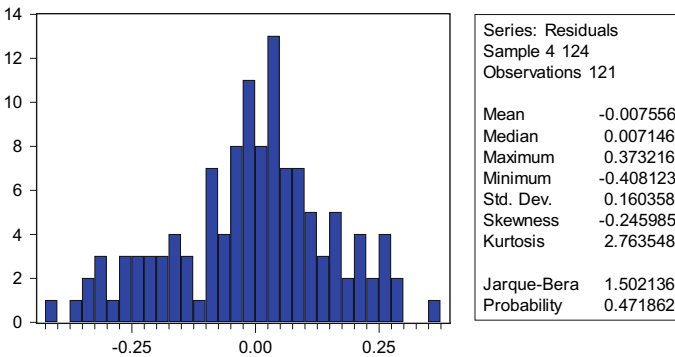


Fig. 2 Histogram of normality test

4.2 Further Analysis

When comparing four models, SARIMA (011)(011)₂ model has taken smallest AIC value, largest R² value. Further DW statistic value of that model is much closer to 2. Q statistic, Normality Test and correlation Test have given insignificant results. Hence SARIMA (011)(011)₂ model can be selected as the most adequate model to capture the trends and seasonal patterns of the paddy production series.

Diagnostic checking of SARIMA (011)(011)₂ Model

When consider the model adequacy of the fitted model, residuals are fairly normally distributed when the few large and small values are not considered. Results of the Correlation LM test is given that no correlations among residuals. Q-statistics of the correlogram of standardized residuals are not significant. Hence there is no serial autocorrelation among residuals of the fitted model.

Table 1 Parameter estimates of SARIMA model

Model	AIC	R2	DW	Q Statistic ^a	Test 1 ^b	Test 2 ^c
SARIMA (011)(010)2	-0.3798	0.4747	1.8767	Sig	Not sig	Sig.
SARIMA (010)(011)2	-0.4210	0.4959	2.8174	Not sig	Not Sig	Sig
SARIMA (011)(011)2	-0.7958	0.6592	1.8811	Not Sig	Not Sig	Not Sig
SARIMA (111)(011)2	-0.7108	0.6249	2.0179	Not Sig	Not Sig	Not Sig

^aHo: No serial correlation up to lag 12 of residual series, H1: There exists serial correlation up to lag 12 of residuals series

^bNormality Test (H0: Residuals are normally distributed, H1: Residuals are not-normally distributed).

^cCorrelation Test (H0: Residuals are uncorrelated, H1: Residuals are correlated).

Model equation of SARIMA (011)(011)₂ Model

$$(1 - B)(1 - B^2) \ln Y_t = (1 - 0.784903B)(1 - 0.967740B^2)e_t \quad (1)$$

Y_t = Paddy production in season t

Validating the requirements for Multivariate Model

The Table 2 illustrates that paddy production(y) and harvested area(y1) logarithmic series have taken equal orders(I(1)) while rain fall (y2) logarithmic series has taken order 0. Therefore only y and y1 variables can be used to build Multivariate time series model.

Test for Cointegration

According to the cointegration rank test at most one cointegration equation exist at 5% level of significance and the cointegration by maximum eigen value indicates same results. Thus there exists a cointegration equations and it implies that variables are cointegrated. A Vector Error correction model has to be fitted including the disequilibrium term. So disequilibrium terms are added to the model as explanatory variables.

The Vector Error Correction Estimates

Diagnostic checking of VEC Model

Residual Portmanteau Tests for Autocorrelations is indicated that residuals of the model are uncorrelated. Residual Portmanteau Tests for Autocorrelations is also concluded the adequacy of the model. Residual plots are indicated that residuals are randomly distributed (not white noise). Apart from the very few data points most of the data points of the correlograms are inside the bandwidth, showed that the auto correlation function support the stationary of the model.

Table 2 Results of unit root test for logarithmic variables (1953–2013)

Variable	Probability of log series	Significant	Order
Paddy production (y)	0.4516	Not Sig.	Cannot identified
Harvested area (y1)	0.2587	Not Sig.	
Rain fall (y2)	0.002	Sig.	I(0)
	Probability of 1st difference of log series		
Paddy production (y)	0.0000	Sig.	I(1)
Harvested area (y1)	0.0000	Sig.	I(1)

Table 3 Results of ECM for LNY & LNY1 (1952–2013)

Variables of error correction model	D(LNY)	D(LNY1)
CointEq 1	0.450530 ^a	0.479370 ^a
D(LNY(-1))	-1.398502 ^a	-0.807413 ^a
D(LNY(-2))	-0.519834 ^a	-0.483595 ^a
D(LNY1(-1))	1.082698 ^a	0.535566 ^a
D(LNY1(-2))	0.796267 ^a	0.760925
C	0.037602 ^a	0.021614

^aSignificant at 5%, LNY = log(y), LNY1 = log(y1).

Table 4 Actual versus forecasted values

Year	Season	Actual value (Ha.000)	Forecast value of SARIMA model (Ha.000)	Forecast value of VEC model (Ha.000)
2014	Yala	1145	1147.086	1412
2014	Maha	2877	2683.481	2456
2015	Yala	1942	1616.341	1764
2015	Maha	2902	2756.867	2433
MAPE			7.16%	15.81%

SARMA model has given the smallest MAPE value.

Model equation of VEC Model

$$d(\ln y) = 0.4505297098 * (\ln y(-1) - 2.723984106 * \ln y1(-1) + 9.17796832) - 1.398502266 * d(\ln y(-1)) - 0.5198337118 * d(\ln y(-2)) + 1.082697997 * d(\ln y1(-1)) + 0.7962672213 * d(\ln y1(-2)) + 0.0376020718 \tag{2}$$

$$d(\ln y1) = 0.4793701563 * (\ln y(-1) - 2.723984106 * \ln y1(-1) + 9.17796832) - 0.8074134735 * d(\ln y(-1)) - 0.4835952083 * d(\ln y(-2)) + 0.5355655228 * d(\ln y1(-1)) \tag{3}$$

Forecasting and Model Comparison

5 Conclusions and Drawbacks

According to the results of two approaches, SARIMA (011)(011)2 model is the most accurate model which can be used to forecast the paddy production of Sri Lanka based on MAPE value. The best model which was developed to paddy production data capture 65.92% of variation ($R^2 = 0.6592$) of the original log series. Since order of the series are equal, paddy production data and harvested area data can be used to develop a multivariate model, however rain fall data cannot be integrated to the model due to uneven order.

The data set consist only 124 data points, if it can increase to a larger value it will be supported to develop long-term forecasting model. The VEC model was fitted using Eviews software and it doesn't has an option to forecast just for next few data points exclusive of all data points of the fitted model. As future works VEC model can be improved with new software.

References

1. Statistics. (n.d.). Retrieved March 2017, from Ministry of Agriculture: <http://www.agrimin.gov.lk/>
2. Paddy Statistics. (n.d.). Retrieved March 2017, from Department of Cencus & Statistics: <http://www.statistics.gov.lk/>
3. Statistics. (2017, March). Retrieved from The Central Bank of Sri Lanka: <http://www.cbsl.gov.lk/>
4. Sivapathasundaram, V., Bogahawatte, C.: Forecasting of Paddy Production in Sri Lanka: A Time Series Analysis. *Trop. Agric. Res.* 21–30 (2012)
5. OG Dayaratna-Banda, J.J.: Sense and Nonsense of Rice Price Controls in Sri Lanka. *Perad. J. Econ.* 2 (2008)
6. Suppiah, R., Yoshino, M. Some agroclimatological aspects of rice production in Sri lanka. *Geogr. Rev. Jpn.* 137–153 (1986)
7. Cooray, T.: Statistical analysis and forecasting of main agriculture output of Sri Lanka: rule-based approach. In: 10th International Symposium. Sabaragamuwa University of Sri Lanka., p. 221 (2006)
8. Kendaragama, K., Bandara, T. (n.d.). Changes in land used patterns in paddy lands. Retrieved March 2017, from, www.goviya.lk
9. Muhammad, M., Abdullah, M.H.: Modelling and forecasting on paddy production in Kelantan under the implementation of system of rice intensification (SRI). *Acad. J. Agric. Res.* 106–113 (2013). <http://dx.doi.org/1.15413/ajar.2013.0112>
10. Sirisena, P.M.T.S., Dammalage, T.: Cultivated paddy area identification and rice yield estimation using free satellite images. Asia Association on remote Sensing. Department of Remote Sensing and GIS, Faculty of Geomatics, Sabaragamuwa University of Sri Lanka (2016)

Chapter 58

The Effects of Awareness Level on Littering Behaviour on Campus: Family Income as Moderator



Mas'udah Asmui, Sharifah Norhuda Syed Wahid,
Suhanom Mohd Zaki, Noorsuraya Mohd Mokhtar
and Siti Suhaila Harith

Abstract Previous studies have revealed that students' awareness on environmental issues is high. In contrast, their positive behaviour towards littering is increasing along some factors, explicitly; facility, campaign, and socio-demographic that were studied with regard to littering awareness level. The socio-demographic variables namely gender, age and religion are literally high related to littering behaviour compared to family income. The objectives of the study are two-folds: first, to ascertain the effects of awareness level on littering behaviour on campus. The second objective is to identify the role of family income as a moderator in the relationship between students' awareness on littering and their littering behaviour. Questionnaires were distributed among students at a selected higher learning institution and the research hypotheses were tested using Structural Equation Modelling. The findings of the study reveal that there is a significant relationship between the students' awareness and littering behaviour on campus. Besides that, family income partially moderates the relationship between the two variables.

Keywords Awareness level · Littering behaviour · Family income

M. Asmui (✉) · S. N. S. Wahid · S. M. Zaki · N. M. Mokhtar · S. S. Harith
Universiti Teknologi MARA, UiTM Cawangan Pahang, Bandar Jengka, Malaysia
e-mail: mas_as@pahang.uitm.edu.my

S. N. S. Wahid
e-mail: sha_norhuda@pahang.uitm.edu.my

S. M. Zaki
e-mail: suhanom@pahang.uitm.edu.my

N. M. Mokhtar
e-mail: noorsuraya@pahang.uitm.edu.my

S. S. Harith
e-mail: ssuhaila@pahang.uitm.edu.my

1 Introduction

Littering is globally recognized as a societal and environmental problem [1, 2]. The social problems related to litter include safety, fire, human health, and indirect health hazards from bacteria, rats, roaches, and mosquitoes that are attracted to litter [3]. Landslides, water pollutions and flash flood are perceived as the most visible sign of environmental problems. It is expected that the amount of solid waste generated in Kuala Lumpur to reach double in the next twenty years; from 3.2 million ton a year today, to 7.7 million tons a year in 2030 [4, 5].

Littering behaviour is affected by many factors such as types of litter, level of education, time and bin location [6–10] as well as littering awareness. Awareness campaign is another factor with could affect littering behaviour [10]. The level of university students' awareness on the issue of waste disposal was very high, yet, their attitudes in waste disposal are a concern, as they did not reach the middle of the scale [11]. Another approach to understand littering is by focusing on the demographic and personal qualities of the person who litters or the "litter bug" [6]. Therefore, the objectives of this study are: (i) to ascertain the effects of awareness level on littering behaviour among university students, (ii) to identify the role of family income as a moderator in the relationship between students' awareness on littering and their littering behaviour.

Gaining a detailed understanding of individuals' different environmentally sensitive behaviours will be important for policy makers as well as researchers who are in search of solutions to the ever-increasing environmental problems that will eventually require human behavioural changes [12]. Therefore, providing information to people on appropriate methods for disposal of waste can be helpful [3, 13].

2 Literature Review

Studies by many researchers have confirmed that littering behaviour tends to be affected by many factors including types of litter, demographic characteristics of litterers, group composition, location, time and environmental indicators [14–17]. Other than that, poor packaging design of commercial products, amount of litter already existed at a particular site, presence and wording of signs referring to litter and the number and or placement and appearance of waste collection bins at the site also contribute to littering habit [14].

Lack of knowledge about the adverse effects on the environment and aesthetic implications has led to littering [18]. Anti-littering campaign via mass media could increase awareness level among residents. Although prior research has shown that such individual-level motivations campaigns typically only produce small changes in behaviour (if any), there is a reason to continue utilizing media messages, and more importantly branding, in litter prevention efforts [6]. At the same time, the environmental education is suggested to be embedded in the school syllabus, as

early as pre-school level, whilst city managers may organized periodical anti-littering campaigns for the public [19, 20]. In addition, passive and active educational activities should also be undertaken where information on the importance of cleanliness is not only conveyed through brochures and distribution of information, but cleaning activities should also be carried out by school children and members of the general public during large events [21]. This is to create a good understanding on the issues related to the environment and an awareness of the consequences of littering [13].

A study conducted in a trade market Malaysia found that family income did not link to littering behaviour [10]. However, a study in San Pablo del Lago, Ecuador, mentioned that income was identified as a predictor of environmental attitudes and behaviour [12]. The higher income group produced 1.54–1.69 kg wastes per month followed by 1.16–1.49 kg per month for middle income group, and 0.93–1.20 kg per month in lower income group [22]. Generally, there is a relationship between family income and litter generation [14]. Higher levels of environmentalism in general are associated with higher income levels [19, 20], whereby higher levels of education are confirmed to be positively associated with environmentalism [12, 23, 24]. This is contrary to a study in San Pablo del Lago where the lower income people had more knowledge about how to dispose of garbage since the waste collection cart did not pass by their areas; they had to find alternative ways to dispose of it [13]. Another finding revealed that the middle and lower socioeconomic group generated more wastes compared to high socioeconomic group [25]. However, the factors of educational level and availability of waste collection cart affected the behaviour of high and low income litterers.

3 Methodology

A set of questionnaire adapted from previous studies [26–28] was constructed for the purpose of data collection. The set was divided into two parts; Part A and Part B. Part A consisted of four sections which focused on facility (four items), campaign (four items), littering awareness (four items) and littering behaviour (five items). A 10-point Likert-scale format ranging from 1 (strongly disagree) to 10 (strongly agree) was used for Part A. Meanwhile, Part B consisted of questions on demographic background of selected respondents. The questionnaire was subjected to a reliability analysis which measured the internal consistency of the items and item-total correlations whereby the reliabilities in the 0.70 range were acceptable and those over 0.80 were good [29, 30]. The scores of the reliability analysis based on a pilot study conducted among 100 respondents indicated that the adapted instrument had good and acceptable reliability with Cronbach's Alpha values of more than 0.70 (facility = 0.890, campaign = 0.790, awareness = 0.883 and behaviour = 0.911).

The same questionnaire was distributed to selected respondents among diploma and degree students at a selected higher learning institution using stratified

sampling technique since the characteristics between the same educational level is homogeneous. The selected students completed the survey during class period in the presence of the researchers to ensure maximum return rate. Attention was given to confidentiality and completing the questionnaire without any help to fulfill the assumption of sample independence. A total of 302 respondents completed the survey.

IBM–Statistical Package for Social Science (IBM–SPSS) software version 24.0 was used to analyze the respondent’s background while Structural Equation Modelling (SEM) through Analysis of Moment Structure (AMOS) software was used to test the hypotheses of this study. If a *p*-value is less than 5% significance level, it means the awareness variable has significant effects to students’ littering behaviour. The following hypotheses will be tested in this study.

- H₁ Awareness has significant effect on positive littering behaviour among students at higher learning institution.
- H₂ Family income moderates the relationship between awareness and positive littering behaviour among students at higher learning institution

Preliminary assumption testing indicated that all the four constructs; facility, campaign, littering awareness and littering behavior were approximately normally distributed since the skewness value is between the acceptable range (−1.5, 1.5) [30]. Meanwhile, SEM analysis needs to test the reliability and validity through unidimensionality, validity and reliability to ascertain or evaluate the fitness of the measurement models [30, 31]. The following Fig. 1 shows the measurement model for confirmatory factor analysis (CFA). Firstly, unidimensionality is achieved since all the paths with standardized regression coefficients have factor loadings at least 0.6 and all measurements are in positive direction. Secondly, validity is accessed through convergent validity, discriminant validity, and constructs validity. Convergent validity is achieved since the average variance extracted (AVE) is at least 0.5, and construct validity is also achieved for each of the fitness indexes for a required level (absolute fit via Root Mean Square of Error Approximation (RMSEA) is not exceeded 0.08, incremental fit via Comparative Fit Index (CFI) is more than 0.9, and parsimonious fit via Chi-Square/Degrees of Freedom (Chisq/df) is less than 3.0. Furthermore, discriminant validity is achieved because the construct is free from redundant items and the highest correlation between exogenous constructs is 0.74, which is not more than 0.85 [30, 31]. Meanwhile, the reliability must satisfy two criteria; composite reliability (CR) achieved for CR at least 0.6 and all AVEs is greater than 0.5.

4 Findings and Discussions

In total, 102 (33.77%) male and 200 (66.23%) female aged between 18 and 23 years old were involved in this study. The result shows that majority of the students’ family income were more than RM3000 (172, 56.95%) and the other 130

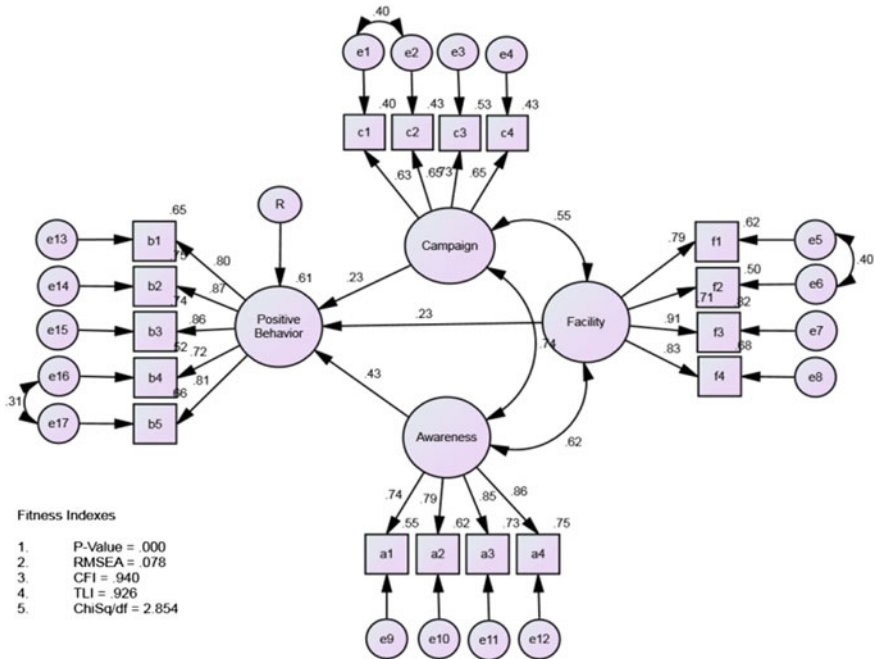


Fig. 1 The Measurement model

(43.05%) were less than RM3000. In addition, this study found that students’ positive littering behaviour is 60.90% affected by campaign activities, facility provided and littering awareness. As stated earlier, the first objective is to ascertain the effect of awareness level on littering behaviour in the campus. Result of the relationship between construct shows that the path between students’ awareness and their good behaviour in littering is also positively significant at 5% level of significance ($\beta = 0.427, p\text{-value} = 0.000$). This significant result shows that awareness become a significant factor regarding to littering [11], as well as facility provided ($\beta = 0.430, p\text{-value} = 0.000$) and campaign activities ($\beta = 0.227, p\text{-value} = 0.012$) [6–11]. Moreover, 61% total variation of students’ positive littering behaviour was influenced by the campaign, facility and awareness as shown in Fig. 1. Meanwhile, the other 39% was predicted by other factors.

The next analysis was to identify the role of family income as moderator on the relationship between students’ awareness on positive littering behaviour and their littering. Students’ awareness has positively significant effect on positive littering behaviour thus, the moderating variable; family income could be tested. Income level gives an impact in littering behaviour since there is existence of a relationship between family income and litter generation [14]. Therefore, the respondents were classified into two different family income groups which were lower and higher

Table 1 The Moderation test for low family income group

	Constrained model	Unconstrained model	Chi-Square differences	Result on moderation
Chi-Square	167.892	163.134	4.758	Significant
Degree of freedom	111	110	1	

Table 2 The moderation test for high family income group

	Constrained model	Unconstrained model	Chi-Square differences	Result on moderation
Chi-Square	219.042	210.579	8.463	Significant
Degree of freedom	111	110	1	

Table 3 Results of the effect of students’ awareness on littering behaviour

Group	Path	Estimate (β)	P
Lower family income	Awareness < —Behaviour	0.561 (0.422)	0.002 ^a
Higher family income	Awareness < —Behaviour	0.508 (0.422)	0.000 ^a

^aSignificance at 5% level of significance.

income groups. The following Tables 1 and 2 show the moderation test results for both family income groups.

The moderation test is significant since the Chi-Square difference between the constrained and unconstrained model is greater than 3.84 [30]. Therefore, it can be concluded that family income either from lower or higher income groups moderated the relationship between students’ awareness and littering behaviour at higher learning institution. This result is consistent with previous study that stated a range of socio-economic factors can affect individual attitudes toward littering in which the monthly income promotes littering behaviour among citizen [14, 22].

Table 3 shows that the type of moderation that interfered in the relationship between the two corresponding variables is partial moderation, since the standardized estimation value for lower and higher family income groups are significant [30].

5 Conclusions and Recommendations

There is no evidence that littering habit differed by sex [18], yet it is found that littering is more common among males [14]. This contradicts to this research result as females of the age of 18–23 years old were the majority of the respondents. In addition, most of the respondents’ family incomes were more than RM3000.00

which indicated that the respondents were in higher income group. Similarly, it is discovered that the wealthier individuals consume more waste than lower income individual [32]. However, both lower or higher family income could affect the students' awareness and their littering behaviour. On the other hand, their positive littering behaviour was high due to lack of littering facility provided, followed by low littering awareness and infrequent campaign activities. This is consistent with a previous study that mentioned inadequate number of trash bins had resulted in overloaded trash around the bins [10].

Ecological management should be improved by adding more trash bins, inserting the elements of environmental education into educational programs and increasing awareness campaigns via mass media. The littering behaviour can be changed through the mediation of the ecological management and litterers jolt controls themselves [10, 26]. Education could nurture behaviour in valuing environmentalism and sustainability among young society. This is due to the important role of the young students and the need for them to be prepared to face future environmental challenges [33–35].

References

1. Ojedokun, O.: Development and structural validation using data from an indigenous (Nigerian) sample. *Management of Environmental Quality: Int J.* **26**(4), 552–565 (2015). <https://doi.org/10.1108/meq-12-2014-0175>
2. Brown, T.J., Ham, S.H., Hughes, M.: Picking up litter: an application of theory-based communication to influence tourist behaviour in protected areas. *J. Sustain. Tour.* **18**, 183–292 (2010)
3. Ayotamuno, J., Gobo, A.: Municipal solid waste management in Port Harcourt Nigeria, obstacles and prospects. *Manag. Environ. Qual.: Int. J.* **15**(4), 389–398 (2004)
4. Mahmud, S.N.D., Osman, K.: The determinants of recycling intention behaviour among the Malaysian school students: an application of theory of planned behaviour. *Procedia-Soc. Behav. Sci.* **9**, 119–124 (2010). <https://doi.org/10.1016/j.sbspro.2010.12.123>
5. Seow, T.W., Md. Jahi, J.M.: Pengurusan sampah sarap di Lembangan Saliran Langat. In: National Seminar on Environmental Issues and Challenges in Malaysia, Universiti Kebangsaan Malaysia, Bangi, 25–26 July (2003)
6. Schultz, P.W., Bator, R.J., Large, L.B., Bruni, C.M., Tabanico, J.J.: Littering in context: personal and environmental predictors of littering behaviour. *Environ. Behav.* **45**(1), 35–59 (2013). <https://doi.org/10.1177/0013916511412179>
7. Steg, L., Vlek, C.: Encouraging pro-environmental behaviour: an integrative review and research agenda. *J. Environ. Psychol.* **29**(3), 309–317 (2009)
8. Beck, R.W.: Literature review-litter, a review of litter studies, attitude surveys and other litter-related literature: report. *Keep Am. Beautiful* (2007)
9. Reams, M.A., Geaghan, J.P., Gendron, R.C.: The link between recycling and litter: a field study. **28**, 92–110 (1996)
10. Asmui, M., Mohd Zaki, S., Syed Wahid, S.N., Mohd Mokhtar, N., Harith, S.S.: Association between litterers' profile and littering behaviour: a chi-square approach. In: Paper Presented at 3rd ISM International Statistical Conference 2016, August (2016)

11. Asmui, M., Mohd Zaki, S., Syed Wahid, S.N., Mohd Mokhtar, N., Harith, S.S.: Islam dan Konsep “Hijau”: Hubungan antara sikap membuang sampah dan tahap kesedaran di kalangan pelajar Institut Pengajian Tinggi Awam. In: Proceeding of Konferensi Akademik, UiTM Cawangan Pahang, (2016)
12. Onel, N., Mukherjee, A.: Analysis of the predictors of five eco-sensitive behaviours. *World J. Sci., Technol. Sustain. Dev.* **11**(1), 16–27 (2014). <https://doi.org/10.1108/wjstsd-08-2013-0031>
13. Aung, M., Arias, M.L.: Examining waste management in San Pablo del Lago, Ecuador: a behavioural framework. *Manag. Environ. Qual.: Int. J.* **17**(6), 740–752 (2006). <https://doi.org/10.1108/14777830610702557>
14. Al-Khatib, I.A., Arafat, H.A., Daoud, R., Shwahneh, H.: Enhanced solid waste management by understanding the effects of gender, income, marital status, and religious convictions on attitudes and practices related to street littering in Nablus-Palestinian territory. *Waste Manag.* **29**, 449–455 (2009)
15. Cialdini, R.B., Reno, R.R.: A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *J. Pers. Soc. Psychol.* **58**(6), 1015–1026 (1990)
16. Gardner, G.T., Stern, P.C.: *Environmental problems and human behaviour*. USA, Allyn & Bacon (1996)
17. Meeker, F.L.: A comparison of table-littering behaviour in two settings: a case for contextual research strategy. *J. Env. Psych.* **17**, 59–68 (1997)
18. Nkwocha, E.E., Okeoma, I.O.: Street littering in Nigerian towns: towards a framework for sustainable urban cleanliness, African research review. *Int. Multidiscip. Res. J. Ethiopio.* **3**(5) (2009)
19. Abdul Latif, S., Omar, M.S., Bidin, Y.H., Awang, Z.: Role of environmental knowledge in creating pro-environmental residents. *Procedia - Soc. Behav. Sci.* **105**, 866–874 (2013). <https://doi.org/10.1016/j.sbspro.2013.11.088>
20. Abdul Latif, S., Omar, M.S., Bidin, Y.H., Awang, Z.: Effect of situational factor on recycling behaviour in determining the quality of life. *J. ASIAN Behav. Stud.* **3**(8), 37–46 (2013)
21. Ong, I.B.L., Sovacool, B.K.: A comparative study of littering and waste in Singapore and Japan. *Resour. Conserv. Recycl.* **61**, 35–42 (2012)
22. World Bank: *What a waste: a global review of solid waste management*. Urban Development & Local Government Unit. Washington, DC: The World Bank (2012)
23. Van Liere, K., Dunlap, R.: The social bases of environmental concern: a review of hypotheses, explanations, and empirical evidence. *Public Opin. Q.* **44**(2), 181–197 (1980)
24. Barr, S.: Factors in influencing environmental attitudes and behaviours. *Environ. Behav.* **39**(4), 435–473 (2007)
25. Khan, D., Kumar, A., Samadder, S.R.: Impact of socioeconomic status on municipal waste generation rate. *Waste Manag.* **49**, 15–25 (2016)
26. Abdul Shukor, F.S., Mohammed, A.H., Awang, M., Abdullah Sani, S.I.: Litter reduction: a review for the important behavioural antecedent approaches. In: 3rd International Conference Business Economic Research Proceeding (2012)
27. Maricopa Association of Governments: *Don't thrash Arizona: Litter evaluation survey*, WWW Document (2012). http://dontrashaz.com/pdf/MAG_2010_Litter-Survey.pdf. Accessed October 2016
28. Wesley High School Otukpo, Benue State Nigeria: *A survey of household solid waste management in Otukpo: a case study of residents around Wesley High School Otukpo, Benue, Nigeria* (2011)
29. Sekaran, U., Bougie, R.: *Research Methods for Business: A Skill Building Approach*. Wiley, London (2010)
30. Zainuddin, A.: *SEM made simple, a gentle approach to learning structural equation modelling*. MPWS Publication Sdn. Bhd, Selangor (2015)
31. Hair, J.F., Black, B.J., Anderson, R.E.: *Multivariate Data Analysis*. Prentice Hall, Upper Saddle River, NJ (2010)

32. Sivakumar, K., Sugirtharan, M.: Impact of family income and size on per capita solid waste generation: a case study in Manmunai North Divisional Secretariat Division of Batticaloa. *J Sci. Univ. Kelaniya* **5**, 13–23 (2010)
33. Jones, N., Roumeliotis, S., Iosifides, T., Hatziantoniou, M., Sfakianaki, E., Tsigianni, E., Thivaïou, K., Billiraki, A., Evaggelinos, K.: Students' perceptions on environmental management of HEIs and the role of social capital: a case study in the University of the Aegean. *Int. J. Sustain. High. Educ.* **14**(3), 278–290 (2013). <https://doi.org/10.1108/ijshe-07-2011-0050>
34. Emanuel, R., Adams, J.N.: College students' perceptions on campus sustainability. *Int. J. Sustain. High. Educ.* **12**, 79–92 (2011)
35. Moody, G.L., Hartel, P.G.: Evaluating and environmental literacy requirement chosen as a method to produce environmentally literate university students. *Int. J. Sustain. High. Educ.* **8**, 355–370 (2007)

Chapter 59

The Trends of Age and Gender Specific Mortality Rates by Ethnic Groups



Saiful Azril Ishak, Syazreen Niza Shair,
Wan Nor Ayunni Wan Ahmad Shukiman, Nurazliyana Mat Radzi
and Nur Salbiah Abdul Rahman

Abstract Continuing increases in life expectancy poses budgetary challenges for the government in particular to ensure pension benefits are sufficient to meet a growing demand from elderly. In this research the mortality and life expectancy at birth trends of each ethnic group including Malay, Chinese and Indian are closely examined by age, gender and ethnic groups from 1991 to 2011. Results showed that the mortality of all ethnic groups, males and females are generally decreased over the years, however we found that the rate of decreases were varied by sub-groups. A comparison between ethnicity indicates Chinese mortality rates are consistently lower than that of Malay and Indian groups, leading to the highest life expectancy at births. Furthermore, the mortality rates of females are lower than males for all ethnic groups.

Keywords Mortality rates · Life expectancy · Pension benefits · Malaysia ethnicity · Trends

S. A. Ishak (✉) · S. N. Shair · W. N. A. W. A. Shukiman · N. M. Radzi · N. S. A. Rahman
Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: sazril@tmsk.uitm.edu.my

S. N. Shair
e-mail: syazreen@tmsk.uitm.edu.my

W. N. A. W. A. Shukiman
e-mail: wnayunni@gmail.com

N. M. Radzi
e-mail: nurazliyanamatrdzi@gmail.com

N. S. A. Rahman
e-mail: nursalbiah@gmail.com

1 Introduction

A decrease in mortality rates has resulted in increasing population survival globally. Like most developed nations, an increase number of developing countries including Malaysia are currently undergoing changes in population age structure so called population ageing. In 2017, the Malaysia population is estimated at 32.0 million with 28.7 million are citizens and 3.3 million are non-citizens [1]. Furthermore, Malaysian population is heavily weighed by the Malays which consists of 50.1% of the total population followed by Chinese, 29.9% and Indians, 6.9% [2]. In view of the diversity of Malaysian population, the ageing process in this country is unique as the rates of decline in mortality are varied by ethnic groups.

The segregation of mortality data by ethnicity suggests that there are differences in mortality rates. This is due to each ethnic group has different lifestyles and historical backgrounds. Since 1957, Malaysia's independence marked a critical ethnicity differences. Most Malays are found working in traditional agricultural sector and the government, while the Chinese dominated the commercial manufacturing, production, mining, construction and Indian worked in rubber estates. In 1969, the government of Malaysia introduced New Economic Policy (NEP) with the aim to ensure equal wealth distribution and equality in the economy for all ethnic groups [3]. This policy has successfully improved the economic status as well as the mortality of population.

The socioeconomic differentials among ethnic groups may have an impact on the mortality patterns. Thus, it is of interest to this research to study the mortality patterns by ethnicity and other variables including age and gender. This research maybe useful in the context of pensions and long-term care in which the length of benefits payment is different according to the survival of each ethnic group. Data for each ethnicity will be segregated by genders and six main age groups: infants, Under 5, adolescents, adults, middle age and elderly. The mortality trends for each group will be closely examined.

2 Methods

Data from Department of Statistic from year 1991 to 2011 were used with division was made by ethnic groups including Malays, Chinese, and Indians. Each ethnicity data was further divided into six (6) group of ages which are Infant (1–12 month), Under 5 (1–years old), adolescent (5–19 years old), adults (20–39 years old), middle ages (40–64 years old) and elderly (65+ years old).

The mortality rates of each ethnic group by age were calculated using *the death/exposure formula*. The formula is as below:

$$q_x = \frac{d_x}{l_x}$$

Where

- q_x = The mortality rate between x and $x + 1$
- d_x = The number of deaths between x and $x + 1$
- l_x = The total exposure at exact group of ages.

Since we want to identify new mortality rates, we also need to calculate new total exposure at exact group of ages. The formula of total exposure, l_x as below:

$$l_x = \frac{\{(1.05) * l_x \text{ male}(\text{groupofages})\} + \{(1) * l_x \text{ female}(\text{groupofages})\}}{2.05}$$

Where

- l_x = The new total exposure at exact group of ages
- $l_x \text{ male}$ = The number of male survives at exact group of ages
- $l_x \text{ female}$ = The number of female survives at exact group of ages
- 1.05 = The ratio rate of male for 1991 until 1999
- 1 = The ratio rate of female for 1991 until 2011
- 2.05 = The total number of ratio male and female.

3 Result and Discussion

3.1 The Mortality Rates of Ethnic Groups by Gender

Based on Fig. 1, the trends of mortality rate by ethnicity show significant decline from year 1991 to 2011. In overall, Chinese has the lowest mortality rate compared to Malay and Indian for both males and females. The comparison between genders shows that female mortality is consistently lower than male and decline at faster rate than male mortality. It is noteworthy that the Malay male mortality pattern is almost constant with the rate of decline is only 2.11%, compared to Chinese male 18.5% and Indian male 16.9%. The minimum rate of decline in mortality among Malay male maybe due to high road traffic deaths occurs in adult group. For females, a significant 19.3% drop in mortality rates recorded for Indian female compared to 10.4% for Malay female and 12.7% for Chinese female.

These findings show that the overall mortality trend for males and females are declining. The declining trends in mortality rates are also occurred in other developing countries such as China, South Korea and Taiwan rates [4]. Interestingly, the Malay mortality decline at slower rate compared to other ethnic groups. It was also found that Indian female mortality decline at faster rate than Chinese and Malay female. Reduction in mortality will directly impact the increase

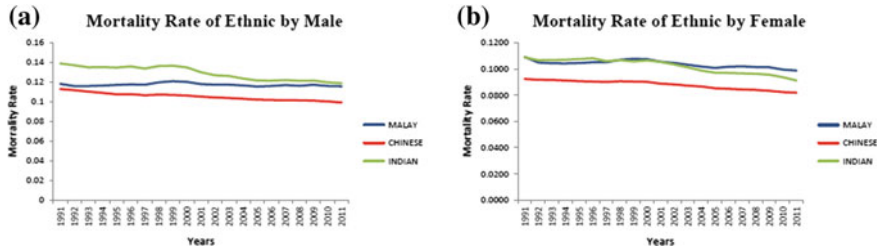


Fig. 1 The mortality rates of Malay, Chinese and Indian by genders: males (a) and females (b), from year 1991 to 2011

in life expectancy. Longer living people require more savings to be made prior to retirement. However, the question whether people live longer in healthy condition or require high demand in healthcare, is debatable.

3.2 *The Mortality Rates of Ethnic Groups by Age*

Overall, the trends of mortality by age and ethnicity show a declining pattern. Chinese still has the lowest mortality rate among all specific age groups followed by Malay and Indian. Figure 2a shows that infants mortality rates decrease significantly for all ethnic groups from year 1991 to 2001 and remain stable after that. More efforts should be done by the government to ensure the mortality rates of infants will continue decreasing in the future. The mortality trend Fig. 2b exhibits a downward trend which indicate a positive mortality improvement over the years for young children ages 1–4 years old. It is noteworthy that the mortality rates of adolescent group in Fig. 2c show no improvement in early years from 1991 to 1997 then decrease continuously years after. The main causes of death for this adolescent group include road injury, HIV, suicide, lower respiratory infections and interpersonal violence which occur among the ages 10 to 19 years [5].

The trends in adult mortality can be seen in Fig. 2d. The mortality rates for Malay and Chinese adults increase in previous years from year 1991 to 2000 and decrease in later years, after 2000. The mortality rates for Indian adults show almost constant pattern with humps occurs throughout the period which might imply a sudden increase in number of death for certain years. During this adult period, people are more active and productive as they have just started working and earned for living. Figure 2e exhibits continuous downward trends with Chinese mortality remains lower than Indian and Malay mortality. However, the rate of decline for this group is slower compared to infants, under 5 and adolescent groups. Finally, the mortality trends by ethnicity for elderly group can be referred to Fig. 2f. This figure shows that the mortality rates of Indian elderly improve substantially faster than Malay and Chinese elderly after year 2000 and becomes better than the Malay elderly after year 2006.

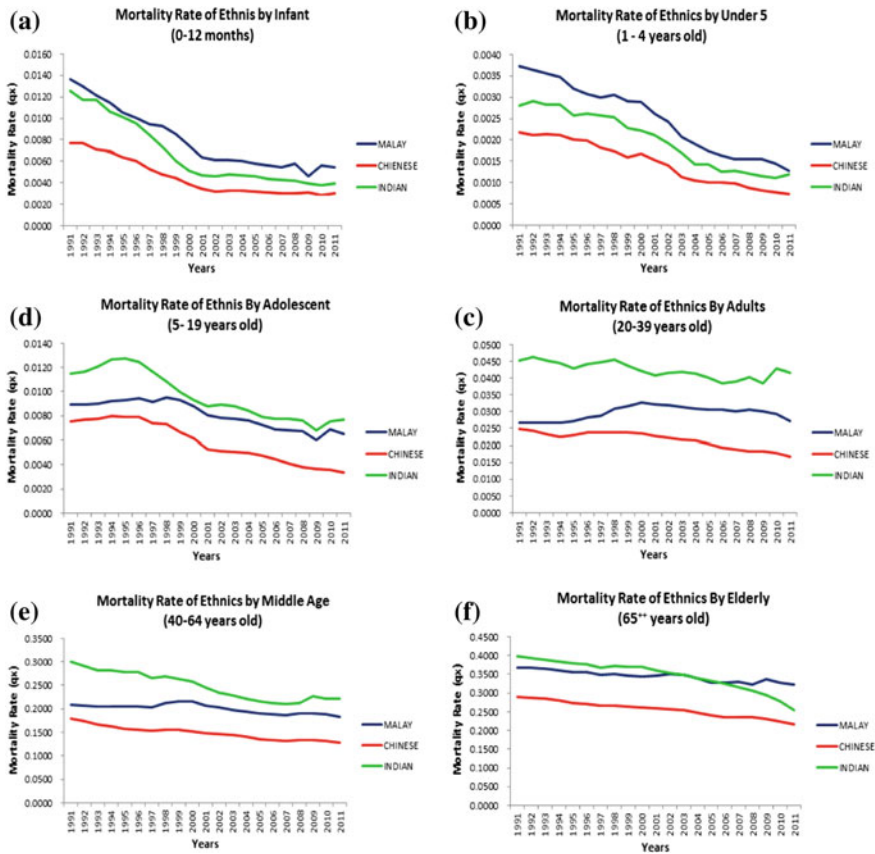


Fig. 2 The mortality rates of Malay, Chinese and Indian by different age groups: Infants (a), Under 5 (b), Adolescent (c), Adults (d), Middle Ages (e) and Elderly (f), from 1991 to 2011

The improvement in mortality generally indicates there are progress in health-care development as results of initiatives from the government to continue upgrading the hospitals and healthcare services and increase the access to urban and rural areas. Apart from that, continuous improvement in medication and medical equipment helps in reducing the mortality rates [6]. Improvement also contributed by the availability of private clinics and hospitals which allowing the access for better healthcare. Based on the results, health care systems need to improve continuously so that the rates of decline in mortality of particular groups such as adults, middle ages and elderlies can be further reduced.

Malaysian is an ageing society hence requires careful estimation of funding to support the living expenses after retirement. There is possibility that money saved during working lifetime is not sufficient to protect against this longevity risks. Thus, awareness about the need to save more for retirement is essential. As the trend for male and female mortality decline over years, this difference indicate that female

will live longer than male. This might affect the issue of equality in income and job availability for female as they also need to accumulate enough wealth for retirement. Further, the effect of inflation might have taken its toll as pension amount will remain constant throughout the years. Indirectly, government need to step up their plan to ensure sufficient funds are available in future to pay for pension benefit.

4 Conclusion

The purpose of this study is to analyze the trends of Malaysian mortality rates by age, gender and ethnic groups. In overall, the mortality rates of sub-groups declined steadily throughout the year 1991–2011. The mortality rates for Chinese were consistently lower than Malay and Indian for both genders and all age groups. The death rates of ethnic groups have generally been decreasing at different rates. The mortality of Malay males decreased at the lowest rate compared to other ethnic males whereas the mortality of Indian females decreased substantially faster than Malay and Chinese females. The differential in mortality can also be seen among different age groups. Significant mortality improvement occurred in infants, under 5 and adolescent groups. Nonetheless, the mortality rates of older age groups such as adults, middle ages and elderly reduced at slower rates than the younger age groups.

References

1. Bujang, M.A., Hamid, A.M., Zolkepali, N.A., Hamedon, N.M., Lazim, S.S., Haniff, J.: Mortality rates by specific age group and gender in Malaysia: trend of 16 years. 1995–2010. *J. Health Inform. Dev. Ctries.* **6**, 521–529 (2012)
2. Ibrahim, R.I., Siri, Z.: Analysis of mortality trends by specific ethnic groups and age group in Malaysia. In: *Proceeding of 21st National Symposium on Mathematical Sciences*. AIP Publishing (2014). <https://doi.org/10.1063/1.4887727>
3. Mean Monthly Gross Household Income by Ethnic Group, Strata and State, Malaysia, 1970–2014, Economic Planning Unit, Prime Minister Department, Malaysia
4. Nagaraj, S., Tey, N.-P., Ng, C., Balakrishnan, B.: Ethnic dimensions of gender differentials in mortality in Malaysia. *J. Popul. Res.* **25**(2), 183–206 (2008)
5. *Population and Demography, current population estimates (2016–2017)*, Department of Statistics, Malaysia
6. World Atlas, Ethnic Group of Malaysia. <http://www.worldatlas.com/articles/ethnic-groups-of-malaysia.html>

Part IV

Application

Chapter 60

A Markov Chain Model for Diabetes Mellitus Patients



**Muhammad Rozi Malim, Faridah Abdul Halim,
Farah Wahidah Md Aris, Nur Musrsyiddah Azizuddin,
Raihannah Othman, Siti Nur Azyyati Rosli
and Siti Fairuz Kamaruzaman**

Abstract Diabetes mellitus is one of the most common chronic diseases that directly contributed to 1.5 million deaths in 2012. Complementary and alternative medicines (CAMs) can be defined as a diverse medical and health care practices or products that are not generally classified as a part of conventional medicine. A number of researchers agreed that it is important for the health and medical practitioners to explore the use of CAMs so that they can educate the patients about the benefits of alternative medicines. The objective of this study is to construct a Markov chain model that represents the status of diabetes mellitus patients after using CAMs in a long run. A dataset from UKM Medical Molecular Biology Institute (UMBI), collected since 2005 with a total of 105,892 participants, was used to construct the model. From the results, this study has concluded that the consumption of CAMs showed a positive effect in curing diabetes mellitus in the long run.

Keywords Markov chain · Diabetes mellitus · Complementary and alternative medicines · Limiting distribution

1 Introduction

Diabetes mellitus is one of the most common chronic diseases in all countries and continues to increase as changing lifestyles lead to reduced physical activity and increased obesity. The term “diabetes mellitus” is defined as metabolic disorder of multiple etiologies characterized by chronic hyperglycemia with disturbance of

M. R. Malim (✉) · F. A. Halim · F. W. Md Aris · N. M. Azizuddin · R. Othman ·
S. N. A. Rosli · S. F. Kamaruzaman
Faculty of Computer and Mathematical Sciences, UiTM, 40450 Shah Alam, Malaysia
e-mail: rozi@tmsk.uitm.edu.my

F. A. Halim
e-mail: faridahh@tmsk.uitm.edu.my

carbohydrate, fat and protein metabolism [1]. It occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin. Hyperglycemia (raised blood sugar) is a common effect of uncontrolled diabetes and over time leads to serious damage to the body's systems, especially the nerves and blood vessels. An estimated 1.5 million deaths were directly caused by diabetes and another 2.2 million were attributable to high blood glucose in 2012. Almost half of all deaths attributable to high blood glucose occur before the age of 70. World Health Organization (WHO) has predicted that diabetes will be the 7th leading cause of death in 2030.

Diabetes mellitus has been classified into several types [1]; Type-1 diabetes (insulin-dependent), Type-2 diabetes (non-insulin dependent), Gestational diabetes, impaired glucose tolerance (IGT) and impaired fasting glycemia (IFG). Type-1 diabetes occurs when the pancreas fails to produce insulin; children and adolescents are the most frequent patients. Type-2 diabetes results from the body's inability to react properly to the insulin produced by the pancreas; it occurs most frequently among adults and the chance of acquiring this type of diabetes increases with age. Gestational diabetes is hyperglycemia with blood glucose values above normal but below those diagnostic of diabetes, occurring during pregnancy. IGT and IFG are intermediate conditions in the transition between normality and diabetes. There are several symptoms that may arise such as thirst, blurring of vision, and weight loss. In a long term of diabetes mellitus, it may effects specific complications of retinopathy with potential of blindness, nephropathy that may lead to renal and features of autonomic dysfunction, including sexual dysfunction. People with diabetes mellitus are in high tendency of having cardiovascular, peripheral vascular and cerebrovascular disease.

Based on the National Centre for Complementary and Integrative Health (NCCIH), complementary and alternative medicines (CAMs) can be defined as a diverse medical and health care practices or products that are not generally classified as a part of conventional medicine [2]. The effectiveness of CAMs is studied separately depending on the type of diseases. Molassiotis et al. [3] conducted a survey on 956 patients from 14 countries out of 33 European countries. The results showed that CAMs are widely used by colon, breast and prostate cancer patients at a rate of 70.2%. In pediatric patients, the rates of using CAMs are equally high (ranging from 32.7% in United Kingdom and 84% in United States). The main reasons that the cancer patients decided to turn to CAMs are their realization in improvements of physical and psychosocial well-being and increasing hope of getting healthier or better in their health condition. A large number of CAMs therapies contain the herbal teas, vitamins, minerals, homeopathy, and relaxation techniques. Although the variety of herbs used is vary by country, the CAMs therapy has become the most popular among cancer patients. Another research has suggested and agreed that the use of CAMs is necessary and became a secular trend in the United States [4]. Some physicians and nurses also put a role in giving information on benefits of CAMs. The researchers agreed that it is important for the health and medical practitioners to explore the use of CAMs so that they can

educate the patients with benefits and being healthy by using this kind of alternative medicines.

One type of CAMs is chromium. Each of us needs to consume chromium in a small amount for daily diet; a trace mineral that could be found in whole-grain breads and some vegetables which helps in the glucose metabolism [2]. It is necessary to consume chromium as it can improve the diabetes control by helping insulin to improve its action in human body [5]. In addition, ginseng and exercise are also categorized as CAMs. Ginseng gives an effect of lowering glucose in the body as it can reduce blood sugar, meanwhile exercise can help patients with diabetes by controlling their weights (related to the factors of diabetes) and lowering the blood sugar with physical activities such as running, cycling and swimming.

The objective of this study is to construct a Markov chain model for the status (absent, present) of diabetes mellitus patients after using CAMs in a long run.

2 Methodology

2.1 Dataset and Description

A secondary dataset was obtained from a medical research institute as the case study. The data were collected since 2005 with a total of 105,892 participants in the age range of 40–65 years. The descriptions of variables are summarized below (Table 1).

2.2 Method of Analysis

A two-state Markov chain model is considered as follows [6]:

Let $P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$, $0 < a, b < 1$, be the transition matrix of a two-state Markov chain (states 0 and 1).

When $a = 1 - b$, the two rows are the same, then states 0 and 1 are independently identically distributed random variables with $P(X_n = 0) = b$ and P

Table 1 Descriptions of variables of dataset

Variable name	Description
Diabetic status (two types of status)	Non-diabetic and diabetic
Age	Age of participants
Types of complementary alternative medicines (CAMs)	Energy therapy; Manipulative body practice; Mind body; Natural products; Others

$(X_n = 1) = a$. When $a \neq 1 - b$, the probability distribution for X_n varies depending on the outcome of the previous stage.

For a two-state Markov chain, by induction, the n -step transition matrix can be stated as follows;

$$P^n = \frac{1}{a+b} \begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b} \begin{bmatrix} a & -a \\ -b & b \end{bmatrix} \tag{1}$$

Note that $|1 - a - b| < 1$ when $0 < a, b < 1$, and thus $|1 - a - b|^n \rightarrow 0$ as $n \rightarrow \infty$, and from (1) we have

$$\lim_{n \rightarrow \infty} P^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix} \tag{2}$$

This tells us that, in the long run, the system will be in state 0 with probability $\frac{b}{(a+b)}$ and in state 1 with probability $\frac{a}{(a+b)}$.

3 Analysis and Results

R software was used for the data analysis. From the dataset, the transition probability of diabetes mellitus status, calculated from the comparison of diabetes mellitus status during recruitment and follow-up appointment after five years, is as follows:

	No	Yes
No	0.997	0.003
Yes	0.056	0.944

Hence, after 10 years (two transitions), 15 years (three transitions), and 20 years (four transitions), we have the following probability matrices:

$$\text{After 10 years, } P^2 = \begin{bmatrix} 0.9942 & 0.0058 \\ 0.1087 & 0.8913 \end{bmatrix}.$$

$$\text{After 15 years, } P^3 = \begin{bmatrix} 0.9915 & 0.0085 \\ 0.1583 & 0.8417 \end{bmatrix}.$$

$$\text{After 20 years, } P^4 = \begin{bmatrix} 0.9890 & 0.0110 \\ 0.2049 & 0.7951 \end{bmatrix}.$$

Therefore, we can conclude that the probability of absent (state 0) of diabetes mellitus after 10 years, 15 years, and 20 years of using complementary alternative medicines (CAMs), given present of diabetes mellitus as the current state, are 0.1087, 0.1583, and 0.2049, respectively. Although the probability values are small, it is enough to show the positive effect of CAMs in curing diabetes mellitus. Comparing the five-year and 10-year periods, the probability has increased from 0.056 to 0.1087 (almost doubled). If we compare 10-year and 15-year periods, the probability has increased from 0.1087 to 0.1583 (an increase of 46%). Finally, a comparison between 15-year and 20-year periods, the probability has increased from 0.1583 to 0.2049 (an increase of 29%).

Using (2), the limiting distribution for the status of diabetes mellitus can be calculated as follows:

$$\lim_{n \rightarrow \infty} P^n = \begin{vmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{vmatrix} = \begin{vmatrix} \frac{0.056}{0.003+0.056} & \frac{0.003}{0.003+0.056} \\ \frac{0.056}{0.003+0.056} & \frac{0.003}{0.003+0.056} \end{vmatrix} = \begin{vmatrix} 0.949 & 0.051 \\ 0.949 & 0.051 \end{vmatrix}.$$

Based on the above matrix, in the long run, the status of diabetes mellitus will be in state 0 with probability 0.949 and the state 1 with probability 0.051, irrespective of the initial status. In other words, in the long-run, approximately 94.9% of the participants (using CAMs) are absent in having diabetes mellitus under the assumptions on the model. However, the term “long-run” poses a question.

4 Conclusion

This study aims to construct a Markov chain model that represents the status of diabetes mellitus patients after using complementary alternative medicines (CAMs) in a long run. A dataset from UKM Medical Molecular Biology Institute (UMBI) was used to construct the model. From the results (model), we can conclude that the use of CAMs showed a positive effect in curing diabetes mellitus (indirectly) in the long run. This is in lines with the results found by other studies [3, 4]. However, the probability of recovering from the disease has increased rather slowly after a long period of time. The term “long-run” raises a question; it could be too long for the patients.

With the fact that CAMs is able to cure diabetes mellitus patients, health and medical practitioners need to do more in-depth medical studies to find ways to make CAMs more effective over a shorter period of time. A further study can be done by clustering CAMs based on their types to observe whether the probability of cure from diabetes mellitus varies across their types.

References

1. WHO: World Health Organization: definition and classification of diabetes mellitus and its complications. Report of a WHO consultation, part 1: diagnosis and classification of diabetes mellitus, Geneva, World Health Organization (1999)
2. Complementary and alternative medicine for diabetes. Retrieved from University of Rochester Medical Center (URMC). <https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=134&ContentID=166> (2016)
3. Molassiotis, A., Fernandez-Ortega, P., Pud, D., Ozden, G.: Use of complementary and alternative medicine in cancer patients: a European survey. *Ann. Oncol.* (2005)
4. Kessler, R., Davis, R., Foster, D.: Long-term trends in the use of complementary and alternative medical therapies in United States. *Ann. Intern. Med.* **135**(4), 262–268 (2001)
5. Complimentary and alternative therapies for diabetes. Retrieved from WebMD. <http://www.webmd.com/diabetes/complementary-and-alternative-diabetes-treatments#1> (2016)
6. Pinsky, M. A., Karlin, S.: An introduction to stochastic modeling, 4th edn. Academic Press (2010). ISBN-10: 0-12-381416-2

Chapter 61

A New Method to Forecast TAIEX Based on Fuzzy Time Series with Trapezoidal Fuzzy Numbers and Center of Gravity Similarity Measure Approach



Siti Musleha Ab Mutalib, Nazirah Ramli, Daud Mohamad
and Norhuda Mohammed

Abstract Various forecasting methods based on fuzzy time series (FTS) have been proposed. However, most of the models used discrete fuzzy sets as a basis for calculating the forecasted values and thus cannot provide the forecasted range under different degree of confidence. A few FTS models used trapezoidal fuzzy numbers (TrFNs) for calculating the forecasted values and produces the forecasted values in term of TrFNs. However, in order to calculate the forecasting accuracy, the forecasted values in terms of TrFNs must be defuzzified. During the defuzzification process, some information has lost from the data which contributed to the inability to grasp the sense of uncertainty that has been kept throughout the forecasting process. In this paper, we propose a new FTS forecasting method based on FTS, TrFNs and center of gravity (COG) similarity measure approach. The COG approach is used to get the forecasting accuracy and to preserve the information that has been kept from being lost. The advantage of the proposed method is that the defuzzification process does not involve.

Keywords Fuzzy time series · Forecasting · Trapezoidal fuzzy numbers · Center of gravity

S. M. Ab Mutalib (✉) · N. Ramli
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
26400 Bandar Jengka, Pahang, Malaysia
e-mail: musleha78@gmail.com

N. Ramli
e-mail: nazirahr@pahang.uitm.edu.my

D. Mohamad · N. Mohammed
Faculty of Computer and Mathematical Sciences, Universiti of Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
e-mail: daud@tmsk.uitm.edu.my

N. Mohammed
e-mail: hudamohammed@pahang.uitm.edu.my

1 Introduction

Forecasting is a process of making prediction about future performance based on the existing historical data. Many methods have been proposed by the previous researchers such as grey Markov model [1], neural network [2] and Box Jenkins model [3]. These methods are categorized as traditional forecasting model whereby they only cater for the data in numeric form, require more historical data and the data must follow normal distribution. In order to cater that problem in traditional method, in 1993, [4] proposed fuzzy time series (FTS) with application to forecast the enrollment of student at University of Alabama. Based on model in [4], many improvement and modification have been made for length [5] and types of fuzzy logical relationship (FLR) [6–8].

In this paper, we propose a new FTS forecasting model with trapezoidal fuzzy numbers (TrFNs) and similarity measure based on COG. The TrFNs is used to represent the linguistic values instead of using discrete fuzzy set as appear in many previous FTS model. Meanwhile, the similarity measure based on center of gravity (COG) is selected to calculate the forecasting performance. The propose model produces the forecasted range under different degree of confidence (DDoC) which is consistent with human intuition. This paper is organized as follows. Section 2 presents the proposed FTS forecasting model based on TrFNs and COG similarity measure approach. Section 3 illustrates a numerical example of the proposed method by using the data of Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX). Discussion and conclusion are presented in Sects. 4 and 5 respectively.

2 A New Method for Forecasting the TAIEX Based on FTS with TrFNs and COG Similarity Measure Approach

In this section, we propose a new method for forecasting the TAIEX based on concept of FTS from [5, 9, 10] and COG similarity measure from [11]. The proposed fuzzy forecasting method is described as follow:

Step 1: Collect the historical data, D_t .

Step 2: Define the universe of discourse, UD . Find the minimum data D_{\min} and the maximum data D_{\max} from the historical data D_t . In order to partition the UD , choose the appropriate positive integer k_1 and k_2 . Thus, the universe of discourse UD is defined as $UD = [D_{\min} - k_1, D_{\max} + k_2]$.

Step 3: Determine the appropriate length of interval l and partition the universe of discourse, UD . The universe of discourse is partitioned by using average based length approach from [12]. The procedures are as follows:

- i. Calculate the absolute difference between values D_{t+1} and D_t whereby $t = 1, 2, \dots, n - 1$ and calculate the average of the difference between data.
- ii. Take half of the average as the length.
- iii. Determine the range and base for the length in (ii) and rounded the length by using the base mapping from [12].
- iv. Calculate the number of interval using formula $m = \frac{(D_{\max} + k_2 - D_{\min} + k_1)}{l}$ where l is the length after rounded.
- v. Partition the universe of discourse UD using the length interval obtained in (iv). Assume there are m intervals which are $u_1 = [d_1, d_2], u_2 = [d_2, d_3] \dots, u_{m-1} = [d_{m-1}, d_m]$ and $u_m = [d_m, d_{m+1}]$.

Step 4: Develop new TrFNs based on the intervals obtained in Step 3(v) to represent the linguistic terms as $A_1 = (d_0, d_1, d_2, d_3), A_2 = (d_1, d_2, d_3, d_4), \dots, A_{m-1} = (d_{m-2}, d_{m-1}, d_m, d_{m+1}), A_m = (d_{m-1}, d_m, d_{m+1}, d_{m+2})$.

Step 5: Fuzzify the historical data D_t . If the value of historical data is located in the range of u_m , then it belongs to fuzzy number A_m .

Step 6: Establish the fuzzy logical relationship (FLR) and FLR group based on FTS concept from [5, 9, 10].

Step 7: Calculate the forecasted value F_t based on [13] and COG similarity measure between forecasted and actual values based on [11].

3 Numerical Example

In the following, an example is used to illustrate the proposed forecasting method to forecast the TAIEX for November 2004 to December 2004.

Step 1: The historical data of TAIEX were taken from [14]. Some of the TAIEX data are shown in Table 1.

Table 1 The historical data of TAIEX for November 2004

Date	Actual	Fuzzy numbers	Date	Actual	Fuzzy numbers
1/11/2004	5656.17	A_1	16/11/2004	5910.85	A_{14}
2/11/2004	5759.61	A_6	17/11/2004	6028.68	A_{20}
3/11/2004	5862.85	A_{12}	18/11/2004	6049.49	A_{21}
4/11/2004	5860.73	A_{12}	19/11/2004	6026.55	A_{20}
5/11/2004	5931.31	A_{15}	22/11/2004	5838.42	A_{10}
8/11/2004	5937.46	A_{15}	23/11/2004	5851.1	A_{11}
9/11/2004	5945.2	A_{16}	24/11/2004	5911.31	A_{14}
10/11/2004	5948.49	A_{16}	25/11/2004	5855.24	A_{11}
11/11/2004	5874.52	A_{12}	26/11/2004	5778.65	A_7
12/11/2004	5917.16	A_{14}	29/11/2004	5785.26	A_8
15/11/2004	5906.69	A_{14}	30/11/2004	5844.76	A_{11}

Step 2: According to the historical data in [14], the minimum data D_{\min} and the maximum data D_{\max} of TAIEX are 5656.17 and 6139.69 respectively. By choosing the two proper integer as $k_1 = 16.17$ and $k_2 = 0.31$, then the universe of discourse UD is defined as $UD = [5640, 6140]$.

Step 3: The average based length with base mapping table is implemented to determine the appropriate length of interval which is 20. By using the length of 20, the number of intervals obtained is 25 which are $u_1 = [5640, 5660]$, $u_2 = [5660, 5680]$... $u_{24} = [6100, 6120]$ and $u_{25} = [6120, 6140]$.

Step 4: Then, the new TrFNs can be defined as $A_1 = (5620, 5640, 5660, 5680)$, $A_2 = (5640, 5660, 5680, 5700)$, ..., $A_{24} = (6080, 6100, 6120, 6140)$, $A_{25} = (6100, 6120, 6140, 6160)$.

Step 5: Fuzzify the TAIEX data. For example, the TAIEX data for 16/11/2004 is 5910.85 and located at range $u_{14} = [5900, 5920]$. Therefore, the corresponding fuzzy number is assigned as A_{14} . The corresponding fuzzy numbers of the TAIEX data is A_{14} shown in Table 1.

Step 6: After fuzzifying the data, the FLR is created and further, the FLR group is generated. The FLR group is presented in Table 2.

Step 7: By using the heuristic rules, the forecasted value F_t is calculated. Then, the actual values and the forecasted values are normalized by dividing each value with 10000. Table 3 shows the similarity measure between the actual and forecasted values.

Table 2 The FLR group of TAIEX

Group 1:	$A_1 \rightarrow A_6$	Group 10:	$A_{15} \rightarrow A_{13}, A_{15} \rightarrow A_{15},$
Group 2:	$A_6 \rightarrow A_{12}$		$A_{15} \rightarrow A_{16}$
Group 3:	$A_7 \rightarrow A_8$	Group 11:	$A_{16} \rightarrow A_{12}, A_{16} \rightarrow A_{16}$
Group 4:	$A_8 \rightarrow A_{11}, A_8 \rightarrow A_{12}$	Group 12:	$A_{18} \rightarrow A_{18}, A_{18} \rightarrow A_{19}$
Group 5:	$A_{10} \rightarrow A_{11}$	Group 13:	$A_{19} \rightarrow A_{18}, A_{19} \rightarrow A_{19},$
Group 6:	$A_{11} \rightarrow A_7, A_{11} \rightarrow A_8,$		$A_{19} \rightarrow A_{23}$
	$A_{11} \rightarrow A_{14}$	Group 14:	$A_{20} \rightarrow A_{10}, A_{20} \rightarrow A_{21}$
Group 7:	$A_{12} \rightarrow A_{12}, A_{12} \rightarrow A_{13},$	Group 15:	$A_{21} \rightarrow A_{20}$
	$A_{12} \rightarrow A_{14}, A_{12} \rightarrow A_{15}$	Group 16:	$A_{23} \rightarrow A_{24}$
Group 8:	$A_{13} \rightarrow A_{14}$	Group 17:	$A_{24} \rightarrow A_{25}$
Group 9:	$A_{14} \rightarrow A_{11}, A_{14} \rightarrow A_{12},$		
	$A_{14} \rightarrow A_{14}, A_{14} \rightarrow A_{15},$		
	$A_{14} \rightarrow A_{19}, A_{14} \rightarrow A_{20}$		

Table 3 Forecasted value of TAIEX for November 2004

Date	Normalize actual value	Normalize forecasted value	Similarity measure
1/11/2004	(0.562, 0.564, 0.566, 0.588)		
2/11/2004	(0.572, 0.574, 0.576, 0.578)	(0.572, 0.574, 0.576, 0.578)	1.0000
3/11/2004	(0.584, 0.586, 0.588, 0.590)	(0.584, 0.586, 0.588, 0.590)	1.0000
4/11/2004	(0.584, 0.586, 0.588, 0.590)	(0.587, 0.589, 0.591, 0.593)	0.9963
5/11/2004	(0.590, 0.592, 0.594, 0.596)	(0.587, 0.589, 0.591, 0.593)	0.9963
8/11/2004	(0.590, 0.592, 0.594, 0.596)	(0.589, 0.591, 0.593, 0.595)	0.9991
9/11/2004	(0.592, 0.594, 0.596, 0.598)	(0.589, 0.591, 0.593, 0.595)	0.9967
10/11/2004	(0.592, 0.594, 0.596, 0.598)	(0.588, 0.590, 0.592, 0.594)	0.9951
11/11/2004	(0.584, 0.586, 0.588, 0.590)	(0.588, 0.590, 0.592, 0.594)	0.9951
12/11/2004	(0.588, 0.590, 0.592, 0.594)	(0.587, 0.589, 0.591, 0.593)	0.9987
15/11/2004	(0.588, 0.590, 0.592, 0.594)	(0.591, 0.593, 0.595, 0.597)	0.9967
16/11/2004	(0.588, 0.590, 0.592, 0.594)	(0.591, 0.593, 0.595, 0.597)	0.9967
17/11/2004	(0.600, 0.602, 0.604, 0.606)	(0.591, 0.593, 0.595, 0.597)	0.9886
18/11/2004	(0.602, 0.604, 0.606, 0.608)	(0.591, 0.593, 0.595, 0.597)	0.9865
19/11/2004	(0.600, 0.602, 0.604, 0.606)	(0.600, 0.602, 0.604, 0.606)	1.0000
22/11/2004	(0.580, 0.582, 0.584, 0.586)	(0.591, 0.593, 0.595, 0.597)	0.9865
23/11/2004	(0.582, 0.584, 0.586, 0.588)	(0.582, 0.584, 0.586, 0.588)	1.0000
24/11/2004	(0.588, 0.590, 0.592, 0.594)	(0.580,0.581,0.583,0.585)	0.9893
25/11/2004	(0.582, 0.584, 0.586, 0.588)	(0.591, 0.593, 0.595, 0.597)	0.9893
26/11/2004	(0.574, 0.576, 0.578, 0.580)	(0.580, 0.581, 0.583, 0.585)	0.9935
29/11/2004	(0.576, 0.578, 0.580, 0.582)	(0.576, 0.578, 0.580, 0.582)	1.0000
30/11/2004	(0.582, 0.584, 0.586, 0.588)	(0.583, 0.585, 0.587, 0.589)	0.9987

4 Discussion

For evaluating the forecasting performance, we develop new approach instead of using the MAPE, MSE or RMSE accuracy. In this FTS forecasting model, the forecasting performance is evaluated based on similarity measure of COG type. Most of FTS forecasting model that used TrFNs, calculate the forecasting accuracy by using MAPE, MSE or RMSE whereby the defuzzification process is involved. During the defuzzification process, some information has lost from the data which contributed to the inability to grasp the sense of uncertainty that has been kept throughout the forecasting process.

For each year, we calculate the degree of similarity between the actual value and the forecasted value. Based on the definition of similarity measure in [11], the value for each fuzzy numbers must in between 0 to 1. Thus, the normalization process is needed. In the TAIEX case study, for each fuzzy number, the value is divided by 10000. The actual value and forecasted value after normalization are shown in Table 3. Based on Table 3, the results show that for each day, the degree of

similarity between actual value and forecasted value is closed to 1. The average degree of similarity for the TAIEX data in Cheng et al. [14] of proposed model is 0.99634. This value indicates that the actual value and forecasted value is closed and similar. This similarity value is used as a forecasting performance to replace the common accuracy. During the calculation of forecasting performance using similarity measure concept, the defuzzification process does not involved.

5 Conclusion

In this paper, we propose a new FTS forecasting model based on TrFNs, second order FTS and COG of similarity measure to forecast the TAIEX data. This new FTS forecasting model produces the forecasted range under DDoC which is consistent with human intuition. The forecasted range under DDoC gives useful knowledge for decision analyst and decision makers in planning strategic decisions. Besides, the forecasting performance in terms of degree of similarity is unique since it preserves the usage of the TrFNs in FTS forecasting whereby the information is preserved through the forecasting procedure. Thus, the fuzzy similarity measure is suitable to evaluate the performance as it can preserve the uncertainty of the data.

Acknowledgements This research is supported by Ministry of Education Malaysia (MOE) and Universiti Teknologi MARA Malaysia (UiTM) under the Research Grant No. 600-RMI/RAGS 5/3 (149/2014).

References

1. Li, X., Chen, W.: A Grey-Markov predication for unemployment rate of graduates in China. In: *Grey Systems and Intelligent Services*, pp. 619–624 (2009)
2. Aladag, C.H., Yolcu, U., Egrioglu, E., Dalar, A.Z.: A new time invariant fuzzy time series forecasting method based on particle swarm optimization. *Appl. Soft. Comput.* **12**(10), 3291–3299 (2012)
3. Mahipan, K., Chutiman, N., Kumphon, B.: A forecasting model for Thailand' s unemployment rate. *Mod. Appl. Sci.* **7**(7), 10–16 (2013)
4. Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series—Part I. *Fuzzy Sets Syst.* **54**, 1–9 (1993)
5. Jilani, T.A., Muhammad, S., Burney, A., Ardil, C.: Fuzzy metric approach for fuzzy time series forecasting based on frequency density based partitioning. *Intl. J. Comput. Electr. Control Inf. Eng.* **4**(7), 39–44 (2010)
6. Chen, S.-M.: Forecasting enrollments based on fuzzy time series. *Fuzzy Sets Syst.* **81**, 311–319 (1996)
7. Kuo, I.-H., Horng, S.-J., Kao, T.-W., Lin, T.-L., Lee, C.-L., Pan, Y.: An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. *Expert Syst. Appl.* **36**(3), 6108–6117 (2009)

8. Cheng, S.-H., Chen, S.-M., Jian, W.-S.: A novel fuzzy time series forecasting method based on fuzzy logical relationships and similarity measures. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, no. 1, pp. 2250–2254 (2015)
9. Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series—part II. *Fuzzy Sets Syst.* **62**, 1–8 (1994)
10. Chen, M.-Y.: A high-order fuzzy time series forecasting model for internet stock trading. *Futur. Gener. Comput. Syst.* **37**, 461–467 (2014)
11. Xu, Z., Shang, S., Qian, W., Shu, W.: A method for fuzzy risk analysis based on the new similarity of trapezoidal fuzzy numbers. *Expert Syst. Appl.* **37**(3), 1920–1927 (2010)
12. Huarng, K.: Effective lengths of intervals to improve forecasting in fuzzy time series. *Fuzzy Sets Syst.* **123**, 387–394 (2011)
13. Cheng, C., Wang, J., Li, C.: Forecasting the number of outpatient visits using a new fuzzy time series based on weighted-transitional matrix. *Expert Syst. Appl.* **34**(4), 2568–2575 (2008)
14. Cheng, S., Chen, S., Jian, W.: Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. *Inf. Sci. (Ny)* **327**, 272–287 (2016)

Chapter 62

A Proposed Conceptual of Derivative Games Based Learning



Ainon Syazana Ab Hamid, Izni Syamsina Saari,
Samsiah Abdul Razak and Aslina Omar

Abstract This concept paper explores certain components that should be combined in the process of designing and developing derivative game. The proposed components include integration of pedagogy, elements of DGBL and ARCS Model in design and development which may be considered in deploying the DGBL approach into the learning of undergraduate calculus in the curriculum setting of Universiti Teknologi MARA Malaysia. The implication of this study can be as a criterion to develop the derivative games and effectiveness of digital games as a motivational tool for students in order to explore and learn the conceptual understanding of differentiation related to function as stated in calculus education syllabus.

Keywords Derivative game · DGBL and ARCS model · Undergraduate calculus

1 Introduction

Due to decreasing performance of Mathematics among students, the teaching and learning of calculus has been hotly debated which has also led to wide series of research conducted on the issue during the last decade. As one of the most crucial

A. S. Ab Hamid (✉) · I. S. Saari
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA,
Melaka Branch, Alor Gajah Campus, 78000 Alor Gajah, Melaka, Malaysia
e-mail: ainonsyazana@melaka.uitm.edu.my

I. S. Saari
e-mail: izni_syamsina@melaka.uitm.edu.my

S. A. Razak
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA,
Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia
e-mail: samsi179@perak.uitm.edu.my

A. Omar
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA,
Johor Branch, Pasir Gudang Campus, 81750 Masai Johor, Malaysia
e-mail: aslina2316@johor.uitm.edu.my

concepts in learning calculus, derivation is a topic that entails students' strong grasp of its basic concept. At the same time, learning derivation concept also assists students integrate with other bodies of knowledge too [1, 2]. Due to high conceptual understanding needed, the derivative is always seen as the root of calculus, making it a complicated topic among students as it also involves function, differentiation, quotient and limit [3, 4]. However, many students struggle to make sense of derivation itself [5] and it is caused by their lack of conceptual understanding [6].

In order to help students to overcome their difficulties in learning concept of derivation involving function, new ways to exploit online games must be formed. According to Prensky [7], online games create a new culture habits and interests of students. Hence, new alternatives online games to understand the difficult concept of derivation must be formed [8]. Therefore, this concept paper will explore certain elements that should be combined in the process of designing and developing derivative game. It is also to verify that derivative online games can be effectively integrated into education system.

2 Digital Game—Based Learning (DGBL)

Digital game-based learning (DGBL) is a motivation of games that combined with curricular contents [9]. DGBL is competitive in nature where students set the acquisition of knowledge development of cognitive skills and simulation forms which may practice their knowledge virtually [10]. Digital game-based learning studies of e-learning environments focus on how to optimize the processing of information contained in multimedia or hypermedia documents [11, 12]. It also called as an interactive pedagogy game that deals with applications that have defined learning outcomes to clarify the difficult subject matter, complex process of understanding with the gameplay, the ability of the player to retain and apply to the real world [13].

3 Implementation of DGBL and Its Effect

Digital learning games are seen as an acceptable form of supplementary learning in Higher Education [14] and an effective method that has the ability to generate new and critical insights on future of development planning in urban networks [15]. Game Based Learning contributes to significantly better performance in student learning who used the games compared to the use of traditional [16, 17]. Results show the use of the game impacted positively on students' knowledge achievement [18]. Frequently-cited arguments held by these researchers for using digital game in education are: can invoke intense engagement in learners [19, 20] and encourage active learning or learning by doing [21]. Empirical evidence also states that games can be effective tool for enhancing learning and understanding of complex subject

matter [22], hence the use of games in learning may impose high level of intrinsic motivation and scores [23] as well as foster collaborations among learners [24].

4 Previous Studies of DGBL in Mathematics

There have been numerous studies conducted pertaining to online games in education especially Mathematics. One of the distinctive past studies is using handheld technology in Mathematics education including calculators, laptops, tablets as well as smartphones [25]. Aydin [26] claimed that studies on approaches of teaching and learning Mathematics for tertiary level should be conducted provided that [27] have concluded that the applications of technologies must result in positive and significant impact towards Mathematics studies.

Another study discovered that the use of software to project dynamic graphics assisted students to visualise the steps involved in mathematics as well as provide meaning to the abstract concept [28]. Moreover [29] showed the findings that students who used the games become more confident of their mathematical abilities and more likely to think that mathematics is also useful outside of the classroom. In addition [30] studied the impacts of a mathematical game on students who learned via games resulted in higher intrinsic motivation, self-efficacy and task involvement.

5 Pedagogy and DGBL

The integration of pedagogical theory of learning and the elements of DGBL are crucial in order to make sure that the design and development of online games perfectly suits the needs of our education. With regard to this, it must be known that there is no specific procedure in joining these two components of pedagogy and DGBL. According to Hamizul and Rahimi [31], three elements that are present in the pedagogy component in DGBL namely (a) Learning objective which is vital in ensuring the accomplishment of learning process, (b) Curriculum needs which may vary from one country to another such as in terms of Outcome Based Education (OBE) as used by Universiti Teknologi MARA in Malaysia as well as other aspects including the alignment of curriculum review and syllabus subject approach for each course. Finally, (c) Learning theory is the process of selecting theories of learning that match with the educational needs to boost the effectiveness of learning process.

5.1 *DGBL Element*

Digital Game-Based Learning (DGBL) refers use of computer games that possess educational value or different kind of software applications that use games for learning and education purposes [32]. Therefore, the game must adhere to learners' needs as well as relevant to their educational needs and also match with various types of learners, user-friendly and hassle-free [33]. Moreover, essential game characteristics that contribute to this engagement are the development of various skills, such as critical thinking and problem-solving skills [34]. Based on Malone [35], in order to ensure online game helps in learning process, the key characteristics of an engaging game should provide (a) rules; (b) goals and objectives; (c) outcomes and feedback; (d) conflict, competition, challenge, or opposition; (e) interaction; and (f) a storyline. Moreover, [36] believes that the key to developing a good game and also a good learning experience is an engaging storyline.

5.2 *ARCS Model*

According to Keller [37], the ARCS model is a system for improving the motivational appeal in which lessons and courses are designed through digital games. There are four categories that characterize human motivations; Attention, Relevance, Confidence, Satisfaction [38] (Table 1).

Table 1 ARCS model [38]

ARCS	Description
Attention	The motivational concern is to stimulate and sustain the learner's attention; the instructor can introduce the learner's curiosity and interest
Relevance	The learner is more likely to be motivated if the content of the instruction responds to his or her perceived needs. Nevertheless, relevance does not have to come from the instructional content
Confidence	Learners are more motivated when challenge is balanced in such a way that the learning process is either too easy or too difficult such that success seems impossible
Satisfaction	Learners feel good about the consequences and if the learners' efforts are consistent with their expectations, they will continue to be motivated towards learning

6 Conclusion

Digital game-based learning plays a great potential for enriching the education of digital natives. A good game can facilitate good learning due to immediate feedback, adjustable difficulty, and gameplay style. It also an interactive pedagogy game that deals with applications and specific learning outcomes to clarify the difficult subject matter, complex process of understanding with the gameplay especially design to target persistent problem areas in undergraduate calculus education. The proposed components which are the integrating pedagogy, DGBL elements and ARCS model in design and development could be used to implement the DGBL approach in undergraduate calculus learning for focusing on University Technology Mara Malaysia curriculum setting. However, future research should also be done especially on the effectiveness of digital games as a motivational tool for students in order to explore and learn the conceptual understanding of differentiation related to function as stated in calculus education syllabus.

Acknowledgements This research is funded by Academic & Research Assimilation (ARAS) that is managed by the Research Management Institute, Universiti Teknologi MARA, Malaysia (600-RMI/DANA 5/3/ ARAS (0171/2016).

References

1. Tall, D.: Conceptual foundations of the calculus. In: Proceedings of the Fourth International Conference on College Mathematics Teaching, pp. 73–88 (1992)
2. Tall, D.: Introducing three worlds of mathematics. *Learn. Math.* **23**(3), 29–33 (2004)
3. Thompson, P.W.: Students, functions, and the undergraduate curriculum. In: Research in Collegiate Mathematic Education, pp. 21–44 (1994)
4. Zandieh, M.: A Theoretical Framework for Analyzing Students' Understanding of the Concept of Derivative. *CBMS Issues in Mathematics Education*, vol. 8, pp. 103–127 (2000)
5. Paramenswaran, R.: On understanding the notion of limits and infinitesimal quantities. *Int. J. Sci. Math. Educ.* **5**, 193–216 (2007)
6. Tall, D.: Looking for the bigger picture. *Learn. Math.* **31**(2), 17–18 (2011)
7. Prensky, M.: *Digital Game-Based Learning*. McGraw-Hill (2001)
8. Papastergiou, M.: Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* **52**, 1–12 (2009)
9. Prensky, M.: Digital game-based learning. *ACM Comput. Entertain.* **1**(1), 1–4 (2003)
10. Erhel, S., Jamet, E.: Digital game-based learning: impact of instructions and feedback on motivation and learning effectiveness. *Comput. Educ.* **67**, 156–167 (2013)
11. Clark, R.E., Mayer, R.E.: *E-learning and the Science of Instruction*, 2nd edn. Jossey-Bass, San Francisco (2008)
12. Mayer, R.E.: Cognitive theory of multimedia learning. In: Mayer, R.E. (ed.) *The Cambridge Handbook of Multimedia Learning*, pp. 31–48. Cambridge University Press, New York (2005)
13. Perrotta, C., Featherstone, G., Aston, H., Houghton, E.: *Game-Based Learning: Latest Evidence and Future Directions*. NFER, Slough (2013)

14. Connolly, T.M., Stansfield, M.H., Hainey, T.: An application of games based learning within software engineering. *Br. J. Edu. Technol.* **38**(3), 416–428 (2007)
15. Mayer, I.S., Carton, L., de Jong, M., Leijten, M., Dammers, E.: Gaming the future of an urban network. *Futures* **36**(3) (2004)
16. Kim, S., Chang, M.: Computer games for the math achievement of diverse students. *J. Educ. Technol. Soc.* **13**(3), 224–232 (2010)
17. Papastergiou, M.: Exploring the potential of computer and video games for health and physical education: a literature review. *Comput. Educ.* **53**, 603–622 (2009)
18. Vahed, A.: The tooth morphology board game: an innovative strategy in tutoring dental technology learners in combating rote learning. In: Conference: 2nd European Conference on Games Based Learning (ECGBL) (2008)
19. Malone, T.W.: Toward a theory of intrinsically motivating instruction. *Cogn. Sci. Multidiscip. J.* **5**(4), 333–369 (1981)
20. Lloyd, P.: Rieber: seriously considering play: designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educ. Technol. Res. Dev.* **44**(2), 43–58 (1996)
21. Garris, R., Ahlers, R., Driskell, J.E.: Games, motivation, and learning: a research and practice model. *Simul. Gaming* **33**(4), 441–467 (2002)
22. Ricci, K., Salas, E., Cannon-Bowers, J.A.: Do computer-based games facilitate knowledge acquisition and retention? *Mil. Psychol.* **8**(4), 295–307 (1996)
23. Liu, M., Horton, L., Olmanson, J., Toprac, P.: A study of learning and motivation in a new media enriched environment for middle school science. *Educ. Technol. Res. Dev.* **59**, 249–265 (2011)
24. Kaptelin, V., Cole, M.: Individual and collective activities in educational computer game playing. 303–316 (2002)
25. Dunn, P., Richardson, A., Oprescu, F., McDonald, C.: Mobile-phone-based classroom response systems: students' perceptions of engagement and learning in a large undergraduate course. *Int. J. Math. Educ. Sci. Technol.* **44**(8), 1160–1174 (2013)
26. Aydin, Y.: The effects of problem based approach on student's conceptual understanding in a university mathematics classroom. *Procedia Soc. Behav. Sci.* 704–707 (2014)
27. Cheung, A., Slavin, R.E.: The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: a meta-analysis. *Educ. Res. Rev.* **9**, 88–113 (2013)
28. Kidron, I., Zehavi, N.: The role of animation in teaching the limit concept. *Int. J. Comput. Algebr. Math. Educ.* **9**(3), 205–227 (2002)
29. Morgan, P., Ritter, S.: An Experimental Study of the Effects of Cognitive Tutor® Algebra I on Student Knowledge and Attitude. Carnegie Learning, Inc., Pittsburgh, PA (2002)
30. Spotnitz, S.: Intrinsic motivation in students with learning disabilities as examined through computer based instruction in mathematics (2001)
31. Hamizul, M.: Nik Mohd Rahimi: design and development of Arabic online games—a conceptual paper. *Procedia Soc. Behav. Sci.* **174**, 1428–1433 (2015)
32. Tang, S., Hanneghan, M.B., El-Rhalibi, A.: Introduction to games-based learning. In: Connolly, T.M., Stansfield, M.H., Boyle, E. (eds.) *Games-Based Learning Advancement for Multisensory Human Computer Interfaces: Techniques and Effective Practices* (2009)
33. Moschini, E.: Designing for the smart player: usability design and user centered design in game-based learning. *Digit. Creat.* **17**, 140–147 (2006)
34. McFarlane, A., Sparrowhawk, A., Heald, Y.: Report on the educational use of games: an exploration by TEEM of the contribution which games can make to the education process (2002)
35. Thomas, W.: Malone: toward a theory of intrinsically motivating instruction. *Cogn. Sci. Multidiscip. J.* **5**(4), 333–369 (1981)

36. Klaila, D.: Game-Based E-Learning Gets Real. Learning Circuits (2001)
37. Keller, J.M.: The use of the ARCS model of motivation in teacher training. In: Shaw, K., Trott, A.J. (eds.) Aspects of Educational Technology Volume XVII: Staff Development and Career Updating. Kogan Page, London (1984)
38. Keller, J.M.: Development and use of the ARCS model of motivational design. J. Instr. Dev. **10**(3), 2–10 (1987)

Chapter 63

AHP Ranking of CSR Human Resource Theme of Takaful Operators



Shahida-Farhan Zakaria and Abd-Razak Ahmad

Abstract A happy employee is a great asset to any organization. Implementing Corporate Social Responsibility (CSR) initiatives on employees is one of the ways to make an employee happy. The aim of this study is to rank the importance of CSR initiatives that had been offered by takaful operators to their employees. A content analysis of the annual reports of 11 takaful operators in Malaysia for the year 2014 was undertaken to examine the CSR initiatives disclosed. There are four groups with a total of 20 CSR activities ranked. The highest priority vector for each group—*human-self, intellect, posterity and wealth*—are *provision of healthy & safe workplace* (0.317), *the existence of employee training and development programs* (0.404), *providing staff home ownership scheme* (0.473) and *policies on the company's remuneration schemes* (0.493), respectively. Takaful operator denoted by TJ ranked highest with an average priority value of 0.315. The findings from this study will contribute as a reference point for Takaful operators to prioritize their CSR initiatives according to Shariah.

Keywords Analytic hierarchy process · Corporate social responsibility · Human resource · Takaful

1 Introduction

Corporate Social Responsibility initiatives are defined as business practices or policies that are based on ethical values and respect for all stakeholders including the employees [2]. For Islamic institutions, the implementation of CSR must fulfill

S.-F. Zakaria (✉) · A.-R. Ahmad
Universiti Teknologi MARA Kedah, Bedong, Malaysia
e-mail: shahidafarhan@kedah.uitm.edu.my

A.-R. Ahmad
e-mail: ara@uitm.edu.my

the objectives of *Shariah*. Darus et al. [3] and Said et al. [5] produce CSR frameworks based on the principles which adhere to Islamic rules and ethics. Darus et al. classify CSR initiatives into four dimensions—*community*, *environment*, *marketplace* and *workplace*—to form an Islamic CSR framework.

Said et al. [5] on the other hand classify CSR initiatives into 5 main groups—*Environment*, *Community*, *Human Resource*, *Energy* and *Product Development*. These are further categorized into *faith*, *human self*, *intellect* and *wealth*. Using the initiatives classified under the *product development* group as listed by Zakaria et al. [7] rank them using analytic hierarchy process (AHP). The results show that the preferred initiatives are ‘*developing products that have safety features*’, ‘*disclosing products that meet safety standards*’, and ‘*training employees in developing safety standards for processes*’. Syed-Noh et al. [6] on the other hand, rank the *community* initiatives of Said et al. The initiatives, in order of preference, are ‘*aiding medical research*, ‘*aiding disaster victims*, ‘*giving donations*, ‘*community activities (within the vicinity)*, and ‘*community outreach program*’. The least important is ‘*sponsoring conferences and seminars*’.

Similarly Ahmad et al. [1] use AHP to rank the preference of *Workplace* CSR initiatives as listed by Darus et al. The results showed that the initiatives that are highly ranked are ‘*fardhu ain tazkirah session*’, ‘*medical benefit for immediate family members*’, ‘*fringe benefit, like entitlement to comprehensive medical benefit or takaful protection*’, ‘*entitlement to special leave to visit the elderly, parents or attending own children’s school activities*’ and ‘*free biennial medical check-ups*’.

The current work extends the work of Said et al. and complements the works of Zakaria et al. [7], Syed-Noh et al. [6] and Ahmad et al. [1]. Using initiatives under the category of *human resource* as classified by Said et al., we not only rank the initiatives, but went a step further by using the average priority vector of each initiative to rank takaful operators in Malaysia in term of their CSR contributions towards their employees. Work by Muhamat et al. [4] concluded that takaful operators in Malaysia are socially responsible and financially profitable.

This paper proceeds as follows. The next section describes the methodological approach, followed by the discussion on the results. The last section concludes.

2 Methodological Approach

The aim of this study is to rank the importance of CSR initiatives that were offered by takaful operators to their employees. The ranking process was done using analytic hierarchy process (AHP). AHP requires respondents to answer pair wise comparison questionnaires between two CSR initiatives at a time. There are five main groups of CSR initiatives that were identified by Said et al. [5]. The groups are *Environment*, *Community*, *Human Resource*, *Energy* and *Product Development*.

This research work ranks the initiatives that fall under the *human-resource* group only. The initiatives refer to practices and policies that are able to raise productivity and improve the working condition and safety of employees [5]. The initiatives are further categorized into 4 smaller sub-groups—*faith, human self, intellect* and *wealth*. Figure 1 shows the ranking process. The process which involve pair-wise comparisons was done within each sub-group. Each initiative will have a priority vector as a result. The higher the priority vector the higher the preference for the initiative is. A content analysis of the annual reports of 11 takaful operators in Malaysia for year 2014 was undertaken to examine the CSR initiatives disclosed. The average sum of the priority vectors for all initiatives collected for each takaful operator will determine how the takaful operators are ranked.

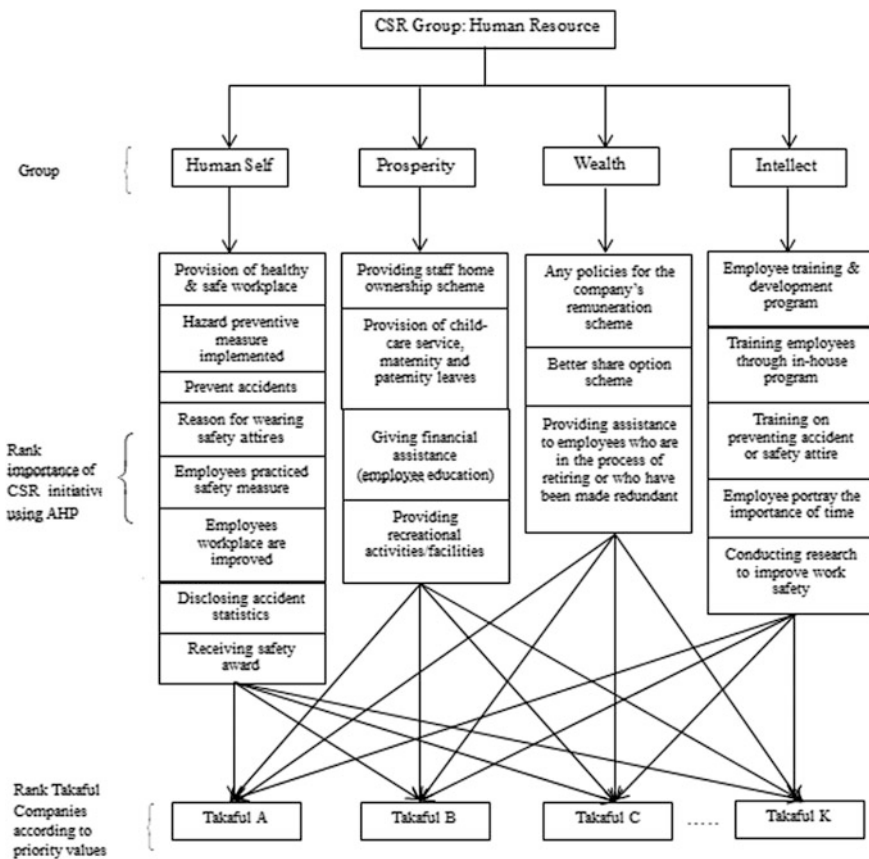


Fig. 1 Flow of the ranking process

3 Results

We aim to evaluate and rank takaful operators in Malaysia based on the sum of the average priority vector calculated for each CSR initiative undertaken by the operators. The first step is to calculate the priority vectors and the result for each of the initiative is given in column three of Table 1.

CSR 1	Provision of healthy & safe workplace	CSR 12	Employee portray the importance of time
CSR 2	Hazard preventive measures are been implemented	CSR 13	Conducting research to improve work safety
CSR 3	How to prevent accidents	CSR 14	Providing staff home ownership scheme
CSR 4	Reason for wearing safety attires	CSR 15	Provision of child-care service, maternity and paternity leaves
CSR 5	Employees practiced safety measure	CSR 16	Giving financial assistance to employee in educational institutions
CSR 6	Employees workplace are improved	CSR 17	Providing recreational activities/facilities
CSR 7	Disclosing accident statistics	CSR 18	Any policies for the company's remuneration scheme
CSR 8	Receiving safety award	CSR 19	Better share option scheme
CSR 9	Existence of employee training & development program	CSR 20	Providing assistance to employees who are in the process of retiring or who have been made redundant
CSR 10	Training employees through in-house programs		
CSR 11	Training on methods of preventing accidents or safety attires		

The results show that *the provision of healthy & safe workplace* (0.317), *the existence of employee training and development programs* (0.404), *providing staff home ownership scheme* (0.473) and the existence of *policies for the company's remuneration schemes* (0.493) are of highest priority for each sub-group *human self, intellect, posterity* and *wealth*, respectively.

In Table 1, the eleven takaful operators are denoted by TA through TK. A content analysis of financial statements was done to identify CSR initiatives undertaken by each takaful operator. A tick ($\sqrt{\quad}$) in the column under each takaful operator indicates that particular takaful operator disclosed the offer of the initiative to its employees. For each tick we sum up its priority vector. The higher the average priority vector the higher the importance of CSR contribution by the operator to its employees. Table 2 provides the average priority vector for each takaful operator. Takaful TJ has the highest average priority value of 0.315.

Table 1 Priority Vector and CSR Initiatives (Human Resource) by Takaful Companies

Sub-group	CSR	Priority vector	Takaful																
			TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK						
Human self	1	0.317	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	2	0.222	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	3	0.165	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	4	0.114	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	5	0.076	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	6	0.050	√	x	x	x	x	x	x	x	x	√	x	x	x	x	x		
	7	0.032	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	8	0.023	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
Intellect	9	0.404	x	x	x	x	√	x	x	x	x	x	x	x	√	x	x		
	10	0.278	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	11	0.177	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	12	0.090	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	13	0.050	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
Posterity	14	0.473	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	15	0.284	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	16	0.170	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
	17	0.073	√	x	x	x	√	x	x	x	x	x	x	x	x	x	x		
Wealth	18	0.493	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		
	19	0.311	x	√	√	x	x	x	√	√	x	x	√	√	√	√	x		
	20	0.196	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
Average priority value			0.154	0.201	0.201	0.123	0.234	0.201	0.123	0.201	0.201	0.123	0.201	0.201	0.315	0.123	0.123		

Table 2 The ranking of Takaful Operators based on human resource CSR initiatives

Takaful	Average priority value
TJ	0.315
TE	0.243
TB	0.201
TC	0.201
TF	0.201
TH	0.201
TI	0.201
TA	0.154
TD	0.123
TG	0.123
TK	0.123

4 Conclusion

A happy employee is a great asset to any organization. Implementing Corporate Social Responsibility (CSR) initiatives on employees is one way to make an employee happy. As one of Islamic financial institutions, takaful companies are expected to implement CSR initiatives that conform to the rules of Shariah. Using analytic hierarchy process, we calculated the priority vector of each CSR initiative and ranked the importance of CSR initiatives that were offered by takaful operators in Malaysia to their employees. The findings from this study contribute as a reference point for Takaful operators to prioritize their CSR initiatives according to Shariah. The highest priority vector for each group—*human-self, intellect, posterity and wealth*—are *provision of healthy & safe workplace* (0.317), *the existence of employee training and development programs* (0.404), *providing staff home ownership scheme* (0.473) and *having policies on the company's remuneration schemes* (0.493), respectively. We then averaged the sum of priority vectors of all CSR initiatives provided by each takaful operators. Takaful operator denoted by TJ ranked highest with an average priority value of 0.315, followed by TE with a value of 0.243.

Acknowledgements The authors would like to express their gratitude to Universiti Teknologi MARA and the Malaysian Ministry of Higher Education (MOHE). This research work is made possible from a research grant from MOHE.

References

1. Ahmad, A., Nasir, I.M., Wan-Abu-Bakar, W., Darus, F., Sidek, N.Z.M.: i-CSR Ranking Workplace Procedia Econ. Financ. Elsevier, November, **31**, 520–524 (2015)
2. Bursa Malaysia CSR Framework: Retrieved December 13, 2012, from http://www.klse.com.my/website/bm/about_us/the_organisation/csr/downloads/csriteup.pdf (2006)

3. Darus, F., Yusoff, H., Naim, D.M.A., Zain, M.M.: Islamic corporate social responsibility (i-CSR) framework from the perspective of Maqasid al-Syariah and Maslahah. *Issues Soc. Environ. Account.* **7**(2), 102–112 (2013)
4. Muhamat, A.A., Jaafar, M.N., Alwi, S.F.S.: Takaful operators' corporate social performance (CSP): an industry perspective, SHS web of conferences, **36** (2017)
5. Said, R., Zainuddin, Y., Haron, H.: The relationship between corporate social responsibility disclosure and corporate governance characteristics in Malaysian public listed companies. *Soc. Responsib. J.* **5**(2), 212–226 (2009)
6. Syed-Noh, S.N., Zakaria, S.F., Rshima Said, R., Sidek, N.Z.M., Ahmad, A.: The ranking of community Shariah compliant CSR activities. In: the 3rd Annual International Conference on Islamic Economics, Pusat Studi Ekonomi Islam, Universitas Sebelas Maret (2016)
7. Zakaria, S.F., Wan-Abu-Bakar, W., Said, R., Syed-Noh, S.N., Ahmad, A.: Ranking islamic corporate social responsibility activities under product development theme using analytic hierarchy process. In: Kor, L.K., (ed.), *Bridging Research Endeavour in Computer and Mathematical Sciences*, UiTM Kedah (2015)

Chapter 64

Assessing Malaysian Teachers' Perception on Computational Thinking Concepts Using SEM



Ung L. Ling, Tammie C. Saibin, Jane Labadin
and Norazila Abdul Aziz

Abstract Computational thinking (CT) concepts are newly introduced concepts in the Malaysian curriculum. This study is therefore designed to investigate Malaysian teachers' perception on the integration of computational thinking skills in their teaching and learning practices. A survey form was designed based on the Technological Acceptance Model (TAM) and was disseminated throughout Malaysia to gauge teachers' perception on CT based on the perceived usefulness of CT, perceived ease of CT integration into teaching and learning practices, teachers' attitude towards CT and their intention to use CT in their classrooms. A total of 166 primary school teachers participated in the survey and the data was analysed using Structural Equation Modelling (SEM). This study managed to predict Malaysian teachers' intention in integrating computational thinking skills in their classroom practices via two significant determinants, namely the perceived ease of integration and positive attitude towards computational thinking. This study is important because it highlights factors affecting teachers' perception on the newly improvised curriculum, and is an effort to support CT delivery in Malaysian classrooms.

Keywords Computational thinking · Technological Acceptance Model (TAM) · Primary school teachers · Perception · Structural Equation Modeling (SEM)

U. L. Ling (✉) · T. C. Saibin
Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM),
88999 Kota Kinabalu, Sabah, Malaysia
e-mail: ungli720@sabah.uitm.edu.my

T. C. Saibin
e-mail: tammi023@sabah.uitm.edu.my

J. Labadin · N. A. Aziz
Institute of Social Informatics and Technological Innovations (ISITI),
Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: ljane@unimas.my

N. A. Aziz
e-mail: anora@unimas.my

1 Introduction

Computational thinking (CT) is as important as reading, writing and counting [1] and it should be integrated into formal curriculum [1–4]. Many researchers agreed that it is a critical skill to have in the 21st century [5–8], since it is a skill that may improve one’s higher order thinking and problem-solving abilities [8–10]. Acknowledging the importance of mastering CT skills, the Malaysia Ministry of Education has integrated the skill into the existing curriculum, and has started implementing the revised Primary School Curriculum (*Kurikulum Standard Sekolah Rendah, KSSR*) in 2017 [11]. This newly revised curriculum has called on teachers, researchers and experts to design a training programme for all teachers to prepare them for potential teaching and learning (TL) challenges. The purpose of this study is to present data that enable all parties to gain a better understanding on the possible dilemma faced by teachers in adapting to and delivering the revised curriculum. This investigation focuses on teachers’ perception in integrating CT skills into their daily TL classroom practices.

2 Background of Study

2.1 Malaysian Computing Scenario and CT

Malaysian education system has been emphasizing on information, communication and technology (ICT) literacy for many years [12, 13]. Teachers have been sent to workshops and training to elevate their ICT skill and assist them in their TL. As technology evolved, researchers shifted their beliefs, students should be taught and trained to think, and should be able to solve problems and eventually be a creator instead of mere end user [6, 14, 15]. When Wing [1] introduced CT concepts in one of her seminar, she defined CT as intellectual and reasoning skills; and she examined how people react and learn to think through the language of computation. Others have defined CT as the tool to visualize problem solving strategies [16]. As CT skills have been integrated in 2017 revised KSSR syllabus, Ministry of Education Malaysia (MOE) has defined CT in the revised curriculum as the ability to think logically and systematically as well as solve problem using computer [17].

In response to a call to ease the delivery of CT concepts, especially to young learners, studies have been carried out to determine the factors that may contribute to the success of TL in this area. Some of the works include the investigation on the learning tools to enhance the TL practices [18–21]. In the case of Malaysia education system, more research needs to be conducted in order to ensure the TL of CT concept is carried out correctly and efficiently. This includes the learning patterns of the learners, technology required, and also how the society accepts CT.

The realisation of curriculum goals involving CT should also include the teachers, as they are the ones who are responsible to deliver the content of CT. Teachers should be trained and be well, and ready to carry out the TL [22].

2.2 Technological Acceptance Model (TAM) and Structural Equation Modelling (SEM)

TAM originated from Ajzen and Fishbein's Theory of Reasoned Action (TRA). According to TRA, behavioural beliefs affect the attitude toward behaviour. Davis [23] then introduced the TAM that shows how users will accept and use a technology. According to TAM, behavioural intentions, attitude and perceived usefulness and perceived ease of the system influenced the actual technology acceptance. Park [24] states that TAM is able to predict 40 to 50% of user acceptance. The TAM has been widely used in various types of information systems. In terms of educational contexts, TAM has been used to investigate students' behaviour towards learning [24, 25].

Structural Equation Modelling (SEM) has been widely used in research of behavioural sciences. SEM often visualised using a graphical path diagram which was invented by Sewall Wright in 1921. SEM provides very general and convenient framework for statistical analysis [26]. SEM model is usually represented in a set of matrix equations. Since this study uses TAM, which is represented by a graphical path diagram, it is thought that working with SEM can measure each of the causal path significantly and reliably.

3 Research Framework and Methodology

This research framework comprises of TAM constructs adopted from [27], consisting of 4 dimensions; namely perceived usefulness (PU), perceived ease of integration (PEI), attitudes towards CT integration (CA) and behavioural intention to integrate CT (BI). However, one item construct is modified to tailor to this study as the research is looking into teachers' technology per skill. In this study, CT is assumed as the newly introduced technology.

This paper reports the results of an investigation that assessed teachers' perception on newly revised curriculum, which integrates of CT skills into their TL practices. In total, 166 in service teachers from primary schools responded to this study. The present study proposes the following hypothesis; H₁: PU, PEI and CA are the factors affecting teachers' behavioural intention (BI) in accepting and implementing CT skills in their TL, H_{2a}: The teachers' perceived usefulness (PU) of CT positively influence their perceived ease of CT integration (PEI) in their TL, H_{2b}: The teachers' perceived usefulness (PU) of CT positively influence their attitude towards CT (CA) in integrating CT in their TL, H_{2c}: Perceived ease of CT

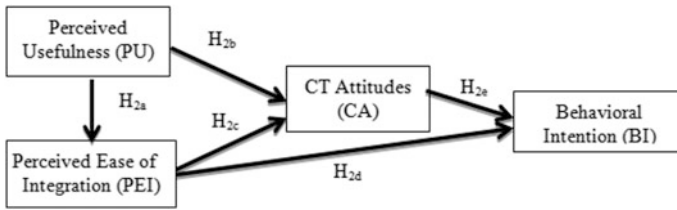


Fig. 1 Proposed research model (H₁)

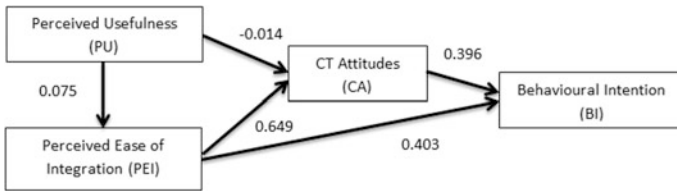


Fig. 2 The output of SEM

integration (PEI) positively influence teachers’ attitude towards CT (CA) in integrating CT in their TL, H_{2d}: Perceived ease of CT integration (PEI) positively influence teachers’ behavioral intention (BI) in integrating CT in their TL and H_{2c}: The teachers’ attitude towards CT (CA) will positively influence teacher’s behavioral intention (BI) in integrating CT in their TL. Figure 1 shows the proposed research model and the relationship matrix.

For the data analysis, the study employed the Structural Equation Modeling (SEM) to test the direct effects, its reliability and validity. The IBM SPSS AMOS 23 was used in the analysis process (Fig. 2).

4 Result

From the analysis result, the model is recursive with the sample size of 166. The data shows that there are 3 dependent variables, CA, PEI and BI, 1 independent variable-PU. Result indicates the ratio of the x² statistic to its degree of freedom is equal to 1, which is less than 5, indicating an acceptable fit, and depicting significant Chi Square Goodness-of-Fit [χ^2 (N = 166, df = 1) = 0.069, p < 0.05], and since the P-value is more than the significance level (0.05), it supports H₁.

All the measurement scales were above 0.70 which is considered good internal consistency. However, the construct item that lies just above the acceptable alpha was perceived usefulness of CT (PU), but since it is still above 0.70, the construct was still taken into consideration.

All hypotheses were supported by the data except for PU→PEU (H_{2a}) and PU→CA (H_{2b}). Teachers' attitude towards CT (CA) was found to be significantly influenced by perceived ease of CT integration ($\beta = 0.649$, C.R. = 9.673, $p < 0.05$), supporting H_{2c} . Behavioural intention (BI) was found to be significantly influenced by perceived ease of CT integration ($\beta = 0.403$, C.R. = 6.200, $p < 0.05$) and teachers' attitude towards CT (CT) ($\beta = 0.396$, C.R. = 6.453, $p < 0.05$), thus supporting H_{2d} and H_{2e} . Squared Multiple Correlations result shows that the variance in PEU by 0.009 or 0.9%, 0.373 or 37.3% in CA and 0.565 or 56.5% variance in BI can be predicted by the model hypotheses. The remaining of the variance of PEU, CA and BI cannot be predicted by this model, which may be due to some external factors that are not considered by this study.

5 Discussion and Conclusion

This study aims to investigate Malaysian primary school teachers' perception on CT concepts, by considering three determinants, namely the perceived usefulness of CT, perceived ease of CT integration and teachers attitude towards newly introduced CT concepts in the revised KSSR syllabus.

All hypotheses were supported except for H_{2a} and H_{2b} . This could be due to the fact that the majority of the respondents have not attended any formal training or workshop related to the newly introduced syllabus, and this is supported by [28] study indicated low percentage of teachers have attended any training related to CT. This causes confusion, misunderstanding of CT among the respondents as well as ignorant about the TL approaches of CT that can be practiced in their classroom. This implies that more work needs to be done by the Ministry of Education to introduce CT concepts to the existing in service teachers as well as pre-service teachers. Result also indicates that perceived ease of CT integration in TL practice, PEI, is a significant determinant of CA and BI, showing better perception and acceptance from the teachers to carry out the TL if the newly CT concepts is easily carried out with the existing curriculum, contributing to the positive attitudes towards CT concepts and eventually improve behavioral intention in integrating it in their classrooms. With the significant influence of these factors, the teachers most likely will integrate CT concepts into their TL practices. The fact that teachers' intention in integrating CT concepts into their TL is influenced by perceived ease of CT concept integration, hence indicating teachers' training plays a big role to equip them especially in TL pedagogy.

This study is an extension work from [28], which is able to support the same finding and an overview on factors that influence teachers' adaptation of CT concept into TL practices. Methodologically, future research will look into TL method specifically on the assessment tools that can be implemented in Malaysian curriculum scenario. We suggest a simple tool that is able to measure the learning

outcome of the learners, especially primary school level, for any means of assessment, since perceived ease of integration of CT in TL plays an important factor in influencing teachers in integrating CT in their daily classroom TL.

References

1. Wing, J.M., *Computational thinking*, in *Magazine Communications of the ACM - Self managed systems CACM* March, 2006, ACM: New York, NY, USA. p. 33–35
2. Guzdial, M.: Education Paving the way for computational thinking. *Communications of the ACM* **51**(8), 25–27 (2008)
3. Qualls, J.A., Sherrell, L.B.: Why computational thinking should be integrated into the curriculum. *Journal of Computing Sciences in Colleges* **25**(5), 66–71 (2010)
4. Lu, J.J. and G.H. Fletcher. *Thinking about computational thinking*. in *ACM SIGCSE Bulletin*. 2009. ACM
5. Buckley, S. *The Role of Computational Thinking and Critical Thinking in Problem Solving in a Learning Environment*. 2012. Kidmore End: Academic Conferences International Limited
6. Barr, D., Harrison, J., Conery, L.: Computational Thinking: A Digital Age Skill for Everyone. *Learning & Leading with Technology* **38**(6), 20–23 (2011)
7. Williams, T. *21st-Century Literacy Requires Computational Thinking*. 2015
8. Mahsa Mohaghegh, M.M., *Computational Thinking: The Skill Set of the 21st Century*. (IJCSIT) *International Journal of Computer Science and Information Technologies*, 2016. **7** (3): p. 1524–1530
9. Fessakis, G., Gouli, E., Mavroudi, E.: Problem solving by 5-6 years old kindergarten children in a computer programming environment: A case study. *Computers & Education* **63**, 87–97 (2013)
10. Richards, J., *Computational thinking: a discipline with uses outside the computer lab?* *Computer Weekly*, 2007: p. 52
11. BERNAMA, *Pemikiran komputasional, sains komputer akan diajar di sekolah tahun depan*, in *Utusan ONLINE*2016, Utusan ONLINE: Putrajaya
12. Chan, F.-M. *ICT in Malaysian schools: Policy and strategies*. in *Workshop on the Promotion of ICT Education to Narrow the Digital Divide, Tokyo*. 2002
13. Bell, G.: Satu keluarga, satu komputer (one home, one computer): Cultural accounts of ICTs in south and southeast Asia. *Design issues* **22**(2), 35–55 (2006)
14. Israel, M., et al.: Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education* **82**, 263–279 (2015)
15. Ramli, R., Yunus, M.M., Ishak, N.M.: Robotic teaching for Malaysian gifted enrichment program. *Procedia - Social and Behavioral Sciences* **15**, 2528–2532 (2011)
16. Barr, V., Stephenson, C.: Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community? *ACM Inroads* **2**(1), 48–54 (2011)
17. Malaysia, K.P., *Kurikulum Standard Sekolah Rendah KSSR (SEMAKAN)*, 2016
18. Weller, M.P., E.Y.-L. Do, and M.D. Gross. *Escape machine: teaching computational thinking with a tangible state machine game*. in *Proceedings of the 7th international conference on Interaction design and children*. 2008. ACM
19. Koh, K.H., et al. *Towards the automatic recognition of computational thinking for adaptive visual language learning*. in *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*. 2010. IEEE
20. Berland, M., Lee, V.R.: Collaborative strategic board games as a site for distributed computational thinking. *International Journal of Game-Based Learning* **1**(2), 65 (2011)

21. Kazimoglu, C., et al.: A Serious Game for Developing Computational Thinking and Learning Introductory Computer Programming. *Procedia - Social and Behavioral Sciences* **47**, 1991–1999 (2012)
22. Koehler, M., Mishra, P.: What is technological pedagogical content knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education* **9**(1), 60–70 (2009)
23. Davis, F.D.: Perceived Usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* **13**(3), 319–340 (1989)
24. Park, S.Y.: An Analysis of the Technology Acceptance Model in Understanding University Students' Behavioral Intention to Use e-Learning. *Educational technology & society* **12**(3), 150–162 (2009)
25. Venkatesh, V., Davis, F.D.: A model of the antecedents of perceived ease of use: Development and test. *Decision Sciences* **27**, 451–481 (1996)
26. Hox, J.J.B., T.M., *An Introduction to Structural Equation Modeling*. *Family Science Review*, 1998. **11**(4): p. 354–373
27. Teo, T., Wong, S.L., Chai, C.S.: A Cross-cultural Examination of the Intention to Use Technology between Singaporean and Malaysian pre-service Teachers: An Application of the Technology Acceptance Model (TAM). *Educational Technology & Society* **11**(4), 265–280 (2008)
28. Ung, L.L., Tammie, C. Saibin, Jane, Labadin and Norazila, Abdul Aziz *Preliminary Investigation: Teachers' Perception on Computational Thinking Concepts*. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 2017

Chapter 65

Chaotic Stochastic Lee-Carter Model in Predicting Kijang Emas Price Movements: A Machine Learning Approach



Siti Nurasyikin Shamsuddin, Nur Haidar Hanafi,
Muhammad Hilmi Samian and Mohd Nazrul Mohd Amin

Abstract Chaos is nonlinear and unpredictable which teaches us to expect the unexpected. Chaos Theory deals with nonlinear things that are effectively impossible to predict or control. There are a lot of important concepts about chaos has been introduced. The deterministic system has produced randomness which is called as “deterministic chaos”. An example of a deterministic chaotic system is the gold price index. In order to analyse the impact of chaos elements in predicting the gold price movements, the Chaotic Stochastic Lee-Carter Model was used in this study with the help of R software. This model is a hybrid of Chaos Theory and Stochastic Lee-Carter Model which was traditionally used in long run forecasts of age specific mortality rates. This research was done by using the data of Kijang Emas gold bullion coin’s selling and buying prices which were retrieved from the Central Bank of Malaysia (BNM). The results showed that Chaos Theory does help in improving the prediction accuracy of the current Stochastic Lee-Carter Model.

Keywords Chaos theory · Lee-Carter model · Data mining · Machine learning · Largest Lyapunov exponent

S. N. Shamsuddin (✉) · N. H. Hanafi · M. H. Samian
Faculty of Computer & Mathematical Sciences, UiTM Negeri Sembilan Kampus,
70300 Seremban 3, Negeri Sembilan Darul Khusus, Malaysia
e-mail: syikin65@ns.uitm.edu.my

N. H. Hanafi
e-mail: haidar54@ns.uitm.edu.my

M. H. Samian
e-mail: hilmi78@ns.uitm.edu.my

M. N. M. Amin
Faculty of Computer & Mathematical Sciences, UiTM, 40450 Shah Alam,
Selangor Darul Ehsan, Malaysia
e-mail: nazrul@tmsk.uitm.edu.my

1 Introduction to Chaos Theory

Chaos Theory deals with nonlinear things that are effectively impossible to predict or control, such as the stock market, weather, our brain states, road traffics and so on. Chaos theory was first introduced by Edvard Lorenz in 1963 and known as a non-linear system [1]. This theory can be described as “the exploration of patterns emerging from apparently random events within a physical or social system” [2]. Since then, chaos theory has been studied widely and several important concepts have been introduced, such as the dimensions, Lyapunov exponents, Fourier transform and Hilbert transform, and attractor reconstruction [3].

2 Stochastic Lee-Carter Model in Predicting Gold Price

Originally, the Lee-Carter model was introduced as a new method of projecting mortality rates by allowing improvements in life expectancy [4]. This model basically is used for time series forecasting. As the Kijang Emas price is also affected by the times, Hanafi et al. [5] used Stochastic Lee-Carter Model to forecast the gold price movements. Findings from this research which based on error comparison shows only small errors occur between actual data and predicted data for selling price indicating a good gold price projection by the model. In order to improve the performance of Stochastic Lee-Carter gold price model in forecasting the gold price movement with minimum error, the authors recommended that the elements of Chaos being injected into the model.

3 Chaos in Gold Market

Both discrete and continuous equations can produce chaos which is usually called “deterministic chaos” and can be considered as some kind of randomness produced by deterministic system. Example of a deterministic chaotic system is the gold price index [6]. There are five different methods used to analyze the trends for gold price over the last 31 years; linear trend analysis, ARMA analysis, Rescaled range analysis, attractor reconstruction and maximal Lyapunov Exponent, detrended fluctuation analysis [7]. However, not all methods give consistent results. Hence, the Chaotic Stochastic Lee-Carter Model is introduced to study the gold price movements to overcome the effect of chaos element.

4 Methodology

4.1 Data Collection

Kijang Emas 1 oz prices starting from January 2002 until December 2016 are collected from the Central Bank of Malaysia website. The data are then cleaned by eliminating any non-transaction days. The cleaned data are only consisted of on average 20 days of transaction for each month. The data are arranged into a rectangle matrix with dimension of 180 months by 20 days.

4.2 The Largest Lyapunov Exponent

See Table 1.

4.3 The Lee-Carter Model

See Table 2.

5 Results

5.1 The Largest Lyapunov Exponent

The value of lambda for Lyapunov Exponent is obtained by using `tseriesChaos` package in R version 3.3.3 and RStudio version 1.0.136. The graph in Fig. 1 shows

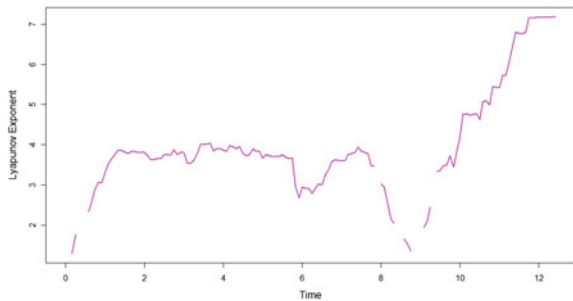
Table 1 R commands for calculating and plotting largest Lyapunov exponent

Command	Description
<code>GPrTS <- read_excel("~/Gold/GPrTS.xlsx")</code>	#Import data as GPrTS
<code>GPrTS.ts <- ts(GPrTS, start = c(2002, 1), end = c(2016, 12), frequency = 12)</code>	#Convert data as time series starting from January 2002 until December 2016
<code>output <- lyap_k(GPrTS.ts, m = 1, d = 1, ref = 180, t = 3, s = 150, eps = 1)</code>	#Estimates the largest Lyapunov exponent using the algorithm of Kantz
<code>lyap(output, 0.73, 2.47)</code>	#Computes the value of lambda

Table 2 R commands for fitting the Lee-Carter model

Command	Description
GoldPrice <- read_excel("~/Gold/GoldPrice.xlsx", col_names = TRUE)	#Import data as GoldPrice
GP <- data.frame(GoldPrice, row.names = 1)	#Set as data frame
LogGP <- as.matrix(log(GP))	#Convert log price into a matrix
a <- colMeans(LogGP)	#Compute the mean log-price for each day, this is the value of betta1
for(j in 1:20) LogGP[, j] <- LogGP[, j] - a[j]	#Subtract the average day pattern from all months
d <- svd(LogGP, 1, 1)	#Compute the singular value decomposition
b <- d\$v/sum(d\$v)	#Normalize the first row of V, this is the value of beta2
k <- d\$u * sum(d\$v) * d\$d[1]	#Overall price change over time, this is the value of kappa2
k1 <- matrix(k, nrow = 12, byrow = TRUE) k1 <- rowMeans(k1) k1 <- matrix(k1)	#Rearrange kappa2 into respective months and find the average for each row, this is the value of kappa2 for the 12 months, respectively

Fig. 1 Lyapunov exponent for 1 oz Kijang Emas price over time



the computed Lyapunov Exponent movements over time for 1 oz Kijang Emas price. The lambda obtained is 0.3403763 which indicates that gold is treated in an unstable and chaotic environment. Nearby points, no matter how close, will diverge to any arbitrary separation.

5.2 The Fitting of Lee-Carter Model

Figures 2 and 3 show the estimated values for β_1 and β_2 . Both factors indicates that the day of transaction has significant effects on the movements of gold price with the greater effect is from β_1 . The gold price will increase as the days progress toward the end of each month.

Figure 4 shows the estimated values of κ_2 for all 180 months in the data frame. Looking to only this graph will not give much of information regarding the effect of month towards the movement of gold price in a year. Thus, the average of κ_2 based on month is calculated and presented in Fig. 5. Based on Fig. 5, it can be seen that from January to September the price wil gradually increase before taking a big dip in price towards the end of each year.

Fig. 2 The value of β_1

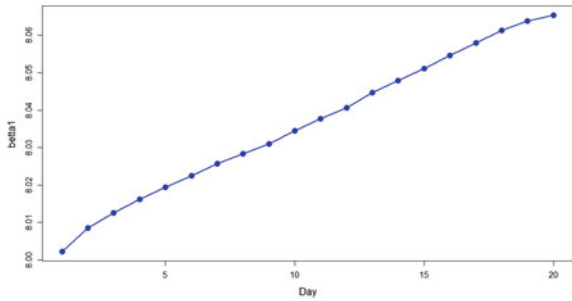


Fig. 3 The value of β_2

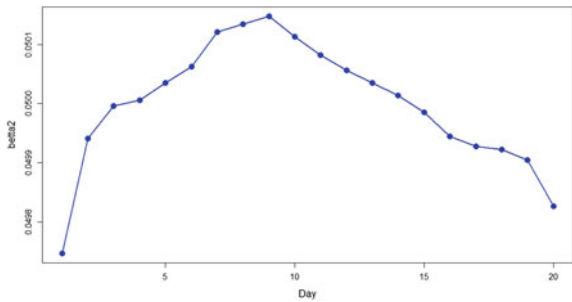


Fig. 4 The value of κ_2

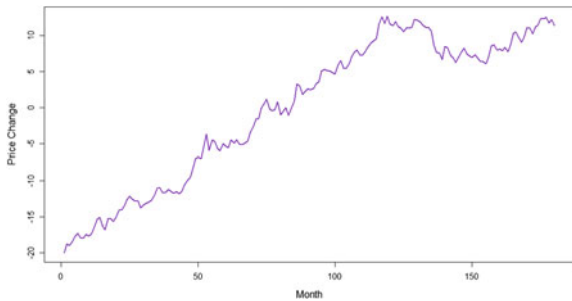
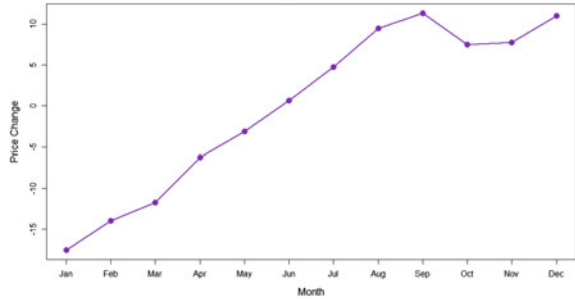


Fig. 5 The value of kappa2 based on month



5.3 Prediction Accuracy

In order to predict the future gold price using the parameters obtained by Lee-Carter, the forecasted values of kappa2 must be obtained. As discussed by previous researchers, the best forecasting method to be used is Holt-Winters model. Commands as in Table 3 are entered into RStudio. The values of 12 months kappa2 are projected for 2017. The plotted values of kappa2 can be seen in Fig. 6.

The equation for Chaotic Lee-Carter model is written as in Eq. 1. The first part of the equation is the original Lee-Carter model where beta1, beta2 and kappa2 are the parameter fitted before. The second part of the equation is the value of percentage size of chaos element existed in the system.

$$\ln m(t, x) = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)} - 100\% * \left[\frac{1}{N\Delta t} \sum \ln \left(\frac{|S(t + \Delta t) - S'(t + \Delta t)|}{|S(t) - S'(t)|} \right) \right] \quad (1)$$

Table 3 Commands for Holt-Winters

Command
<code>fit <- HoltWinters(k, gamma = FALSE)</code>
<code>forecast(fit, h = 12)</code>
<code>plot(forecast(fit))</code>

Fig. 6 Forecasted values of kappa2

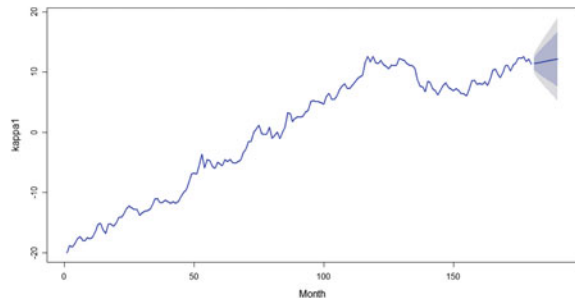


Table 4 MAPE

Model	MAPE (%)
Chaotic Lee-Carter model	0.4870961
Lee-Carter model	0.8227745

Table 5 Forecasted value of Kijang Emas for Day 1–7 in January 2017

Day	Price (RM)
1	5278.53
2	5323.41
3	5348.40
4	5368.39
5	5387.61
6	5405.42
7	5426.65

Table 4 show the mean absolute percentage error for both models. Based on the comparison, it is clearly that the Chaotic Lee-Carter model performs better than the normal Lee-Carter model. Thus, Chaotic Lee-Carter model will be used to forecast the price of Kijang Emas for the first week of January in 2017. The forecasted values are shown in Table 5.

6 Conclusions and Recommendation

The results obtained from this research are consistent with the findings in my previous researchers. The results showed that Chaos Theory does help in improving the prediction accuracy of the current Stochastic Lee-Carter model with significant improvement in the value of mean absolute percentage error. It is recommended that future researchers would try fitting the gold price to other models in the family of Generalized Lee-Carter model.

Acknowledgements This study is conducted with the support from Universiti Teknologi MARA’s iRAGS grant 600-RMI/IRAGS 5/3 (19/2015).

References

1. Mahmoudabadi, A., Andalibi, A.: The assessment of applying chaos theory for daily traffic estimation. In: Proceeding of the 2014 International Conference on Industrial Engineering and Operations Management (2014)
2. Griffiths, D.E., Hart, A.W., Blair, B.G.: Still another approach to administration: chaos theory. *Educ. Adm. Q.* 27(3), 430–451 (1991)

3. Liu, Z.: Chaotic time series analysis. In: *Mathematical Problems in Engineering* Volume. Hindawi Publishing Corporation (2010)
4. Lee, R.D., Carter L.R.: Modelling and forecasting U.S. mortality. *J. Am. Stat. Assoc.* **87**, 659–671 (1992)
5. Hanafi, N.H., et.al.: Investigation of Kijang Emas gold bullion coin price movement: application of stochastic Lee-Carter model and chaos theory (2016)
6. Haniyas, M., Magafas, L., Stavrinides, S.: Chaotic analysis of gold price index. *J. Eng. Sci. Technol. Rev.* **8**(1), 16–18 (2015) (Special Issue on Econophysics)
7. Alan, N., et al.: Chaotic trend possibility in the gold market. *Chaotic Model. Simul. (CMSIM)* **1**, 75–81 (2013)

Chapter 66

Efficiency of General Takaful Industry in Malaysia: A Two-Stage Data Envelopment Analysis



Hui Shan Lee, Fan Fah Cheng, Annuar Md Nassir,
Nazrul Hisyam Ab Razak and Wai Mun Har

Abstract The efficiency level in general Takaful industry is analysed from 2011 to 2015. Two-stage Data Envelopment Analysis (DEA) which is a mathematical programming method will be used to identify the factors that influence the efficiency level of general Takaful industry. DEA enables efficiency scores for each general Takaful operator in Takaful industry to be approximated relative to a dominant set of efficient firms with similar characteristics. The factors from insurance firm-specific perspective and corporate governance perspective are to be examined. The findings show that the turnover and size factors are positively and significantly influence the efficiency of general Takaful, but total net income shows a negative relationship with cost efficiency. Meanwhile, the portion of Muslim directors in the board of directors is found to have influence the general Takaful operators' efficiency. The results recommend that a wider pool of experienced and talented Muslim executive directors fitted with Islamic finance knowledge is able to manage general Takaful operators well with the achievement of cost efficiency.

Keywords Efficiency · Data envelopment analysis · General takaful · Corporate governance

H. S. Lee (✉) · W. M. Har
Universiti Tunku Abdul Rahman, Petaling Jaya, Malaysia
e-mail: hslee@utar.edu.my

W. M. Har
e-mail: harwm@utar.edu.my

F. F. Cheng · A. Md Nassir · N. H. Ab Razak
Universiti Putra Malaysia, Seri Kembangan, Malaysia
e-mail: chengfanfah@upm.edu.my

A. Md Nassir
e-mail: annuar@upm.edu.my

N. H. Ab Razak
e-mail: nazrul@upm.edu.my

1 Introduction

Islamic insurance is a financial institution that provides risk hedging services, produces liquidity, diversifies financial losses and facilitates long term investment in an economy, as recorded by Lee et al. [1]. Takaful (Islamic insurance) is the Shariah-compliant alternative to conventional insurance that avoids prohibited concepts to conform to Islamic jurisprudence, in which risks are distributed by the members in each Takaful risk pool [2]. Takaful emphasises the need for transactions to be prohibited from Haram trade or business-related activities and to be governed by Islamic finance law to set a higher standard for investments and promote greater accountability and risk mitigation. Recently, the global spread of Islamic finance has integrated the financial systems of Muslim countries that serve to strengthen the global financial markets. The importance of Takaful as a significant component of global Islamic finance requires more attention and research. Therefore, research on Takaful industry is needed and its key issues should be addressed in order to accomplish its long-term growth [3]. Malaysia is a multi-faith country with Islam being the largest practised religion, the investigation of general Takaful in Malaysia is important. This paper examines the cost efficiency and the impacts of firm-specific factors and corporate governance on efficiency of Takaful insurance operators in one of the Islamic emerging countries, Malaysia.

Global insurers are interested on emerging markets because they find it as a business opportunity and they are able to grow more prosperous. Interestingly, Malaysia is an emerging country with a larger population of Muslim that makes the Malaysian Takaful industry as one of the fast growing markets in the world. Ismail et al. [4] identify that Malaysia is the largest Takaful market in South East Asia region, with a 62% market share and followed by Indonesia at 33% in 2015.

The performance of Takaful operator is driven by many factors, for example size of the firms, turnover from the business and total income. As measures of financial performance is broad, and allows debate from various perspectives, this study is confined to examine one key element of performance of an insurance industry, that is, efficiency. Al-Amri [5] and Goud [6] suggest that if the complicated Takaful model is not fully developed and tested, Takaful operation systems would behave poorly. Hence, this paper raises the question “what are the firm-specific factors that impact the efficiency of Takaful insurance?”

Furthermore, a good governance system could assist Takaful operators to improve operational efficiencies [2]. According to agency theory, the entrustment of managerial responsibilities by owners (principal) and managers (agents) involves the existence of mechanism to monitor the performance of managers to ensure that they use their delegated powers to the best interests of the principals. Since Takaful industry is also part of the global financial system, examining the issues of corporate governance mechanism in order maintain investors' confidence has become unavoidable. This is because directors' fiduciary responsibilities in Takaful operators not only limited to shareholders and policyholders, but they are also subject to oversight by the Shariah supervisory board of the company and the regulators.

Hence, the board composition of Takaful operators is worth to be examined given the unique characteristics of Takaful market.

The novelty of this article is that it uses data on general Takaful (Islamic insurance) operators to look for evidence on firm-specific factors and board composition issues on the efficiency level. The contribution of this paper could help the readers to understand the business function of the Islamic corporate finance in general and more specifically to understand the emergent sector of the insurance industry by investigating the economic efficiency of Takaful operators.

2 Literature Review

The insurance industry encounters many challenges, such as solvency risks, a shifting regulatory environment and competition. Determining the efficiency of the insurance industry is clearly imperative in identifying how insurers will respond to these challenges by recognising the related input and output factors. Pioneer study measures the cost efficiency in the non-life insurance industry [7]. Later, Mahlberg and Url [8] and Biener et al. [9] employed data envelopment analysis (DEA) developed by Charnes et al. [10] in assessing the efficiency of insurance firms. This method does not impose any functional form on the data, allowing the method to use multiple inputs and outputs. The results of these studies show that firm size, ownership structure, mode of business and human capital are important factors affecting firm performance. In our study, there are total three firm-specific factors will be examined, namely turnover of the general Takaful, size of the general and total income generated by general Takaful industry.

Moving the efficiency study in Takaful industry, the studies are relatively fresh as the majority of them exist from the year 2010 onwards. A study by Kader et al. [11] investigate the cost efficiency of non-life Takaful insurance firms in ten Islamic countries from 2004 to 2006 and find that board size, product specialisation and firm size have positive impacts on the cost efficiency. Additionally, cost efficiency is not influenced by the separate roles of Chief Executive Officer and Chairman. This study is later being extended by Kader et al. [2] that explore the relationship between cost efficiency and board composition in non-life Takaful firms from 2004 to 2007 in 17 Islamic countries. The findings suggest that the levels of cost efficiency in Takaful are similar to the efficiency in developed non-life insurance markets. They confirm that the impacts of board composition on the efficiency depend on the firm-specific factors.

A recent exploratory study by Al-Amri [5] focused on the efficiency in Gulf Cooperation Council (GCC) countries. Their studies are focusing on technical efficiency with is different with the studies by Kader et al. [2, 11] that focus on cost efficiency. Al-Amri [5] finds that the Takaful industry in GCC is moderately efficient and suggests that improvement on technical efficiency is needed. Malaysia is experiencing an issue of the lack of skilled human resources, particularly with talent drain occurring in Takaful industry [4]. We believe that the problem of lack of

talents will affect the decision made by the members of board of director. Thus, we examine both cost with the incorporation of firm-specific characteristics and corporate governance structure as the factors influencing the inefficiencies.

3 Data and Methodology

This study will examine a total of 8 general Takaful operators in Malaysia spanning the period from 2011 to 2015. The firm-level data were collected from the annual reports of the Takaful operators.

To examine the factors that influence the efficiency of Takaful insurance in Malaysia, this study will conduct non-parametric frontier data envelopment analysis (DEA) in the first stage to obtain a DEA score. Then, in the second stage, panel regression will be employed with the DEA score as the dependent variable and the insurance firms' specific factors and corporate governance factors as the independent variables. The DEA approach with variable returns to scale (VRS) is used to identify the input-oriented technical efficiency of each Takaful operator.

In the second stage of the panel regression, the equation used is panel linear regression. The independent variables consist of firm specific factors and corporate governance factors. Following the literature by [9], the firm specific factors are firm size (SIZE), international diversification with foreign participation (FOREIGN) and provision for outstanding claims (PROVC). The corporate governance factors are the size of board of directors (NOBOARD), the ratio of the non-executive in the board of directors (RATIONE) and the ratio of Muslim directors in the board of directors (RATIOMUSLIM) [2].

$$\ln CE_{it} = \beta_0 + \beta_1 \ln \text{TURNOVER}_{it} + \beta_2 \ln \text{SIZE}_{it} + \beta_3 \ln \text{TI}_{it} + \beta_4 \text{NOBOARD}_{it} + \beta_5 \text{RATIONE}_{it} + \beta_6 \text{RATIOMUSLIM}_{it} + \varepsilon \quad (1)$$

4 Results and Discussions

General Takaful operators exhibit an average CE of 75.2%, indicates that there is room for improvement in cost efficiency. Note the minimum board of directors consist of 5 members and maximum is 9 members and averagely the general Takaful operators will involve of 7 members in the board of directors. The ratio of Muslim directors to the total board of directors is 0.68575, suggest that majority of the members are Muslim. The ratio of non-executive directors is very high and reports a value of 0.9555, indicates that the decision making for the general Takaful operators are made by non-executive (outside) directors.

Table 1 presents the second stage of panel regression results to identify the influence of firm specific and corporate governance factors to the cost efficiency general Takaful operators in Malaysia. Our explanation will focus on Model 1 because its R^2 is the highest. Model 2 to 9 are to show the influence of the variables to the cost efficiency when some of the variables are not included. In model 1, the coefficient of TURNOVER is positive and significant at 5%. The result indicates that the more contribution of Takaful fund generated from total assets will improve the cost efficiency. The SIZE variable also reports a significant influence towards cost efficiency. The result suggests that the larger Takaful operators are able to achieve economic of scale thus improves the cost efficiency. Moving to total net income, the result shows significance negative relationship towards cost efficiency at 5% significant level. This finding suggests that when the Takaful operators able to gain higher income, they are less motivated to reduce cost thus leads to cost inefficiency. Turning to corporate governance variables, among the NOBOARD, RATIOMUSLIM and RATIONE, only RATIOMUSLIM depicts significant influence over cost efficiency. RATIOMUSLIM shows positive significant level at 5% to the cost efficiency, suggest that Muslim directors are more skilful to manage Takaful operators. This is because they are trained well on the knowledge on Takaful related information. The board of director member size has negative impacts towards the cost efficiency. This study suggests that large board may be less cohesive and more complicated to coordinate. Thus, smaller board size is more appropriate to the Takaful operators in Malaysia. This is because lesser members of board could minimise the chance of free riding by individual directors and improve the decision-making outcomes that leads to improvement in cost efficiency. Nevertheless, the result is unable to prove a significant outcome. RATIONE depicts a negative coefficient. It implies that the higher the proportion of non-executive directors in board, it will deteriorate the cost efficiency. This result proposes that outside directors could have excessive prudence, risk aversion and have higher concerns about compliance with Shariah principle could dampen the cost efficiency of the Takaful operators. However, the result is not significant.

Overall, from firm specific factors, TURNOVER and SIZE suggest positive relationship with cost efficiency but TI suggest negative nexus with cost efficiency. From corporate governance factors, only RATIOMUSLIM indicates positive significant relationship with cost efficiency.

5 Conclusions

This study employs Data Envelopment Analysis (DEA) to study the efficiency of 8 general Takaful operators in Malaysia spanning the period from 2011 to 2015. There are few important findings emerge from this research. We found that when the general Takaful operators gain higher income, they are less motivated to improve cost efficiency. This situation should be avoided to ensure that the Takaful industry in Malaysia could grow healthily and competitive as the Takaful industry

Table 1 Panel regression results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ln TURNOVER	0.740** (0.294)	0.592* (0.290)		0.594* (0.292)	0.728** (0.290)	0.581* (0.289)	0.726** (0.298)	0.584* (0.293)	0.722** (0.288)
ln SIZE	0.877** (0.411)	0.679 (0.413)		0.725* (0.420)	0.810* (0.405)	0.737* (0.415)	0.811* (0.414)	0.756* (0.423)	0.879** (0.405)
ln TI	-0.885** (0.404)	-0.682 (0.403)		-0.742* (0.413)	-0.816** (0.397)	-0.748* (0.407)	-0.818* (0.407)	-0.775* (0.417)	-0.894** (0.398)
NOBOARD	-0.0452 (0.0959)		-0.0316 (0.0904)	0.0647 (0.0821)			0.00512 (0.0891)	0.0399 (0.0874)	
RATIOMUSLIM	1.163* (0.637)		0.674 (0.628)		0.975* (0.553)		0.959 (0.626)		1.017* (0.549)
RATIONE	-1.354 (1.032)		-1.333 (1.076)			-1.051 (0.970)		-0.893 (1.043)	-1.159 (0.932)
CONSTANT	2.433 (1.431)	1.374* (0.766)	1.794 (1.369)	1.178 (0.810)	0.888 (0.789)	2.540* (1.320)	0.880 (0.815)	2.244 (1.488)	2.153 (1.283)
N	40	40	40	40	40	40	40	40	40
R-squared	0.276	0.143	0.078	0.162	0.228	0.177	0.229	0.184	0.270
Firms	8	8	8	8	8	8	8	8	8

Note ** indicate significance at 0.05 level, * indicate significance at 0.1 level. Standard errors are reported in parentheses

in other countries especially countries with higher Muslim population. Furthermore, the general Takaful operators are efficient in utilising its total assets to generate contribution from the Takaful participants. Next, since size has positive impacts towards cost efficiency, the general Takaful operators in Malaysia could consider expanding their branches to achieve economic of scale. Second, corporate governance factors do influence the cost efficiency in general Takaful operators with higher proportion of Muslim members in the board of directors. This is because they are capable to manage more complicated Takaful products compared to conventional products. Finally, Takaful operators should be more transparent to disclosure more information relating to their operations in order to facilitate investors and other external stakeholders to draw well-informed opinions on Takaful operators.

We believe that this study could have potentially important policy implications. Regulators in Takaful markets could strengthen the supervisory and regulatory oversight of the Takaful industry with an effective Shariah framework and ensure that the Takaful operators are complying with the Shariah principles and to have a more transparent reporting practice relating to Shariah compliance information. This is to ensure that the Takaful market could support the integral feature in the Islamic banking and finance by strengthening the development in Takaful market.

Acknowledgements The authors are grateful to the support from Universiti Tunku Abdul Rahman, Universiti Putra Malaysia and Fundamental Research Grant Scheme from Ministry of Higher Education (FRGS/1/2017/SS01/UTAR/03/1).

References

1. Lee, H.S., Low, K.L.T., Chong, S.C., Sia, B.K.: Influence of secondary and tertiary literacy on life insurance consumption: case of selected ASEAN countries. Geneva Pap. Risk Insur. Issues Pract. **43**, 1–15 (2018). <https://doi.org/10.1057/s41288-017-0050-7>
2. Kader, H.A., Adams, M., Hardwick, P., Kwon, W.J.: Cost efficiency and board composition under different takaful insurance business models. *Int. Rev. Financ. Anal.* **32**, 60–70 (2014). <https://doi.org/10.1016/j.irfa.2013.12.008>
3. Mayenkar, S.S.: Global Takaful industry poised to be at \$25.5b by 2020 | GulfNews.com, <http://gulfnews.com/business/economy/global-takaful-industry-poised-to-be-at-25-5b-by-2020-1.1596226> (2015)
4. Ismail, F., Hassan, A.A.M., Jaffer, S., Alajaji, K., Unwin, L., Ten, Y.Y., Jamil, S.: Global Takaful Report 2017 (2017)
5. Al-Amri, K.: Takaful insurance efficiency in the GCC countries. *Humanomics* **31**, 354–371 (2015). <https://doi.org/10.1108/H-05-2014-0039>
6. Goud, B.: World Takaful Report (2016). http://www.takafulprimer.com/main/downloads/ms_5860.pdf
7. Cummins, J.D., Weiss, M.A.: Measuring cost efficiency in the property-liability insurance industry. *J. Bank. Financ.* **17**, 463–481 (1993). <https://doi.org/10.1177/107755877803500107>
8. Mahlberg, B., Url, T.: Single market effects on productivity in the German insurance industry. *J. Bank. Financ.* **34**, 1540–1548 (2010). <https://doi.org/10.1016/j.jbankfin.2009.09.005>

9. Biener, C., Eling, M., Jia, R.: The structure of the global reinsurance market: an analysis of efficiency, scale, and scope. *J. Bank. Financ.* **77**, 213–229 (2017). <https://doi.org/10.1016/j.jbankfin.2017.01.017>
10. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **2**, 429–444 (1978). [https://doi.org/10.1016/0377-2217\(78\)90138-8](https://doi.org/10.1016/0377-2217(78)90138-8)
11. Kader, H.A., Adams, M.B., Hardwick, P.: The cost efficiency of takaful insurance companies. *Geneva Pap. Risk Insur. Issues Pract.* **35**, 161–181 (2010). <https://doi.org/10.1057/gpp.2009.33>

Chapter 67

Enhancement of DNA Gel Images



CT Munnirah Niesha Mohd Shafee, Ahmad Khudzairi Khalid
and Zarith Sofiah Othman

Abstract The use of image enhancement is to process an image so that result is more suitable than the original image for specific application. The objective of enhancing the image is to improve the standard or the quality of the image from the original one. DNA gel image is one of the digital medical images prove to be corrupted by some degree of noise. It is due to the presence of corruption present in a transmission and an acquisition of many effects. This type of image need to be enhanced before it can be used for analysis or diagnose. This paper compares three different techniques of image enhancement which are used to enhance the DNA gel images namely Enhancement of DNA Gel Image using Thresholding, Shifting, and Filtering Techniques or Method 1, Enhancement of DNA Gel Image using Background Subtraction Technique or Method 2, and Enhancement of DNA Gel Image using Improved Background Subtraction Method or Method 3. The evaluation of the result is done based on the calculation result of Peak Signal to Noise Ratio (PSNR) value. The experimental results show that the third method of image enhancement is a better method to be applied as it shows a higher PSNR value compared to the other which means it improves the image better.

Keywords DNA gel image · Enhancement · PSNR

C. M. N. M. Shafee (✉)

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
Cawangan Johor, Kampus Segamat, 85000 Segamat, Johor, Malaysia
e-mail: ctmun518@johor.uitm.edu.my

A. K. Khalid · Z. S. Othman

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
Cawangan Johor, Kampus Pasir Gudang, 81750 Masai, Johor, Malaysia
e-mail: ahmad4829@johor.uitm.edu.my

Z. S. Othman

e-mail: zarithsofiah@johor.uitm.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_67

1 Introduction

In recent years, the role of image in medical area has rapidly increased. The technologies in medical imaging also have been improved and many innovations have been undertaken to make the medical data useful. Furthermore, these innovations and the creation of the new methods and technologies in medical field will develop the new trends for the future of medical management. These digitized medical data are used in many areas of medical area such as revolutionizing diagnosis, therapy planning and follow-up, and clinical as well as biomedical research.

However, the rapid increase of technologies also brings several disadvantages. According to [1], digital images are corrupted by some degree of noise due to the presence of corruption present in transmission and acquisition by many effects. Therefore, the best enhancement techniques or methods are required to make the image valid and still reliable.

Deoxyribonucleic acid (DNA) image is an example of medical image. Basically, X-ray crystallography is used to capture the DNA image before electron microscope was introduced. Nevertheless, image taken using X-ray crystallography only defects the DNA itself and produced more noises compared to the other one.

The presence of noises may lead to errors in analyzing data process. In order to make the analysis of DNA gel image more efficient and effective, digital image processing techniques are introduced to enhance the DNA gel images. In this paper, three methods have been developed to enhance the DNA image. The enhanced images are diagnosed using Peak Signal to Noise Ratio (PSNR) where the method with the highest PSNR value is the best method that enhances the image.

2 Literature Review

Gel Electrophoresis (GE) is widely using a technique to separate DNA according to their size and weight [2–4]. The result of GE is a series of band which is position based on the molecular weight. Image that has been captured in GE usually contain anomalies such as noise and large smear on a strong non-uniform background [5]. Besides, the faint's band and the color of the band near to the image background also become a problem in analysis stage of the image [6, 7].

Research by Sengar et al. [8] stated that digital medical image such as DNA gel image great problem is the present of noise. Ye et al. [9] and Sotaquira [10] have supported this statement by saying that noise occurred in almost all medical images and Gaussian noise, Speckle noise, Salt and pepper noise, Rician noise and Brownian noise are noises which are found in medical images. The medical images need to be enhanced to improve the quality of the image in order to ensure the information get from the data analysis is reliable for the diagnosis and treatment.

There are a lot of methods used to enhance the DNA gel image from the previous researches. One of the research by Kaabouch et al. [11] proposed an enhancing algorithm which consists of four basic steps, namely thresholding, shifting, filtering and data processing. The author stated that the result of the combination techniques in the proposed algorithm eliminates defect due to the noise in average quality image.

The other research used combination techniques in order to enhance the DNA gel image by Archana et al. [12]. In this research, the combination of contrast stretching and histogram equalization is used as the enhanced method. It is said that the combination of the techniques in the proposed method produced a better result on the quality of image whether in subjective or objective evaluation.

Combination of morphological operation technique also said as one of the famous method of enhancement of DNA gel image. Helmy and El-Tawel [13] proposed the combination of morphological operation and the filtering operation to estimate the background image and to eliminate the noise in image.

3 Methodology

Three enhancement methods that have being designed in this research. Two of them are adapted from previous work by Kaabouch et al. [11] and Helmy and El-Tawel [13]. Meanwhile, the third method is the new proposed method and has been tested to produce better result than the earlier two methods. Basically, there are several steps involve in each method before the final enhance image is evaluated.

The first method is enhancement of DNA gel image using thresholding, shifting and filtering. Thresholding is a process of converting a greyscale image to a bi-level image using an optimal threshold. The purpose of thresholding an image is to extract pixels from the image which represent an object such as text or other line image data. After the image has been threshold, shifting activity take place. In this activity, the minimum level of the profile is shifted to zero. This helps in increasing the contrast of the image. The shifted image then will be filtered using the combination of top-hat and bottom-hat filtering to highlight the faint band and improved the image's quality.

The second method is enhancement of DNA gel image using background subtraction technique which it is used to account the differences between lanes and also the length of each lane. The first step in background subtraction is to estimate the background of the image by using the morphological closing operation with the 10 pixels circular type structuring element. Image resulted from this operation then being subtracted from the original image to extract the image's details. After the background subtraction, the image is filtered with the combination of top-hat filter with circular structuring element of ten pixels and 5×5 median filter for better quality image.

Enhancement of DNA gel image using improved background subtraction method is the last method. This method is almost similar with the second method. For the first step, the same operation of background subtraction is applied. The different between these methods with the second method is in the filtering part. The combinations of three filters are proposed to improve the quality of the image. After background removal, the image is first filtered by 2-D Gaussian filter to remove the Gaussian noise in the image. Kaur and Sharma [9] have stated that the noises that mostly occur in medical images are Additive White Gaussian Noise, salt and pepper noise and speckle noise. Because of the Gaussian noise is listed as one of mostly occur noise in medical image, so it is practical in applying the Gaussian filter to the image to enhance it. After the Gaussian filtering process, the top-hat and median filtering with the 3×3 dimension took place.

After applying the enhancement method, the evaluation phase needs to be carried out to evaluate all method that has been applied. The type of evaluation in this research is based on the imperceptibility evaluation of the image. Peak signal to noise ratio (PSNR) is used to evaluate the effectiveness of the enhancement method. The calculation of PSNR value is done for every enhanced image. Based from the research review that has been done, it is said that the method with the higher PSNR value indicates the best enhancement method [14].

4 Finding and Discussion

Three Experiments and the testing are done on all three methods with 25 of 256×256 grayscale images and in JPEG format. The PSNR is used to evaluate the enhanced image and the calculation result of Peak Signal to Noise Ratio (PSNR) of each image is recorded in Table 1. Figure 1 shows the comparisons PSNR results of all three methods.

From the result of Table 1 and Fig. 1, we can see the result of PSNR value for Method 3 is the highest among all methods. Average PSNR value for all 25 tested images in Method 3 is 31.6260 decibels (dB) while in Method 1 and Method 2 are 8.1399 dB and 23.5964 dB. Due to this high value of PSNR, the enhanced image of Method 3 indicates that the image contains less noise and has a better perception than the other two methods. This means that the enhanced image of Method 3 is more imperceptible and has a better quality then the other two Method. Based on this result, and by analyzing the result of each method, it has been observed that Method 3 produces a better enhance image as compare to the other two methods.

Table 1 PSNR values of three methods

No image	PSNR of method 1	PSNR of method 2	PSNR of method 3
GI_001	9.4945	21.5319	29.7745
GI_002	8.3019	22.3003	28.3901
GI_003	8.1135	22.5900	31.9702
GI_004	8.4368	25.8970	30.3760
GI_005	8.1065	20.5921	29.4094
GI_006	7.7385	23.7799	31.9823
GI_007	8.0629	25.9036	32.6387
GI_008	8.4911	24.4912	30.1469
GI_009	8.3349	24.6490	30.3893
GI_010	8.3511	21.4640	30.6372
GI_011	8.6021	21.7647	30.8131
GI_012	8.5112	24.1191	33.6305
GI_013	7.4864	22.8685	31.0742
GI_014	8.0299	23.6299	32.2286
GI_015	7.9559	30.1188	39.9297
GI_016	7.0048	28.4772	38.2180
GI_017	6.8559	27.1626	38.5998
GI_018	8.4322	23.1275	31.7474
GI_019	7.2710	24.6795	31.2596
GI_020	8.1715	21.9697	28.2887
GI_021	9.1634	24.1502	30.1776
GI_022	11.1605	23.7118	33.6206
GI_023	6.4473	19.7514	26.5941
GI_024	6.2112	18.6651	26.2085
GI_025	8.7649	22.5163	32.5457

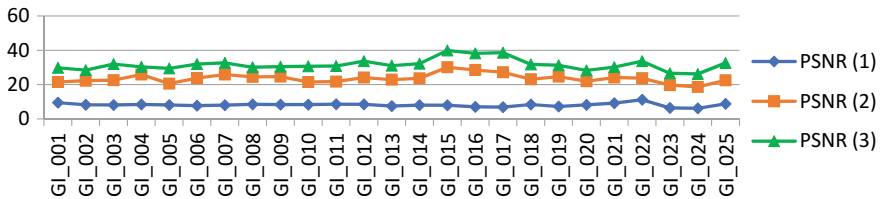


Fig. 1 PSNR value for three methods

5 Conclusion

As a conclusion, this research has experiment these three different methods to enhance the DNA gel images. Two of the methods which are Method 1 and Method 2 are adapted from the previous research. The last method which is Method 3 improved the second method with a new combination of filters. The methods have been successfully tested and all the enhanced images have been evaluated by using PSNR. The experimental result shows that PSNR value for Method 3 is higher compared to Method 1 and Method 2. This means that Method 3 capable to remove noise and enhance the DNA gel image better than the other two methods.

References

1. Bailey, D.G., Christie, B.C.: Processing of DNA and protein electrophoresis gels by image analysis. In: 2nd New Zealand Conference on Image and Vision Computing, August 1994, pp. 221–228. Massey University, New Zealand (1994)
2. Quelhas, P., Marcuzzo, M., Mendonça, A.M., Campilho, A.: Cell nuclei and cytoplasm joint segmentation using the sliding band filter. *IEEE Trans. Med. Imaging* **29**(8), 1463–1473 (2010). <https://doi.org/10.1109/TMI.2010.2048253>
3. Park, S.C., Na, I.S., Kim, S.H., Lee, G.S., Oh, K.H., Kim, J.H., Han, T.H.: Lanes detection in PCR gel electrophoresis images. In: 2011 IEEE 11th International Conference on Computer and Information Technology (CIT), August 2011, pp. 306–313. IEEE (2011). <https://doi.org/10.1109/cit.2011.89>
4. Lin, C.Y., Ching, Y.T., Yang, Y.L.: Automatic method to compare the lanes in gel electrophoresis images. *IEEE Trans. Inf. Technol. Biomed.* **11**(2), 179–189 (2007)
5. Caridade, C.M.R., Margal, A.R.S., Mendonga, T., Pessoa, A.M., Pereira, S.: An automatic method to identify and extract information of DNA bands in Gel Electrophoresis Images. In: Engineering in Medicine and Biology Society, EMBC 2009. Annual International Conference of the IEEE, September 2009, pp. 1024–1027. IEEE (2009). <https://doi.org/10.1109/iembs.2009.5332462>
6. Mao, H., Xie, M.: Lane detection based on Hough transform and endpoints classification. In: 2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP), December 2012, pp. 125–127. IEEE (2012). <https://doi.org/10.1109/icwamtip.2012.6413455>
7. Kaur, M., Sharma, R.: Restoration of medical images using denoising. *Int. J. Sci. Emerg. Technol. Latest Trends* **5**(1), 35–38 (2013)
8. Sengar, R.S., Upadhyay, A.K., Singh, M.: Robust Pre-processing and Post-Processing Methods for 2D Gel Electrophoresis Images using Non-Separable Quincunx Wavelet
9. Ye, X., Suen, C.Y., Cherié, M., Wang, E.: A recent development in image analysis of electrophoresis gels. In: Vision Interface, May 1999, vol. 99, pp. 19–21 (1999)
10. Gutierrez, M.A.S.: On the use of distance maps in the analysis of 1D DNA gel images. In: 2009 International Conference on Digital Image Processing, March 2009, pp. 172–176. IEEE (2009). <https://doi.org/10.1109/icdip.2009.58>
11. Kaabouch, N., Schultz, R.R., Milavetz, B.: An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement. In: 2007 IEEE International Conference on Electro/Information Technology, May 2007, pp. 26–31. IEEE (2007). <https://doi.org/10.1109/eit.2007.4374496>

12. Goel, S., Verma, A., Kumar, N.: Gray level enhancement to emphasize less dynamic region within image using genetic algorithm. In: 2013 IEEE 3rd International Advance Computing Conference (IACC), February 2013, pp. 1171–1176. IEEE (2013)
13. Helmy, A.K., El-Tawel, G.S.: Semiautomatic detection of lanes and bands in DNA gel electrophoresis images. *J. Biomed. Sci. Eng.* **6**(01), 76 (2013). <https://doi.org/10.4236/jbise.2013.61010>
14. Rashwan, S., Sarhan, A., Faheem, M.T., Youssef, B.A.: Fuzzy watershed segmentation algorithm: an enhanced algorithm for 2D gel electrophoresis image segmentation. *Int. J. Data Min. Bioinform.* **12**(3), 275–293 (2015)

Chapter 68

Factors Affecting Entrepreneurial Intention Among IKN Students



Faridah Abdul Halim, Muhammad Rozi Malim, Siti Iliyana Hamdan,
Atika Salehan and Farhana Syahirah Kamaruzzaman

Abstract The unemployment rate among graduates has been steadily increasing over the years and some have decided to engage in entrepreneurship. Government initiatives such as entrepreneurship pro-grams and funds have been created to attract youth to entrepreneurship. However, becoming an entrepreneur is a big challenge as it requires certain characteristics that influence an individual's entrepreneurial intention. This study explores the factors affecting the entrepreneurial intention among students at National Craft Institute (IKN), Rawang. It covers demographic profiles, educational supports, attitudinal factors and behavioral factors, and how these factors influence the entrepreneurial intentions among IKN students. A questionnaire was distributed to students of six different programs using simple random sampling. The data were analyzed using correlation analysis and multiple linear regression. Results revealed that there is a relationship between educational supports, attitudinal factors, and behavioral factors in developing entrepreneurial intention among students. This study also found that there is a significant difference in entrepreneurial intention between the three factors, with educational supports being the most influential.

Keyword Entrepreneurship · Entrepreneurial intention · Educational supports · Attitudinal factors · Behavioral factors

1 Introduction

The rising rate of unemployment among graduates has become a national issue in Malaysia. In May 2015, the Prime Minister Department revealed that a total of 161,000 graduates out of 400,000 were unemployed [1]. Due to limited employ-

F. A. Halim (✉) · M. R. Malim · S. I. Hamdan · A. Salehan · F. S. Kamaruzzaman
Faculty of Computer and Mathematical Sciences, UiTM, 40450 Shah Alam, Malaysia
e-mail: faridahh@tmsk.uitm.edu.my

M. R. Malim
e-mail: rozi@tmsk.uitm.edu.my

ment opportunities and dissatisfaction in salaries, many graduates have been motivated to build their own career as entrepreneurs. Based on a study on 4,673 Malaysian youths, only 63.3% have entrepreneurial interest and potential [2]. These youths continue their interest and desire to acquire skills that are entrepreneurial in nature [3].

Entrepreneurship is an attitude that reflects an individual's motivation to identify and pursue an opportunity to produce new economic success [4]. According to several literatures, entrepreneurs have their own personality traits, creativity, curiosity, socioeconomic characteristics, risk taking and particular nature of business activities [5, 6]. Chan et al. [3] claimed that youth who have been identified as entrepreneurs are able to survive in an economic crisis since they have decided on entrepreneurship at the expense of other careers. However, the number of youth entrepreneurs in Malaysia is still low due to their lack of knowledge and exposure to entrepreneurship.

The National Craft Institute (IKN) is responsible for producing skilled graduates in the craft field (*batik*, weaving, wood, metal, ceramic, rattan, and bamboo). The craft industry is able to provide many job opportunities for those who are interested and creative. This study aims to analyze the factors affecting the entrepreneurial intention among IKN students; i.e., (i) to examine the relationship between educational supports, attitudinal factors, behavioral factors and entrepreneurial intention among students, and (ii) to identify the most influential factor.

2 Methodology

This study is based on the theoretical model by Mumtaz et al. [7]. The factors that assumed to affect the entrepreneurial intention are educational supports, behavioral factors, and attitudinal factors. The items for educational supports are syllabus and pedagogy, while behavioral factors consist of risk taking and creativity. Personality traits, curiosity and locus of control are items for attitudinal factors.

The methods of analysis are correlation and multiple linear regression. Students of IKN Rawang, Selangor are considered as the case study. A descriptive study is used to describe the demographic profiles and factors affecting entrepreneurial intention, and a causal study is used to identify causes and effects of the factors. The total number of students is 365. A questionnaire adopted from Mumtaz et al. [7] was distributed to all students, and only 123 were returned. The questionnaire consists of six sections: A (demographic profiles), B (entrepreneurial intention), C (business information), D (educational supports), E (attitudinal factors), and F (behavioral factors). The data are analyzed using SPSS.

2.1 Reliability Analysis

Cronbach's alpha is the most common measurement of internal consistency or reliability to see how closely related a set of items are as a group. Generally, alpha coefficient value ranges from 0 to 1. The closer the Cronbach's alpha coefficient is to 1, the greater the internal consistency of items in the scale.

2.2 Descriptive Statistics

Demographic profiles are represented by frequencies and percentages. Since the entrepreneurial intention factors (education supports, behavioral factors, and attitudinal factors) are placed on a Likert scale, the analysis is usually done using the sample mean and sample standard deviation [7–9]. Correlation is used to determine the relationship between two variables. The correlation $r = 0$ means no correlation and $r = +1$ means a perfect positive relationship. This would realize the first objective; to examine the relationship between educational supports, attitudinal factors, behavioral factors and entrepreneurial intention among students.

2.3 Multiple Linear Regression

The model: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$; where y = entrepreneurial intention, x_1 = educational supports, x_2 = behavioral factors, x_3 = attitudinal factors, β_i = i th parameter of regression line, ε = random error. This would realize the second objective.

Significance of Model: Testing the validity of the model using Analysis of Variance; $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. Rejection of H_0 implies that the model is significant.

Significance of Independent Variables: Using a t -test. $H_0: \beta_k = 0$ against $H_1: \beta_k \neq 0$. If a coefficient were found to be insignificant (H_0 is not rejected), the variable would also be insignificant in the presence of other variables in the model.

Coefficient of Determination: $R^2 = \frac{SSR \text{ (Sum of Square Regression)}}{SST \text{ (Sum of Square Total)}}$; the proportion of the total variation in Y associated with the use of X variables. However, we prefer looking at the adjusted R^2 , $R_{adj}^2 = \frac{SSE/(n-k)}{SST/(n-1)}$, k = number of parameters and n = sample size. It measures the proportion of the variation in Y that is explained by the $(p - 1)$ predictor variables after considering the sample size n and the number of independent variables.

3 Analysis and Results

3.1 Reliability Test

Table 1 shows the Cronbach’s alpha coefficients for all variables. All values exceed 0.7, and it can be concluded that there is high internal consistency of the items in the questionnaire. The widely accepted social science cut-off is at least 0.70 [11].

3.2 Descriptive Statistics

A total of 123 questionnaires were administered but 10 invalid. The remaining 113 were then coded using SPSS to analyze the profiles of respondents, and all items in Section B to Section F. The respondents comprise both males (40.7%) and females (59.3%), and the majority being 20 year-old (28.3%). Figure 1 shows a majority of the respondents (55.8%) are not from entrepreneurial families with 31.9% females, while 44.2% are from entrepreneurial families with 27.43% females. The high

Table 1 Summary of Cronbach’s alpha

Items	Number of questions	Cronbach’s alpha
Entrepreneurial intention	4	0.819
Educational supports	8	0.818
Attitudinal factors	10	0.814
Behavioral factors	10	0.825

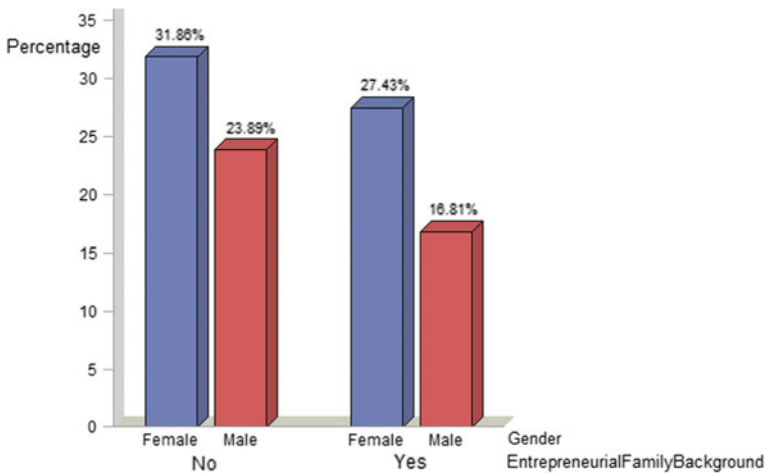


Fig. 1 Entrepreneurial family backgrounds according to gender

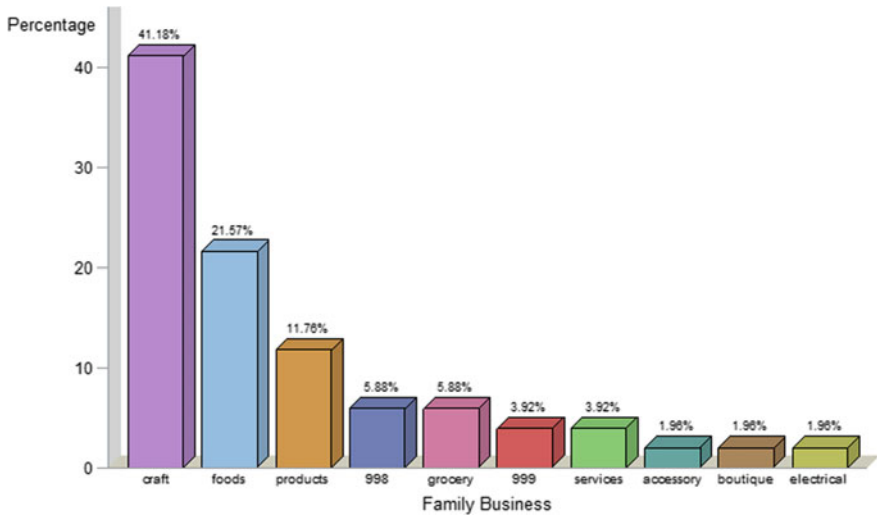


Fig. 2 Types of family business

percentage of females was expected since they are the dominant group. Approximately 30% of the respondents from entrepreneurial families have fathers as entrepreneurs.

Figure 2 shows that 41.18% of the family businesses are involved in the craft sector, followed by the food sector with 21.57%.

Figure 3 demonstrates that the craft sector is the most popular choice among respondents, collecting a total of 93.81%.

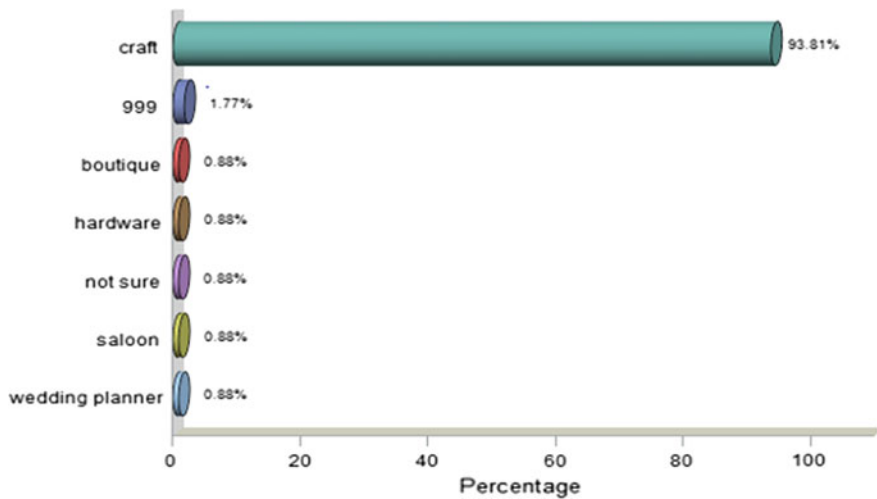


Fig. 3 Types of business students intend to do

Table 2 shows that the majority of the respondents (69.91%) expects to begin their own businesses within 3–5 years, while 20.35% expect to start within 1 to 2 years. Table 3 shows the expected pressures that might be faced when owning a business; 80.53% agreed that “inadequate funds to operate business” poses the highest pressure. Table 4 shows the main reason that respondents chose to become entrepreneurs is largely due to their interest in business (67.3%).

Table 5 shows that 42.48% have been involved in entrepreneurial activities at least once, while 39.82% have had the experience ranging between two to four times.

According to Table 6, only 20.4% of the students are involved in family businesses. Table 7 shows that the mean of entrepreneurial intention is 4.07; i.e., most of the respondents are agreed on the idea of being an entrepreneur. The means for the three factors (behavioral, attitudinal, and educational) are approximately the same, 3.78, 3.61, and 3.77, respectively; i.e., most are agreed with the factors. All independent variables (factors) have almost the same standard deviations.

Table 2 Expectation to begin business

Duration	Frequency	Percentage
Within 1–2 years	23	20.35
Within 3–5 years	79	69.91
Within 6–10 years	11	9.73

Table 3 Pressures that will be faced in business

Type of pressure	Frequency	Percentage
Inadequate knowledge on accounting and tax	72	63.72
Inadequate funds to operate business	91	80.53
Lack of computer skills	34	30.09
Problem in marketing, products and services	57	50.44
Poor internal control of the business	58	51.33
Incomplete record keeping	29	25.66
Lack of motivation	41	36.28

Table 4 Reasons to become an entrepreneur

Reasons	Frequency	Percentage
Interested to do business	76	67.3
Influence and encouragement by family members	9	8.0
Influence and encouragement by friends	8	7.1
To help family business	8	7.1
Desire to make quick profit	8	7.1
Others	4	3.6

Table 5 Entrepreneurial activities

Oftenness in entrepreneurial activities	Frequency	Percentage
Once	48	42.48
2–4 times	45	39.82
5–7 times	9	7.96
More than 7 times	6	5.31

Table 6 Businesses conduct related to family business

Conducting business	Is this a family business	Frequency	Percentage
Yes	Yes	23	20.4
	No	11	9.7
No	–	79	69.9

Table 7 Descriptive statistics of variables

Variables	Mean	Std. Deviation
Entrepreneurial intention	4.0708	0.70984
Behavioral factors	3.7796	0.56350
Attitudinal factors	3.6062	0.57170
Educational supports	3.7743	0.58793

The relationship between the three factors and entrepreneurial intention were examined. The results of Spearman Test are shown in Table 8.

H_1 : There is a significant relationship between behavioral factors and entrepreneurial intention; $r = 0.4684$ ($p < 0.05$)—a weak positive correlation.

H_2 : There is a significant relationship between attitudinal factors and entrepreneurial intention; $r = 0.4442$ ($p < 0.05$)—a weak positive correlation.

H_3 : There is a significant relationship between educational supports and entrepreneurial intention; $r = 0.5078$ ($p < 0.05$)—a moderate positive correlation.

Table 8 Analysis of spearman correlation

	Entrepreneurial intention	Behavioral	Attitudinal	Educational
Entrepreneurial Intention	1.000			
Behavioral factors	0.4684*	1.000		
Attitudinal factors	0.4442*	0.5545*	1.000	
Educational supports	0.5078*	0.3690*	0.3551*	1.000

*Correlation is significant at $p < 0.05$

All correlation coefficients are significant. Thus, the hypotheses H_1 , H_2 , and H_3 are accepted. The results also show that there is a relationship between the three factors towards developing entrepreneurial intention.

3.3 Multiple Linear Regression

A multiple linear regression model is used for testing the response variable (entrepreneurial intention) with predictor variables (educational supports, attitudinal factors, and behavioral factors), and hence to identify the most influential factor.

H_4 : Educational supports is the most influential that leads to entrepreneurial intention.

H_5 : Behavioral factors is the most influential that leads to entrepreneurial intention.

H_6 : Attitudinal factors is the most influential that leads to entrepreneurial intention.

Tables 9 and 10 display the results of multiple linear regression for entrepreneurial intention with the three factors. $R^2 = 0.4548$ means that 45.48% of variations in the entrepreneurial intention are explained by the three factors. Based on Table 10, since the significance value is 0.0001 (<0.05), there is a significant difference in the means of entrepreneurial intention between factors. Thus, the model is significant.

The final model is: $\hat{Y} = 0.5673 + 0.4460X_1 + 0.2578X_2 + 0.2409X_3$;

where \hat{Y} = Entrepreneurial Intention, X_1 = Educational Supports

X_2 = Behavioral Factors, and X_3 = Attitudinal Factors.

The highest standardized beta value of 0.4460 reveals that educational supports is the most influential factor that leads to entrepreneurial intention among students. Thus hypothesis H_4 is accepted.

Table 9 Multiple linear regression between variables and entrepreneurial intention

Variable	Beta	Sig.
Constant	0.5673	
Educational supports	0.4460	0.014*
Attitudinal factors	0.2409	0.007*
Behavioral factors	0.2578	0.000*

* $p < 0.05$

$R^2 = 0.4548$

Adjusted $R^2 = 0.4394$

Standard error = 0.4748

Table 10 Analysis of variance

Sources	Sum of squares	df	Mean square	<i>F</i>	Sig.
Regression	19.9317	3	6.6439	29.48	0.0001
Residual	23.8916	106	0.2254		
Total	43.8233	109			

4 Conclusions

This study provides information on the factors affecting entrepreneurial intention among IKN students. The majority of respondents want to be entrepreneurs. However, the respondents' preference to work for 3–5 years before venturing into their own businesses illustrated low levels of interest at the idea of becoming entrepreneurs immediately after graduation. With regards to the factors that influence individual entrepreneurial intention, there is a relationship between attitudinal factors, behavioral factors, and educational supports towards developing entrepreneurial intention among students. The respondents' behavior and attitude towards entrepreneurship is weakly related to their entrepreneurial intention, while the educational supports is found to have a moderate relationship. Nevertheless, the multiple regression analysis showed that the three factors affecting entrepreneurial intention were significant, thus proving that these factors do influence graduates' entrepreneurial interest, with educational supports being the most influential.

Further studies on entrepreneurial intention are strongly recommended by examining other factors such as the culture of entrepreneurship; e.g., hardworking, discipline, and integrity. In addition, the curriculum at IKN should be improvised and revised to allow students' creative and innovative thinking skills to be initiated. On top of that, IKN should provide a platform for students to translate their ideas into reality.

References

1. The Malaysian Insider: Graduates among 400,000 currently unemployed in Malaysia. <http://www.themalaysianinsider.com/malaysia/article/graduates-among-400000currently-unemployed-in-malaysia-says-minister> (2015). Accessed 13 May 2015
2. IPPBM: Indeks Belia Malaysia, Institut Penyelidikan Pembangunan Belia Malaysia, Kementerian Belia dan Sukan Malaysia (2008) <http://www.iyres.gov.my/index.php?lang=bm>
3. Chan, K.L., Geraldine, S.S., Bahiyah, A.H.: Malay youth entrepreneurship in Malaysia: an empirical update. *GEOGRAFIA Online Malays. J. Soc. Space* 5(2), 55–67 (2009)
4. Van Gelderen, M., Brand, M., Van Praag, M., Bodewes, W., Van Gils, A.: Explaining entrepreneurial intentions by means of the theory of planned behaviour. *Career Dev. Int.* 13(6), 538–559 (2008)
5. Bjerke, B.: *Understanding Entrepreneurship*. Edward ELgar, Cheltenham (2007)

6. Abu Bakar, S.G., Salwa, K., Amina, M.N.: An assessment of students' entrepreneurial intentions in Tertiary Institution: a case of Kano State Polytechnic, Nigeria. *Int. J. Asian Soc. Sci.* **4**(3), 434–443 (2014)
7. Mumtaz, B.A.K., Munirah, S., Halimahton, K.: The Relationship between educational support entrepreneurial intentions in Malaysia Higher Learning Institution. *Procedia Soc. Behav. Sci.* **69**, 2164–2173 (2012)
8. Zaharah, G., Asmahani, I., Fakrul, A.Z.: Factors affecting entrepreneurial intention among UniSZA students. *Asian Soc. Sci.* **9**(1), 85 (2013)
9. Karimi, S., Harm, J.A., Biemans, Lans, T., Mulder, M., Chizari, M.: Role of entrepreneurship education in developing students entrepreneurial intentions (2012). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152944
10. Meulenbroek, V.M. Causation and effectuation: the influence of family back-ground on the entrepreneurial decision making process in emerging countries, Bachelor thesis, International Business Administration, University of Twente (2014)
11. Nunnally, J.C., Bernstein, I.H. The assessment of reliability. In: *Psychometric Theory*, vol. 3, pp. 248–292. McGraw-Hill, New York, NY (1994)

Chapter 69

Improving the Food Manufacturing System by Using Simulation and DEA



Noor Fatin Kamarudin, Ruzanita Mat Rani
and Faridah Abdul Halim

Abstract This paper presents the application of simulation and Data Envelopment Analysis (DEA) in improving the food manufacturing system. Simulation and DEA are used to improve the system by identifying improvement models and determine the best improvement model. The simulation model is used to generate inputs and outputs of improvement models. DEA-BCC model with output orientation is used to determine the efficient improvement model which can maximize output with the given input. Then, cross efficiency is used to rank the efficient improvement models and select the best improvement model. The IP 4 is the best improvement model where the model suggested to relocate the workers. The application methods and the results given can assist the management of the company to make better decisions and can provide ideas to other SME companies for improving the performance of the food manufacturing system.

Keywords Simulation · DEA · Cross efficiency · Improvement model

1 Introduction

Small and Medium Enterprise (SME) sub-sector plays a vital role in the Malaysia economy and is considered as the backbone of industrial development in the country. An enterprise is considered as SME based on the annual sales turnover or number of full-time employees. SME in the manufacturing sector is defined as an

N. F. Kamarudin · R. M. Rani (✉) · F. A. Halim
Centre for Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
e-mail: ruzanita@tmsk.uitm.edu.my

N. F. Kamarudin
e-mail: noorfatinkamarudin@ymail.com

F. A. Halim
e-mail: faridahh@tmsk.uitm.edu.my

enterprise with full-time employees not exceeding 200 employees or with an annual turnover of not exceeding RM50 million.

SMEs in the food manufacturing sector are involved in activities such as processing and production of raw materials of food. There are many categories of food products produced by SMEs in Malaysia such as fishery products, processed fruits and vegetables, cocoa based products, beverages, frozen food and others. Shah and Ahmad (2015) summarized that the most important sector in the Malaysia economy is the food manufacturing sector. This is because the food manufacturing sector help other sectors such as service and agriculture sectors. For example in order to get the source of raw materials for production from agriculture in the production process.

Based on the Economic Census 2011, the food manufacturing sectors are comprised of 5,723 establishments or 15.1% of the total number of SMEs in the manufacturing sector at 37,861 establishments. The percentage shows that the food manufacturing sub-sector is the second highest after the wearing apparel sub-sector. The food manufacturing system of SMEs place more importance on food productivity [1]. The improvement of food manufacturing system to a company is to reduce the cost of production as well as to enhance the quality of products [2].

The remaining of this paper is organized as follows. Section 2 covers the background information about the food manufacturing system, the development of simulation model, DEA-BCC output orientation model and cross efficiency. The experimental results are deliberated in Sect. 3. Finally, Sect. 4 encompasses conclusions.

2 Materials and Methods

The food manufacturing system of product called CHOKEE is chosen in this study. This product is produced by SME company in Klang. The production line of the CHOKEE product has been divided into four processes. The first step in producing CHOKEE product is premix process. The premix process is to formulize the raw materials. Next is mixing process. The mixing process is to mix all raw materials. After mixing process is done, then filling process will be carried out. The filling process is the process in which the product is packed into packet of 1 kg. The final step is packaging process. In this process the packet of 1 kg will be placed in the box. Each box has 12 packets of 1 kg. The entities that are used in CHOKEE product production line is the powder of 1 kg packet.

In this study, the first stage is to develop simulation model of the actual system. Secondly, the improvement models will be identified. Then, to develop simulation models of the improvement models. Simulation model is used in order to obtain the inputs and outputs (performance criteria).

The simulation model is widely used to improve the performance of the food manufacturing system. The second stage is to develop the DEA model. The DMUs are the improvement models. Inputs and outputs obtained from the simulation

model will be used in the second stage. The DEA-BCC output orientation model will be used to maximize the outputs with the given inputs. The cross efficiency will be used to rank the improvement models.

Simulation model is developed using ARENA software to envisage the actual system. The simulation model of the actual system is developed based on data collection of daily activities on the production line. The data provided by the company is from 8.00 am until 5.00 pm for nine working hours. All the data obtained will be analyzed by Input Analyzer in order to identify the best distribution. The distribution obtained will be used in developing simulation model. The food manufacturing system in Fig. 1 is simulation model developed using ARENA software. Then, each simulation model must go through verification and validation process in order to check the similarity between simulation model and the actual system [3]. The difference must not exceed 10% to declare that the model is valid [4].

DEA is used to measure the relative efficiency of homogenous decision making units (DMUs). The evaluation is based on the inputs used and outputs produced. DEA is used as a decision making tool in determining the best decision among possible alternatives. BCC model output orientation is used in the study. The BCC model output orientation (1) is shown below [5].

$$\theta_0 = \min \sum_{i=1}^m z_i u_{ik0} - z_0$$

$$\text{All is s.t } \sum_{r=1}^s x_r y_{rk0} = 1, \sum_{r=1}^s x_r y_{rk} - \sum_{i=1}^m z_i u_{ik} + z_0 \leq 0 \quad (1)$$

$$z_0 \text{ free}, x_r, z_i \geq \varepsilon, r = 1, \dots, s, i = 1, \dots, m_k = 1, \dots, n$$

where θ_0 is the relative efficiency of DMU_0 . DMU_0 is DMU under evaluation, k is the DMU index, r is the output index, i is the input index y_{rk} is the value of the r th output for the k th DMU, u_{ik} is the value of the i th input for the k th DMU, x_r is the weight given to the r th output, z_i is the weight given to the i th input, and DMU_0 is efficient if $\theta_0 = 1$, DMU_0 is inefficient when $\frac{1}{\theta_0} < 1$. To choose the best DMU, Cross

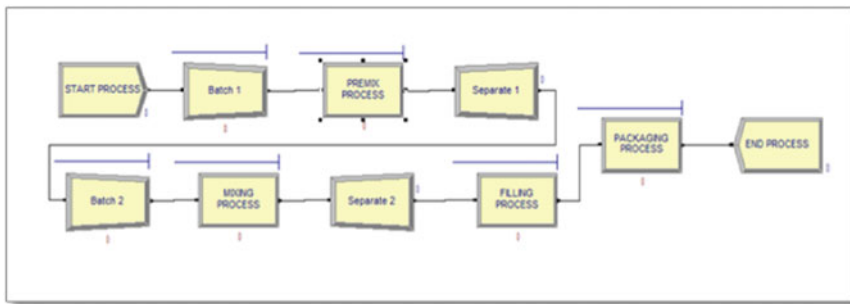


Fig. 1 Simulation model of CHOKEE product production line

efficiency will be used to rank the efficient DMUs. The efficient DMUs is ranked using cross efficiency. DMU in the first rank will be the best DMU. Cross efficiency (2) that is calculated based on input and output weights of efficient DMUs [6].

$$E_{vt} = \frac{\sum_{i=1}^m z_{iv} u_{it}}{\sum_{r=1}^s x_{rv} y_{rt}} \quad v, t = 1, 2, \dots, n \quad (2)$$

where E_{vt} is the score for DMU_t using the optimal weights selected by DMU_v, y_{rt} is the value of the r th output for DMU_t, u_{it} is the value of the i th input for DMU_t, x_{rv} is the weight given to the r th output for DMU_v and, z_{iv} is the weight given to the i th input for DMU_v. The average of E_{vt} can be calculated using this formula

$$\bar{E}_t = \frac{1}{n} \sum_{v=1}^n E_{vt} \quad (3)$$

3 Experimental Result

The results of the simulation model which consist of the average processing time, the average waiting time and the average total processing time. The average processing time for the mixing process is 22.1762 min. According to the management of the company, the processing time of the mixing process is targeted for not more than 20 min. This is to maintain the quality of CHOKEE product. The average waiting time of the premix process is 5.6489 min and packaging process is 6.3607 min, respectively. In the premix process, no machine is used. All task in the premix process is done by workers. In the packaging process only one worker is involved. Hence, it will cause a long waiting time of each process. Therefore, the management of the company has decided to reduce the waiting time in the premix and packaging process.

Each process has a different number of workers. Two workers are assigned to the premix process as Workers 1 and 2. Next process is the mixing process in which three workers labeled as Workers 3, 4, and 5 are assigned. Workers 6 and 7 are assigned to the filling process. Only one worker is assigned to the packaging process which is Worker 8.

Based on the results of the simulation model, average worker utilization is unbalanced. The workers who are involved in the premix and packaging process, their average worker utilization percentage are almost 100%. It can be seen that the bottlenecks of the system occurred in the premix and packaging process that caused the entity to queue longer compared to the mixing and filling processes.

In this study, 14 improvement models are identified. IP 1: One worker from mixing process is transferred to the premix process and the total workers in premix

process is three. IP 2: One worker from mixing process is transferred to the filling process and the total workers in filling process is three. IP 3: Two workers from mixing process is transferred. One worker to the premix process and one worker to the filling process, respectively. IP 4: One worker from mixing process is transferred to the packaging process and the total workers in packaging process is two. IP 5: One worker from premix process and one worker from filling process are transferred to the packaging process and the total workers in packaging process is three. IP 6: One worker from premix process and one worker from filling process are transferred to the packaging process and the time between arrival is reduced by 5 min. IP 7: The time between arrival is reduced by 5 min. IP 8: The time between arrival is reduced by 8 min. IP 9: The processing time of filling process is reduced by 3 s. IP 10: The processing time in filling process is reduced by 3 s while processing time in packaging process is reduced by 4 s. IP 11: One worker from premix process is transferred to the filling process and the processing time in filling process is reduced by 3 s while the processing time in packaging process is reduced by 4 s. IP 12: One worker from premix process is transferred to the filling process, the processing time is reduced by 8 min and the processing time in filling process is reduced by 3 s while the processing time in packaging process is reduced by 4 s. IP 13: One worker in filling process is reduced and the total of workers for the whole process is 7 workers. The processing time in filling process is reduced by 3 s while the processing time in packaging process is reduced by 4 s. IP14: The worker in the premix, filling, and packaging process is added. Each process has three workers. This means that the total of workers for the whole process are 12 workers. The processing time in filling process is reduced by 3 s while in the processing time in packaging process is reduced by 4 s.

In this study, the decision making units (DMU) are improvement models (IP 1, IP 2, IP 3, IP 4, IP5, IP 6, IP 7, IP 8, IP 9, IP 10, IP 11, IP 12, IP13, IP 14). DEA can be used as a decision making tool in determining the best alternative [7]. The average total production time, the average number of entities in the system, and the number of workers are the inputs. The outputs are average total production and the average worker utilization. Table 1 shows the efficiency score of each improvement model.

Based on cross efficiency score, IP 4 is selected as the best improvement model. According to IP 4, reallocation of worker is performed due to the bottleneck that occurred in the packaging process. Thus, two workers are allocated to the mixing

Table 1 The efficiency score of each improvement model

Improvement model (DMU)	Efficiency score	Improvement model (DMU)	Efficiency score	Improvement model (DMU)	Efficiency score
IP 1	0.9905	IP 6	0.4576	IP 11	0.4411
IP 2	0.9052	IP 7	0.9112	IP 12	0.4411
IP 3	0.5572	IP 8	0.9112	IP 13	1.0000
IP 4	1.0000	IP 9	0.8538	IP 14	1.0000
IP 5	0.4576	IP 10	1.0000		

Table 2 Comparison between actual simulation model and improvement model 4

Model	Average total output (packets)	Average WIP (packet)	Average worker utilization (%)
Actual simulation model	1763	487	81.37
IP 4	1804	445	78.25
Difference (%)	2	9	4

process and two workers are allocated to the packaging process. By implementing this improvement, the company's management can reduce the cost of hiring additional workers. IP 4 also scores the highest on the average worker utilization. IP 4 can produced 1804 packets of CHOKEE with 445 packets of CHOKEE still in the system within 9 h of operation.

Table 2 shows that the average processing time, average waiting time, and average total production time of IP 4 decreases compared to the actual simulation model. IP 4 shows an increase of 2% in the total output and also reduces the number of entities which are still in the system by 8%. The average worker utilization increases by 4% when workers are reallocated to the bottleneck area. As a consequence, IP 4 is selected as the best improvement model since the modifications implemented can improve the actual model to become more efficient and productive.

4 Conclusions

In conclusion, the objectives of this study are achieved. The first objective of this study is to identify the activities that can cause bottlenecks in the food manufacturing system. The activities that can cause bottlenecks in the operating system are identified in the packaging process. The second objective is to identify improvement models. In this study the improvement models are identified based on discussion and agreement with the management of the company. The 14 improvement models have been suggested. The third objective is to determine the efficient improvement models. This study has identified the efficient improvement models. IP 4 is the best improvement model which suggested to reallocate one worker from the mixing process to the packaging process in order to overcome the bottleneck problem.

Acknowledgements This study was made possible by the continuous support from Universiti Teknologi MARA Grant No. 600-IRMI/DANA 5/3/LESTARI (0130/2016).

References

1. Dora, M., Kumar, M., Van Goubergen, D., Molnar, A., Gellynck, X.: Operational performance and critical success factors of lean manufacturing in European food processing SMEs. *Trends Food Sci. Technol.* **31**(2), 156–164 (2013)
2. Gunday, G., Ulusoy, G., Kilic, K., Alpkan, L.: Effects of innovation types on firm performance. *Int. J. Prod. Econ.* **133**(2), 662–676 (2011)
3. Sargent, G.R.: Verification and validation of simulation models. In: *Proceedings of the Winter Simulation Conference*, pp. 166–183 (2010)
4. Anderson, D.R., Sweeney, D.J., Williams, T.A.: *An Introduction to Management Science*, pp. 619. Thomson South-Western, Ohio (2005)
5. Banker, R.D., Charnes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manag. Sci.* **30**(9), 1078–1092 (1984)
6. Sexton, T.R., Silkman, R.H., Hogan, J.A.: Data envelopment analysis: Critique and extensions. *New Dir. Progr. Eval.* **1986**, 73–105 (1986)
7. Mat Rani, R., Ismail, W.R., Ishak, I.: An integrated simulation and data envelopment analysis in improving SME food production system. *World J. Model. Simul.* **10**(2), 136–147 (2014)

Chapter 70

Modelling Multi-dimensional Contingency Tables: LASSO and Stepwise Algorithms



Nur Huda Nabihan Md Shahri and Susana Conde

Abstract This study identifies an efficient method for the main interaction between categorical variables in multi-dimensional contingency tables. LASSO and Stepwise Algorithms are cross-validation methods for finding the penalty coefficient and model selection method. The methods used the Akaike Information Criterion (AIC) and p -value as indicators for selecting parameters to be included in the model. The aims of the study are to review the literature related to multi-dimensional contingency tables with log-linear models and high dimensional tables; to analyse the obesity dataset from Locksmith (GSK) GP Research Database from around the year 2000, where the dataset is composed of $p = 10$ binary comorbidities in $n = 5000$ patients using the models; and lastly, to compare the results obtained from the models. Stepwise Algorithms is an appropriate method for finding the parsimonious interaction structure between the categorical variables. The method defines a continuous shrinking operation that can produce coefficients which are exactly zero.

Keywords Multi-dimensional • Contingency table • LASSO • Stepwise algorithms

1 Introduction

The term contingency table is connected with cross-classified categorical data [1]. The observations or units of a sampled population are cross-classified according to each of several categorical variables such as disease, gender, age, or species. The early history of contingency tables began in the year 1900 focusing only on measuring associations and hypergeometric analyses for lower-dimensional

N. H. N. Md Shahri (✉)
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
e-mail: huda_nabihan@tmsk.uitm.edu.my

S. Conde
University of Glasgow, Glasgow, UK

(two-way) tables [2]. In two-way contingency tables, we can find the dependence structure between the variables using the Pearson chi-squared test for independence and deviance. The observed value of the statistic can be compared with a $\chi^2(n - p)$ distribution where n is the number of observations and p is the number of variables [3]. The Pearson chi-squared statistic is equivalent to maximum likelihood estimation in estimating parameters.

However, contingency tables are constructed with categorical variables. The history of categorical data analysis started in the 19th century with the main focus on developing log-linear models for modelling the counts and maximum likelihood estimation [4]. It has been developing rapidly in the past 50 years. Its application in social science has become routine in recent years [5]. When we cross-classify more than two variables, the table is called a multi-dimensional table. If the dimensions of the table are high, then it becomes difficult to find the interaction structure between the variables for various reasons including computation problems (many parameters to be estimated) and no knowledge a priori whether the estimates exist (for example, if the table contains zeros or low frequencies, this kind of problem may arise). Such tables pose special problems for analysis and interpretation [6].

2 Methodology

A dataset with $p = 10$ binary comorbidities from Locksmith (GSK) GP Research Database is applied in this study. Comorbidity is a medical condition that coexists with a primary disease. We are interested in finding the interaction pattern between the comorbidities. used. These are coded 1 for present and 0 for absent. The comorbidities are alcohol use, arthralgia, back pain, blood pressure, contraception NOS, coughing, adequate exercise, normal cervix smear, smoking, and weight.

2.1 Contingency Table

We will focus on contingency tables constructed with binary variables. The dimensions of the tables will be at least 2×2 or higher, the latter arising when we classify the observations of a sample with more than two variables. The entries in the cells for these data are called frequencies or counts.

Cross-classified categorical data analysis is one of the multivariate analysis techniques. In this study, the binary variables are coded 0 and 1. We will assume a sample formed with n independent and identically distributed observations of p binary variables X_1, X_2, \dots, X_p . We will construct a p -dimensional contingency table so it will be composed of $q = 2^p$ cells. We assume that the frequencies in the cells, denoted as Y_1, Y_2, \dots, Y_q , follow a multinomial random variable.

2.2 Log-Linear Models

Let X_1, X_2, \dots, X_p of p binary variables, which may represent diseases, recurrence, and others, be coded as 1 for present and 0 for absent and be observed in n observations. We consider a contingency table formed with these binary variables [7] and denote Y_{i_1, i_2, \dots, i_p} for the random variable that represents the frequency in the cell (i_1, i_2, \dots, i_p) where $(i_1, i_2, \dots, i_p) \in B = \{0, 1\} \times \{0, 1\} \times \dots \times \{0, 1\}$ so that $i_j = 0, 1$. We consider a bijective mapping between B and $E = \{1, 2, 3, \dots, 2^p\}$ [8] such that the leftmost subscript varies the fastest.

A log-linear model becomes

$$\log E(Y_i) = \theta^{C0} + y_i^T \theta, i = 1, 2, \dots, q \tag{1}$$

where $E(Y_i)$ is the expectation of the random variable Y_i , θ^{C0} , θ is a parameter vector with dimension k , and is the i th row of a design matrix X with dimensions $q \times k$. For example, for a 2^3 table, a saturated log-linear model will contain eight parameters: the constant, three main effects, three two-way interactions, and one three-way interaction.

2.3 Hierarchical Models

In a hierarchical model, when a higher-order (interaction) term is included in the model, then all the related lower-order terms are also included [8]. We can say that the higher-order interaction term measures a deviation from the lower-order terms [8]. For instance, if the two-way (first-order) interaction C1C2 is included, then the main effects C1 and C2 are also included. Moreover, hierarchical models can be defined in the following way: if an interaction θ_a is zero, then all higher-order interactions θ_b are also zero for all $b \subseteq a$ [9].

2.4 Maximum Likelihood Estimation (MLE)

The joint probability density function $Y = (Y_1, Y_2, \dots, Y_q)$ is viewed as a function of π and the likelihood is calculated as:

$$L(\pi_1, \pi_2, \dots, \pi_q | y) = \prod_i^q \frac{\pi_1, \pi_2, \dots, \pi_q}{y_1! \dots y_p! (q - y_1 - \dots - y_p)!} \pi_1^{y_1} \dots \pi_p^{y_p} (1 - \pi_1 - \dots - \pi_p)^{n - y_1 - \dots - y_p} \tag{2}$$

The aim of the Maximum Likelihood Estimation (MLE) is to find the parameter values that maximise the likelihood function, L , and to find the set of values of the unknown parameters that make the MLE, L as large as possible. Alternatively, we can maximise the log-likelihood function instead of the likelihood function. It is because the values that maximise L will also maximise the value of $\log L$. The logarithmic transformation is referred to as a monotone transformation [5].

2.5 Lasso

The LASSO is a regression method that involves penalising the absolute size of the regression coefficients based on the value of the tuning parameter, λ . To do this, the LASSO will eliminate variables to zero, thus performing an automatic variable selection.

Moreover, the LASSO will shrink the value of nonzero coefficients. It is similar to ridge regression when penalising the absolute values of the coefficients introduces shrinkage towards zero.

The LASSO is a popular tool for sparse linear regression, and it is most commonly used when dealing with a huge number of variables p and when p exceeds the number of observations [10]. Moreover, it is used to find a parsimonious pattern as the data has many zeros, and it may occur that the maximum likelihood estimates of some parameters do not exist (i.e. they are infinity).

The LASSO is used for linear regression to estimate parameter and variables selection. It is also a constraint optimisation problem. Let (x_i, y_i) , $i = 1, 2, \dots, N$, where $x_i = (x_{i1}, \dots, x_{ip})^T$ are the predictor variables, and y_i are the responses. The LASSO can be defined using the least square function as shown below [11]:

$$\hat{\beta}^\lambda = \arg \min_{\beta} \left[\sum_i (Y - X\beta)_i^2 + \lambda \sum_j |\beta_j| \right] \tag{3}$$

with notation:

$Y = (Y_1, Y_2, \dots, Y_n)$ is the response vector

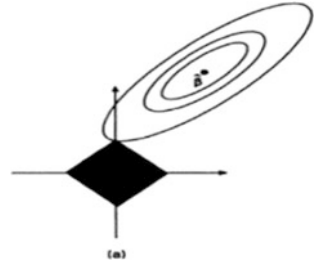
$\hat{\beta}^\lambda$ is the penalised least square estimator

$Y = X\beta + \varepsilon$

λ is the regularisation parameter and it will be estimated.

In some cases, the MLE is undesirable in regression due to large variability; p is large relative to n and lacks interpretability. Perhaps the most common solution to this problem is to find a subset of important variables. As we mentioned above on

Fig. 1 Estimation picture for the LASSO



the characteristics of the LASSO, the estimated LASSO is equal to zero when $t = 0$. Therefore, the MLE does not exist in the case of zero count in the contingency table.

Likewise, we can have a zero count in a contingency table and more variables than in the observations. In this case, the LASSO can be the more appropriate estimation method as some parameters may be zero, and then it becomes ‘a sparse model’ (Fig. 1).

2.6 Cross-Validation

The K-fold cross-validation is a model validation technique for assessing the accuracy and validity of a statistical model by generalising an independent dataset into roughly equal-sized parts. The independent dataset is divided into two parts: the test data from the first set of data and training data from all other sets. One observation is taken out at a time. Modelling of the data uses one part only which is from the training data. We then repeat this process K-1 more times, each time using a different part of the dataset as the test data, and the remainder being the training data.

Each time the process is repeated, we calculate the predictive negative log-likelihood score proportional to the out-of-sample negative log-likelihood or prediction errors. Each prediction error is calculated using λ , and the mean of the prediction errors is referred to as the cross-validation score. The value of λ is defined by the minimum values of the cross-validation score. This score is on the same scale when varying the number of observations and hence, can be used to compare different contingency tables with different number of cells entries [9]. Thus, to estimate the value of β , we need to perform an algorithm transformation using the λ .

Subsequently, the model selected for this part is used to predict the values in the training set of data. The goals of cross-validation are to avoid the problem of overfitting the model(s) to a single dataset and to define the model in the test data that shows good predictive accuracy. In this study, we are going to use the 5-fold cross-validation technique.

2.7 Stepwise Algorithms

For **backward elimination**, we start with all variables included in the model. Then we test the deletion of each variable using the chosen model's comparison criterion. In our case, we are using the Wald Test (p -value) Akaike Information Criterion (AIC). We add the variable (if any) that improves the model the most by being deleted and repeat this process until no further improvement is possible.

On the contrary, in the **forward selection**, we start with no variable in the model. Then we test the addition of each variable using the chosen model's comparison criterion. Again, in our case, we are using the Wald Test (p -value) AIC. We add the variable (if any) that improves the model the most, and we repeat this process until no improvement of the model is possible.

The AIC is used as a measurement tool, computed and compared to find a model with the minimum value of the AIC. The model with the minimum AIC is chosen and considered as the best model. When applying the AIC, our aim is to compare the goodness of fit of various models and choose the one that is most parsimonious.

Then we construct a contingency table with the 10 binary variables. Let $q = 2^{10} = 1024$ be the number of cells. We will name the cells $c_1, c_2, c_3, \dots, c_{10}$ in order. The first cell in the table is indexed as $Y_1 = (0, 0, \dots, 0)$ and its frequency represents the number of patients with all comorbidities absent. Using the order commented before, the next cell is $Y_2 = (1, 0, \dots, 0)$ and its frequency represents the number of patients who have 1 comorbidity present and all the other comorbidities absent, and so on.

3 Results

We use **Yates' code design matrices**, and hence, the columns are orthogonal. Two vectors are orthogonal when the scalar product of the vectors is equal to one. Thus, we have to transform the design matrix by changing the values of zero to -1 .

In our general dataset, we have the contingency table with 100 variables. However, we are interested in finding the interaction pattern only for a 10-dimensional table as mentioned above.

We are going to perform model selection using log-linear models to find the interaction pattern. The best model to fit the data will be identified using the *glm* command, the LASSO method, and Stepwise Regression backward and forward elimination.

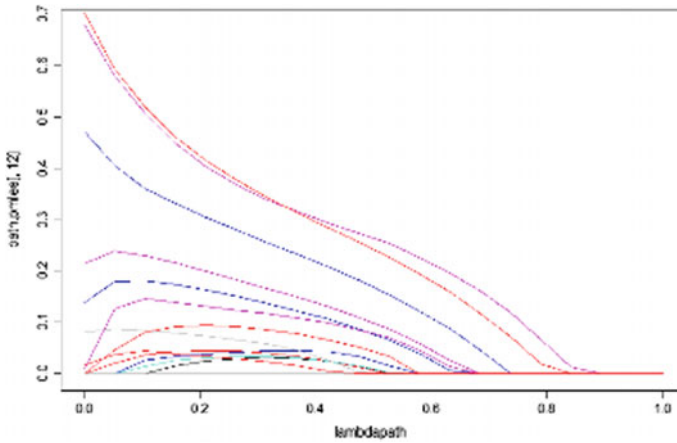


Fig. 2 LASSO shrinkage

Table 1 Stepwise algorithms

Method	No. of variables	Null deviance	Residual deviance	AIC
LASSO	34	38,705.9	889.91	1,902.4
Stepwise backward (<i>p</i> -value)	36	38,705.9	582.25	1,604.7
Stepwise: Backward (AIC)	29	38,706	1,600	2,596.5
Stepwise: Forward	34	37,710	1,596.1	2,604.9

From the simulation of the LASSO package, we find that:
 $\lambda = 0.006013$ and negative likelihood score = -7.505172 .

The **refitted** model is:

$$Y = c1 + c2 + c3 + c4 + c5 + c6 + c7 + c8 + c9 + c10 + c1c7 + c1c8 + c1c9 + c1c10 + c2c3 + c2c6 + c2c7 + c2c9 + c2c10 + c3c5 + c3c6 + c3c7 + c4c8 + c4c9 + c4c10 + c5c6 + c5c7 + c5c8 + c5c9 + c6c7 + c6c10 + c7c8 + c7c9 + c7c10 + c9c10$$

The plotted graph shown in Fig. 2 indicates that the value of λ is close to zero when the estimated value β approaches infinity.

From the analysis, we can summarise the output in the table: (Table 1).

4 Conclusion and Discussion

An efficient method was developed for identifying interaction patterns of categorical variables. A few variables showed interaction with each other. This study was based on the data of diseases provided by the pharmaceutical company which had

given us general and amenable ways for regularisation analysis in handling high dimensional, sparse contingency tables. The sampling scheme maximum likelihood estimator in a log-linear model from a contingency table can be said as a multinomial distribution function. This is because it includes 10 variables with 2^{10} random variables. In this study, we can say that the LASSO is not an appropriate method for finding a parsimonious interaction structure in the obesity data compared to the classical method, namely the stepwise algorithms. In fact, the LASSO is not a hierarchical model where one variable is included in the interaction term; instead, none of the main effects is included whereas the objective is focused on the main effect. Nevertheless, the new method (LASSO) is suitable for shrinkage and selection for regression and generalised regression problems. Even though we faced some difficulties due to the large dataset which we had about 10 variables and a contingency table with many zeros, the LASSO defines a continuous shrinking operation that can produce coefficients of exactly 0.

References

1. Pearson, K.: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinb. Dublin Philos. Mag. J. Sci.* **50**(302), 157–175 (1900)
2. Heyde, C.C., Seneta, E.: *IJ Bienaymé: statistical theory anticipated*, vol. 3. Springer Science & Business Media (2012)
3. Dobson, A.J., Barnett, A.: *An introduction to generalized linear models*. CRC Press (2008)
4. Fienberg, S.E., Rinaldo, A.: Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J. Stat. Plan. Inference* **137**(11), 3430–3445 (2007)
5. Powers, D., Xie, Y.: *Statistical Methods for Categorical Data Analysis*. Emerald Group Publishing (2008)
6. Fienberg, S.E.: *The Analysis of Cross-Classified Categorical Data*. Springer Science & Business Media (2007)
7. O’Flaherty, M., MacKenzie, G.: Algorithm AS 172: direct simulation of nested Fortran DO-LOOPS. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **31**(1), 71–74
8. Bishop, Y.M., Fienberg, S.E., Holland, P.W., Light, R.J., Mosteller, F.: Book review: discrete multivariate analysis: theory and practice. *Appl. Psychol. Meas.* **1**(2), 297–306 (1977)
9. Dahinden, C., Parmigiani, G., Emerick, M.C., Bühlmann, P.: Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinform.* **8**(1), 476 (2007)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* 267–288 (1996)
11. Tibshirani, R.J.: The lasso problem and uniqueness. *Electron. J. Stat.* **7**, 1456–1490 (2013)

Chapter 71

Microwave Tomography: A Numerical Study of Solving Linear Equations in the Non-linear Inverse Scattering Problem



Latifah Mohamed, Nur Adyani Mohd Affendi, Azuwa Ali
and Nooraihan Abdullah

Abstract Microwave tomography is an inverse scattering problem which is formalized by determines the position and dielectric properties distribution of the unknown object from the measured scattered field. The quantitative image obtained from microwave tomography approach provides information directly correlated to the internal structure and composition of the examined object, which is necessity in biomedical and geo-surveying applications. Recently, studies on microwave imaging (MWI) for early breast cancer detection have attracted attentions. The breast cancer detection based on microwave tomography relies on large differences in dielectric properties between healthy and malignant tissues. In MWI, the examined object is successively illuminated by transmitting antenna and the resulting electromagnetic field is measured by the receiving antennas. The image reconstruction process in microwave tomography involves forward problem and inverse problem. For the forward problem, Method of Moment (MoM) is used to obtain the measurement data, which are related to the scattered field resulting from the interaction between the known incident field and the scatterers inside the imaging region. The measurement data is then used in the inverse problem, where Distorted Born Iterative Method (DBIM) is proposed to reconstruct the dielectric

L. Mohamed (✉) · N. A. M. Affendi · A. Ali
School of Electrical Systems Engineering, Universiti Malaysia Perlis, Arau, Malaysia
e-mail: latifah@unimap.edu.my

N. A. M. Affendi
e-mail: nuradyani@unimap.edu.my

A. Ali
e-mail: azuwa@unimap.edu.my

N. Abdullah
Institut Matematik Kejuruteraan, Universiti Malaysia Perlis, Arau, Malaysia
e-mail: raihan@unimap.edu.my

properties distribution of the object by solving the non-linear inverse scattering problem. The simulation results show that the ill-posedness of the non-linear problem can be reduced and a more stable solution can be performed by choosing the appropriate solving techniques of the system of linear equations.

Keywords Microwave tomography · Inverse scattering · Method of moment · DBIM · Non-linear

1 Introduction

Recently, studies on the early detection of breast cancer by microwave imaging (MWI) have attracted considerable interest among researchers [1]. This imaging system has advantages such as low cost, being non-invasive and provides high image resolution, thus it is potential for an early cancer detection. The two major approaches of microwave imaging today are radar-based imaging and tomographic methods. In radar-based imaging, strong scatterers is found inside an object, meanwhile the latter approach generates the cross-sectional slices of the dielectric properties.

The breast cancer detection based on MWI relies on large differences in electromagnetic properties between normal and malignant tissues (cancer). In MWI, the examined object is successively illuminated by transmitting antenna and the resulting electromagnetic field is measured by the receiving antennas.

In this report, a compact sized imaging sensor with the multi-polarization configuration as proposed in [2, 3] is considered. The image reconstruction process in microwave tomography involves forward problem and inverse problem. For the forward problem, Method of Moment (MoM) is used to obtain the measurement data, which are related to the scattered field resulting from the interaction between the known incident field and the scatterers inside the imaging region. The measurement data is then used in the inverse problem, where Distorted Born Iterative Method (DBIM) [4, 5] is proposed to reconstruct the dielectric properties distribution of the object by solving the non-linear inverse scattering problem. The appropriate solving techniques of the system of linear equations are investigated.

2 Imaging Algorithm

2.1 Forward Problem

Solving the forward problem gives the computed field at the receiving points. Several different methods can be used to implement the wave equation into the forward problem. Method of Moment (MoM) is the conventional approach for solving volume integral equations in the frequency domain.

Suppose the inhomogeneous medium with a finite volume S is embedded in homogeneous background medium with relative permittivity ϵ'_b and conductivity σ_b . The total field e_v is assumed to be the sum of the incident field e_v^i (without an object) and the scattered field e_v^s (caused by the object), according to Eq. 1.

$$e_v(r) = e_v^i(r) + e_v^s(r) \quad (1)$$

where the notation v indexing the total number of views which is a multi-view process by rotating the receivers. The scattered field e_v^s on the receivers in form of a convolution in the integral formulation as written in Eq. 2.

$$e_v^s(r) = \iint_S G(r, r') e_v(r') \chi(r') dr' \quad (2)$$

where the term $G(r, r')$ represent the three-dimensional (3-D) free-space Green's function, $e_v(r')$ is the total field inside the object region, and $\chi(r')$ is the contrast of the complex permittivity. The index r represent the observation points (i.e. the receiving antennas), meanwhile r' represents the source point inside the object region. By inserting Eq. 2 into Eqs. 1, 3 gives the equation for the total field.

$$e_v(r) = e_v^i(r) + \iint_S G(r, r') e_v(r') \chi(r') dr' \quad (3)$$

2.2 Inverse Problems

The inverse problem is formalized by finding the position and complex permittivity distribution of the unknown object from the measured scattered field. The examined object is fine discretized into voxels. In each voxel, the complex properties are estimated, which consist of the relative permittivity and conductivity distribution. The inverse problem involves three major steps at each iteration:

- (1) Minimize the difference or error between the measured scattered field and the calculated scattered field.
- (2) From (1), find the Jacobian matrix, J (the derivate-matrix of the computed scattered field with respect to the complex contrast in the object).
- (3) Update the contrast distribution of the object under investigation.

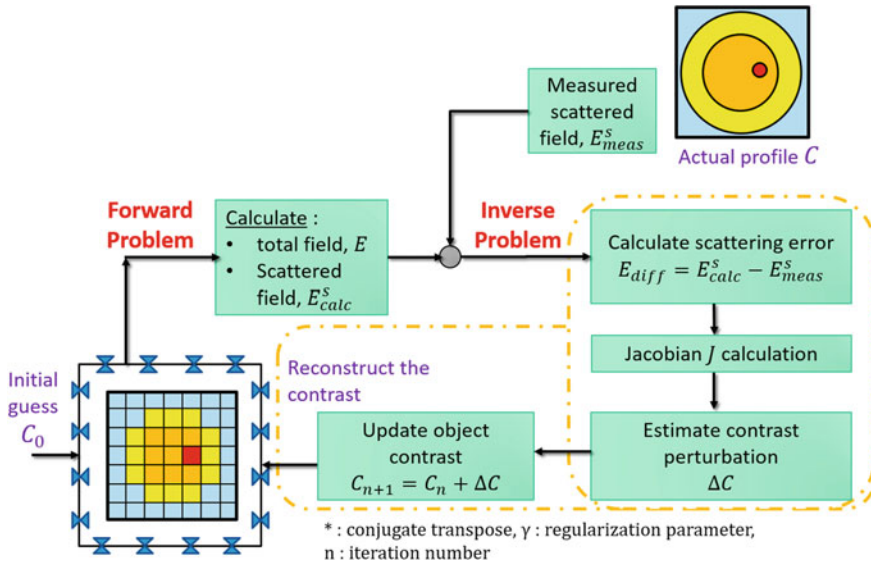


Fig. 1 Flowchart of microwave tomography-based breast imaging

This process continues until the error is sufficient small, thus the reconstructed image of the object is the complex permittivity map used in the direct problem, depicted in Fig. 1. In this report, the electromagnetic inverse scattering problem is solved Distorted Born Iterative Method (DBIM) [4, 5].

Distorted Born Iterative Method (DBIM). Based on Eq. 4, for a measurement at an observation location \mathbf{r} at a given frequency, the scattered electric field can be expressed by an integral equation

$$\begin{aligned}
 E^s(\mathbf{r}) &= E^t(\mathbf{r}) - E^i(\mathbf{r}) \\
 &= k_0^2 \int_S \overline{G}^b(\mathbf{r}|\mathbf{r}') E^t(\mathbf{r}') [\varepsilon^*(\mathbf{r}') - \varepsilon_b(\mathbf{r}')] d\mathbf{r}'
 \end{aligned}
 \tag{4}$$

The total field E^t in S depends on the multiple scattering interactions between the features of the complex permittivity. Therefore, the scattered field E^s is nonlinearly related to the contrast function due to the product $E^t(\mathbf{r}')[\varepsilon^*(\mathbf{r}') - \varepsilon_b(\mathbf{r}')]$ in the inte-grand of Eq. 4. Here, the distorted Born iterative method (DBIM) is employed, where the non-linear relation can be linearized using the Born approximation. At each iteration of the DBIM, the total field within the actual permittivity in S is approximated by the total field in the background medium $\varepsilon_b(\mathbf{r})$ i.e. E^b replace E^t in Eq. 4. The approach requires computation of the fields at the antennas and inside S for each iteration of the background medium. DBIM method begins with an initial

guess for the background permittivity that may include any available a priori information about the object permittivity [4–6]. In this case, the initial guess was chosen to be close to those of adipose tissues. DBIM does not require inverse matrix calculation for the computation of the Jacobian matrix.

2.3 Solving Linear Equations Techniques

Empirical Method. The contrast perturbations or solution, ΔC in Fig. 1 can be determined by Newton-Kantorovich method. The computation of the Jacobian matrix J is a central part of the optimization process. J is the derivative matrix containing the scattered field dependence of the contrast inside the object. The regularization parameter γ can be determined by empirically according to the convergence of the process [7]. The computation of the contrast step may be derived from Eq. 5 with the known Jacobian matrix.

$$\Delta C = [J^*J + \gamma I]^{-1} J^* E_{diff} \quad (5)$$

L-Curve Method. Another technique can be used to solve the linear equations is using Tikhonov method. Any regularized solution must lie on or above the Tikhonov L-curve. Hence, γ corresponding to the L-curve's corner is chosen as the appropriate regularization parameter. The L-curve almost always has a characteristic L-shaped appearance with a distinct corner when plotted in log-log scale [8, 9].

Conjugate Gradient Stabilized Method (Bi-CGSTAB). This algorithm is rather stable, converge smoother and produces more accurate solutions [10]. The bicgstab command of MATLAB at the requested tolerance and number of iterations is used. For a linear equation of $Ax = b$, the unknown x is determined by $x = \text{bicgstab}(A, b, \text{tol}, \text{maxit})$, where tol and maxit specifies the tolerance of the method and the maximum number of iterations, respectively. Here, we assumed that the regularization parameter determined by empirical method is empirically good.

3 Simulation Model and Results

Figure 2a is the actual model of breast and the cancer is marked by an arrow for clarity. Here, 10% of the volume ratio of breast model is occupied by fibro-glandular tissues that are distributed randomly within the adipose tissue.

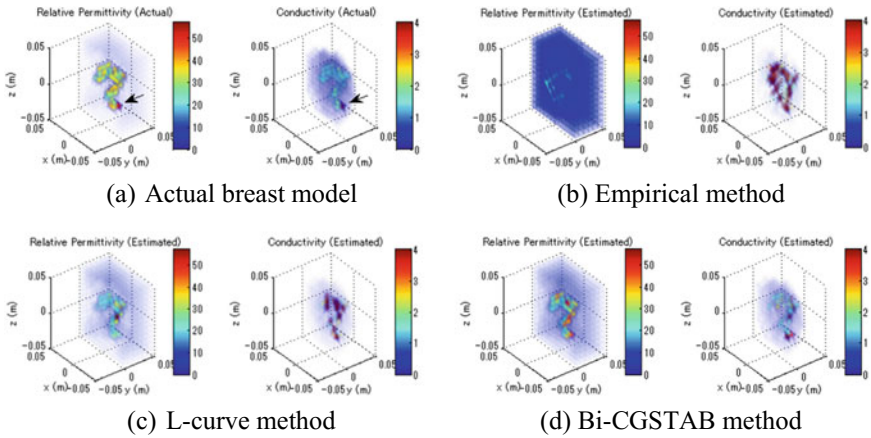


Fig. 2 3-D distribution using different solving techniques

The existence of chest wall is omitted to clearly show the contrast between tissues in breast model. Figure 2b–d show the results of the 3-D reconstructed images after 250 iterations. It is observed that the L-curve method perform better reconstruction in relative permittivity compared to empirical method. It is understood that to choose an appropriate regularization parameter at every iteration of DBIM is difficult. However, the image reconstruction is adequate when using Bi-CGSTAB. The fibro glandular tissues were approximately reconstructed and the cancer presence is distinct for relative permittivity and conductivity.

The comparison on performance metrics for 100 iterations using three solving techniques of nonsymmetrical linear systems is shown in Fig. 3. It is observed that the quality factor performance of Bi-CGSTAB is superior to empirical and L-curve method. Similarly, the root mean square error (RMSE) of Bi-CGSTAB is gradually decreasing along with iteration number compared to other two methods.

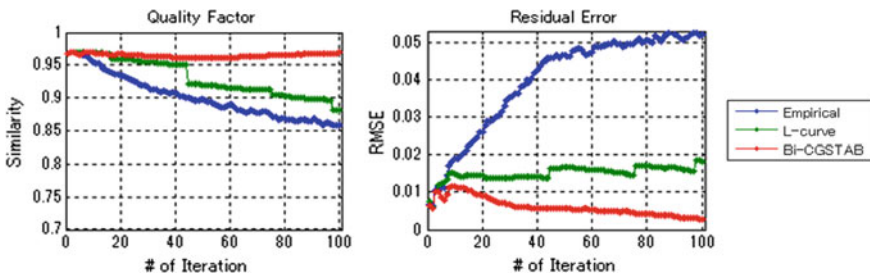


Fig. 3 Performance metrics between three solving techniques

4 Conclusion

The appropriate solving techniques of the system of linear equations are investigated to reconstruct the dielectric properties distribution of the object by solving the non-linear inverse scattering problem. The simulation results show that the ill-posedness of the non-linear problem can be reduced and a more stable solution can be performed by choosing Bi-CGSTAB method as the solving techniques of the system of linear equations.

References

1. Nikolova, N.K.: Microwave imaging for breast cancer. *IEEE Microw. Mag.* **12**(7), 78–94 (2011)
2. Ono, Y., Mohamed, L., Kuwahara, Y.: Optimization of configuration on imaging sensor for microwave tomography. In: Proceedings of the 2016 IEICE Society Conference, 2016, vol. C-2-79, no. ISSN 1349-144X
3. Mohamed, L., Ozawa, N., Ono, Y., Kamiya, T., Kuwahara, Y.: Study of correlation coefficient for breast tumor detection in microwave tomography. *Open J. Antennas Propag.* **3** (December), 27–36 (2015)
4. Shea, J.D., Kosmas, P., Van Veen, B.D., Hagness, S.C.: Contrast-enhanced microwave imaging of breast tumors: a computational study using 3D realistic numerical phantoms. *Inverse Probl.* **26**(74009), 1–22 (2010)
5. Shea, J.D., Kosmas, P., Hagness, S.C., Van Veen, B.D.: Three-dimensional microwave imaging of realistic numerical breast phantoms via a multiple-frequency inverse scattering technique. *Med. Phys.* **37**(8), 4210–4226 (2010)
6. Winters, D.W., Van Veen, B.D., Hagness, S.C.: A sparsity regularization approach to the electromagnetic inverse scattering problem. *IEEE Trans. Antennas Propag.* **58**(1), 145–154 (2010)
7. Joachimowicz, N., Pichot, C., Hugonin, J.-P.: Inverse Scattering: an iterative numerical method for electromagnetic imaging. *IEEE Trans. Antennas Propag.* **39**(12), 1742–1752 (1991)
8. Hansen, P.C., O’Leary, D.P.: The use of the L-curve in the regularization of discrete Ill-Posed problems. *SIAM J. Sci. Comput.* **14**(6), 1487–1503 (1993)
9. Hansen, P.C.: *Regularization Tools—A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems* (2008)
10. van der Vorst, H.A.: BI-CGSTAB: a fast and smoothly converging variant of Bi-Cg for the solution of nonsymmetric linear systems*. *SIAM J. Sci. Stat. Comput.* **13**(2), 631–644 (1992)

Chapter 72

Multimedia Learning Tools for Autism Children



Jasmin Ilyani Ahmad, Suhailah Mohd Yusof
and Noor Hasnita Abdul Talib

Abstract Autism Spectrum Disorder (ASD) is an impact development of brain that impaired the social interactions and communication skills. Children with ASD show difficulties in their communication, interactions, engagement with other people or play activities. This kind of children will have lack capabilities of learning in school. The strength of ASD children may include strong visual-spatial skills, non-verbal problem-solving skills, visual and auditory memory. Hence, the objective of this paper is to identify the need of learning tools that can cater this problem. The quantitative method is applied which is by using questionnaire. The technique used for collecting the data is random sampling technique. From the findings, most of respondents tend to need multimedia learning tools as to help the children learning process. Therefore, the appropriate treatment and special education tools are needed to ensure the autism children can learn and function normally.

Keywords Autism · Symptom · Communication · Multimedia · Courseware

1 Introduction

Autism is a mental disorder in which the autistic people have impairments in social functioning, communication, and behavior. Autism spectrum disorder (ASD) includes neurodevelopmental abnormalities characterized by impaired ability to communicate and interact socially and by restricted and repetitive patterns of behavior, interests, and activities [1]. Furthermore, ASD characteristics can be

J. I. Ahmad (✉) · S. M. Yusof · N. H. A. Talib

Department of Computer and Mathematical Sciences, Universiti Teknologi, MARA Kedah
Branch, Johor Bahru, Malaysia

e-mail: jasmin464@kedah.uitm.edu.my

S. M. Yusof

e-mail: suhailah_my@kedah.uitm.edu.my

N. H. A. Talib

e-mail: nhasnita@kedah.uitm.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference
on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_72

manifested differently between individuals. According to Lloyd, MacDoald and Lord [2], based on varying degrees of autistic people behaviors, they can be diagnosed as autism, Asperger syndrome, Pervasive Developmental Disorder (PDD) or Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). Due to the increasing awareness of ASD cases, there is a need for obtaining information about how ASD children can be assisted primarily in education. This is because, there is a difference between each individual in which there are students who can learn along with normal students while for those who are more unique they need to study in a specialized schools [3]. This paper aims to study the need of multimedia learning tools that can be used by the ASD children. The remainder of the paper is organized as follows: Sect. 2, covers the overview of autism, Sect. 3 describes the research methodology, findings discussed in Sect. 4 and the last section concludes this work.

2 Overview of Autism

2.1 *Categories and Symptoms of Autism*

This section discussed about the categories and symptoms of autism spectrum disorder among the autistic children. Autistic disorders can occur in the spectrum from mild, moderate and severe ranges based on diverse of dissimilar symptoms. Autism has been categorized into several types aimed at assessing the impairments severity level of ASD children so that they can be helped with the best treatment to equip themselves with the necessary skills [4, 5]. The symptoms may be not the same among those children, but generally can be categorize as lack of communication, social interaction and showing repetitive behaviors are considered as the commonly core symptoms for them. Symptoms under communication shows that during their early ages, they do not interact with others even with their parents. They preferred to be alone as they look different from other children. It is very difficult when they do not understand by others due to lack of communication skill. Moreover, autistic children cannot interpret face expression or body language of other people. This situation will make them feel too far from others even though they are sitting in the same classroom. Due to communication and social skills, it tends to make them slow learners to catch up their lessons. On the other hand, the symptoms of autism may be look different for different autistic children. It depends on how serious the autism for the children. Mild autism can be very slow and passive but some with serious can act aggressively and harm themselves and other peoples around them. Some of them have the symptoms like self-injurious behavior that can be classified under repetitive behaviors [6]. Table 1 shows three basic types of autism; classic, Asperger's syn-drome and Pervasive developmental disorder or atypical autism.

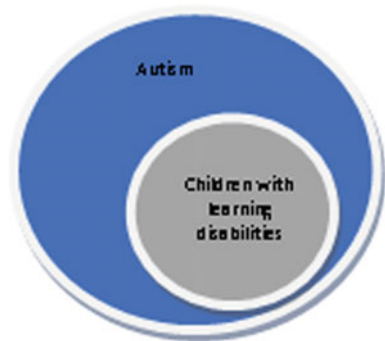
Table 1 Autism types and symptoms

Autism type	Symptoms
Classic	Lack of interest in other people
	Displays little or no emotions
	Has difficulty communicating
	Exhibits repetitive behaviors
Asperger’s syndrome	High-functioning autism’
	Have normal to exceptional intelligence and language skills
	Have difficulties with social interaction and communication
	May show repetitive behaviors
Pervasive developmental disorder	Children who do not meet the full criteria for classic autism or Asperger’s syndrome
	May have difficulties with social interaction and communication, and/or display repetitive behaviors

2.2 Learning Disabilities Among Autistic Children

Autism is not learning disability, but the autism itself can affect the learning process. The learning process among autistic children can be more challenging as they have a problem in social and communication skills. The teaching and learning approach on autistic kids may be different on children with learning disabilities. For instance, group discussion among children may increase their understanding of the lesson, however by using the same technique it will not help the children with autism. The teaching and learning approach may be different depends on how

Fig. 1 Autism and learning disability



serious the autism is. Some of the autistic children have mild and some with serious autism. Teachers and parents should apply appropriate approach to cater those children. Therefore, the therapies and intervention will be benefit for them to reduce autism symptoms and increase their skills and abilities in their life [4]. Figure 1 shows the situation where children with learning disabilities may also have autism symptom disorder, but not all autistic children is people with learning disability.

2.3 Features of Multimedia Courseware Suitable for Autistic Children

Autistic children that usually lack of social and communication skills will just ignore other people around them for instance their parents, teachers and friends. They cannot pay attention and focus on what other peoples are saying or acting. Due to this problem, technology approach or computer based intervention [5] can be adapted in teaching and learning process instead of using classic technique like chalk and talk for autistic children. According to Lau and Lian [6] multimedia software is better to serve as cognitive tool to the autistic children. The element that should be included in multimedia courseware for autistic children are text and graphics that can give benefit for them in mastering reading skill [7]. Other than that, an interactive multimedia courseware is needed that can improve their communication skills [8]. According to Garcia-villamisar and Dattilo [9], interactive multimedia courseware also can help autistic children in recognizing and interpreting face expression and body language that they may be lack to. Furthermore, the computerized game also can improve face recognition skills among autistic children [10]. They can start interacts with the system and at the same time can improve their social skills [11]. In order to cater the challenges with autistic children in learning process, the user interface of the courseware also play an important role. Any graphical user interface is mostly appropriate however it should be realized that rich multimedia interface can lead poor result in user experience [12].

3 Research Methodology

In this section, quantitative method is applied specifically by using data collection techniques which is through questionnaire. The questionnaire is used in order to identify and analyze the findings. In general, probability sampling is chosen since the sample has a known probability of being selected. As there is many type of probability sampling, the researchers have applied simple random sampling

technique to select the sample of population of respondents. However, the focus group is still inside the state of Kedah, Malaysia. Out of 30 people, only 29 respondents answered the survey questions successfully. This gives a response rate of 96.7%. The questionnaire is divided into several parts. However, only two parts are discussed in detail. Part A comprises questions regarding demographic information while Part B focuses on learning tools that can be applied for autism children.

4 Findings and Discussion

There are 65.5% of the respondents that involved in this study are 35 years old and above where the remainder are in between 25–34 years old. Other than that, about 76% of the total respondents are female and only 24% are male. On top of that, about 82.8% of the respondents are among professional and management group whereas the remainder comes from support group and housewives. From the results, it shows that almost all the respondents agree with the statement that the appropriate tool is needed in assisting autistic children in learning process which is 96.6%. Referring to Fig. 2, about 89.7% of the respondents choose multimedia courseware as a tool to assist them in learning, followed by flash card which is 62.1% and soft book with only 58.6%.

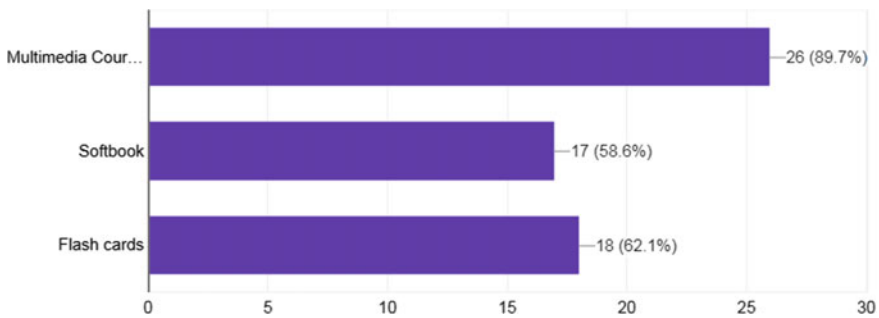


Fig. 2 Tools suitable for autism children

5 Conclusion

The autism spectrum disorder (ASD) generally impaired the communication and social skills among the autistic children. Thus, the approach of learning can be different from normal children perspective. Instead of having face-to-face classroom, the autistic children need alternative tool in assisting them to be more focus to their lesson. Therefore, multimedia courseware is needed with attracted graphical user interface, text, images, sound, animation and video that can attract the autistic children to focus as to assist them in their learning process.

References

1. Eroglu M.S., Toprak, S., Urgan O, M.D., Ozge, E., Onur, M.D., Arzu Denizbasi, M.D., Haldun Akoglu, M.D., Cigdem Ozpolat, M.D., Akoglu, E.: DSM-IV diagnostic and statistical manual of mental disorder **33** (2012)
2. Lloyd, M., MacDonald, M., Lord, C.: Motor skills of toddlers with autism spectrum disorders. *Autism* **17**(2), 133–146 (2013)
3. Khare, R., Mullick, A.: Research tools to learn about the needs of children with autism. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **57**(1), 506–510 (2013)
4. Goldsmith, T.R., Leblanc, L.A.: Use of technology in interventions for children with autism. **1** (2), 166–178 (2004)
5. Ramdoss, S., et al.: Use of computer-based interventions to teach communication skills to children with autism spectrum disorders: a systematic review, 55–76 (2011)
6. Lau, F., Lian, J.M.: Adapted design of multimedia-facilitated language learning program for children with autism
7. Omar, S., Bidin, A.: The impact of multimedia graphic and text with autistic learners in reading. **3**(12), 989–996 (2015)
8. Curzio, O.: An interactive multimedia system for treating autism spectrum disorder an interactive multimedia system for treating, October, 2016
9. Garcia-villamizar, D., Dattilo, J.: Effects of an interactive multimedia computer program on the facial emotions processing and social adaptation in adults with ASD **1**, 1–6 (2016)
10. Tanaka, J.W., et al.: Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let's Face It! program **8**, 944–952 (2010)
11. Lozano, J., Ballesta, J., Alcaraz, S.: Software for teaching emotions to students with autism spectrum disorder, 139–147 (2011)
12. Pavlov, N.: User interface for people with autism spectrum disorders **2014**, 128–134 (2014)

Chapter 73

Relapse Cases Among Drug Addicts Using Logistic Regression Modeling



Siti Fairus Mokhtar, Fazillah Bosli, Norashikin Nasarudin
and Fathiyah Ahmad@Ahmad Jali

Abstract The objective of this study is to use this statistical method to determine the factors which are considered to be significant contributors for relapse to happen. Logistic regression analysis is an important tool used in the analysis of the relationship between various explanatory variables and nominal response variables. There are eight predictors in this study. The predictors involved are gender, race, religious, age, level of education, type of drug, reason to drug, and technique to drug. The dependent variable is the status of the drug addict either relapses or not. Four hundred samples were randomly selected from National Anti-Drug Agency (NADA) in Kedah. The finding of the study revealed age and type of drug (Opiat) is highly significant. The coefficient of age and type of drug (Opiat) is 0.114 and 2.360. The older age increase the probability of drug addict to repeat. Type of drug indicates that drug addicts who use Opiat increase the probability to repeat compared to drug addict who use ATS (Amphetamine Type Stimulant (ATS)). The findings are beneficial to reduce number of drug addict.

Keywords Drug factor relapse · Logistic regression analysis

1 Introduction

Drug abuse among Malaysians is a major issue and has contributed to serious social problems in Malaysia. Drug abuse is a situation when a drug is taken out of any medication and socially disapproved for the use. While, drug addiction means the

S. F. Mokhtar (✉) · F. Bosli · N. Nasarudin · F. Ahmad@Ahmad Jali
Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: fairus706@kedah.uitm.edu.my

F. Bosli
e-mail: fazillah@kedah.uitm.edu.my

N. Nasarudin
e-mail: norashikin116@kedah.uitm.edu.my

F. Ahmad@Ahmad Jali
e-mail: fathiyah@kedah.uitm.edu.my

continued use of drugs, which leads to dependence on the drug because the addict anticipates pain or discomfort if they withdraw from the use of the drug [1].

Farhana [2] in her article, stated that the Deputy Prime Minister, Datuk Seri Dr Ahmad Zahid Hamidi, declared that a total of 131,841 drug addicts have been registered in Malaysia between January 2010 and February 2016 based on report done by the National Drug Information System, National Anti-Drug Agency (NADA). Four main categories affecting tendency for drug abuse including environmental factors, family factors, personal factors and social factors were identified based on previous study conducted on people attending an addiction treatment centre [3]. Based on interviews conducted among drug addicts and administrative staffs at Malaysian Private Rehabilitation Centre indicated that the top two significant factors contributing to drug abuse are peer influence and curiosity [4]. Meanwhile, study by Foo et al. [5] revealed that drug abuse usually caused by a combination of several factors such as curiosity, tension release and betrayal of spouse instead of just one factor.

A study by Tam and Foo [4] mentioned about the rate of relapse case is very high and consistently increasing over fifty percent for the past decades. Relapsed addiction means, usage, intake or misuse of psychoactive substances after one had received drug addiction treatment and rehabilitation, physically and psychologically [6]. Lack of strong self-efficacy to avoid temptations, hurdles and challenges in life is the main cause of relapsed addictions amongst drug addicts in Malaysia [7]. Majority of the drug addict on relapse cases perceived low to medium level of support from their community [8]. The social, environmental and personal factors as well as level of education significantly affect the fresh and relapse drug abusers [7].

A holistic treatment approach with a combination of cognitive-behavioural, medical, social, and spiritual components were needed by relapse cases among drug addict [9]. Previous study by using logistic regression showed that the age and self-motivation are two predictors that influence relapse to happen [10]. Thus, this study aims to determine the factors which are considered to be significant in contributing to relapse case in small areas in Kedah by using Logistic regression analysis.

2 Methodology

In this study, data were obtained from National Anti-Drug Agency (NADA) which involved four hundred samples were studied. The sample was drawn from area Kedah. There are seven predictors in this study. The predictors involved are gender, race, religious, age, level of education, type of drug, reason to drug, and technique to drug. The dependent variable is the status of the drug addict either relapses or not.

Logistic Regression is used to describe data and to explain the relationship between one dependent binary variable and more independent variables [11]. The reason of using Binary Logistics Regression is because the outcome variable is a

categorical variable. It is common practice to assume that the outcome variable, denoted as Y , is a dichotomous variable having either relapsed or non-relapse as the outcome.

Mathematically logistic regression estimates a multiple linear regression functions defined as:

$$\log it(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_KX_K \tag{1}$$

Fitting a model makes it useful in prediction and analysis of new data sets. Firstly, random allocated data two a training group (70%) and a testing group (30%). After fitting the models on the training group, they were applied to the testing group. The fitted model was then used to predict the samples belonging to different classes of the response variable. Finally, IBM SPSS Statistics was used to evaluate the fitness of the models.

3 Result and Discussion

There are six predictors which are extremely not statistically significant. They will not contribute to the overall model performance. The remaining two predictors are age and types of drugs are statistically significant with p -value equal to 0.000 and 0.002 respectively. Thus model which included age and type of drugs as the predictors will be run. Types of drugs are Amphetamine Type Stimulants (ATS) and Opiat. The result is reported in Table 1.

The regression coefficient for age and type of drug (Opiat) is 0.114 and 2.360 respectively. The β 's coefficient analysis is described in Table 1. Wald test is used to determine the contribution of each predictor. The predictor variable with p -value (Significance) less than 0.05 contributes significantly to the predictive ability of the model. Age and type of drug (Opiat) is highly significant with p -value equal to 0.000 and 0.002 respectively. The coefficient of age and type of drug (Opiat) is 0.114 and 2.360 which carries positive sign. This sign indicates that the older age increases the probability of drug addict to repeat. The coefficient for type of drug indicates that drug addicts who use Opiat increase the probability to repeat compared to drug addicts who use ATS.

Table 1 also shows the odds ratio information for each predictor. The odds ratio for age is 1.120. This indicates that for each unit increased in age the odds ratio

Table 1 Model estimation

Variables	B	Std. error	Wald	Significance	Exp(B)
Intercept	-5.667	0.956	-5.928	0.000	0.003
Age	0.114	0.018	6.187	0.00	1.120
ATS	0.469	0.935	0.502	0.616	1.599
Opiat	2.360	0.776	3.041	0.002	10.589

increased the probability to repeat by 1.120 times compared to younger. Meanwhile, the odds ratio for type of drug (Opiat) is 10.589. This indicates those drug addicts who used Opiat are 10.589 times more likely to repeat compared to drug addict who used ATS.

Tables 2 and 3 compares the predicted values with the observed values for training data and validation data. In this case, overall 73.68% of the drug addicts are correctly classified using training data. This model is able to classify correctly 72.18% of the drug addicts as repeating drug addict and 75% was also correctly classified as new drug addict. However, overall 69.57% of the drug addicts are correctly classified using validation data. This model is able to classify correctly 75.47% of the drug addicts as repeating drug addict and 64.52% has been classified correctly as new drug addict (Table 4).

The Cox & Snell R square and Nagelkerke R Square give general information about the amount of variation in the dependent variable explained by the model. The value of Cox & Snell R square and Nagelkerke R Square is 0.290 and 0.387 respectively; suggesting that between 29.0 and 38.7% of the variability is explained by this set of variables.

Table 2 Classification table (training data)

Predicted				
Type of drug addict				
		New	Repeat	Percentage correct
Observed	New	96	37	72.18
	Repeat	38	114	75.00
Overall percentage				73.68

Table 3 Classification table (validation data)

Predicted				
Type of drug addict				
		New	Repeat	Percentage correct
Observed	New	40	13	75.47
	Repeat	22	40	64.52
Overall percentage				69.57

Table 4 Goodness of fit

Cox & Snell R Square	Nagelkerke R Square
0.290	0.387

4 Conclusion

A logistic regression was employed in this study to find the most significant predictors toward relapse to happen. There are two predictors that influence relapse to happen which are age and type of drug. Age and type of drug (Opiat) is highly significant. The coefficient of age and type of drug (Opiat) is 0.114 and 2.360 which carries positive sign. This sign indicates that the older age increases the probability of drug addict to repeat. The coefficient for type of drug indicate that drug addicts who use Opiat increase the probability to repeat compared to drug addict who use ATS.

The odds ratio for age is 1.120. This indicates that for each unit increased in age the odds ratio increased the probability to repeat by 1.120 times compared to younger drug addict. Meanwhile, the odds ratio for type of drug (Opiat) is 10.589. This indicates those drug addicts who used Opiat 10.589 times are more likely to repeat compared to drug addicts who used ATS. OPIAT is a drug that has a high potential to cause addiction and can affect one's thoughts and actions [12]. This type of drug is also classified as depressant. It is a type of substance that affects the central nervous system. The findings are beneficial to reduce number of drug addict.

This research was supported by Tn. Khairul Farrizan b Tn. Abdullah Zawawi who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper.

References

1. Karofi, U.A.: Drug abuse and criminal behaviour in Penang, Malaysia: a multivariate analysis. *Bangladesh e-Journal Sociol.* **2**(2), 90–116 (2005)
2. Nokman, F.S.: More than 130,000 drug addicts in Malaysia to date, figures show. *New Straits Times*. Retrieved from <http://www.nst.com.my>, 19 April 2016
3. Amiri, M., Taheri, Z., Hosseini, M., Mohsenpour, M., Davidson, P.M.: Factors affecting tendency for drug abuse in people attending addiction treatment centres: a quatitative content analysis. *J. Addict. Res. Ther.* **7**(2), 1–4 (2016)
4. Tam, C.L., Foo, Y.C.: A qualitative study on drug abuse relapse in Malaysia: contributory factors and treatment effectiveness. *Int. J. Collab. Res. Intern. Med. Public Health* **5**(4), 217–232 (2013)
5. Foo, Y.C., Tam, C.L., Lee, H. L.: Family factors and peer influence in drug abuse: a study in rehabilitation centre. *Int. J. Collab. Res. Intern. Med. Public Health* **4**(3), 190–201(2012)
6. Mohamed, M.N.: Peranan & Penglibatan Keluarga dan Masyarakat Dalam Pencegahan Penagihan Berulang. *Jurnal PERKAMA*. Bil. 6. ISSN 0127/6301. Terbitan Persatuan Kaunseling Malaysia (1996)
7. Ibrahim, F., Kumar, N.: Factors effecting drug relapse in Malaysia: an empirical evidence. *J. Asian Soc. Sci.* **5**(12), 37–44 (2009)
8. Ibrahim, F., Kumar, N.: The influence of community on relapse addiction to drug use: evidence from Malaysia. *Eur. J. Soc. Sci.* **11**(3), 471–476 (2009)
9. Chie, Q.T., Tam, C.L., Gregory, B., Hoang, M.D., Khairuddin, R.: Substance abuse, relapse, and treatment program evaluation in Malaysia: perspective of rehab patients and staff using the mixed method approach. *Front. Psychiatry* **7**(90), 1–13 (2016)

10. Ismail, M.T., Alias, S.N.S.: Binary logistic regression modelling: measuring the probability of relapse cases among drug addict. In: Proceedings of the 21st National Symposium of Mathematical Sciences (SKSM21). AIP Conference Proceeding, vol. 1605, pp. 792–797 (2014)
11. Hair, J.F., Black, B., Babin, B., Anderson, R.E., Tatham, R.L.: Multivariate Data Analysis, vol. 6. Prentice-Hall, Upper Saddle River, NJ (2006)
12. Wegman, M.P., Altice, F.L., Kaur, S., Rajandaran, V., Osornprasop, S., Wilson, D., Wilson, D.P., Kamarulzaman, A.: Relapse to opioid use in opioid-dependent individuals released from compulsory drug detention centres compared with those from voluntary methadone treatment centres in Malaysia: a two-arm, prospective observational study. *Lancet Glob. Health* **5**, e198–207 (2017)

Chapter 74

The Influence of Fixed Rhythm Auditory Icon on Food Intake Mimicry



Suzilah Ismail, Norhayati Yusof and Hanif Baharin

Abstract Mimicry is defined as the activity of copying the behaviour or speech of other people. Mimicry can play an important role in influencing human behaviour which is very useful in persuasive technology. Thus in this study, an experiment was conducted to determine the effect of fixed rhythm auditory icon on food intake mimicry based on gender. Thirty males and thirty females were involved by screening their food intake patterns individually in eating an apple without listening to any sound. Next, a fixed rhythm auditory icon which is represented by a sound of biting an apple for every ten second was introduced while each participant was eating an apple during the experiment. Food intake mimicry is defined as eating that occurs within five seconds of hearing the fixed rhythm auditory icon. The findings showed that both males and females were influenced by the fixed rhythm auditory icon with an average of 48 and 50% of food intake mimicry respectively. However, there was insignificant different between male and female average mimicry percentage, which indicated both gender has similar tendency of food intake mimicry. It was noticeable that food intake mimicry started only after the participants have engaged in half of the experiment time. Perhaps by prolonging the eating time, the mimicry percentage will increase because the person is already adapting to the fixed rhythm. Thus, these findings indicated that auditory icon can play a role in influencing good eating behaviour by imposing eating slowly which is advantageous for health.

Keywords Mimicry · Food intake · Persuasive technology · Fixed rhythm · Auditory icon

S. Ismail (✉) · N. Yusof

School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

e-mail: halizus@uum.edu.my

N. Yusof

e-mail: norhayati@uum.edu.my

H. Baharin

Institute of Visual Informatics (IVI), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

e-mail: hbaharin@ukm.edu.my

© Springer Nature Singapore Pte Ltd. 2019

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference*

on Computing, Mathematics and Statistics (iCMS2017),

https://doi.org/10.1007/978-981-13-7279-7_74

1 Introduction

Mimicry is an imitation of other people behaviour or speech. This paper highlight the tendency of human to mimic each other during interaction which is useful in persuasive technology where interactive devices can be embedded in influencing people behaviours and attitudes [1, 2]. According to Lakin et al. [3], people tend to mimic their partners when they are interacting with each other by copying the speech accent, posture, gesture and mannerism. Human mimicry functions to create friendly and trustworthy ambience while socializing [4]. Persuasive technology may take advantage of mimicry to persuade good behaviour such as in food intake.

Eating slowly or fast has impact on health. The beneficial of eating slowly are good for digestion and can reduce food and calorie intake [5, 6] while eating fast can lead to overeating and obesity [7]. Previous study conducted by Zin et al. [8] revealed auditory icons that represent eating cause food intake mimicry in human. They conducted an experiment involving thirteen participants' eating apples while listening to a sound loop of auditory icons of apple biting. In this study we improvise the experiment by taking gender into consideration. The purpose is to determine the effect of fixed rhythm auditory icon on food intake mimicry based on gender.

2 Methodology

In this study, the experiment was conducted on thirty male and thirty female students age of 19. It started by screening their food intake patterns individually in eating an apple without listening to any sound. Next, a fixed rhythm auditory icon which is represented by a sound of biting an apple for every ten second was introduced while each participant was eating an apple during the experiment. Food intake mimicry is defined as eating that occurs within five seconds of hearing the fixed rhythm auditory icon. The experiment was recorded using a good video camera in ensuring accurate data is collected by counting the number of biting for each participant. Food intake mimicry, m is presented by percentages as outline by the following equation.

$$m = \frac{\text{no. of biting occurs within five seconds of hearing the rhythm}}{\text{total number of biting}} \times 100 \quad (1)$$

Exploratory Data Analysis (EDA, i.e. graphical and numerical) and supported by normality test were used in determining the correct measure of central tendency (i.e. either mean or median) in determining the average % food intake mimicry to represent males and females. Next, based on the findings obtained regarding the normality of the data, inferential statistics was conducted in testing the significant difference of the two genders.

3 Findings and Discussions

Figures 1 and 2, display Exploratory Data Analysis (EDA) results by showing the graphical and numerical summary of % food intake mimicry for both genders. The histogram and QQ plots showed the data are approximately normal supported by mean and median close to each other and the skewness and kurtosis near to zero (0). This is confirmed by normality test which revealed insignificant p-value for both genders. Thus both data are normally distributed; therefore mean are used to represent the average % food intake mimicry for males and females which have the values of 48% and 50% respectively.

Figure 3 shows similarity of the two box plots for the percent mimic between males and females. The t-test result confirmed that there are no significant different between the average percent mimic of males and females. This indicated that both genders have the same percentage of food intake mimicry.

The findings showed that both males and females were influenced by the fixed rhythm auditory icon with an average of 48 and 50% of food intake mimicry respectively. However, there was insignificant different between male and female average mimicry percentage, which indicated both gender has similar tendency of food intake mimicry. It was noticeable that food intake mimicry started only after the participants have engaged in half of the experiment time.

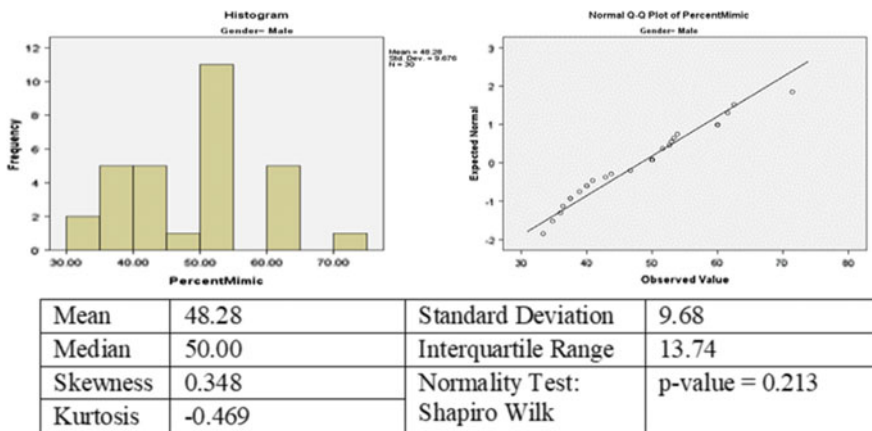


Fig. 1 EDA (graphical & numerical) and normality test for males percent mimic

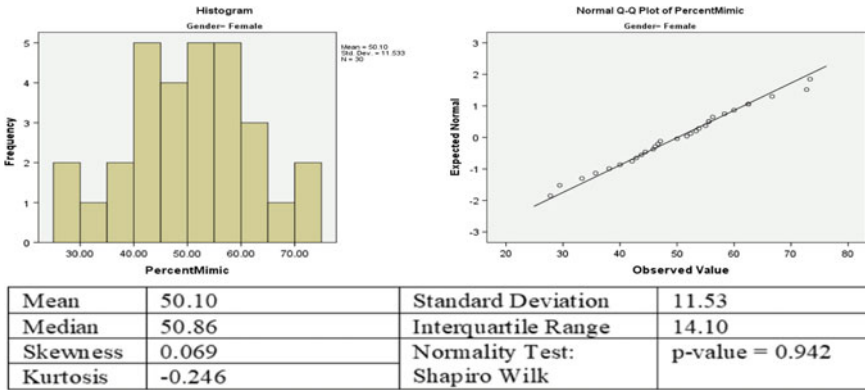


Fig. 2 EDA (graphical & numerical) and normality test for females percent mimic

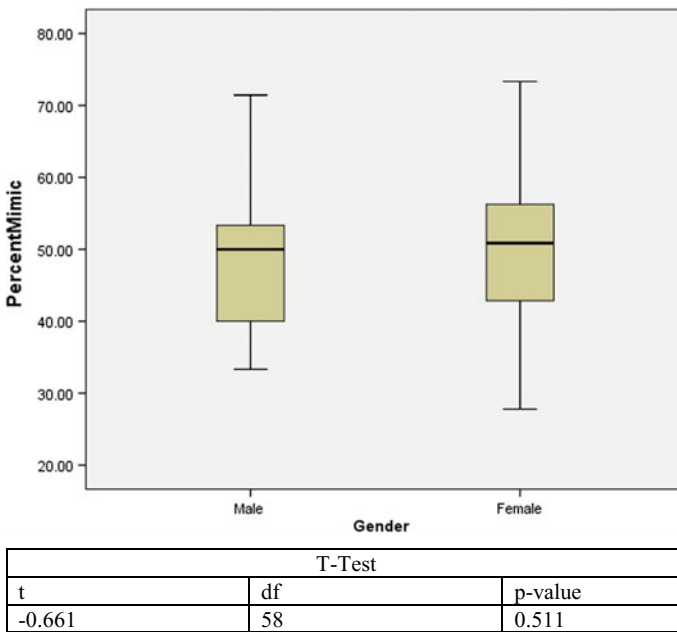


Fig. 3 Box plot & t-test of percent mimic for males and females

4 Conclusions

Perhaps by prolonging the eating time, the mimicry percentage will increase because the person is already adapting to the fixed rhythm. Thus, these findings indicated that auditory icon can play a role in influencing good eating behaviour by imposing eating slowly which is advantageous for health.

References

1. Fogg, B.J.: A behaviour model for persuasive design. In: Proceedings of the 4th International Conference on Persuasive Technology, vol. 40 (2009a)
2. Fogg, B.J.: Creating Persuasive Technologies: An Eight Steps Design Process. vol. 44. Persuasive (2009b)
3. Lakin, J.L., Jefferis, V.E., Cheng, C.M., Chartrand, T.L.: The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverbal Behav.* **27** (3), 145–162 (2003)
4. Luo, P., Ng-Thow-Hing, V., Neff, M.: An examination of whether people prefer agents whose gestures mimic their own. Paper presented at the intelligent virtual agents (2013)
5. Stuart, R.B.: Behavioral control of overeating. *Behav. Res. Ther.* **5**, 357–365 (1967)
6. Andrade, A.M., Greene, G.W., Melanson, K.J.: Eating slowly led to decreases in energy intake within meals in healthy women. *Am. Diet. Assoc.* **108**(7), 1186–1191 (2008)
7. Zandian, M., Ioakimidis, I., Bergström, J., Brodin, U., Bergh, C., Leon, M., Södersten, P.: Children eat their school lunch too quickly: an exploratory study of the effect on food intake. *BMC Public Health* (2012)
8. Zin, N., Baharin, H., Rosli: Can auditory icons induce food intake mimicry? In: Proceedings of the 5th International Conference on Computing and Informatics ICOCI 2015 (2015)