Azizur Rahman *Editor*

Statistics for Data Science and Policy Analysis

EXTRAS ONLINE

Springer

Statistics for Data Science and Policy Analysis

Azizur Rahman
Editor

Statistics for Data Science and Policy Analysis

Springer

Editor
Azizur Rahman
School of Computing and Mathematics
Charles Sturt University
Wagga Wagga, NSW, Australia

ISBN 978-981-15-1734-1 ISBN 978-981-15-1735-8 (eBook)
https://doi.org/10.1007/978-981-15-1735-8

© Springer Nature Singapore Pte Ltd. 2020
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

To all who have supported, contributed and participated in the ASPAC2019.

Preface

This book constitutes the refereed proceedings of the Applied Statistics and Policy Analysis Conference 2019 (ASPAC2019), held on 5–6 September 2019, in Wagga Wagga, NSW, Australia.

The overwhelming growth of data and its users is a reality, which has put new thoughts amongst the research community to devise new ideas for giving data-driven evidence-based policy decisions at local, state, national and international levels. In recent years, applied statistics and data science have received renewed interest from a broad range of stakeholders ranging from governments to corporations and end users of data and its analysis or modelling tools. As a result, applied statistics and data science research such as data mining and policy analysis have been placed high as a national priority in many countries including Australia. In this data-centric world with vastly growing demand situation, there is a need to ensure that reliable statistical and modelling solutions that address important and emerging policy issues at both public and private institutions are disseminated timely and widely amongst the research and industry community.

The conference aims to foster and promote international research collaborations and exchange ideas between data scientists, applied statisticians and data modellers who will detail the latest innovations in research to gather and disseminate information from small to big data settings and from policy analysts who will describe how they use existing information from increasing big data environments and indicate areas in which there need to be methodological and technological developments. Another aim is to provide a forum for presenting current research work in the form of keynote, invited and contributed sessions and establish connections between researchers at tertiary institutions and working in industry in Australasia and overseas. Panel discussions during the conference have provided a great, relaxing and enjoyable atmosphere in which participants can share their scholarly achievements in their academic field as well as communicate with others.

The theme of ASPAC2019 is "Effective policy through the use of big data, accurate estimates and modern computing tools and statistical modelling". The keynote sessions involve increasing access to government microdata for research purposes to micro-geospatial targeting for precision public policy. The conference also has a range of invited sessions and panel discussion which are specially focusing on industry's research issues and modelling opportunities. It illustrates at "how we can use the different modelling tools and statistics techniques to inform policy decision making processes in a better way". Around 55 authors and some local researchers took part in the conference which created an excellent discussion in the parallel sessions and provided great stimulus for future research. At the conference banquet ceremony, five PhD students and early career researchers received ASPAC2019 Travel Awards for their excellent papers. An additional five papers' authors received some partial travel support for presenting their good quality research papers at the conference. In addition, the solid academic atmosphere and beautiful environment of Charles Sturt University also made the conference successful, comfortable and enjoyable. Taking this opportunity, we would like to acknowledge the support we received from the Australian Government's Department of Infrastructure, Regional Development and Cities, Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Statistical Society of Australia (SSA), Brock University, University of Canberra and Charles Sturt University for the conference organisation.

The 26 revised full papers presented were carefully reviewed and selected for this proceedings book volume out of 60 submissions from 16 countries distributed across the globe. Each paper was anonymously peer-reviewed by at least two reviewers, who gave review feedback, objective and/or helpful revision comments to the authors where necessary. So, the outcome has made this volume of a very high quality. The papers are organised into six topical sections: applied statistics and Bayesian modelling, agricultural statistics and policy analysis, data science and image processing statistics, health statistics and social policy, small area estimation and spatial microsimulation and business analytics and managements policy analysis. Furthermore, the key issues at the ASPAC2019 covered diverse areas of current popularity and state-of-the-art research in Statistics, Data Science and Policy Analysis.

Wagga Wagga, NSW, Australia Azizur Rahman
September 2019

Organisation

ASPAC2019 was organised by Statistics and Data Mining Research Group within the School of Computing and Mathematics and Data Science Research Unit in the Faculty of Business, Justice and Behavioural Sciences at the Charles Sturt University and supported by Brock University (Ontario, Canada); Australian Government's Department of Infrastructure, Regional Development and Cities, Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Statistical Society of Australia (SSA), University of Canberra and Charles Sturt University.

Program Chairs
Associate Professor Azizur Rahman, *Charles Sturt University*
Professor Syed Ejaz Ahmed, *Brock University*

Organising Committee
Associate Professor Azizur Rahman (Chair), *Charles Sturt University*
Professor Raymond Chambers, *University of Wollongong*
Professor Syed Ejaz Ahmed, *Brock University*
Professor Terence Mills, *La Trobe University*
Professor Denny Meyer, *Swinburne University of Technology*
Associate Professor Md Rafiqul Islam, *Charles Sturt University*
Associate Professor Ben Phillips, *Australian National University*
Dr. Nizamul Islam, *The Luxembourg Institute of Socio-Economic Research (LISER)*
Dr. Ryan Ip, *Charles Sturt University*
Dr. Siu-Ming Tam, *Australian Bureau of Statistics and University of Wollongong*
Associate Professor Shuangzhe Liu, *University of Canberra*
Dr. Jahar Bhowmik, *Swinburne University of Technology*
Dr. Md Musfequs Salehin, *Charles Sturt University*
Dr. Cathy Honge Gong, *Australian National University*
Dr. Dmitry Demskoy, *Charles Sturt University*
Dr. Sumonkanti Das, *Maastricht University*
Associate Professor Md Zahidul Islam, *Charles Sturt University*
Dr. Anwaar Ulhaq, *Charles Sturt University*

Prof. Marcin Szymkowiak, *Poznań University*
Dr. Md Asaduzzaman, *NSW-DPI and CSU*
Associate Professor Andrew Robinson, *University of Melbourne*
Dr. Maheswaran Rohan, *NSW Department of Primary Industries (DPI)*
Dr. Mizanur Khondoker, *University of East Anglia, UK*

Program Committee
A. Rahman, Prof. J. Hicks *(CSU)*, Prof. S.E. Ahmed *(Brock)*, M. R. Islam, Prof. T. Mills *(La Trobe University),* Dr. S. Das *(Maastricht University)*, R. Ip, Dr. C. H. Gong *(ANU)*, M. M. Salehin, Dr. R. Yunus *(University of Malaya),* Dr. Z. Li *(CSU),* Prof. D. Meyer *(Swinburne)*, Dr. A. Kabir *(CSU)*, A/Prof. X. Huang *(CSU)*, Dr. L. Zheng *(CSU),* Dr. S. Rehman *(CSU)*, Dr. J. Bhowmik *(Swinburne),* Dr. N. Islam *(LISER, Luxembourg)*, Prof. M. Szymkowiak *(Poznań University of Economics and Business),* Dr. S-M. Tam *(ABS and UoW)*, A/Prof. B. Phillips *(ANU),* A/Prof. S. Liu *(UC)*, Dr. M. Asaduzzaman *(NSW-DPI and CSU),* Dr. M. Rohan *(NSW-DPI),* A/Prof. A. Robinson *(Melbourne)*, Prof. R. Chambers (NIASRA, Wollongong) and Dr. M. Khondoker *(UEA, UK).*

Awards Committee
Associate Professor Azizur Rahman (Chair), Associate Professor Md Rafiqul Islam, Dr. Lihong Zheng and Professor Dennis Trewin.

Publicity Team
Dr. Anwaar Ulhaq (Co-chair), Dr. Ashad Kabir (Co-chair), Dr. Sabih Rehman, Dr. Dmitry Demskoy, Dr. Ryan Ip and Azizur Rahman.

Acknowledgements

We would like to take this opportunity to thank all the participants, all of whom worked exceptionally hard to ensure this conference was a success. We want to express our sincere thanks to prominent professors, Dennis Trewin, Former Australian Statistician (from 2000 to 2007) at the Australian Bureau of Statistics (ABS), Australian National University (ANU); Robert Tanton, Director of the National Centre for Social and Economic Modelling (NATSEM) at the Institute for Governance and Policy Analysis, University of Canberra; Chang-Tsun Li, Department of Computer Science, University of Warwick, UK, and Director of Research, Centre for Cyber Security Research and Innovation, Deakin University; David Balding, Director of the Melbourne Integrative Genomics (MIG) and Professor of Statistical Genomics, University of Melbourne; Susan Wilson, Senior Statistical Consultant in Stats Central, University of New South Wales, and Emeritus Professor in the Mathematical Sciences Institute, ANU; and S (Subu) V Subramanian, Director of a Harvard University-wide Initiative on Applied Quantitative Methods in Social Sciences and Founding Director of Graduate Studies for the Interdisciplinary PhD Programme in Population Health Sciences, Harvard University, USA, for kindly accepting the invitation to give keynote speeches and panel discussions. Special thanks are also due to Professor John Germov, Charles Sturt Provost and Deputy Vice Chancellor (Academic) and the Acting Vice Chancellor, for his valuable time and speech to open ASPAC2019, as well as Professor Syed Ejaz Ahmed, Dean of Science and Mathematics at Brock University, for his dedicated time, advice and excellent service as the conference program co-chairs.

We would like to thank all the members of the Organising Committee, Program Committee, Publicity Team, Awards Committee and Referees for the peer reviews who have provided excellent supports to ASPAC2019. Thanks are also due to all of the valuable sponsors (i.e. Bureau of Infrastructure, Transport and Regional Economics in the Department of Infrastructure, Regional Development and Cities, ARC Centre of Excellence for Mathematical and Statistical Frontiers, Statistical Society of Australia, Brock University, National Centre for Social and Economic Modelling (NATSEM) and Institute for Governance and Policy Analysis at UC and Charles Sturt University) and partners (i.e. Springer and Science Community

Organisation) for their significant contributions. We also appreciate Springer for the publication of the proceedings volume. Specially, we would like to take the opportunity to thank Lucie Bartonek, Marielle Klijn, Prasad Gurunadham and Dr. Celine Chang from Springer for their invaluable advice, assistance and all of the excellent professional communications and services in the publication process. This book will give excellent exposure of the papers accepted for publication.

A special thank goes to Prof. Dennis Trewin for his great service to the ASPAC2019 Awards Committee; Drs. Dmitry Demskoy, Lihong Zheng, Ryan Ip and A/Prof. Md Rafiqul Islam for their outstanding local support; Amna Sabih and Emma Humphries (who serves as a new Casual Admin Officer for some time of the conference) for their utmost professionalism and enthusiasm in providing administrative supports; A/Profs. Md Zahidul Islam, Md Rafiqul Islam and Tanveer Zia for chairing sessions; Prof. Tracey Green (Executive Dean, Faculty of Business, Justice and Behavioural Sciences) and Prof. Mark Morison (Associate Dean Research in BJBS) for their constant support, advice and encouragements; and Prof. Irfan Altas (Head of School of Computing and Mathematics) for his endless support, advice, leadership and stimulus in this conference.

Finally, we would like to give our sincere thanks to everyone who has contributed in some other forms and to all delegates for attending the conference this year – we hope you enjoyed ASPAC2019.

Wagga Wagga, NSW, Australia Azizur Rahman
September 2019

Contents

Part I
Applied Statistics and Bayesian Modeling

Chapter 1
Applied Bayesian Modeling for Assessment of Interpretation Uncertainty in Spatial Domains

Scott McManus, Azizur Rahman, Ana Horta, and Jacqueline Coombes

Abstract In the mining industry, code compliant reporting standards for public announcements have been developed setting minimum standards for public reporting of exploration results and mineral resources. These include an assessment of the quality and confidence in the data and work carried out since public reporting aims to provide information that is *material, transparent and competent* to investors.

There are four phases required to estimate an mineral resource (preparation, investigation, model creation and validation), and estimation is highly dependent on the accuracy of the preparation stage which is a result of the quality of the geological interpretation given for the mineralization process and current spatial location. Performance of feasibility studies in mining projects has been poor, with a 50% failure rate, 17% of failures are attributable to issues in geological interpretation. This interpretation seeks to spatially define geologically homogenous areas in the resource (spatial domains), corresponding to a single statistical population with a single orientation, where possible. In the estimation workflow, the creation of the spatial domain presents a challenge in terms of assessing the uncertainty in the geological interpretation often due to the manual and subjective interpretation used to guide its creation as well as in spatial domains with several mineralization overprint events.

The proposed work investigates a Bayesian method using multivariate quantitative data combined with qualitative data to assess the interpretation uncertainty of

S. McManus (✉)
Charles Sturt University, Port Macquarie, NSW, Australia
e-mail: smcmanus@csu.edu.au

A. Rahman
School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW, Australia

A. Horta
Charles Sturt University, Albury, NSW, Australia

J. Coombes
Snowden Technologies, Perth, WA, Australia

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_1

classification of borehole intervals to a spatial domain defined by a 3D 'wireframe' or 'rock type' model interpretation using either implicit or explicit modeling techniques.

Keywords Bayesian · Uncertainty · Spatial domain

1.1 Introduction

In Australia, the reporting of mining results of public companies is governed by the Joint Ore Committee 2012 [1] reporting code which sets minimum standards for public reporting of results of mineral exploration and resource estimation. These include an assessment of the quality and confidence in the data and work carried out to obtain an estimate of mineral resources since public reporting aims to provide information that is *Material, Transparent and Competent* to investors. At each step in the workflow required to estimate a mineral resource (Preparation, Investigation, Model Creation & Validation) [2], the reporting standards request the author to determine and state the confidence level associated to the data, geological understanding, interpretation and estimation. The accumulated confidence from each component of the phases, expressed in either qualitative or quantitative terms is often presented as a list of evidence to justify the classification of the mineral resource into differing levels of increasing geological confidence, namely Inferred, Indicated or Measured [3].

The standards do not specify how the estimation or assessment of uncertainty is to be carried out and therefore it is the responsibility of the 'Competent'/ 'Qualified' person authoring the report to ensure that the methods and techniques used are appropriate.

McCarthy [4] found that the performance of Feasibility Studies in mining projects has been poor, with a 50% failure rate with Geology Interpretation, Resource and Reserve estimation issues directly contribute to 17% of the failure. There has been a recent call for improving the quality of public reporting by ensuring all risk is quantified for all components of feasibility studies [5] to allow public investors and industry professionals to make better-informed decisions.

When considering the types of uncertainty in earth models [6], measurement error (sampling and analytical sources of error), interpretation and processing of raw measurements (compositing, correcting, turning spectra into elemental composition), spatial uncertainty and estimation uncertainty (from kriging variance, kriging probability and simulation) can be easily assessed in a quantitative way and these are usually communicated during reporting. Interpretation of the geological setting is the most problematic, namely, quantifying interpretation uncertainty in allocating or classifying borehole intercepts into a spatial domain.

Spatial domains are created early in the mineral estimation process, during the Preparation phase and are derived from the geological interpretation. Data related to economic grade, presence of multi-elements, and qualitative and quantitative geological attributes are used to manually construct (explicit) or used by specialized software to automatically construct (implicit) spatial domains usually represented as a 3D Wireframe mesh. Implicit modeling techniques utilize either Dual Kriging [7], Discrete Surface Interpolation [8] or geostatistical clustering [9].

Interpretation uncertainty of spatial domains can be assessed if models are not deterministic [10]. But as expert knowledge is a key component in the creation of spatial domains, their models are deterministic and only updated when new data is collected. Therefore, quantitative uncertainty is rarely assessed, instead subjective qualitative statements are used.

Hence, current industry practice is that total uncertainty assessment of the spatial domain is expressed in qualitative terms of how well physical evidence: a) supports a homogenous grade population; and b) models the interpreted geology. This process evaluates the univariate statistics for the spatial domain to investigate how well it defines a single grade element population representing the resource, and uses 3D visualization to ensure it makes 'geological' sense [3].

To obtain an uncertainty measure of qualitative interpretation data and to test if these provide an accurate interpretation of the spatial domains, Bayesian modelling is the most appropriate approach due to its ability to use different types of data sources. So far in the mining industry, Bayesian approaches have not been adopted since there is a traditional reliance in expert opinion.

This work reports results using a Bayesian approximation methodology to assess the interpretation uncertainty of borehole intercepts classified into spatial domains which are then used to create 3D wireframes or numerical models of the morphology of the geological and mineralization interpretation. The methodology can be used to assess interpretation uncertainty of spatial domains in early-stage projects or used to assess or audit spatial domains from mature projects created using implicit or explicit methodologies where a lack of robust variograms make spatial uncertainty methods problematic.

The methodology presented in this paper allows for the use of qualitative and quantitative multivariate data or a mixture of data as well as expert knowledge to assess the interpretation uncertainty of intercepts assigned to spatial domains. The case study is a gold project that has been closed for 10 years in South East Queensland, Australia. A potential buyer for the project wanted to assess the risk of the project from two key factors; the qualitative geological logging as research has shown that traditional visual core logging (qualitative data) can provide biased or poor quality data [9], and the spatial domains as both are critical foundations of any future mine planning and financial decisions. The previous geological team were no longer available and the only laboratory analyses were for gold. It is unclear what data was used to guide the spatial domain interpretation due to lack of documentation, but the assumption is that gold values and expert knowledge guided by the visually logged qualitative data where used. Limited borehole core has been kept and was preserved in a core library, so nondestructive sampling was required. There was not enough borehole data to ensure robust variography and so the method presented here was used to assess the interpretation uncertainty of the spatial domains using Portable XRF (pXRF) measurements as the quantitative data to compare and validate the qualitative categorical data interpretation.

1.2 Materials and Methods

1.2.1 Samples

Gympie Gold Mines has provided full access to their historical resource database, which includes, reports, GIS, databases, for both qualitative and quantitative exploration, resource, reserve and production stages of their project. Historical borehole core was destroyed with 13 partial holes rescued and donated to the Geological Survey of Queensland's Zillmere core library in Brisbane (GSQ). Of these boreholes, five intersect mineralized structures, which were coded as spatial domains defining gold mineralization in 3D space. Two holes intersect the Inglewood structure, two the Partridge structure and one structure called 'Peel Off'. During the life of the project, only gold was analyzed for, along with subjective (qualitative) visual core logging of the following geological elements; descriptive lithology, alteration intensity, mineralization intensity and quartz veining. From 9 remaining holes (of which 5 have spatial domain intersections), there are 375 assay samples.

The core varies from being HQ (63.5 mm diameter) to NQ (47.6 mm diameter) in size and is either whole, half or quarter core. Some of the core is broken or missing. Table 1.1 shows the summary of boreholes, the depth downhole sampled, the number of samples within that interval and if it includes a spatial domain intercept. The 20 cm samples were combined into 1 m samples for 235 samples total to match the length of the qualitative visual logged data.

1.2.2 Spatial Domains

Samples have been interpreted as belonging to a spatial domain (Inglewood, Partridge, Peel off or not mineralized) by the mining company, based on geological subjective (qualitative) logging, subjective (quantitative) logging, chemical analysis

Table 1.1 Samples taken every 20 cm downhole

Hole	From (m)	To (m)	No of samples	Spatial Domain
G023	302.2	320.6	92	Inglewood
G135	108.7	203	463	None
G135 W1	370.6	416.8	230	Inglewood
G141	128.45	151.6	119	Partridge
G162	133	149.8	88	Partridge
G215	435	447	61	None
G224	188	206	86	Peel off
G250	100	98	36	None
G256	330	337	4	None

for gold and spatial relationships related to current geological, genetic and structural models using an explicit methodology. The non-spatial domain material covers many different lithological and structural types and so can have some variability. The spatial domains have similar lithology, chemistry, and structural types and are grouped on the basis that they are spatially related and form a population of similar or homogenous material that hosts related gold mineralization.

1.2.3 Handheld X-Ray Fluorescence Measurement

As the core is currently stored in the GSQ core library future sampling and analysis must be non-destructive. pXRF provides an efficient, cost-effective and non-destructive method of obtaining systematic light and heavy multielement measurements [11] that may be suitable for calculating lithology, alteration, structure and homogeneous mineralized spatial domains and therefore useful for assessing the uncertainty in classifying intercepts to a spatial domain. Sampling and QA/QC Protocols were developed guided by a mining industry pXRF workflow [12]. Measurements were averaged to produce a homogenous sample [13] to match the gold sample intervals as well as subjective categorical geological visual logging, for comparison.

1.2.4 Statistical Analysis Methods

1.2.4.1 Variable Selection

The following elements were measured using the pXRF, Al, Si, P, S, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Rb, Sr, Y, Zr, Nb, Mo, Pd, Ag, Cd, Sn, Sb, Ba, W, Hg, Pb, Bi, Th and LE (a combined result for light elements not able to be measured by the instrument) in percent. In this present study, the main aim was to find the covariates that indicate the presence of a particular spatial domain from the measured pXRF data and the qualitative categorical visual borehole logging. Within the R statistics program [14] Principal Component Analysis, Multinomial, Stepwise General Linear and RandomForest models were utilized to select important variables. The variable selections were evaluated using Akaike information criterion (AIC), Bayesian information criterion (BIC), Kappa and model performance from a training data (60% of total data) set using predictions from a validation dataset (Leave one out Cross validation) [15] using a multinomial model. This was completed for three sets of variables to enable assessment of the utility of pXRF and qualitative data in assessing uncertainty of the spatial domains.

1. pXRF measurements only
2. qualitative, categorical visual logging only
3. combined pXRF measurements and qualitative, categorical visual logging

Logistic regression is most frequently used to predict the relationship between a dichotomous (binary) outcome variable and a set of covariates but with a few modifications, it may also be used when the outcome variable is polytomous [16]. For a polytomous dependent variable ‘Y’, consider the categories are coded as 0, 1, 2 and 3 such that,

$$\mathrm{Y} = \begin{cases} 0 \text{ if ith intercept is Inglewood;} \\ 1 \text{ if ith intercept has no mineralized spatial domain;} \\ 2 \text{ if ith intercept is Partridge; and} \\ 3 \text{ if ith intercept is Peel Off.} \end{cases}$$

Since each of our dependent variables has four categories, then polytomous logistic regression model is used in this study to assess the influence of the variables on spatial domain selection. Equation (1.1) is the generalized form of the multinomial model for more than 3 categories (or in our case 4), j = 0,1,2,3

$$P\left(Y_i = k\right) = \frac{Exp\left(\sum_{j=0}^{p} X_{ij}\beta_{jk}\right)}{\sum_{m=1}^{n} Exp\left(\sum_{j=0}^{p} X_{ij}\beta_{jm}\right)} \tag{1.1}$$

With the resultant likelihood function being Equation (1.2)

$$l\left(\beta\right) = \prod_{i=1}^{n} \left[\pi_0(x_i)^{y_{0i}} \pi\left(x_i\right)^{y^{1i}} \pi_2(x_i)^{y_{2i}}\right] \tag{1.2}$$

For each of the three sets of data combinations the following variables were selected for the reduced multinomial model;

- pXRF Only: Nb, K, Zr, Rb, Fe & Al (from 33 pXRF element concentrations)
- Categorical Data Only: Quartz Veins per meter, Veins per meter, Intensity of Primary Carbon Alteration, Intensity of Calcite Veining, Intensity of Non-Sedimentary Pyrite Alteration and Intensity of Sedimentary Pyrite. (from 36 Categorical qualitative visually logged data)
- Combined data: Quartz Veins per meter, Veins per meter, Intensity of Primary Carbon Alteration, Intensity of Calcite Veining, LE and Ti. (from the combined dataset of 33 pXRF element concentrations and 36 Categorical qualitative visually logged data)

1.2.4.2 Bayesian Methodology

Bayesian approximation is a method of apportioning credibility across parameters by sampling from simulations of the likelihood equation, predicting from the model and then assessing uncertainty through simulation and sampling of the posterior distribution. The creation of spatial domains involves quantitative and qualitative data as well as accessible prior geological information from literature and/or expert experimental knowledge [17], thus, Bayesian methods might be more appropriate to use for determining interpretation uncertainty in the creation of spatial domains.

Bayesian approximation and simulation have been used to assess spatial uncertainty [18] in a soil context and prediction uncertainty of classification of real and fake ivory products in the field of Forensic Science [19].

Taking the multinomial model from Sect. 1.2.4.1, a prior distribution is specified on the parameters credibility is re-allocated across the parameter values using simulation to produce an accurate approximation of Bayesian posterior distribution. Stan with the BRMS [20] R connection was chosen as it has a built-in softmax function. Stan makes use of the Hamiltonian Monte Carlo (HMC) [21] method of generating Monte Carlo steps to approximate Bayesian regression.

A weakly informative prior Normal(0,8) was selected [22], review of trace and density plots suggested there had been good chain mixing and that all features of the posterior distribution had been accurately represented from the sampling, supported by autocorrelation of less than 5 with good sampling efficiency and effectiveness, with Rhat values less than 1.05.

From the likelihood derived from the multinomial likelihood function in Equation (1.2), the data were centered and scaled. An arbitrary run in time of 1000 iterations was used followed by 10,000 iterations for 3 chains using the BRMS function BRM with a 'categorical' response distribution [20]. The MCMC output of all the parameters passed convergence tests [23]. After assessing the skewness of the posterior distributions the median from the posterior draws was used to distribute credibility over the model parameters.

The BRMS Predict function, using the predicted parameters from the simulations, was used to determine the probability of each sample belonging to each of the 4 classes, the class with the highest probability was selected for the classification. The median value from the posterior distribution was selected for the measure of uncertainty for each of the classifications after assessing the skewness of the distributions. Testing of the Bayesian models was via Leave one out and K-folds, the Pareto K diagnostic was 0.7 or less for 92.3% of the data suggesting that the models are appropriate. The final model presentation was by Bayesian Interval plots and Confusion Matrix model metrics [24] were created from the BMRS predict function output.

A series of regression models for each of the three multinomial models were determined and the credibility distributed over the parameters in the form of the following outcomes for each;

$$\lambda_k \text{where } \log\left(\frac{\emptyset_k}{\emptyset_r}\right) = \log\left(\frac{\exp\left(\beta_{0,k} + \beta_{1,k}x\right)}{\exp\left(\beta_{0,r} + \beta_{1,r}x\right)}\right) = \beta_{0,k} + \beta_{1,k}x. \tag{1.3}$$

Drawing from the posterior distributions from the simulations, point predictions were determined for each of the 235 borehole intervals.

1.3 Results

Confusion Matrix Accuracy and Precision for each of the spatial domains using the three different multinomial models are shown in Tables 1.2 and 1.3 includes the results of the combined model metrics for each of the three multinomial models using; pXRF, Categorical and a combined dataset.

Fig. 1.1 through Fig. 1.3 are plots of the Bayesian prediction with an Uncertainty envelope around each prediction. The predictions are in order of sampling down each of the holes. The Inglewood class has the lowest accuracy in all three sample sets as well as the largest amount of uncertainty. The Inglewood structure is variable and contains different features in different boreholes, and so its certainty rarely reaches 70%, whereas the Partridge zone is always quite high, as it is less

Table 1.2 Bayesian accuracy and precision per category for each model

	pXRF	Categorical	Combined
Inglewood	73.0%	86.5%	88.2%
none	76.5%	93.5%	92.1%
Partridge	94.6%	100.0%	95.0%
Peel off	60.2%	94.0%	94.4%

Table 1.3 Bayesian model results metrics

	Accuracy	Kappa	Precision	NIR
pXRF	88.1%	0.58	69.9%	82.1%
Categorical	96.2%	0.87	91.7%	82.1%
Combined	95.7%	0.86	94.7%	82.1%

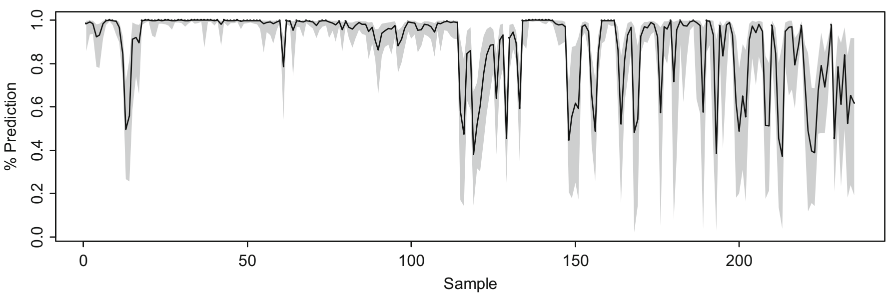


Fig. 1.1 Median prediction value with Bayesian uncertainty interval for ‘pXRF Data’

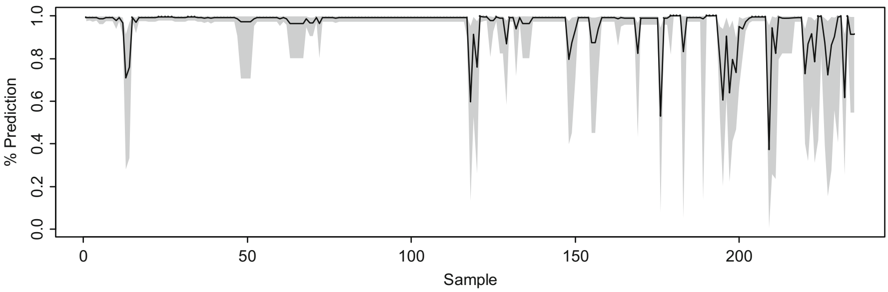


Fig. 1.2 Median prediction value with Bayesian uncertainty interval for 'Categorical model'

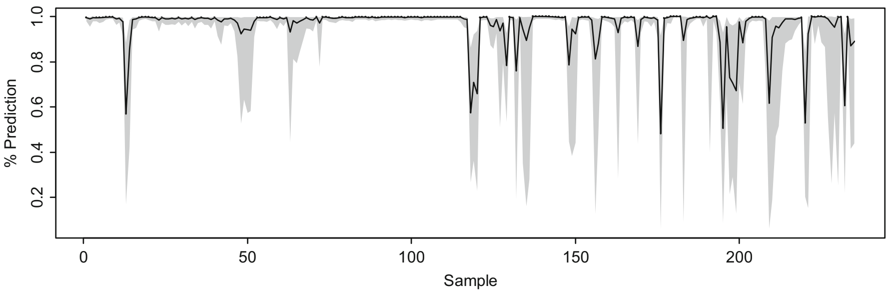


Fig. 1.3 Median prediction value with Bayesian uncertainty interval for 'All model'

mixed and more homogenous and has a higher contrast chemically, physically and structurally to the unmineralised rock classified as 'None'. The class 'None', is usually sub-economic grade and often has a high certainty of around 98%, but it can visually have a mixture of veins and unmineralised rock, and so the alteration from the veining and proximity of the structure to say one of the economic classes like Inglewood can affect the uncertainty of the prediction. This is seen in the geochemistry of the pXRF values with depletions and enrichments in certain elements, and adjacent intervals to mineralized spatial domains can have a lower certainty value of 48% due to the heterogeneities of mixed rock and the chemical effects of wall rock alteration. The Peel Off class had the least amount of samples but it was different to the Inglewood and Partridge classes, chemically it can be falsely predicted as Inglewood when only using the pXRF dataset model, but is defined when using only the categorical dataset or the combined pXRF and Categorical data, where it had high accuracy, precision and certainty in the point predictions.

When comparing the three models from the three different datasets, the pXRF data has the least accuracy and highest uncertainty, the Categorical dataset has the highest accuracy and lower uncertainty and the combined model of geological logging and pXRF data has higher accuracy, the highest precision and certainty. The kappa statistics are greater than 0.8 where Categorical data is included. All three models are better than using no model at all with accuracy greater than the 'No

Information Rate'. The combined model had the highest accuracy when predicting the more chemically complex Inglewood and Peel Off spatial domains.

1.4 Discussion

The use of all three models using pXRF, Categorical and Combined data all have high accuracy, but the pXRF data alone has lower uncertainty when predicting the spatial domains that have higher heterogeneity, and where structural components with similar geochemistry might well be better able to be discriminated by geologists rather than pXRF measurements. This can happen in structurally controlled mineral deposits or mineral deposits that may have overprinting ore types that need to be separated for resource estimation or mine planning.

In terms of the case study, the qualitative categorical visual logging and explicit (expert knowledge) interpretation by the site geologists has provided an accurate dataset with high confidence interpretation when assigning borehole intervals to spatial domains. The use of a multinomial prediction model and Bayesian approximation through simulation shows that this dataset has performed better than the prediction from models using only pXRF measurements, which is likely due to the reduced light elements available to the pXRF instrument.

The Inglewood domain was the most challenging spatial domain to assess due to the heterogeneities in geological fabric within its structure, which, was expressed in broad uncertainty intervals. The Combined model of qualitative visual logging and pXRF data provided the highest accuracy for that spatial domain. The frequentist multinomial model also showed high accuracy (88.2%), the benefit of the Bayesian method is that we can also state, that despite the high accuracy in point prediction that there is a high degree of uncertainty in the interpretation.

In lieu of laboratory multielement data being available, pXRF was able to provide cost effective quantitative data with which to compare with the qualitative data and validate the quality of the Categorical interpretation.

In terms of public reporting in a JORC context, a quantitative measurement of uncertainty using the qualitative data can now be assessed and reported, as well as propagated with uncertainties from other phases of resource estimation process to provide a combined assessment of uncertainty.

Acknowledgments The author would like to acknowledge the support provided by Australian Government Research Training Program Scholarship, Gympie Gold Mines, Australian Bauxite Ltd, Qld DNRME's Exploration Data Centre, Zillmere, CSU Spatial Analysis Unit, the CSU Indigenous Support Centre, Port Macquarie, CSU School of Computing and Statistics and the CSU School of Environmental Science.

References

1. JORC: The JORC Code 2012 Edition, Joint Ore Reserves Committee of The Australasian Institute of Mining and Metallurgy, Australian Institute of Geoscientists and Minerals Council of Australia (JORC) (2012)
2. Coombes, J.: The Art and Science of Resource Estimation, Australia (2008)
3. Coombes, J.: I'D Like to be OK with MIK, UC?, Australia (2016)
4. McCarthy, P.L.: Managing risk in feasibility studies. In: Edwards, A.C. (ed.) Mineral Resource and Ore Reserve Estimation: The AusIMM Guide to Good Practice. Monograph/Australasian Institute of Mining and Metallurgy 30, pp. 13–18. Australasian Institute of Metallurgy and Australasian Institute of Mining and Metallurgy, Carlton (2014)
5. Noppé, M.: A framework for presenting and benchmarking resource projects. In: AusIMM Project Evaluation Conference Ausimm, Brisbane (2016)
6. Caers, J.: Modeling Uncertainty in the Earth Sciences. Wiley, Hoboken (2011)
7. Galli, A., Murillo, E., and Thomann, J.: Dual kriging-its properties and its uses in direct contouring. Verly, G. et al., pp. 621–634 (1984)
8. Mallet, J.-L.: Discrete smooth interpolation in geometric modelling. Comput. Aided Des. **24**(4), 178–191 (1992)
9. Fouedjio, F., Hill, E.J., Laukamp, C.: Geostatistical clustering as an aid for ore body domaining: case study at the Rocklea Dome channel iron ore deposit, Western Australia. Appl. Earth Sci. **127**(1), 15–29 (2018)
10. Lark, R.M., et al.: A statistical assessment of the uncertainty in a 3-D geological framework model. Proc. Geol. Assoc. **124**(6), 946–958 (2013)
11. Arne, D.C., Mackie, R.A., Jones, S.A.: The use of property-scale portable X-ray fluorescence data in gold exploration: advantages and limitations. Geochem. Explor. Environ. Anal. **14**(3), 233–244 (2014)
12. Fisher, L., et al.: Resolution of geochemical and lithostratigraphic complexity: a workflow for application of portable X-ray fluorescence to mineral exploration. Geochem. Explor. Environ. Anal. **14**(2), 149–159 (2014)
13. Bourke, A., Ross, P.-S.: Portable X-ray fluorescence measurements on exploration drill-cores: comparing performance on unprepared cores and powders for 'whole-rock' analysis. Geochem. Explor. Environ. Anal. **16**, 147–157 (2015)
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2018)
15. Gelman, A.: Bayesian Data Analysis, 3rd edn. ed. Chapman & Hall/CRC Texts in Statistical Science, ed. A. Gelman. CRC Press, Boca Raton (2013).
16. Hosmer, D.W.: Applied Logistic Regression. 3rd edn. ed. Wiley Series in Probability and Statistics, ed. S. Lemeshow and R.X. Sturdivant. Wiley, Hoboken (2013)
17. Rahman, A., et al.: An assessment of the effects of prior distributions on the Bayesian predictive inference. Int. J. Stat. Probab. **5**(5), 31 (2016)
18. Horta, A., et al.: Geostatistical data integration model for contamination assessment. Math. Geosci. **45**(5), 575–590 (2013)
19. Buddhachat, K., et al.: Distinguishing real from fake ivory products by elemental analyses: a Bayesian hybrid classification method. Forensic Sci. Int. **272**, 142–149 (2017)
20. Bürkner, P.-C.: brms: an R package for Bayesian multilevel models using Stan. J. Stat. Softw. **80**(1), 1–28 (2017)
21. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo (2011)
22. Team, S.D.: Stan Modeling Language: User's Guide and Reference Manual. Version (2018)
23. Brooks, S., et al.: Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton (2011)
24. Rahman, A., Nimmy, S., and Sarowar, G.: Developing an automated machine learning approach to test discontinuity in dna for detecting tuberculosis. In: International Conference on Management Science and Engineering Management. Springer (2018)

Chapter 2
Computing Robust Statistics via an EM Algorithm

Maheswaran Rohan

Abstract Maximum likelihood is perhaps the most common method to estimate model parameters in applied statistics. However, it is well known that maximum likelihood estimators often have poor properties when outliers are present. Robust estimation methods are often used for estimating the model parameters in the presence of outliers, but these methods lack a unified approach. We propose a unified method using EM algorithm to make statistical modelling more robust. In this paper, we describe the proposed method of robust estimation and demonstrate it using the example of estimating the location parameter. Well known real data sets with outliers were used to demonstrate the application of proposed estimator. Finally, the proposed estimator is compared with standard M-estimator. In this talk, the location case was considered for simplicity, but directly extends to the robust estimation of parameters in a broad range of statistical models. Hence this proposed method aligns with the classical statistical modelling, in terms of a unified approach.

Keywords M-estimator · Mixture distributions · Improper distribution

2.1 Overview of Estimating Statistical Model Parameters

One of the methods for estimating statistical model parameters is a likelihood approach. Often, the maximum likelihood estimator is not optimal if the data generating process deviates slightly from the assumed true model [13]. A long history of data deviation from the true model can be found in the literature ([1, 2] and page 70, Gelman et al. [5]). For example, a data set of the speed of light (page 70, Gelman et al. [5]) was reported by Simon Newcomb in 1882. Figure 2.1 displays a density plot of the speed of light data, and shows that two observations were quite atypical by the virtue of being far from the bulk of the data.

M. Rohan (✉)
NSW Department of Primary Industries, Wagga Wagga, NSW, Australia
e-mail: maheswaran.rohan@dpi.nsw.gov.au

© Springer Nature Singapore Pte Ltd. 2020
A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_2

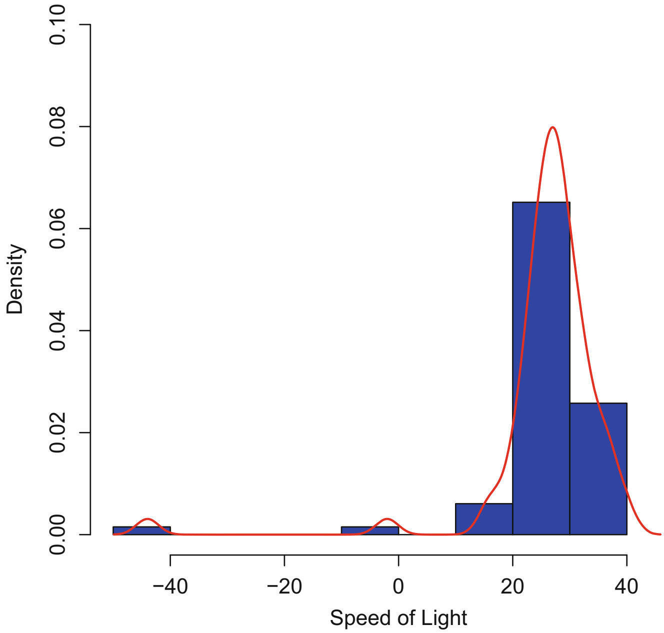


Fig. 2.1 Density plot of the speed of light

Due to the non-robustness of the maximum likelihood estimates in the presence of outliers, many robust alternative methods have been developed to have good estimates for the parameters of the model. Robust methods of statistical modelling provided reliable good parameter estimates in terms of maintaining high efficiency and stability on the presence of outliers. Construction of the main foundation for the modern robust statistics theory began in 1970s. This major work was compiled and reported by Andrew et al. [3]. However, it had failed to win over the mainstream statistics in those days, because robust estimators were not in closed form, which can not be simply calculated by hand. To address the closed form, iterative processes were used to estimate the parameters. Later, Holland and Welsch [8] revealed that they are iterative re-weighted least squares estimators where the weights are allocated using the known tuning parameter, with small weights corresponding to outlying observations. Presently, there is a substantial literature on robust statistics and with increasingly high power computer access,robust statistics is now comprehensively moving to mainstream statistics.

The shortcoming of the methodology of computing robust statistics is often a non-unified approach. We propose a unified method using an EM algorithm to make statistical modelling more robust. In this paper, we describe the proposed method of robust estimation and demonstrate it using the example of estimating the parameters. This simple example directly extends to the robust estimation of parameters in a broad range of statistical models [15].

2.2 Model

Let $y_1, \ldots, y_n$ be independent random variables with a common distribution $F(y, \theta) = F(y - \theta)$, where θ is an unknown location parameter. Consider the model (2.1)

$$y_i = \theta + \epsilon_i \ , i = 1 \ldots n \tag{2.1}$$

where the errors $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with common distribution. The problem is to estimate the parameter θ in the presence of outliers. If the common distribution is normally distributed with mean zero, the optimal estimate for θ in the absence of outliers is the mean $\bar{y}$. The sample mean $\bar{y}$ may be arbitrarily disturbed by a single outlier, as any data value $y_i \to \pm\infty$ then $\bar{y} \to \pm\infty$. Huber [9] found the method of estimating the location parameter θ in the presence of outliers, but extending this method to the whole class of generalised linear model (GLM) was not feasible.

We seek in this paper to find a unique method that is less sensitive to extreme observations while easily extended for the whole family of GLM [15]. Our strategy is to formalize the neighborhood of F, given in (2.2)

$$h(y, \theta) = \lambda f(y, \theta) + (1 - \lambda) g(y) \tag{2.2}$$

where

- The probability density function f is the first derivative of F.
- $0 \leq (1 - \lambda) \leq 1$ captures the proportion of contaminated data. It is common to fix the λ in the robust literature. For example $\beta-$trimmed mean, where β is fixed. We will often choose λ to be 0.95 or similar.
- g may be defined as an improper parameter free distribution over the real line. A detailed discussion of choosing g will be addressed in Sect. 2.5.

Note that this additional component, g would have no intrinsic interest and does not represent a serious attempt to model the outliers. In fact, the finite mixture form is being used as a mathematical tool to compute weights for each data point.

2.3 Robust Estimation Based on Our Algorithm

Our proposed robust statistic is defined by (2.3)

$$\hat{\theta} = \arg\max_{\theta} L_o(\theta) \tag{2.3}$$

where $L_o(\theta)$, given in (2.4), is the observed likelihood function of θ

$$L_o(\theta) = \Pi_{i=1}^{n} h(y_i, \theta) = \Pi_{i=1}^{n} \{\lambda f(y_i, \theta) + (1 - \lambda) g(y_i)\} \tag{2.4}$$

It is complicated to maximize the observed likelihood, $L_o(\theta)$, so we use the complete likelihood $L_c(\theta)$, because we could obtain a maximum likelihood estimate for θ if the complete data $X = (y_1 \ldots, y_n, z_1, \ldots, z_n)$ was observed [14], where z_i' s are defined below.

$$\hat{\theta} = \arg\max_{\theta} L_c(\theta) \tag{2.5}$$

where

$$L_c(\theta) = \Pi_{i=1}^{n} \left[[\lambda f(y_i, \theta)]^{z_i} \times [(1 - \lambda) g(y_i)]^{1 - z_i} \right] \tag{2.6}$$

and

$$z_i = \begin{cases} 1 \text{ if } y_i \in f \\ 0 \text{ if } y_i \in g \end{cases} \tag{2.7}$$

Since z_i's are treated as unobserved random variables, the EM algorithm [4] was employed to estimate the parameter θ along with missing observations $z_i, i = 1, \ldots, n$

The E-Step of the algorithm gives

$$\hat{z}_i = \frac{\lambda f(y_i, \hat{\theta})}{\lambda f(y_i, \hat{\theta}) + (1 - \lambda) g(y_i)} = z(y_i, \hat{\theta}) \tag{2.8}$$

And M-step of the algorithm gives

$$\sum_{i=1}^{n} \hat{z}_i (y_i - \hat{\theta}) = 0 \tag{2.9}$$

The proof of result (2.9) is given in Appendix A. The equations (2.8) and (2.9) are the most interesting parts of our proposed estimator, because equation (2.9) is a weighted form of the maximum likelihood equation and the weight for observation y_i is allocated by $\hat{z}_i$, which is usually down-weighted from the bulk of data. This is very similar to the M-estimator approach.

Since z_i does depend on θ, we iterate the E-step of (2.8) and the M-step of (2.9) until the updates no longer change the parameter estimates. For a given initial value for θ, say $\hat{\theta}^{(0)}$, the new estimate $\hat{\theta}^{(1)}$ is obtained by

$$\hat{\theta}^{(1)} = \frac{\sum_{i=1}^{n} \hat{z}_i^{(0)} y_i}{\sum_{i=1}^{n} \hat{z}_i^{(0)}}$$

where $\hat{z}_i^{(0)} = z(y_i, \hat{\theta}^{(0)})$ for $i = 1, \ldots, n$. At the mth process of the iteration, we can write

$$\hat{\theta}^{(m+1)} = \frac{\sum_{i=1}^{n} \hat{z}_i^{(m)} y_i}{\sum_{i=1}^{n} \hat{z}_i^{(m)}} \tag{2.10}$$

where $\hat{z}_i^{(m)} = z(y_i, \hat{\theta}^{(m)})$ for $m = 0, 1 \ldots$. If the limit of sequence of $\left\{\hat{\theta}^{(m)}\right\}_{m=0}^{\infty}$ exists, the proposed estimate, $\hat{\theta}$, for θ is taken as the limit

$$\hat{\theta} = \lim_{m\to\infty} \hat{\theta}^{(m)}$$

2.4 Standard Error of $\hat{\theta}$

One of the main criticisms of the usage of the EM algorithm is that the uncertainty of estimates are difficult to assess. Louis [12] described the method, given in (2.12), for finding the observed information matrix when data are absent. However the observed information matrix $I_o(\theta, y)$ could be obtained directly especially for the scalar case [14, p. 120], where $y = (y_1, \ldots, y_n)$ is a vector.

$$\begin{aligned}
I_o(\theta, y) &= -\frac{\partial}{\partial\theta}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n} z_i(y_i - \theta)\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} z_i - \sum_{i=1}^{n}\left(\frac{\partial z_i}{\partial\theta}\right)(y_i - \theta)\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} z_i - \sum_{i=1}^{n} z_i(1 - z_i)\left(\frac{y_i - \theta}{\sigma^2}\right)(y_i - \theta)\right]
\end{aligned}$$

$$I_o(\hat{\theta}, y) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{z}_i - \frac{1}{\sigma^4}\sum_{i=1}^{n} \hat{z}_i(1 - \hat{z}_i)(y_i - \hat{\theta})^2 \tag{2.11}$$

This result is exactly the same form of the Louis method.

$$I_o(\theta, y) = \mathcal{I}_c(\theta, y) - \mathcal{I}_m(\theta, y) \tag{2.12}$$

where $\mathcal{I}_c(\theta, y)$ is the conditional expected complete-data information matrix given observed data; and missing information matrix $\mathcal{I}_m(\theta, y)$ is a correction for the information lost to the missing data. All these notations are clearly defined in [14]. The standard error of $\hat{\theta}$ is $\sqrt{I_o(\hat{\theta}, y)^{-1}}$.

2.5 Choice of g

The selection of an improper distribution of g has been discussed by a number of authors. Hennig [6] proposed to add the improper distribution, accounting for "noise", to improve the mixture model estimation. Hennig and Coretto [7] discussed how to choose the "noise" component. Longford and D'Ursob [11] describe the mechanism of mixture models associated with improper distribution, and explain a method for selecting an appropriate improper function.

Following steps are suggested to define the improper distributions g (constant) for the proposed algorithm:

- Step 1: Compute the range (r) of the data.
- Step 2: Compute the constant (c) as the inverse of the range of data $(c = r^{-1})$.
- Step 3: Define $g = c$ on $\mathcal{R}^+$ if data is non-negative; otherwise define on real line $\mathcal{R}$.

2.6 Simulation Studies

For numerical illustration, we randomly generated a data set of 80 observations from $\mathcal{N}(\theta = 10, \sigma = 3)$, which we consider to be correct, contaminated by 20 observations from $\mathcal{N}(\theta = 10, \sigma_1 = 30)$. The advantages of using the contaminated generated data are (i) we know the true value of the location parameter, $\theta = 10$, which can be compared with the estimated value directly, (ii) we also know the other parameters' true values such as the scale parameter, $\sigma = 3$ and the proportion of the non-contaminated observations $\lambda = 0.8$. The g for the proposed method is defined as in Sect. 2.5.

We report the proposed estimates for various values of $\lambda = 0.8$, 0.9 and 0.95 with the standard error, based on the equation (2.12). Later, we compare the proposed estimate to common existing methods, (i) Huber estimate, and (ii) Tukey estimate. The tuning constants, such as $k = 1.345$ for Huber, and $k = 4.685$ for Tukey, were chosen to gain 95% as efficiency as the sample mean under the normal distribution [10]. Bootstrapping with one thousand samples was used to compute the standard errors for Huber and Tukey estimates, but the method explained in Sect. 2.4 was used for computing standard error of the proposed estimate. Finally, we generated K = 500 data sets with a sample size of 100 from the model described at the beginning of the section and computed all mentioned estimates including the proposed method for all five hundred data sets. The results are displayed in Fig. 2.2.

The estimates of θ with various λ values and the MLE are given in Table 2.1. These are very similar with small standard errors (around 0.34, Table 2.1), and all values of $\hat{\theta}$ are very close to the true value, 10. As a result, we can say that $\hat{\theta}$ is a better estimate than the maximum likelihood estimate, $\bar{y} = 8.813$.

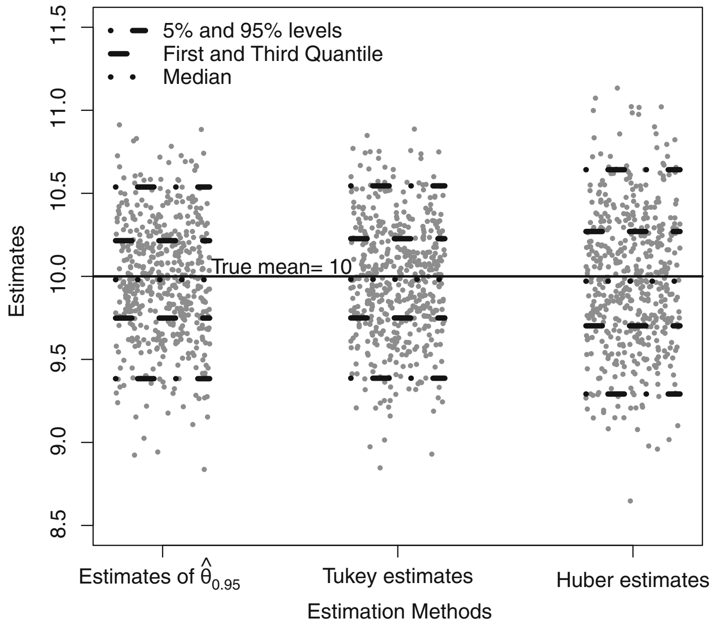


Fig. 2.2 Estimates for 500 data sets: (**a**) Proposed estimate $\hat{\theta}_{\lambda=0.95}$, (**b**) Tukey estimates with k = 4.685 and (**c**) Huber estimates with k = 1.345

Table 2.1 Location estimates (MLE and Mixture) for the generated data

		Estimate	Std. error
MLE		8.813	1.406
Mixture estimates	$\hat{\theta}_{0.8}$	9.666	0.349
	$\hat{\theta}_{0.9}$	9.684	0.342
	$\hat{\theta}_{0.95}$	9.708	0.340
Tukey		9.729	0.373
Huber		9.426	0.473

$\hat{\theta}_{\lambda^*}$ is an estimate for θ by the mixture approach when $\lambda = \lambda^*$

The Tukey and Huber estimates with standard errors for the data are given in the third and fourth rows of Table 2.1 respectively. As anticipated, the estimate $\hat{\theta}_{0.95}$ is very similar to the Huber and Tukey estimates, but closer to Tukey estimate.

All estimates including $\hat{\theta}_{0.95}$ are computed for the five hundred data sets and the results are displayed as scatter plots in Fig. 2.2. All three plots show that estimates seem to be roughly symmetric around the true value of 10 with small variance. The mean (sd) of five hundred estimates for $\hat{\theta}_{0.95}$, Tukey and Huber are 9.98 (0.36), 9.98 (0.35) and 9.98(0.41) respectively. They are near to the true value 10. However, the proposed estimate $\hat{\theta}_{0.95}$ gave a better estimates with 85% of the estimated values lying between 9.5 and 10.5, the Tukey and the Huber methods resulted in 84% and 78% of estimates falling on their range respectively.

In summary, the proposed estimator performs similarly to existing robust estimators. However, the proposed mixture method is a general approach to the modification of maximum likelihood estimates for robustness and is equivalent of a tuning constant, $c/\lambda - c$, may be chosen intuitively. The mixture approach does not seem to be considered in current robust estimation literature. This mixture likelihood framework was already developed further by [15].

2.7 Applications

This section applies the proposed method using real data sets: (i) the speed of light, (ii) determinations of copper in wholemeal flour data (chem), and (iii) nickel content of the Canadian syenite SY-3 (Abbey). The method is compared to other estimates, including MLE after removing outliers (MLE$_{-outlier}$) for all real data sets, are computed for a visual comparison of estimates.

2.7.1 Speed of Light

Simon Newcomb collected data experimentally between July 1882 and September 1882 to calculate the speed of light by measuring the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back (Fig. 2.1). Gelman et al. [5] analysed the data in the Bayesian frame work described in pages 77 and 160.

The function g may be defined on the real line as a constant of 0.012 ($= 1/(40 - (-44))$). To define the scale parameter (σ), a robust estimate of the median absolute deviation (MAD) was computed from the data before computing the location estimate. It is 4.4.

The results for MLE, MLE$_{-outlier}$ and the proposed estimate with $\lambda = 0.95$, Huber estimate and Tukey estimate are given in the second column of Table 2.2. As expected, the robust estimates, including the proposed estimates, are quite different from the sample mean (26.212). However, the proposed estimate is very similar to Huber and Tukey estimates. The standard error(SE) of the proposed estimate (0.59) is less than half the SE of the MLE (1.32).

2.7.2 Chem Data

This data set has 24 determinations of copper (μg g^{-1}) in wholemeal flour and it was used to explain why not to reject outliers [2]. In software R, it is called "chem" data and it is freely available in "MASS" package [16]. Figure 2.3a displays chem data and shows that one observation, 28.95, deviates from the bulk of the data.

Table 2.2 Location estimates with standard errors for various data sets

Methods	Speed of light (SE)	Chem (SE)	Abbey (SE)
MLE	26.212 (1.32)	4.280 (1.08)	16.006 (3.82)
$MLE_{-outlier}$	27.750 (0.64)	3.208 (0.14)	11.043 (0.84)
Mixture estimate with $\lambda = 0.95$	[a]27.734 (0.59)	[b]3.122 (0.12)	[c]10.994 (0.90)
Tukey	[a]27.638 (0.65)	[b]3.144 (0.36)	[c]10.705 (1.48)
Huber	[a]27.380 (0.66)	[b]3.216 (0.15)	[c]11.437 (1.32)

Scale parameter σ for data sets, speed of light, chem and abbey taken as Median Absolute Deviation from the Median (MAD): 4.45[a], 0.53[b] and 4.45[c] were used respectively

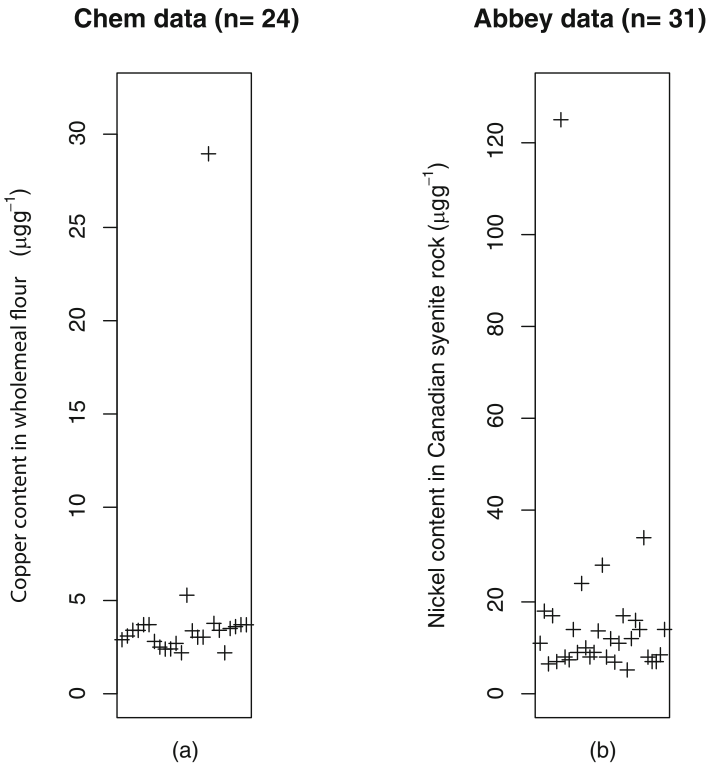


Fig. 2.3 (**a**) Chem data (**b**) Abbey data

To apply our method, $g = 0.038 (= 1/(28.95 - 2.2))$, $\lambda = 0.95$ and scale parameter $\sigma = \text{mad(chem)} = 0.53$ were selected before computing the proposed estimate. Results are given in the third column of Table 2.2. Clearly, MLE is the most vulnerable to outliers, but MLE without outliers (3.208) provides a good estimate of the center of the data. The proposed estimate (3.122) is a good representation of the center of the data, and similar to other robust estimates. The SE for all mentioned estimates, except MLE (1.08), ranges from 0.12 to 0.36.

2.7.3 Abbey Data

Abbey [1] presented 31 observations of nickel content (μg g^{-1}) in SY-3, a Canadian syenite rock. It is known as "abbey" in software R and can be obtained from "MASS" package. Figure 2.3b shows the abbey data and a number of observations clearly stand out from the bulk of the data.

Before computing the proposed estimate, we defined $g = 0.0083 (= 1/(125 - 5.2))$, $\lambda = 0.95$ and scale parameter $\sigma = \text{mad(abbey)} = 4.45$. Results are given in the fourth column of Table 2.2. MLE without three outliers (11.043) performs better than MLE (16.006) in-terms of finding the center of the data. The proposed estimate (10.994) does represent a good estimate of the center of the data and its standard error (0.90) is less than a quarter it was with the outliers present (3.82). The proposed estimate, including standard error is almost equivalent to the robust estimates.

2.8 Conclusion

In the robust statistics literature, the mixture model was often used to generate 'contaminated data' for evaluating robust estimators. In contrast, the finite mixture form is used here as a mathematical tool to estimate the parameters of the formal model in the presence of outliers.

This paper has illustrated the estimation of location parameter and shown that this method works very well with examples of generated contaminated data and real data sets. Extensions of the novel principle for a wider range of statistical models are straightforward [15]. It shows robust statistics may be computed in a unified framework using the EM algorithm. Further, this method is easy to apply in practical situations.

Finally, the simple examples in this paper show that ignoring or removing outliers delivers a markedly different result. This practice of ignoring outliers may be very costly (Kendal, page 110, 1991; [13]), so that the decision/policy makers should give pay careful attention to the process which developed data they are basing their decision or policy on.

Acknowledgements I would like to express sincere thanks to Dr Murray Jorgensen (AUT University) for his valuable discussion throughout my research and advice of this manuscript. I also acknowledged Dr Iain Hume (NSW DPI) for comments on the manuscript.

Appendix A: Proof of Equation (2.9)

$$
\begin{aligned}
l_c(\theta) &= \log L_c(\theta) \\
&= \sum_{i=1}^{n} \left[\hat{z}_i \log(\lambda) + \hat{z}_i \log f(y_i - \theta) + (1 - \hat{z}_i)\log(1 - \lambda) + (1 - \hat{z}_i)\log(g)\right] \\
&= \sum_{i=1}^{n} \hat{z}_i \log f(y_i - \theta) + \text{constant}
\end{aligned}
$$

For MLE, $l_c(\theta) = \sum_{i=1}^{n} \hat{z}_i (y_i - \theta)^2 + \text{constant}$ is to be maximized with respect to θ,

$$\frac{dl_c(\theta)}{d\theta} = 0$$

$$\sum_{i=1}^{n} \hat{z}_i (y_i - \theta) = 0.$$

References

1. Abbey, S.: Robust measures and the estimator limit. Geostand. Newslett. **12**, 241 (1988)
2. Analytical Methods Committee: Robust statistics how not to reject outliers. The Analyst **114**, 1693–1702 (1989)
3. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W.: Robust Estimates of Location: Survey and Advances, vol. 279. Princeton University Press, Princeton (1972)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. (B) **39**, 1–38 (1977)
5. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall, London (1995)
6. Hennig, C.: Robustness of ML estimators of location-scale mixtures. In: Innovations in Classification, Data Science, and Information Systems, pp. 128–137. Springer, Heidelberg (2005)
7. Hennig, C., Coretto, P.: The noise component in model-based cluster analysis. In: Data Analysis, Machine Learning and Applications, pp. 127–138. Springer, Berlin (2008)
8. Holland, P.W., Welsch, R.E.: Robust regression using iteratively reweighted least-squares. Commun. Stat. Theor. Methods **A6**(9), 813–827 (1977)
9. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
10. Huber, P., Ronchetti, E.M.: Robust Statistics, 2nd edn. Wiley, New York (2009)

11. Longforda, N.T., D'Ursob, P.: Mixture models with an improper component. J. Appl. Stat. **38**(11), 2511–2521 (2011)
12. Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. B **44**, 226–233 (1982)
13. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics. Wiley, West Sussex, England (2006)
14. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley, Hoboken, New Jersey (1996)
15. Rohan, M.: Using Finite Mixtures to Robustify Statistical Models. Ph.D. Thesis, The University of Waikato (2011)
16. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)

Chapter 3
Determining Risk Factors of Antenatal Care Attendance and its Frequency in Bangladesh: An Application of Count Regression Analysis

Kakoli Rani Bhowmik, Sumonkanti Das, and Md. Atiqul Islam

Abstract Standard Poisson and negative binomial regression models are the common count regression analysis tools for modelling the number of antenatal care (ANC) visits. Two-part (zero and count) models like zero-inflated and hurdle regression models are recommended for modelling ANC visits with excess zeros. The intra-cluster correlation (ICC) can be accounted by incorporating cluster-specific random intercepts in the corresponding standard and two-part models. The existence of excess zeros in the distribution of ANC visits in Bangladesh raises the issue of identifying a proper count regression model for the number of ANC visits covering the issues of overdispersion, zero-inflation, and ICC in determining the risk factors of ANC use and its frequency. The data have been extracted from the 2014 Bangladesh Demographic and Health Survey. The hurdle negative binomial regression model with cluster-specific random effects at both zero- and count- parts is found as the best fitted model. Women who have poor education status, live in poor households, have less access to mass media, and belong to Sylhet and Chittagong divisions are less likely to use prenatal care and to have more ANC visits. In addition, women who live in rural areas, depend on other family members' decision for taking health care, and have unintended pregnancies had lower tendency to more ANC visits. The findings recommend incorporation of random community effects along with overdispersion and zero-inflation in modelling the ANC data of Bangladesh, and model selection should be model-driven rather than data-driven since practically assumption of structural zeros is tough to meet.

K. R. Bhowmik (✉)
Leiden University Medical Centre, Leiden, The Netherlands
e-mail: k.r.bhowmik@lumc.nl

S. Das
Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh

Quantitative Economics, Maastricht University, Mastricht, The Netherlands
e-mail: s.das@maastrichtuniversity.nl

Md. A. Islam
Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_3

Keywords Hurdle model · Negative binomial model · Prenatal care · Random effects · Uniformity test · Zero-inflation

3.1 Introduction

In Bangladesh, a significant proportion of women do not take any ANC visits during their pregnancy period. During the last two decades, this proportion reduced from 50% in 1994 to 20% in 2014 [1]. Consequently, there are great chance of excess zeros (zero-inflation) in the distribution of the number of ANC visits in Bangladesh. However, many studies on determining the risk factors of ANC visits in Bangladesh utilize the standard Poisson regression (PR) and negative binomial regression (NBR) models based on the independence assumption of the number of ANC visits and ignoring the excess zeroes [2, 3]. In such situations, hurdle and zero-inflated regression models (HR and ZIR) known as two-part model are applied for accounting these excess zeros [4]. In few studies, ZIR and HR models have been developed for the frequency of ANC visits to overcome the issues of overdispersion and zero-inflation [5, 6]. Even in some studies, the PR or NBR model has been developed using either non-zero ANC visits or ANC visits without examination of zero-inflation [7–8].

Since complex cluster-sampling design is employed for household level surveys in Bangladesh, it is very common to have intra-cluster correlation (ICC) among the response values. Consequently, it is necessary to account for the ICC along with the issues of overdispersion and zero-inflation in modelling the ANC visits for Bangladeshi pregnant women. Guliani, Sepehri and Serieux [9] explored the determinants of ANC use and the frequency of ANC visits utilizing ANC data of 32 developing countries employing two-part Hurdle model incorporating the cluster effect in model development. No study has been found considering the underlying issues of zero-inflation and the ICC in modelling the frequency of ANC visits of Bangladeshi women. The overdispersion and correlation problem can be partially solved by incorporating cluster/subject specific random effects in the standard PR and NBR models, but the issue of zero-inflation may be remained. The ZIR and HR models are required to extend for accounting this correlation by assuming cluster specific random effects at either count process or both count and zero processes [10, 11].

In this study, a proper count regression model of ANC visits in Bangladesh has been explored covering all the three issues of overdispersion, zero-inflation, and ICC by employing the possible count regression models: the standard PR and NBR models, zero-inflated PR and NBR (ZIPR, ZINBR) models, hurdle PR and NBR (HPR, HNBR) models, mixed effects PR and NBR (MPR, MNBR), mixed effects zero-inflated (MZIPR, MZINBR) and hurdle (MHPR, MHNBR) models. Since types of excess zeros are not possible to determine, both ZIR and HR type models are examined in this study. The extra-benefit of using ZIR and HR models over the standard PR and NBR models is that risk factors of no prenatal care use as well as the determinants of the frequency of prenatal care use can be identified simultaneously. Thus, the main objectives of this study are two-fold: (i) developing a proper count regression model for the frequency of ANC visits in Bangladesh covering the issues of overdispersion, zero-inflation, and ICC; and (ii) determining the risk factors of no ANC use, and also the determinants of the frequency of ANC visits. This study focuses to determine individual-, household-, regional-, and community-level risk factors on a woman's decision to use prenatal care and the frequency of that use after controlling for the unobserved community level factors.

3.2 Methods

3.2.1 Data Description

The data of this study have been obtained from the nationally representative 2014 Bangladesh Demographic and Health Survey (BDHS) [1]. Two-stage stratified sampling design was implemented to select 600 clusters (393 from rural and 207 from urban areas) with probability proportional to the enumeration area size at first stage and then 30 Households per cluster were drawn with an equal probability systematic procedure at second stage. In this study, a total of 4493 ever married women who gave birth in the three years preceding the survey have been analysed. Information of ANC visits are collected only for the last birth even though a woman had two or more live births within the given period. The number of ANC visits (non-negative integer) is the target response variable for which a proper count regression model is aimed to identify in this study. A number of explanatory variables at individual-, household-, regional-, and community levels have been considered for modelling the frequency of ANC visits based on some recent studies on ANC utilization [3, 12, 13]. The bivariate relationship of the explanatory variables with the number of ANC visits was examined at first by developing simple PR model for each of the explanatory variables. A number of individual level variables like education status of woman and her husband, woman's access to mass media exposure, woman's contribution in the decision of her health care utilization, woman's desire of pregnancy along with household wealth status, place of residence, and regional settings have been considered as explanatory variables in the study for developing regression models of the number of ANC visits.

3.2.2 Statistical Models

Let y_{ij} denotes number of ANC visits of i^{th} women living in j^{th} cluster, and the vector $\boldsymbol{X}_{ij}$ denotes the corresponding values of the considered explanatory variables. Assuming independence of ANC visits of i^{th} women, the PR and NBR model can be defined as:

$$\log\left(\mu_{ij}\right) = \beta_0 + \boldsymbol{\beta} \boldsymbol{X}_{ij}^T$$

where μ_{ij} is the expected number of ANC visits as a function of explanatory variables, β_0 is the overall intercept, and $\boldsymbol{\beta}$ is the vector of regression coefficients. The PR and NBR models differ based on the assumed distribution of y_{ij} in respective models. The NBR model is an extension of the PR model incorporating the dispersion parameter in PR model. In PR, the response variable is assumed to follow Poisson distribution with $E(y_{ij}) = \mu_{ij} = \ var\ (y_{ij})$, while in NBR the response variable is assumed to follow negative binomial distribution with $E(y_{ij}) = \mu_{ij}$ and $\text{var}\left(y_{ij}\right) = \mu_{ij} + \mu_{ij}^2/\theta$, where θ is the shape parameter which controls the dispersion. An acceptable way of accommodating non-independence of observations is to develop mixed effects model, which is also known as multilevel modelling. A simple mixed effects PR/NBR model can be developed by incorporating cluster-specific random effects in the standard PR/NBR models as:

$$\log\left(\mu_{ij}\right) = \left(\beta_0 + b_{oj}\right) + \beta X_{ij}^T$$

where, b_{oj} stands for random intercept at cluster level and assumed to follow a normal distribution with constant variance.

The zero-inflated and hurdle extensions of the PR and NBR models are the most prominent and effective models for handling excess zeros in a count data [4]. Both the zero-inflated and Hurdle models have two components (count component and zero component) and separate explanatory variables can be used in the components. Again let $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ are vectors of known explanatory variables used for developing zero- and count-component models respectively. Then the standard zero- and count-component models can be written as

$$logit\left(\varphi_{ij}\right) = \gamma_0 + \gamma Z_{ij}^T \text{ and } \log\left(\mu_{ij}\right) = \beta_0 + \beta X_{ij}^T$$

where μ_{ij} is the probability of zero counts from the binary process (binary logistic model). The difference in these two types of model can be explained by the distribution of y_{ij}. The number of ANC visits is modeled using ZIR model with y_{ij} as:

$$P\left[y_{ij} = 0\right] = \varphi_{ij} + \left(1 - \varphi_{ij}\right) f\left(y_{ij}\right) \text{ and } P\left[y_{ij} \geq 1\right] = \left(1 - \varphi_{ij}\right) f\left(y_{ij}\right)$$

where $f(.)$ is density of either Poisson or negative binomial distribution. While in Hurdle model the number of ANC visits is modeled assuming

$$P\left[y_{ij}=0\right]=\varphi_{ij} \text{ and } P\left[y_{ij}\geq 1\right]=\left(1-\varphi_{ij}\right)f\left(y_{ij}\right)/\left\{1-f\left(y_{ij}=0\right)\right\}$$

where $f(.)$ is the density of standard truncated-at-zero distribution and φ_{ij} is the probability of an observation being zero. The mixed effects ZIR and HR models can be expressed by adding a cluster-specific random component b_{0j} with β_0 and another extra cluster-specific random component random component c_{0j} with γ_0 in zero-component as:

$$logit\left(\varphi_{ij}\right)=\left(\gamma_0+c_{0j}\right)+\gamma Z_{ij}^T \text{ and } \log\left(\mu_{ij}\right)=\left(\beta_0+b_{0j}\right)+\beta X_{ij}^T.$$

The mixed models with extra random effects at zero component are denoted hereafter as MZINBR.ERE (for example) for the MZINBR model. The mixed effects ZIR and HR models are developed by maximum likelihood method approximating the integrals over the random effects with an adaptive Gaussian quadrature rule [14]. For mixed effects ZIR and HR models, “GLMMadaptive” package of R [15] has been employed for the numerical approximation through adaptive Gaussian quadrature. Other R-packages used for developing different version of PR and NBR models are “MASS”, “pscl”, and “lme4”.

This mixture of two types models in a ZIR/HR model allows to interpret answers of two questions (i) which factors influence whether a pregnant woman will attend ANC or not and (ii) which factors can predict the number of times she may visit ANC? Moreover, different explanatory variables can have different impacts at each decision process of two questions.

3.2.3 Model Selection

Considering the data characteristics and different assumptions of the utilized regression models, the comparison of the fitted models is not straightforward. However, several model selection criteria have been utilized to select the most appropriate model: (i) likelihood ratio (LR) test for comparing nested models (the complex model can be constructed by adding variables to simpler model), (ii) Vuong test for comparing non-nested models (model with different distribution function) [16], (iii) Akaike’s information criteria (AIC) for comparing the competing non-nested models (though, it does not indicate whether one model is significantly better than another model), (iv) significance of the dispersion parameter, (v) significance of the zero-inflation, (vi) goodness-of-fit of the model (H_0: fitted model suits well for the data) which is referred as uniformity test, and (vii) significance of cluster-specific random effects. The tests of overdispersion, zero-inflation, and uniformity are conducted based on residual diagnostics for hierarchical (multi-level / mixed) regression models available in DHARMa (Diagnostics for Hierarchical Regression

Models) package [17]. Thus, a step by step comparison procedure is followed providing priority to uniformity test and the significance of the cluster-specific random effects for selecting the final model.

3.3 Results and Discussion

The distribution of the number of ANC visits shown in Fig. 3.1 is positively skewed with very low mean (2.75) and median number of ANC visits (2.0). The distribution of the number of ANC from any source shows that about 22% women did not take any ANC visits and only 32% pregnant women took ANC at least 4 times during their pregnancy period. It is observed that both mean and median frequency of ANC visits significantly vary with administrative region, place of residence, women's education, partners' education, women's mass media exposure, status of women's pregnancy wanted, household wealth status, and household decision maker on healthcare. These significantly associated explanatory variables are used in the model development process.

The policy makers, stakeholders, and donors always explore the risk factors of no prenatal use or lower frequency of prenatal use for setting up the new strategies with the goal of improving the current situation of maternal health care. A proper count regression model of the number of ANC visits incorporating multiple factors help to identify the core risk factors. One-part regression models (such as PR and NBR) and two-part regression models (such as ZIR or HR) are developed with and

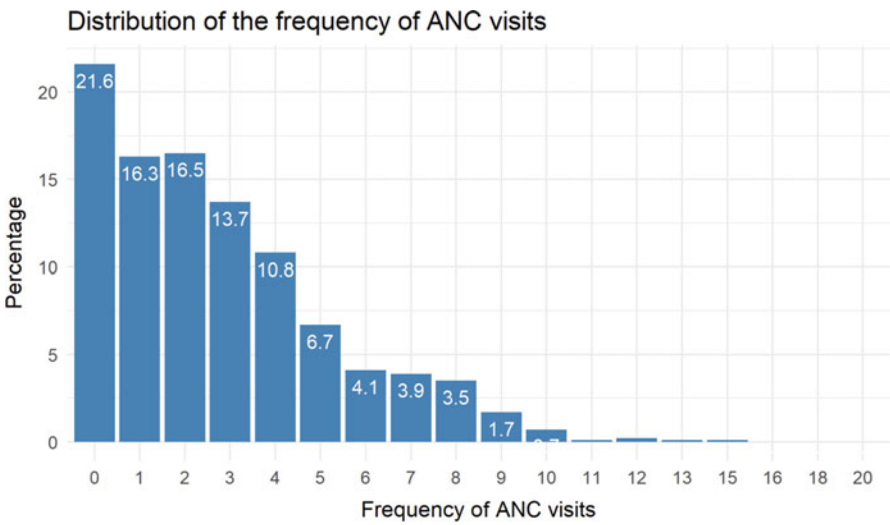


Fig. 3.1 Distribution of the number of antenatal care (ANC) visits among pregnant women in Bangladesh, BDHS 2014

without consideration of ICC and compared to examine whether there are actually two generating process exists in the frequency of ANC visits.

A fixed set of explanatory variables have been used for developing count regression models of ANC frequency for comparison purpose, except the explanatory variables in the zero-component for ZIR and HR models. The comparison of the standard PR and NBR models with their mixed effects models MPR and MNBR shown in Table 3.1 indicates that PR and MPR models fail to capture the overdispersion, while both NBR and MNBR accounts overdispersion but all models are unable to account the issue of zero-inflation. Among these models NBR model seems better for the considered data according to DHARMa uniformity test with very lower p-value (0.066), however AIC, log-likelihood, and LR test indicate inclusion of random intercept is required for the ANC data. Thus the ZIR and HR models without and with random intercepts were developed. The results of Vuong tests for non-nested models shown in Table 3.2 indicates that either ZINBR or HNBR can be considered as better model for accounting excess zeroes. Table 3.2 also reflects that the over-dispersion is captured better by the NBR based models than the PR based models. The results of LR tests for nested models shown in Table 3.3 indicate that cluster-specific random intercepts should be considered in the NBR based models. Random effects are also found important for both count-

Table 3.1 Akaike's information criteria (AIC), log-likelihood, likelihood-ratio (LR), dispersion, zero-inflation and uniformity tests of PR, NBR, MPR and MNBR models

Model	AIC	log-likelihood (df)	LR test (p-value)	Dispersion test (ratio & p-value)	Zero-inflation test (ratio Statistic & p-value)	Uniformity test (D-Statistic & p-value)
PR	19274.66	−9613.33 (24)	712.52 (0.000)	1.358 & 0.00	2.001 & 0.00	0.114 & 0.00
MPR	18565.31	−9257.66 (25)		1.096 & 0.004	1.685 & 0.00	0071 & 0.00
NBR	18314.42	−9132.21 (25)	216.47 (0.000)	0.937 & 0.000	1.208 & 0.00	0.019 & 0.066
MNBR	18099.96	−9023.98 (26)		0.913 & 0.002	1.223 & 0.00	0.022 & 0.026

Table 3.2 Vuong tests for the non-nested PR, ZIPR, HPR, NBR, ZINBR and HNBR models

Model 1	Model 2	Test statistic (AIC corrected)	p-value	Better model
PR	NBR	−11.852	< 2.22e-16	NBR
PR	ZIPR	−12.316	< 2.22e-16	ZIPR
PR	HPR	−12.222	< 2.22e-16	HPR
ZIPR	HPR	0.532	0.297	ZIPR/HPR
NBR	ZINBR	−7.820	< 2.639e-15	ZINBR
NBR	HNBR	−7.719	< 5.882e-15	HNBR
ZINBR	HNBR	0.024	0.490	ZINBR/HNBR
ZIPR	ZINBR	−7.829	2.461e-15	ZINBR
HPR	HNBR	−7.766	4.053e-15	HNBR

Table 3.3 Likelihood Ratio (LR) tests for the nested ZINBR, MZINBR, MNZINBR.ERE, HNBR, MHNBR, and MHNBR.ERE models

Model 1	Model 2	LR test statistic	DF	p-value
ZINBR	MZINBR	128.40	1	< 2.2e-16
MZINBR	MZINBR.ERE	66.71	2	< 0.0001
HNBR	MHNBR	86.59	1	< 2.2e-16
MHNBR	MHNBR.ERE	109.81	2	< 0.0001

Table 3.4 DHARMa Uniformity Test of MZINBR, MNZINBR.ERE, MHNBR, and MHNBR.ERE models and the corresponding count-part (σ_C^2) and zero-part (σ_Z^2) variance components

Model	MZINBR	MZINBR.ERE	MHNBR	MHNBR.ERE
D-Statistic	0.016	0.022	0.011	0.018
p-value	0.199	0.025	0.653	0.121
σ_C^2	0.074	0.060	0.064	0.066
σ_Z^2	–	1.264	–	0.582
ρ_c	0.022	0.018	0.019	0.020
ρ_z	–	0.278	–	0.150
Log-likelihood	−8911.048	−8877.694	−8932.08	−8877.176

P < 0.05 indicates the model doesn't fit well for the count data

and zero-part models in both cases of ZINBR (MZINBR and MZINBR.ERE) and HNBR (MHNBR and MHNBR.ERE) models.

Since there are four possible candidates as the best model for the ANC data, the DHARMa's uniformity test has been done for finding the best suitable model. Table 3.4 shows that ZINBR with random effects at count-part (MZINBR) confirms uniformity with the observed count data but LR test in Table 3.3 shows that this model still requires random intercepts at the zero part (MZINBR.ERE). However, the MZINBR.ERE failed the uniformity test. On the other hand, HNBR with random intercepts at count-part (MHNBR) and HNBR with random intercepts at both count and zero parts (MHNBR.ERE) passed the uniformity test. Thus, MHNBR.ERE can be considered as the best model among the possible best candidate models for the ANC data of Bangladesh. Also, MHNBR and MHNBR.ERE models show comparatively lower cluster-specific variance components (as well as lower ICC) than MZINBR and MZINBR.ERE models provide (Table 3.4). Informal diagnoses of the cluster-specific residuals through Q-Q plot, histogram and distribution shown in Fig. 3.2 confirm that the cluster-specific residuals obtained from both count- and zero-parts are normally distributed.

According to the selected HR model with random effects at both count and zero components (MHNBR.ERE) shown in Table 3.5, division, place of residence, household economic status, women media exposure, women and their partner's education status, women's contribution in health care decision, and desire for pregnancy have highly significant effects on either zero prenatal care use or the frequency of prenatal care use. The count-part model shows the effects of the

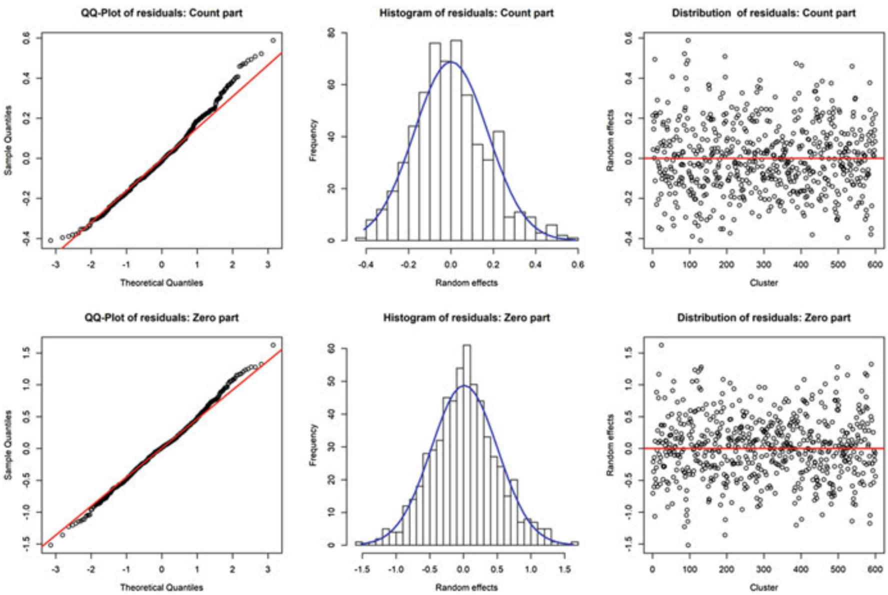


Fig. 3.2 Model diagnostics of both count and zero parts of the fitted mixed effect hurdle model through Q-Q plot, histogram and distribution of the cluster-specific residuals

considered factors on the frequency of ANC visits represented as incidence rate ratio (IRR), while the zero-part model shows the effects of the considered factors on the women decision of taking no ANC represented as odds ratio (OR). Since both parts have cluster-specific random effects, the estimated parameters represent the effects of individual-, household-, regional-, and community-level characteristics on ANC attendance and frequency of ANC use after controlling the unobserved community level factors. It is noted that regression coefficients of other models are not presented here, only their summary statistics are stated for comparison purpose.

From the count-part model, it is observed that women residing in Khulna (IRR:1.182) and Rangpur division (IRR:1.239), women from richer (IRR: 1.155) and richest (IRR: 1.306) households, women having access of mass media at least once a week (IRR:1.156), women attending secondary (IRR: 1.213) and higher education (IRR:1.430), and women with desire of pregnancy (IRR:1.232) had significantly higher IRR of attending ANC visits while women residing in rural area (IRR: 0.840) and women without power on health care decision (IRR:0.870) had significantly lower IRR after accounting the variation at the cluster level.

On the other hand, the regression coefficients of the zero-part model indicate that mothers residing in Khulna (OR: 0.496) and Rangpur (OR:0.629) divisions, women from middle (OR:0.633), richer (OR:0.401) and richest (OR:0.197) households, women having access of mass media at least once a week (OR:0.567), women having primary (OR: 0.725), secondary (OR: 0.450) and higher education (OR: 0.243), and women's partners having secondary (OR:0.628) and higher education

Table 3.5 Estimated incidence rate ratio (IRR) of having ANC visits and odds ratio (OR) of not attending any ANC visit (with 95% CI and p-values) from the hurdle negative binomial regression with random intercept at both count- and zero-part (HNBR.ERE) models, BDHS 2014

		Count-part (Number of ANC visits)				Zero-part (No ANC attendance)			
Factors	Category	*IRR*	95% CI		p-value	*OR*	95% CI		p-value
Region	Barisal ^{R}								
	Chittagong	0.907	0.80	1.03	0.133	1.226	0.83	1.82	0.309
	Dhaka	0.998	0.88	1.13	0.981	0.796	0.53	1.20	0.272
	Khulna	1.182	1.04	1.35	0.012	0.496	0.31	0.78	0.002
	Rajshahi	1.008	0.88	1.15	0.907	0.900	0.59	1.37	0.627
	Rangpur	1.239	1.09	1.41	0.001	0.629	0.41	0.96	0.032
	Sylhet	0.931	0.81	1.07	0.308	1.468	0.98	2.20	0.061
Place of Residence	Urban ^{R}								
	Rural	0.840	0.78	0.90	0.000				
Wealth Status	Poorest ^{R}								
	Poorer	1.048	0.95	1.15	0.348	0.804	0.64	1.01	0.065
	Middle	1.026	0.93	1.14	0.615	0.633	0.48	0.83	0.001
	Richer	1.155	1.04	1.28	0.007	0.401	0.29	0.55	0.000
	Richest	1.306	1.16	1.47	0.000	0.197	0.13	0.31	0.000
Mass Media exposure	Not at all ^{R}								
	Less than once a week	1.100	0.99	1.22	0.064	0.766	0.57	1.03	0.082
	At least once a week	1.156	1.08	1.24	0.000	0.567	0.45	0.71	0.000
Mother's education	Illiterate ^{R}								
	Primary	1.088	0.98	1.21	0.130	0.725	0.57	0.92	0.009
	Secondary	1.213	1.09	1.35	0.001	0.450	0.35	0.59	0.000
	Higher	1.430	1.25	1.63	0.000	0.243	0.14	0.43	0.000
Pregnancy wanted	No ^{R}								
	Yes	1.232	1.12	1.36	0.000				
Decision on health care	Woman alone ^{R}								
	Woman & husband	0.969	0.90	1.05	0.418				
	Husband alone	0.920	0.85	1.00	0.051				
	Other	0.870	0.78	0.97	0.012				
Partner's education	Illiterate ^{R}								
	Primary	0.985	0.91	1.07	0.722	0.969	0.78	1.20	0.772
	Secondary	1.005	0.92	1.10	0.911	0.628	0.48	0.82	0.000
	Higher	1.092	0.98	1.21	0.103	0.474	0.30	0.75	0.001
Intercept		1.906	1.59	2.29	0.000	1.090	0.75	1.59	0.654

^{R}refers to reference group

(OR:0.474) had significantly lower odds of not attending ANC visit. The values of cluster-specific variance components at count-part (σ_c^2 =0.07 and ρ_c=0.02) and zero-part (σ_Z^2 =0.58 and ρ_z=0.15) indicate significant variation in the number of ANC visits due to between-cluster heterogeneity.

3.4 Conclusions

The selected HR model confirms that two processes generate the frequency of the ANC visits in Bangladesh: one process generates zero ANC visits and the other generates frequency of the ANC visits. The significance of cluster-specific variance component at both zero- and count-part indicates that community (cluster) has a significant effect in the variation of both woman's decision on prenatal care use and the number of ANC visits. The application of multilevel modelling in this study has allowed to account community-level variations in the number of ANC visits, although the maximum variations originated mainly from women-, household-, and regional level factors. The significant cluster-level variation also indicates that the goal of reducing maternal death could be reached if heterogeneity in the prenatal care use and its frequency could be reduced at the community level.

The study has shown the necessity of considering community effects (ICC) along with overdispersion and zero-inflation in modelling the ANC data of Bangladeshi women and hence, for identifying risk factors of not attending in any ANC visit as well as the number of ANC visits. Though only random intercept model has been covered in this study, further investigation can be done for checking significance of any random slope in the model. Also, the variability at higher administrative units (such as district and sub-district) can also be examined by developing three or four level models [18]. In addition, selection of either ZIR or HR models should be taken carefully since assumption of all structural zeros is tough to meet in real world data. It is better to take decision based on model rather than based on types of zeros.

The study findings suggest that women living in Khulna and Rangpur divisions are more positive to attend and have more ANC visits during their pregnancy period. The inequalities in attending and taking adequate ANC visits by women education and household economic condition suggest that government still need more initiatives focusing to the illiterate and less educated women as well as who live in poor households. It is also observed that partners' education has significant influence on their wives attendance to have ANC but not in the frequency. Though the access to mass media is very cheap to all people, still a big proportion of women are not concerned about their maternal health and ignore taking any care during their pregnancy. While two important factors "desire to pregnancy" and "contribution to health care decision" are found to have significant influence on the frequency of ANC visits. These two factors indirectly linked with women empowerment, which helped women to seek more medical care during their pregnancy. Lower tendency to have more ANC visits among rural women suggest that either they are still conservative to seek frequent ANC visits during their pregnancy or do not have persistent access to have adequate medical care. These findings of this study might help the policy makers to find out which socio-economic and demographic groups should be provided priority in regards to encouraging women to attend and to take more ANC from medically trained personnel during their pregnancy period. The study findings also suggest that beside the improvement of women academic education and household economic condition, women should be motivated to change their attitude to seek medical care during their pregnancy by improving their maternal education.

References

1. NIPORT, Mitra and Associates: ICF International. Bangladesh Demographic and Health Survey 2014: Bangladesh and Rockville. Maryland, NIPORT, Mitra and Associates, and ICF International, Dhaka (2016)
2. Amrin, A.: An analysis of the status of antenatal care in Bangladesh. Int. J. Sci. Res. Methodol. **5**(2), 49–57 (2016)
3. Islam, M.M., Masud, M.S.: Determinants of frequency and contents of antenatal care visits in Bangladesh: assessing the extent of compliance with the WHO recommendations. PLoS One. **13**(9), e0204752 (2018)
4. Staub, K.E., Winkelmann, R.: Consistent estimation of zero-inflated count models. Health Econ. **22**(6), 673–686 (2013)
5. Hossain, B., Hoque, A.A.: Women empowerment and antenatal care utilization in Bangladesh. J. Dev. Areas. **49**(2), 109–124 (2015)
6. Ali, M.M.: Microeconometric analysis on determinants of antenatal care in Bangladesh: a finite mixture modelling approach. Indian Econ. J. **60**(3), 91–107 (2012)
7. Yusuf, O.B., Ugalahi, L.O.: On the performance of the Poisson, negative binomial and generalized Poisson regression models in the prediction of antenatal care visits in Nigeria. Am. J. Math. Stat. **5**(3), 128–136 (2015)
8. Zegeye, E.A., Mbonigaba, J., Dimbuene, Z.T.: Factors associated with the utilization of antenatal care and prevention of mother-to-child HIV transmission services in Ethiopia: applying a count regression model. BMC Womens Health. **18**(1), 187 (2018)
9. Guliani, H., Sepehri, A., Serieux, J.: Determinants of prenatal care use: evidence from 32 low-income countries across Asia, Sub-Saharan Africa and Latin America. Health Policy Plan. **29**(5), 589–602 (2013)
10. Hall, D.: Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. **56**, 1030–1039 (2000)
11. Yau, K.K., Lee, A.H.: Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Stat. Med. **20**, 2907–2920 (2001)
12. Rahman, M., Islam, R., Islam, A.Z.: Rural-urban differentials of utilization of ante-natal health-care services in Bangladesh. Health Policy Dev. **6**(3), 117–125 (2008)
13. Ali, N., Sultana, M., Sheikh, N., Akram, R., Mahumud, R.A., Asaduzzaman, M., et al.: Predictors of optimal antenatal care service utilization among adolescents and adult women in Bangladesh. Health Serv. Res Manage. Epidemiol. **5**, 2333392818781729 (2018)
14. Pinheiro, J.C., Bates, D.M.: Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comput. Graph. Stat. **4**(1), 12–35 (1995)
15. Rizopoulos, D.: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature (GLMMAdaptive). https://CRAN.R-project.org/package=GLMMadaptive (2019).
16. Vuong, Q.H.: Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 307–333 (1989)

17. Hartig, F.: DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.2.3 (2019)
18. Das, S., Rahman, A., Ahamed, A., Rahman, S.T.: Multi-level models can benefit from minimizing higher-order variations: an illustration using child malnutrition data. J. Stat. Comput. Simul. **89**(6), 1090–1110 (2019)

Chapter 4
On Propensity Score Methodology

Paul Dewick and Shuangzhe Liu

Abstract In an observational study, researchers are constantly required to distinguish the effects caused by the assignment of treatment. Propensity score methodology is one way to determine the effects of, and their probabilities, given a vector of observed covariates, which is particularly popular in the fields of medical, pharmaceutical and social sciences. However, there are mixed views for the best methodologies to use and an overall understanding of the propensity score methodology. Also, there is minimal literature for propensity score methods being used within the broader scientific community. Propensity score methodology can be suited to determine effects caused by, not only treatment of pharmaceutical medication, but for "treatment" of some external event, proposed event or interaction within the wider community. For example, the effect on a regional community due to business closure, or a road by-pass would be a reasonable case of how propensity score methods can be further used within the wider scientific community.

The main objective of this paper is to demonstrate how propensity score methodology can be used to answer questions on effects caused by an external event or interaction on a community. The propensity score methodology will be given a framework that can be followed, explained and reported that will help allow for robust decision making, planning and policy decisions to be undertaken.

Keywords Propensity scores · Methodology · Framework · Policy

4.1 Propensity Score Methods – Estimating Causal Effects

Propensity score methods within the literature contains many variations and methodologies that are based on the methods contained within this paper. This paper highlights the most commonly used methods. Propensity score methodology

P. Dewick (✉) · S. Liu
University of Canberra, Canberra, Australia
e-mail: Paul.Dewick@Canberra.edu.au; Shuangzhe.Liu@Canberra.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_4

was introduced in 1983. The central role of the propensity score in observational studies for causal effects, Rosenbaum and Rubin [1] gives the mathematical theory that underpins propensity score methodology. Propensity score methods are based on the Rubin's causal model [1], which is a methodology to estimate the causal effects of treatment between two groups, a treated and a non-treated group and the estimated probability of treatment effect from an observational non-randomized study. Propensity score methods [2] are a group of strategies that aim to reduce selection bias by balancing differences between treated and non-treatment groups on observed covariates.

4.1.1 A Framework to Model Propensity Score Methods

The literature that surround propensity score methods can at times be confusing as propensity score terminology and process are not clearly defined, which can lead to a misunderstanding of the methodology. A framework to model propensity score methods is given in this paper is to guide the researcher through the process and to identify areas where the methodology can be used and also give an understanding of where the other methods may "fit in" with the propensity score methods.

As the propensity score literature is usually described using the same format, by focusing on the outcomes of the propensity score methodology will allow for better understanding of the processes involved which will give the process a basic framework that will allow for further inclusions of more complex propensity score methodologies.

4.2 Methods to Estimate the Effect of Treatment Using Propensity Scores

The methodology used to estimate the effects of treatment by using propensity score methods is shown in Fig. 4.1, which uses the covariates from a treatment and non-treatment observational, non-randomized study.

There are four [3] commonly used methods to calculate the effect of treatment using propensity score methodology that adjust for confounding variables in observational studies which are:

1. Matching on the propensity scores (Matching),
2. Inverse probability weighing,
3. Subclassification on the propensity scores (Subclassification or Stratification), and
4. Propensity scores and covariance adjustment (Covariate adjustment).

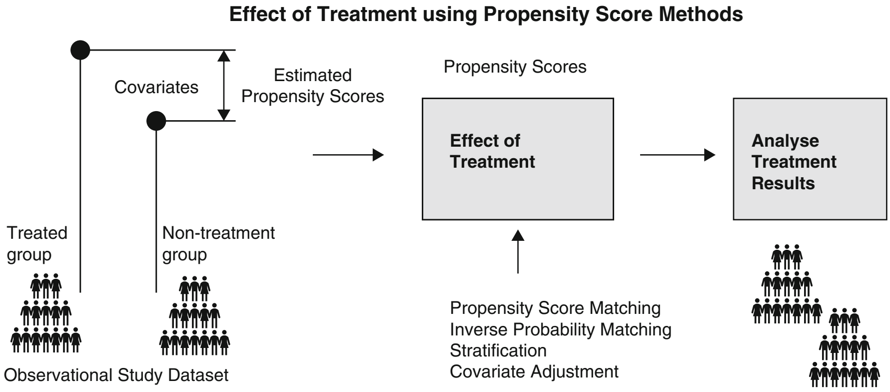


Fig. 4.1 Diagram of estimating the effect of treatment using propensity score methods

From these four propensity score methods available, literature shows that many applied researchers favor the use of matching (matching on the propensity scores) due to the ease [3] it can be undertaken and understood by both researcher and report audience. Matching is a procedure [4] that involves dropping, repeating or grouping observations from an observed data set to reduce covariance imbalances between the treatment and non-treatment groups that were not avoided during data collection.

Often mentioned within the literature are the propensity score matching algorithm classes which define the properties which all matching methods possess. (some methods may also belong to more than one class). These matching algorithms are defined in two classes which are either Equal Percent Bias Reducing (EPBR) or Monotonic Imbalance Bounding (MIB). The EPBR class balances the matched data [5] by reducing the mean-imbalance on one variable that can improve the imbalance on all variables by a proportional amount. The MIB is a class of matching method [5] which produces subsets on the bases of a given vector of tuning parameters (such as caliper), one for each covariate. As a result, the number of matched units is a function of the tuning parameter.

In practice, MIB methods [5] sometimes generate too few matched observations (this may be considered an important advantage). These methods require assumptions to be made about the data [5]. Whilst some of these assumptions do hold when using propensity score methods, some assumptions do not hold true in real life therefore the use of the matching method chosen must be checked to avoid misspecification.

Most matching algorithms are EPBR, however the CEM methodology is a MIB class algorithm. Should the dataset be appropriate for a CEM model, then it therefore should achieve good balance, however, the other EPBR methods may require several iterations of different types of matching methods to achieve good balance.

4.2.1 Matching on the Propensity Scores (Matching)

Whilst matching is the most common method to undertake a propensity score analysis, with many matching methodologies available to estimate the treatment effect. The most common methods are;

1. Full Matching,
2. Nearest Neighbour Matching (Greedy),
3. Coarsened Exact Matching (CEM),
4. Optimal Matching,
5. Genetic Matching, and
6. Limited Exact Matching.

The matching methodologies will not be discussed here as there is plenty of coverage within the literature. Refer to [6–8].

4.2.1.1 Matching "with" or "without" Replacement

A key issue associated with matching is whether matching should be undertaken "with replacement" or "without replacement". Matching with replacement [9] can yield better matches because non-treatment's that are similar to many treated units can be used multiple times. However, it may only select a few unique non-treated units. As a result, inference can become more complex [7] because the matched non-treatment is no longer independent, some are in the matched sample more than once and this needs to be accounted for in the outcome analysis which is beyond the scope of this paper.

Another issue when matching with replacement [7] it is also possible that the treatment effect estimate will be biased on just a small number of non-treated units. With that said, matching with replacement can be useful [7] when there are a limited number of non-treatment units with values similar to those in the treated group, therefore researchers should use matching with replacement with caution.

4.2.2 Inverse Probability Weighing

Inverse probability weighting [10] uses the whole dataset but reweights individuals to increase the weights of those who received unexpected exposures. This procedure can be thought of as producing additional observations for those parts of the target population from which there were few observations. It effectively generates a pseudo population with near perfect covariate balance between treatment groups.

4.2.3 Subclassification

Subclassification on the propensity score is a method [6] of adjustment for confounding variables that involves, stratifying subjects into mutually exclusive subsets based on their estimated propensity score. Subjects are then stratified into subsets based on previously defined thresholds of the estimated propensity score.

4.2.4 Covariance Adjustment

With covariate adjustment (regression adjustment), the propensity scores are first estimated and then the estimated propensity scores are regressed on the treatment. This methodology [11] is used at times in combination with one of the approaches, matching, stratification, or weighing to remove any residual differences between treatment group.

4.2.5 Diagnostics of Matching, Subclassification and Covariate Adjustment Methods

An important step in using propensity score methods is the diagnostic quality of the matching, inverse probability weighing, subclassification or covariate adjustment. Achieving good balance (common support, overlap) is critical as achieving this will demonstrate that the model has been correctly specified and the diagnostic methods will be able to determine the estimated effect of treatment. Gaining good balance may require several iterations. However, balance checking [12] is one of the most problematic aspects of the propensity score methodology.

4.2.5.1 Balance

To achieve good balance, matching, inverse probability weighting, subclassification or covariate adjustment may need to be undertaken multiple times using various methods to achieve good balance, because better the balance, the bias within the model will be reduced. However, should the methodology fail to produce good balance using the methods suggested here, then the model may not be suitable to allow for a competent estimate of treatment effect to be undertaken, therefore propensity score methods are not suitable for the data that is being modeled.

4.2.5.2 Assessing Balance

Whatever methodology is used to estimate the effect of treatment, the diagnostics are the same for assessing balance, with balance defined as any method that aims to equate (or balance) the distribution of covariates in the treated and non-treatment groups [7, 13], it is common within the literature to suggest that significance tests or a better method, a standardized difference in means [7] be undertaken to asses balance.

Common diagnostics [14] include, t-tests of the covariates, Kolmogorov-Smirnov tests and other comparisons of distributions. However, as the balance property is from a given sample and not a super-population significance testing is not appropriate [7, 13]. The matching may appear to result in better balance solely due to the decrease in sample size compared with the initial unmatched sample [14]. Assessment of balance should be undertaken using methods that asses balance of the sample and not some hypothetical population also, the sample size should not affect the value of the statistic.

A more robust methods to assess the balance includes [6] standardized differences (standard difference), comparing higher-order moments, propensity score summary statistics and empirical quantile-quantile plots for each variable. The standardized difference [14] does not depend on the unit measurement, it satisfies the criteria of its property of the sample and does not depend upon the size of the sample. As propensity score methods are undertaken using a non-randomized observational study, you cannot expect to achieve perfect balance for all measured baseline variables between treated and non-treatment subjects in the matched sample.

With many covariates [7] difficult to carefully examine numeric diagnostics for each graphical diagnostics can be helpful for getting a quick assessment of the covariate balance. A first step is to examine the distribution of the propensity scores in the original and matched groups, this is also helpful for assessing the balance.

For weighing or subclassification [7], plots such as this can show the dots with their size proportional to their weight. For continuous covariates, you can also examine quantile-quantile (QQ) plots, which compare the empirical distributions of each variable in the treated and non-treatment groups. QQ plots [7] compare the quantiles of variable in the treatment group against the corresponding quantiles in the non-treatment group. If the two groups have identical empirical distributions, all points would lie on the 45-degree line. For weighing methods, weighted boxplots can provide similar information. A plot of the standardized differences of means [7] gives a quick overview of whether balance has improved for individual covariates.

As shown in this paper, there is a wide variety of matching methods available and there is little guidance on the best methods, Stuart [7] has developed some guidelines, which are methodology to access the balance;

1. The method that yields the smallest standardized differences of means across the largest number of covariates,
2. The method that minimizes the standardized differences of means of a few particularly prognostic covariates, and
3. The method that results in the fewest number of "large" standardized differences of means (greater than 0.25).

4.2.5.3 Estimating Treatment Results

The treatment effect needs to be estimated form the matched covariates. There are several methods available, the most common method to estimate the treatment effect is to calculate the Average Effect of the Treatment (ATT), and the Average Treatment Effect (ATE), where;

1. ATT – Effect for those in the treatment group, or the average gain from treatment of those who were actually treated and,
2. ATE – Effect on all individuals, treatment and non-treatment, or the treatment effect in the population.

The ATT and the ATE can be obtained from the matching, inverse probability weighing, subclassification or covariate adjustment process. However, not all of these methodologies estimate the ATT or the ATE, and some will calculate both. Prior to selection of the matching process, a decision needs to be made on which is the most important for the study.

The difference in the calculation of the ATT and ATE [2] is that the ATE requires adequate balance (common support) for both treated and non-treated units, whereas the ATT only requires that the distribution of propensity scores of the treated is contained within distribution of the scores of untreated units.

Prior to choosing the propensity score methodology to estimate the treatment effect [11] consideration needs to be undertaken to whether the ATT or ATE is more relevant and also, what is feasible, given the data. Propensity score methods often focus on estimating the ATT and not the ATE [13] as the closest non-treated matches are selected for the treated units, and unmatched units are often excluded from the analysis. As a guide, a matching method implementation by Stuart [7];

1. If estimating the ATE, matching is usually undertaken by IPTW or full matching,
2. If estimating the ATT and there are many more non-treatment units (x3), nearest neighbor without replacement should be undertaken and,
3. If estimating the ATT and there are not (or not many) more non-treatment than treated units, appropriate choices would be subclassification or full matching.

4.2.5.4 Alternate Methods to Estimate the Treatment Effect

An alternate approach to estimate the treatment effect is to undertake a regression [15]. An advantage of regression is that it provides a level of "double robustness" by adjusting for any remaining covariate imbalance. The outcomes of the treated and non-treated groups are then compared using a regression model that controls for all covariates used in matching, plus a treatment indicator variable. The coefficient associated with this indicator is interpreted as the treatment effect.

4.2.6 Other Issues Identified Associated with Propensity Score Methodologies

Other issues that have been identified within the literature is misspecification. Misspecification can lead to an increase in bias and variance. The magnitude of bias depends on model misspecification.

4.2.6.1 Misspecification of the Dataset

Misspecification of the dataset can be where the dataset contains variables that are fully (or nearly fully) predictive of treatment assignment. Propensity score methods must have strong ignorability [7] which implies that treatment assignment is independent of the potential outcome, given the covariance.

Generally, poor performance [7] is found of methods that use a relatively small set of "predictors of convenience" including variables that are un-associated with treatment assignment, as they will be of little influence in the propensity score model. However, excluding a potentially important confounder can be very costly in terms of increased bias. Therefore, the covariates that are chosen for the model must be carefully assed.

4.2.6.2 Misspecification of the Matching, Subclassification and Covariate Adjustment Methods

Misspecification of the matching, subclassification or covariate adjustment method used to estimate the treatment effect will introduce high levels of bias which will have a detrimental effect on the accuracy of the propensity score model. The correct specification of the propensity score is indicated by whether estimating the treatment effect achieved good balance between the treated and non-treated groups on the relevant covariates.

4.3 Probability of Treatment Using Propensity Score Methods

After the effect of treatment has been estimated, assuming that the model was correctly specified, estimating the probability of treatment is undertaken by further analysis on the matched, inverse probability weighing, stratified or covariance adjusted dataset by using regression methods or decision trees to determine the probability of treatment of a unit that comes from the treatment group as shown in Fig. 4.2.

Other variations for calculating the probability of treatment using propensity score methodology have been suggested within the literature but again are beyond the scope of this paper.

4.3.1 Probability of Treatment Using Regression Methods

The estimation of probability of treatment is undertaken after matching on the estimated propensity scores has been undertaken. The propensity score is frequently calculated using a logistic regression model [6] with exposure to the treatment as the dependent variable. This is undertaken as a separate step after matching, subclassification, or covariate adjustment has been successfully undertaken, with a balanced and correctly specified model, with the researcher generally having a choice of several different methods available, which are;

1. Logistic Regression,
2. Probit Regression, and
3. Generalized Boosted Regression.

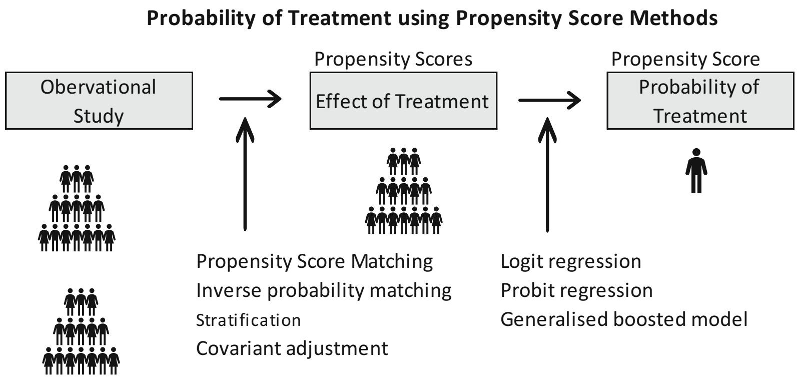


Fig. 4.2 Diagram of estimating the probability of treatment using propensity score methods

4.3.1.1 Logistic Regression

Logistic regression is the technique that is most associated with propensity score analysis as it is the simplest and easiest to understand and implement. The logistic expression is used [9] to determine the probability of membership into the treatment group.

Logistic regression is more versatile and better suited for modelling [16] most situations as it does not rely on the assumption that the independent variables are normally distributed. With logistic regression, a mathematical model of a set of explanatory variables is used to predict using a logit transformation of the dependent variable.

4.3.1.2 Probit Regression

The probit and logistic regression models tend to produce similar results [17] and that the choice between the two is largely one of convenience and convention. For models that have more than 2 categories [17], the logistic model may be preferred because the corresponding probit model is too computationally demanding. For panel data, you can only estimate a fixed effects model with logit, not with probit.

4.3.2 Probability of Treatment Using Decision Trees

A regression tree [18] uses a recursive tree fitting algorithm to estimate a function describing the relationship between a multivariate set of independent variables and a single dependent variable, such as treatment assignment.

4.3.2.1 Generalized Boosted Model

Generalized boosting [18] is a general automated, data adaptive algorithm that can be used with a large number of covariates and predict treatment assignment. It is a multivariate non-parametric regression technique [18] to estimate propensity scores.

Generalized boosted models are based on decision trees [9] that create a complex model by combining multiple simple models using iterative algorithms. Through this iterative process, these models will then include interactions and polynomial terms that will produce a better model without external guidance.

The specific fit of the model [9] is dependent on parameters such as the number of trees, the interaction depth, the fraction of data used for training and the threshold to stop iterations. Some evidence suggests that [9] generalized boosted models may outperform logistic regression.

4.3.3 *Estimated Probability of Treatment Results*

Unlike estimating the treatment effect, the estimated probability of treatment is the estimate directly calculated from the regression and decision trees methods described above, as the model has already been determined to be correctly specified.

4.4 Propensity Score Methods Applied to Regional Centres

The propensity score methods that are given in this paper can be used to help guide policy decisions by determining the effect of a business closure within a regional center. This procedure would be not different to the propensity score methods explained, where, a regional center that may have a business shut down would be the treatment, and similar town that suffered the same fate (similar to the treatment town) would be the non-treatment regional center.

Quasi-experiment designs would be suitable for this type of study [19], as it identifies a comparison group that is similar as possible to the treatment group in terms of baseline characteristics. The comparison group captures what would have been the outcome if the closure of the town had not been implemented. Therefore, the closure can be said to have caused any difference in outcome between the treatment and non-treatment groups. If there are reliable measures of all confounding covariates, then propensity score methods can enable the identification of causal effects.

Therefore, determining the effects of business shutdown in a regional city to determine if there will be an effect on a regional community and the probability that a business will be affected is viable using propensity score methods.

Evaluation of funded programs is an important policy issue to determine the effects of a given program's effectiveness and efficiency. Propensity score methods have been used for this very purpose to assess the impact of policy intervention, where a program was implemented [20].

4.5 Concluding Remarks

Although not as frequently used outside the fields of medical and pharmaceutical studies and at times the literature can be confusing, propensity score methodology is starting to gain some traction in other areas of research that may have interests in causal effects.

Propensity score methodology can be integrated into other studies, to confirm or help justify policy decisions or external causal effects on communities. Further uses may be identified within the literature. As an aim of this paper is to give an overview of propensity score methods and to highlight the basic structure of the analysis, it will provide a researcher insight to identify propensity score methods that will suit a particular situation.

References

1. Rosenbaum, R.P., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika. **70**(1), 41–55 (1983)
2. Leite, L.W., Jimenez, F., Kaya, Y., Stapleton, L.M., MacInnes, J.W., Sandbach, R.: An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. Multivar. Behav. Res. **50**, 265–284 (2015). https://doi.org/10.1080/00273171.2014.991018
3. Austin, P.C., Stuart, E.A.: Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Stat. Method. Res. **26**(6), 2505–2525 (2015). https://doi.org/10.1177/0962280215601134
4. Kosuke, I., King, G., Stuart, E.A.: Misunderstandings between experimentalists and observationalists about causal inference. J. R. Stat. Soc. A. **171**(2), 481–502 (2008)
5. Stefano, I.M., King, G., Porro, G.: Multivariate matching methods that are monotonic imbalance bounding. J. Am. Stat. Assoc. **106**(493), 345–361 (2011). https://doi.org/10.1198/jasa.2011.tm09599
6. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Covariate Behav. Res. **46**, 399–424 (2011). https://doi.org/10.1080/00273171.2001.568786
7. Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci. **25**(1), 1–21 (2010). https://doi.org/10.1214/09-STS313
8. Stuart, E.A., Rubin, D.B.: Matching methods for causal inference: designing observational studies, DRAFT. To appear in best practices in quantitative methods (Osbourne, J., Ed.). Sage, Thousand Oaks (2007)
9. Olmos, A., Govindasamy, P.: A practical guide for using propensity score weighing in R, practical assessment. Res. Eval. **20**(13), 1–8 (2015). ISSN 1531-7714
10. Markus, E.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G.W., Pocock, S.J.: Comparison of propensity score methods and covariate adjustment, evaluation in 4 cardiovascular studies. J. Am. Coll. Cardiol. **69**(3), (2017). ISSN 0735-1097
11. Inacio, C.S., Chen, M.Y., Paxton, E.W., Namba, R.S., Kurtz, S.M., Cafri, G.: Statistics in brief: an introduction to the use of propensity scores. Clin. Orthop. Relat. Res. **473**, 2722–2726 (2014). https://doi.org/10.1007/s11999-015-4239-4
12. Hill, J.: Commentary, discussion of research propensity-score matching: comments on a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003′ by Peter Austin. Stat. Med. **27**, 2055–2061 (2008). https://doi.org/10.1002/sim.3245
13. Sanni, A.M., Groenwold, R.H.H., Klungel, O.H.: Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. Am. J. Clin. Nutr. **104**, 247–258 (2016)
14. Austin, P.C.: A critical appraisal of propensity-score matching in medical literature between 1996 and 2003. Stat. Med. **27**, 2037–2049 (2008). https://doi.org/10.1002/sim.3150
15. McMurry, L.T., Hu, Y., Blackstone, E.H., Kozower, B.D.: Propensity scores: methods, considerations, and applications in the journal of thoracic and cardiovascular surgery. J. Thorac. Cardiovasc. Surg. **150**(1), 14–19 (2015)
16. NCSS Statistical Software: Chapter 321, Logistic Regression. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pd. Last accessed 2019/7/23

17. Williams, R.: Alternatives to logistic regression (Brief Overview), pp. 1–5. University of Notre Dame (2018). https://www3.nd.edu/~rwilliam/
18. McCaffrey, F.D., Ridgeway, G., Morral, A.R.: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. RAN.. http://citeseerx.ist.psu.edu. Last accessed 2019/06/23
19. Yongnam, K., Steiner, P.: Quasi-experimental design for causal inference. Educ. Psychol., **51**(395–405), (2016). https://doi.org/10.1080/00461520.2016.1207177
20. Michalek, J.: Counterfactual impact evaluation of EU rural development programms – propensity score matching methodology applied to select EU member states, volume 2: a regional approach Report, RC Scientific and Policy Repots, EUR 25419 EN (2012). https://doi.org/10.2791/8228

Chapter 5
Asking Good Questions to Understand Voluntary Enrolments in Mathematics

Ning Li

Abstract It is of national concern that participation in higher-level mathematics subjects in senior secondary schools has been declining over the last few decades in Australia. As a gateway subject for tertiary studies in Science, Technology, Engineering, and Mathematics (STEM), the persistent decline can impact the country's economy in the long term. Understanding the causes of the decline can inform practice and help shed light on possible solutions. This paper describes the design and validation of a survey instrument for measuring factors that influence students' decisions to continue or discontinue studying mathematics beyond Year 10. Taking a social cognitive perspective, the instrument investigates motivation in subject selection under the assumption that what people think, believe and feel affects how they behave. An initial form of the instrument was developed and piloted to 564 Years 10 & 11 students. The responses were then used to analyze the reliability, factorial structure, and discrimination of the form. Psychometric evidences support the formation of a reduced form on scales of self-concept, self-efficacy, subjective value, anxiety and learning experience in mathematics. The refined form has reliable internal consistency and a clear structure.

Keywords Instrument · Survey item analysis · Self-efficacy in mathematics

5.1 Introduction

Mathematics is not compulsory in senior secondary schools in Australia. Students can opt out of mathematics in Years 11 and 12. When they decide to continue, they have the options to select between elementary, intermediate, and advanced level mathematics subjects. The elementary level subjects are not intended for providing a foundation for tertiary studies that involve mathematics. The higher-

N. Li (✉)
Australian Mathematical Sciences Institute, Melbourne, VIC, Australia
e-mail: Ning.li@amsi.org.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_5

level subjects, however, are often required for admission into university courses where mathematics is an integrated part of the discipline.

Each year between 2006 and 2016, about 90% boys and 80% girls in Year 12 studied some form of mathematics [1]. Although this rate appears to be high and stable, the majority enrollments were in elementary level only, which resulted in a major leak in the STEM pipeline. Over time, students have been shifting towards elementary mathematics. The enrolments at advanced-level mathematics have continued to decline over the last few decades [2–4]. In 2016, only 12% of the boys and 7% of the girls in Year 12 studied advanced mathematics in the country [1].

The diminishing interest in higher-level mathematics received widespread attention. Substantial funds from the government and industries were allocated to promote increases in participation. A considerable body of relevant research was conducted in Australia and in the world. In the USA, the expectancy-value theory has been developed and used to assert that women are less likely to pursue mathematically intensive careers due to lower expectancies and lower perceived values in mathematics as compared to men [5], whereas the mindset theory [6] suggests that students who endorse a fixed mindset on their mathematical abilities are more susceptible to reduced mathematics performance. In the UK, researchers [7] explicitly seek to answer the question that "Is the UK an outlier?" in mathematics participation, and review factors that encourage or discourage participate in post-16 mathematics [8]. In Australia, many studies explored possible reasons for the low participation [9, 10, 11, 12, 13, among others]. The "Maths? Why Not?" project [14] analyzes teachers' views on factors that influence students' decisions in the investigation of the institutional influences at school level. A recent study in Western Australia reports students' dissatisfaction with mathematics as one of the main reasons for the avoidance of higher-level mathematics [15].

Another large-scale 5-year project, started from 2016, is the CHOOSE**MATHS** partnership between the Australian Mathematical Sciences Institute (AMSI) and the BHP Foundation. The project exists as a reflection to the Chief Scientist's [16] call for action to encourage female STEM participation. CHOOSE**MATHS** aims to build mathematical capability and increase participation of school students, particularly, girls and young women across the STEM pipeline from classroom to industry. The project works with 120 schools across the country, provides professional development training, and works with students on intervention lessons. Research is one of the four components of CHOOSE**MATHS**, and the present study is part of the CHOOSE**MATHS** research. The focus of this research is to build an instrument for gathering direct measures on students' motivation in subject selection. The intended level of detail in the measurements is expected to enrich the broad conclusion of students' dissatisfaction reported by many previous studied in the literature. The lack of a ready-to-use instrument at the desired level of investigation reinforces the need for this work.

5.2 Motivation Model

What motivational factors are responsible for explaining the various behaviors in voluntary mathematics enrolments? In searching for a theoretic base to guide the design, we acknowledge the inherent value of using multiple theoretical orientations to explain the differences in behavior choices. By adopting a broad social cognitive approach, our development draws on the understanding that a mathematics enrolment choice of a student encompasses the student's self-perceived ability to perform mathematical tasks, the motivation to employ that ability [17], and the emotion related to mathematics study [18]. Without being able to generate a feeling of satisfaction in the interaction with mathematics, it seems unlikely that a student will choose mathematics. Without seeing the value of mathematics, the student would be equally unlikely to choose mathematics even though the student is highly capable. Hence, mathematical capacity, self-perceived competence beliefs, emotions associated with mathematics, and the value attached to mathematics each plays a key role in the decision making.

The mathematics enrolment choice motivation (MECM) instrument invented in this paper is based on a hierarchical model that integrates the self-efficacy theory [19] and the expectancy-value theory [5]. The central constructs in the model are (a) direct reasons to keep or drop mathematics, (b) self-concept in mathematics and self-efficacy in mathematics, (c) perceived value of mathematics, (d) anxiety associated with mathematics (maths anxiety); and (e) learning experience in mathematics. The hierarchy is formed to reflect sources of self-efficacy [18]. In the hierarchy, block (a) sits on the top and is based on blocks (b), (c) and (d) that are in turn supported by block (e) (Fig. 5.1).

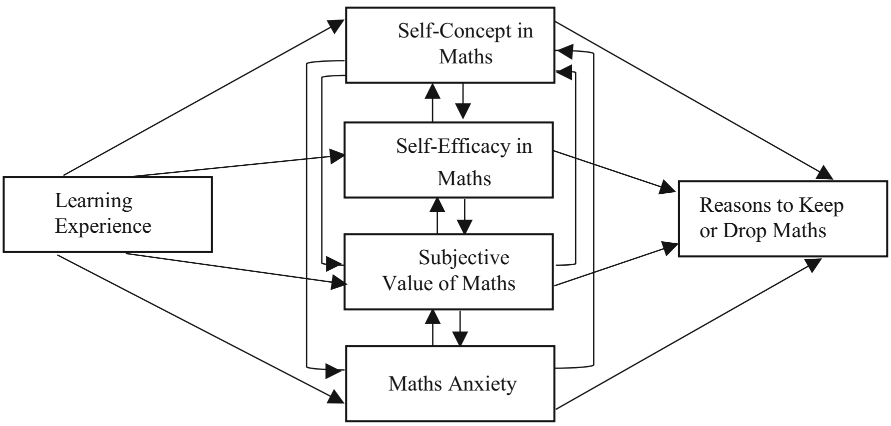


Fig. 5.1 A mathematics enrolment choice motivation (MECM) model

5.3 Model Components

5.3.1 *Self-Concept and Self-Efficacy in Mathematics*

Self-concept is one's perception of self [20] [21]. Self-concept in mathematics is one's judgement of self with regarding mathematical capability [22]. It is generally posited that self-concept is relatively stable and it influences the choice and direction of behaviour. Negative self-concept can undermine confidence, bias self-evaluation, and impair achievement.

Self-efficacy refers more specifically to one's perception on capability to produce the desirable outcome [23]. According to Bandura, efficacy expectation is developed as an operational mechanism in self-appraisal and self-regulation. Self-efficacy forms through mastery, vicarious, verbal persuasion, and emotive-based procedures. Mathematics self-efficacy is one's judgement on competence to perform mathematical tasks successfully [24]. It produces performance outcomes through cognitive, motivational and selection processes [25]. According to the social cognitive theory, given appropriate level of skills and adequate incentives, self-efficacy is a major determinant for the choices people make, effort to expend, and duration to sustain the effort. Thus, self-efficacy evaluation and expectations have been frequently used as a template for predicting the occurrence and persistence of behaviour.

The two self-constructs overlap with each other, with the boundary between them not always clear. Both require self-evaluation, but one has a focus on oneself as a person and the other has a focus on one's capability to perform a specific task [26].

5.3.2 *Subjective Task Values*

The cognitive self-evaluation in self-efficacy comprises performance-efficacy expectancy and performance-outcome expectancy [19]. The former is one's estimate that a given behavior will produce certain outcomes and the latter is the conviction that one can successfully execute the behavior needed to produce the outcome. Both expectancies are essential in motivational mechanism. Outcome-expectancy alone will not produce the desired performance if the component capabilities are missing. The perception about one's ability to perform a task alone will not motivate an action either, if one perceives no benefits or rewards.

Eccles *et al* [5] further consolidated the component performance-outcome expectancy in Bandura [19] and expanded it to formally introduced the term subjective task value (STV). In her theory, academic choices are linked to STV in addition to outcome expectation. Eccles's STV included four components: intrinsic value or enjoyment, utility value, attainment value, and cost or any negative aspects.

5.3.3 Maths Anxiety

Maths anxiety is the tendency to feel anxious when attempting to solve mathematical problems [27]. It is one of the extreme emotions developed as a reflection to the success or failure in mathematics-related behavior. Mathematics-related emotions in turn act as a feedback system to the motivational process [28]. On average, girls are found to experience a higher level of maths anxiety, which is believed to be associated with their lower academic achievement [5]. In the present study, maths anxiety is measured to reflect the extreme negative emotion formed through a long-term interaction with mathematics over a period, rather than as a temporary affective arousal.

5.3.4 From Constructs to Items

A question or a statement in a questionnaire is called a survey item or an item. Development of the initial pool of items aims to write down statements that tap on all aspects of a construct. To achieve content validity, it should be avoided that only a proportion of the construct was measured [29, 30]. The development of MCEM instrument largely relied on the previous research. In particular, the initial pool has been inspired by many existing instruments [31–40]. Ideas were learnt or criticised, sometimes without taking any item and sometimes items being modified. Most items were newly created to reflect the desired goal of investigation, by incorporating into the pool responses to the question 'Why did you study maths?' or 'Why didn't you study maths?' from a small number of Year 11 students. The initial pool was then undergone through face evaluation by experts.

Two sub-scales were designed for self-efficacy. One relates to the situations that were hypothesized to be necessary for initiating the decisions to enrol in mathematics. Selected situations reflect students' self-regulated behaviours in managing mathematics study. The other sub-scale relates to mathematics curriculum topics recommended by the Australian Curriculum, Assessment and Reporting Authority (ACARA) for Years 10 & 11. These questions were framed to measure the perceived competence, so that participants did not actually solve the problems. Following recommendations of Bandura [41], the items in both sub-scales were worded as *can* do.

The instrument consisted scales of self-concept, self-efficacy, maths anxiety, STV and learning experience. A 7-point Likert ruler was used for all the items, labelled as *1 = Strongly Disagree* to *7 = Strongly Agree*, or *1 = Not at all Confident* to *7 = Very Confident* as appropriate. For page limits, only some of the scales will be reported here.

5.4 Pilot Study and Validation

5.4.1 Participants

The initial form of the instrument was tested on 275 Year 10 and 289 Year 11 students in CHOOSE**MATHS** participating schools from New South Wales, Western Australia and Victoria between July and November in 2018. Prior to the pilot study, ethics clearance has been obtained from relevant state government and Catholic Education authorities. Due to the heavy study load for Year 11 students and a tight timeline, there was not much scope for sample selection. Participation in the survey was voluntary and anonymous. Of the participants, 39% are boys, 57% are girls, and 4% are others.

The survey was implemented online. Participants answered it independently in a group setting in school, with an average completion time of 15 min. Demographic information and specific subjects enrolled in for each student were also collected. All Years 10 and 11 students in the participating schools were invited to undertake the survey, irrespective of whether they selected to study mathematics or not. However, a large number of non-mathematics students did not undertake the survey, which resulted in an inflated mathematics participation rate in the sample (96%) and also made it unaffordable to test the questionnaire items designed for non-mathematics students.

5.4.2 Procedure

The content validity of the instrument has relied on the design stage that incorporated the mainstream motivation theories in the literature. The purpose of the following procedure is to test the internal consistency and factorial structure of the items for each scale, and to shorten the instrument form, ideally to within an average of 10 min.

A construct is usually multi-faceted. Consequently, multiple items are necessary in constructing a scale to measure the construct. The summative score that a respondent obtains from all the items of a scale constitutes the score of the scale, sometimes called the test score. Ideally, a scale designed to measure a trait has a minimum number of items that are sufficiently wide to cover all the aspects of the trait. These items should also be agreeable to measure the same designated construct, but meanwhile diverse enough to represent the various facets without too much overlapping. Such a status can be achieved through psychometric analysis. In the following, internal consistency (IC), factor analysis (FA), and item response theory (IRT) were used in combination to test and refine the instrument.

To assess the IC of a scale, apart from the Cronbach's α on all items of the scale, labelled as 'Overall' in Table 5.1, three other quantities were used for each item: 'Item-test' correlation, 'item-rest' correlation, and the Cronbach's α without the

Table 5.1 Inter-item correlations, the Cronbach's α, and discrimination coefficients of items for the scale of self-concept in mathematics

Item	Item Text	Item-Test Corr	Item-Rest Corr	α	Discr Coef	Final α	
`cmpoth`	Compared to other students in my class I am good at maths	0.815	0.770	0.918	2.872	0.857	*
`cmpslf`	Compared to my other school subjects I am good at maths	0.834	0.784	0.917	2.891	0.847	*
`maeasy`	Work in maths classes is easy for me	0.825	0.779	0.917	2.737		
`dowell`	I have always done well in maths	0.738	0.670	0.923	1.829	0.872	*
`hoples`	I'm hopeless when it comes to maths	0.863	0.823	0.915	−3.390	0.850	*
`confus`	I am always confused in my maths class	0.773	0.717	0.920	−2.328		
`nobrai`	I don't have the brain to learn maths	0.737	0.665	0.923	−2.178		
`2hard`	Maths is too hard for me	0.803	0.753	0.919	−2.748		
`expwel`	I expect to do very well in any maths class I take	0.643	0.554	0.929			
`lernhi`	I am confident that I can learn maths at a higher level	0.764	0.698	0.921	1.845	0.872	*
	Overall			0.928		0.885	

*: Indicate items that have been retained in the final scale

item ('item-delete' α). The item-test correlation assesses the consistency between an item and the scale, and the item-rest correlation assesses the consistency between an item and other items. They are displayed in the 3rd & fourth columns of Table 5.1 respectively. The item-delete α, in the next column of the table, quantifies the consistency of the scale when an item is excluded. The criteria of low 'item-test' and/or low 'item-rest' correlations and high 'item-delete' α were used to judge the 'badness' of an item. Following the common practice in the literature (see, for example, [42]), a Cronbach's α value between 0.8 and 0.9 was considered very good in the present study.

Exploratory FA was also used to examine the scale structure and linear combinations of the items that contain most information. The eigenvalues, the uniqueness of each item (i.e., the proportion of variance for the item that is not associated with the common factor, 1- communality), and AIC and/or BIC were used to filter items for exclusion. Also used were the discrimination coefficient of each item estimated from the graded response model (GRM) [43] and the associated test characteristic curve (TCC) in the investigation of how sensitive an item or a set of items as a whole were in differentiating the varying level of the construct under study.

5.4.3 Results

Self-Concept in Mathematics The item `expwel` 'I expect to do very well in any maths class I take', in Column 2 of Table 5.1, was firstly removed due to high item-delete α and low item-test correlation. Then the least informative items, as estimated from the GRM, in each cluster of highly correlated items were removed to achieve a monotonical TCC. The retained items, asterisked in the final column of Table 5.1, had $\alpha = 0.885$ with two principal components: an overall self-evaluation of mathematical capability, captured by `cmpoth cmpslf hoples`, and a self-appraisal and self-inference along the time dimension, captured by `dowell lernhi` – the belief of having done well in the past and being able to do well in the future.

Self-Efficacy in Mathematics This scale consisted of two sub-scales: situations-specific self-efficacy and topic-specific self-efficacy. The former refers to situations related to mathematics learning and latter refers to mathematics curriculum topics expected of Year 10 & 11 students.

For the situation-specific scale, correlation analysis led to immediate removal of one item in the initial pool. The high α value, 0.934, indicated redundancy in the remaining items. After several iterations, the less discriminant item within each group of highly correlated items or the items that had overwhelmingly agreeable responses were removed. The 7 items retained, listed in Table 5.2, had the α = 0.874, minimum discrimination 1.59, and a monotonic TCC. FA revealed that the scale consisted of 2 principal components. The first principal component, measuring the degree of persistence in mathematics learning, accounted for 62% of the total

Table 5.2 Retained items for self-efficacy in mathematics

Item	Item Text
Situation-specific (α = 0.874)	
`smotiv`	Motivate yourself to do school work in maths
`sdead`	Finish your maths assignments by the deadline
`sclari`	Clarify doubts in maths in class
`smarks`	Obtain the marks you want if you try hard
`scontin`	Continue working even if have trouble learning the material
`sothint`	Study maths when there are other interesting things to do
`steach`	Live up to what your teachers expect of you in maths
Topic-specific (α = 0.907)	
`ttiles`	Calculate the number of square meters of tiles needed to cover a floor
`tinves`	Calculate the final value of an investment using compound interest
`tpytha`	Find the length of an unknown side using Pythagoras
`ttrigo`	Using Trigonometry to find the size of an unknown angle of a triangle
`tparab`	Determining the x-intercepts and turning points of parabolas
`tsimpl`	Simplify expressions that have surds in them
`troot`	Determine if the square root of a given number is rational

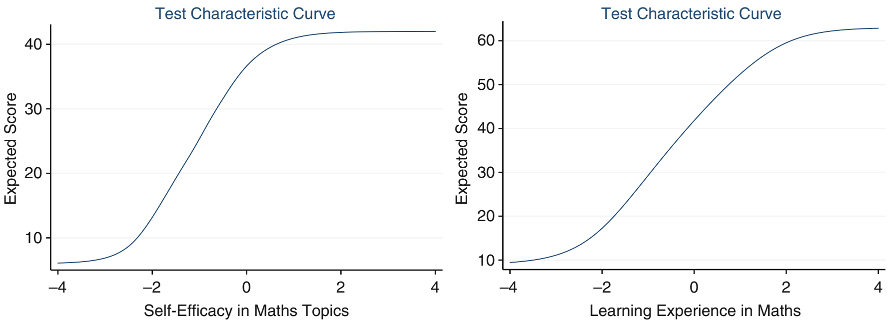


Fig. 5.2 (*Left*) TCC of self-efficacy in mathematics topics after excluding items `logri`. (*Right*) TCC of learning experience in mathematics from the 9 items retained

variability. The second principal component measured the confidence in managing study, obtaining desired marks and living up to expectations.

For the topic-specific self-efficacy sub-scale, one item was removed because it had the lowest 'item-rest' correlation and lowest discrimination corresponding to a nearly flat characteristic curve over the entire latent trait continuum. The remaining items had the Cronbach's $\alpha = 0.907$ and a monotonical TCC. However, the TCC, displayed on the left graph of Fig. 5.2, was flat beyond 2 in the latent trait continuum. It indicated that the sub-scale was unable to adequately distinguish the difference between the high response categories. Thus, two 'harder' new items: 'Solve a pair of simultaneous equations using the elimination or substitution method' and 'Calculate and interpret the probabilities of various events associated with a given probability distribution, by hand in simple cases' have been added to the sub-scale, to make the sub-scale to be more sensitive to high self-efficacy in mathematics.

Learning Experience The learning experience scale initially consisted of 25 items. They were designed to capture a range of mathematics learning experiences perceived to be important in shaping student motivation in learning mathematics and subject selection. The graded response model was fit to all the items using data from the pilot study. The information function for each item was used to compare the discrimination powers. After several iterations of discarding the least informative items within each highly correlated cluster, 9 items were retained, listed in Table 5.3.

The retained items were highly consistent with $\alpha = 0.879$, minimum discrimination 1.4 for each item, and excellent TCC (as shown in the right-hand side of Fig. 5.2). These items accounted for 91.5% of the variation in the original 25-item score. FA revealed that the items loaded onto 3 principal components. The first main component of students' experiences in learning mathematics reflected the teacher practice and teaching quality. The second was students' own achievements in problem solving, helping peers, and doing well in mathematics. The last principal component represented students' experiences on received recognition and expectation from teachers and parents. The data showed that the successful or

Table 5.3 Retained items for learning experience in mathematics

Item	Item Text
Learning experience ($\alpha = 0.879$)	
`solve`	I solved problems that most of my classmates did not
`friend`	I helped my friends with maths class work or assignments
`didwell`	I did very well in maths
`feedbk`	Teachers gave me helpful feedback on my maths work
`relate`	Teachers showed how maths is related to everyday life
`hands`	There were many hands-on activities to help me learn maths
`tprais`	My teachers praised me for doing well in maths
`pprais`	My parents/caregivers praised me for doing well in maths
`texpec`	My teacher(s) expected me to continue studying maths into year 12

Table 5.4 Initial and retained (marked with asterisks*)* items for enjoyment in mathematics

Item	Item Text		α	α'
`satisf`	I get a great deal of satisfaction out of solving maths problem		0.882	
`likem`	I like maths	*	0.873	0.782
`answer`	Enjoy answering questions in maths class		0.878	
`playm`	I enjoy playing maths activities		0.882	
`perfect`	I continue working on maths tasks until everything is perfect	*	0.884	0.818
`pltake`	I plan to take as much maths as I can during my education	*	0.881	0.761
`elemen`	I enjoy learning elementary level maths		0.905	
`advan`	I enjoy learning advanced level maths such as Specialist Maths	*	0.888	0.834
`nohard`	I don't enjoy solving hard maths problems		0.896	
`glad`	I am glad I don't have to do maths any more		0.89	
`rather`	I would rather to do a maths assignment than to write an essay		0.886	
`dull`	Maths is dull and boring		0.888	
Scale			0.895	0.842

The α and α' denote the internal consistency for the initial and the retained items respectively

unsuccessful experiences in these aspects appear to be the most sensitive measures to capture differences that shape motivation in mathematics learning.

Enjoyment of Mathematics Correlation analysis of the 12 items on intrinsic interest in mathematics, listed in Table 5.4, led to the removal of `elemen nohard`. The resulting α value, 0.908, indicated redundancy in the remaining items and also suggested possible removal of some items. After a few iterations of removing the least informative items within each cluster of highly correlated items, 4 items, asterisked in Table 5.4, were retained at the end. The internal consistency of the retained items, 0.842, is satisfactorily high and the corresponding TCC was monotonic. The refined scale score correlated highly with the original 12-item score (correlation coefficient was 0.917), indicating that the retained items represented the initial items well.

Utility Value of Mathematics Nine items were initially proposed for the utility value of mathematics (Table 5.5). With an internal consistency 0.835 and a monotonically

Table 5.5 Retained items for utility value, attainment value, and cost of mathematics

Item	Item Text
Utility value ($\alpha = 0.835$)	
`best`	Maths is one of my best subjects
`think`	Studying maths helps develop problem-solving skills and critical thinking
`4uni`	I need to do Year 12 maths to get into the university course of my choice
`avoid`	I want to avoid studying maths at university
`nocaree`	I don't see any careers I can do with maths
`matjob`	I want to find a job that involves the use of maths
`ATAR`	Doing maths is not a good strategy for my ATAR
`societ`	Maths is not as helpful to the society as other school subjects
`enough`	Year 10 maths should be enough for my future life

Table 5.6 Initial and retained (asterisked) items for attainment value and cost of mathematics

Item	Item Text	
Attainment value ($\alpha = 0.820$)		
`dowell`	Doing well in maths is very important for me	*
`goodst`	A good student should learn more maths	*
`parent`	My parents think maths is very useful	*
`nobenef`	I don't see any benefit of doing advanced level maths	
Cost ($\alpha = 0.883$)		
`hardA`	It is harder to get Grade A in maths than in other school subjects	*
`harder`	Learning maths is harder than learning other school subjects	*
`another`	I can get a much better score in another subject than in maths with the same effort	

increasing TCC, all the items for the utility value were retained. The items loaded onto two principal components, explaining 59% and 47% of the total variation respectively. The first main component, composing all the negatively worded items, reflected the considerations from the perspective of future life, personal and societal benefits that students attached to the value of mathematics. The second component, composing the remaining 4 items, reflected the considerations from personal development perspective.

Attainment Value of Mathematics Four items were initially proposed for the attainment value of mathematics, given in Table 5.6. After removing the least informative item, corresponding to the lowest curve in Fig. 5.3 ('I don't see any benefit of doing advanced level maths'), the remaining items had satisfactorily high internal consistency $\alpha = 0.820$ and monotonic TCC as displayed on the left graph of Fig. 5.4.

Cost of Mathematics Learning The cost reflects negative aspects on studying mathematics. It may include performance anxiety, fear of failure or of success, effort to succeed, and lost opportunities. Focusing on perceived *relative* disadvantages, this scale in our instrument had 3 items initially (given in Table 5.6). Correlation analysis of the 3 items led to the removal of `another`. The remaining 2 items, with

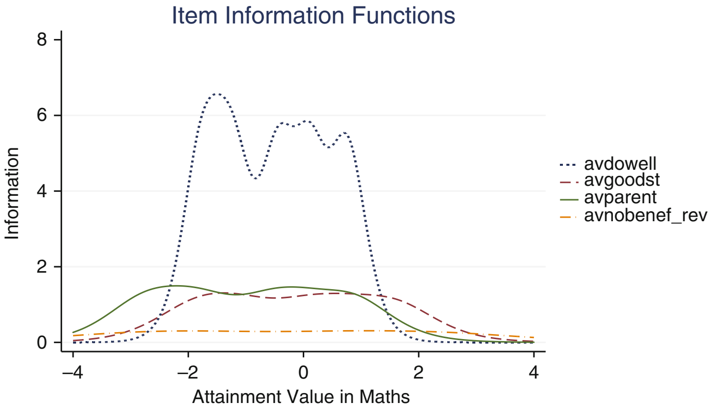


Fig. 5.3 Item information function for the 4 items of attainment value of mathematics

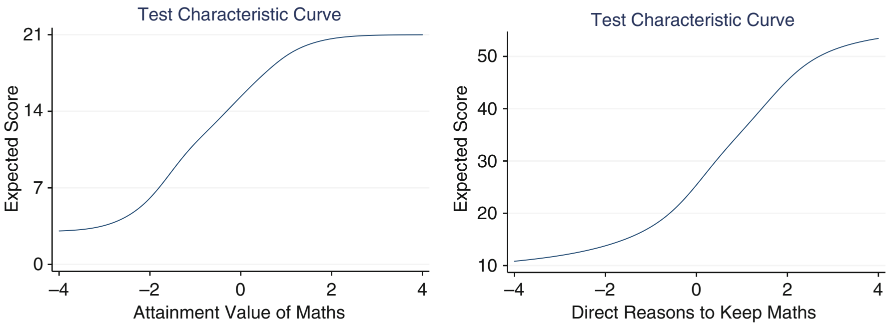


Fig. 5.4 (*Left*) TCC for the 3 retained items of attainment value in mathematics, and (*Right*) TCC for the items of direct reasons to keep mathematics

Table 5.7 Initial and retained (marked with asterisks*)* items for maths anxiety

Item	Item Text	
`homewk`	I get very nervous doing maths homework by myself	*
`notime`	I feel nervous when I don't have enough time to solve maths problems in class	*
`called`	I worry that I will be called to answer questions in maths class	*
`prepar`	I worry that I would not have enough time to prepare for maths tests	*
`mark`	I worry that I would get poor marks in maths tests	
`ability`	I feel uneasy because it seems I have no maths ability	*
`whylrn`	I feel uneasy because I do not know why I learn maths	*

an internal consistency 0.883, measure the perceived cost of learning mathematics as compared with the effort and outcome of learning other school subjects.

Maths Anxiety Listed in Table 5.7 are the 7 items initially proposed for measuring maths anxiety. Items `prepar` and `mark` were highly correlated, and `mark` was

Table 5.8 Items of direct reasons for keeping mathematics in Year 11

Item	Item Text
Reasons for Keeping Maths ($\alpha = 0.771$)	
`friend`	I want to go with the same option in course selection as my friends
`requir`	I need one more subject to make up the total number of subjects required by
`challe`	I chose maths to challenge myself
`4parent`	I chose maths to please my parents/caregivers
`random`	I just chose maths at random without much thinking
`popula`	I selected maths because it is a popular choice
`teachr`	I like my maths teacher(s)
`sister`	My sisters or brothers did very well in maths and I think I can do well too

less informative, hence was removed. After excluding `mark`, the remaining items had a high internal reliability $\alpha = 0.871$, with a monotonically increasing TCC that was displayed on the right-hand side of Fig. 5.2. The remaining items loaded on two principal components, with 3 items for each: (a) the anxiety associated with conducting mathematics tests, solving mathematics problems and answering mathematics problems in class; and (b) the anxious feeling associated with less confronting mathematical situations.

Direct Reasons for Keeping Mathematics Listed in Table 5.8 were 8 items for extra reasons to keep or drop mathematics in the initial form. They were proposed as a supplementary to other items in the instrument. The internal consistency, as measured by the Cronbach's α, was 0.771 in this case. One way to increase the relatively low α value is by dropping some less correlated items. However, the reduced scale exhibited a flat TCC over the latent continuum $(-4, -2)$. This indicated that more relevant items are needed for a better TCC and a higher internal consistency. However, because these items, by design, are to be used together with other scales of the instrument, we are content with this α value in this case. Hence, all the initial items were retained.

Direct Reasons for Dropping Mathematics Listed in Table 5.9 were the 10 items proposed under the sub-title of direct reasons for discontinuing with the study of mathematics beyond the compulsory period. Similar to the rational for setting up the items for direct reasons for keeping mathematics, these items were proposed as a supplement to the self-concept and self-efficacy in mathematics, mathematics learning experience, subjective value of mathematics, and maths anxiety. Notice that the items of the reasons for dropping mathematics were designed for non-mathematics students. Since there were not many non-mathematics students participated in the pilot study, the internal consistency on this group of the items was only calculated from 18 respondents. Based on this small sample, these items were highly consistent ($\alpha = 0.867$). However, the graded response model could not be fitted on such a small sample. Hence, the discrimination properties could not be examined for this scale in this case.

Table 5.9 Items of direct reasons for dropping mathematics in Year 11

Item	Item Text
Reasons for dropping maths ($\alpha = 0.867$)	
`badpre`	I didn't do very well in my previous maths class and I think it will get harder
`othsubj`	Other subjects were easier to me
`exam`	I wasn't sure that I would pass the exams
`contrl`	I do not feel in control of my maths study outcome if I chose maths
`clash`	There is or will be a time clash between maths class and my favourite subject
`little`	I plan to work in an area that requires little maths
`recomm`	My teacher didn't recommend it for me
`friend`	My good friends seem don't think I can do Year 11 maths well
`encrag`	I didn't get much encouragement from my parents/caregivers/teachers
`course`	The course I want to study at university doesn't require a maths more than Year 10

5.5 Remarks

A new instrument for measuring motivation in voluntary mathematics enrolments is developed, under the assumption that perceptions of success capability, value of mathematics, and feelings associated with mathematics are crucial to the choice behaviour. The preliminary evidences show that the instrument is reliable, discriminant, and has clear structure. The estimated completion time for the target population is 7 min without collecting demographic and subject details, and 10 min with the extras.

This development involved an application of motivation theories for the design, and an application of statistical methods for evaluation. During the analysis, a certain degree of subjectivity is inevitable. Sometimes, an exclusion of a different item may lead to a scale with equally good properties. The validation involved depends on the specific data set. Psychometric properties should be re-examined in future data even though the primary purpose of a study is not on the scale properties.

References

1. Li, N., Koch, I.: Gender report: participation, performance and attitudes towards mathematics. Australian Mathematical Sciences Institute, Melbourne (2017)
2. Forgasz, H.: Australian year 12 mathematics enrolments: patterns and trends – past and present. Australian Mathematical Sciences Institute, Melbourne
3. Kennedy, J., Lyons, T., Quinn, F.: The continuing decline of science and mathematics enrolments in Australian high schools. Teach. Sci. **60**(2), 34–46 (2014)
4. Barrington, F., & Evans, M. (2017). Year 12 mathematics participation in Australia 2007–2016. https://amsi.org.au/publications/year-12-mathematics-participation-australia-2007-2016/
5. Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., Midgley, C.: Expectancies, values, and academic behaviors. In: Spence, J.T. (ed.) Achievement and Achievement Motivation, pp. 75–146. W. H. Freeman, San Francisco (1983)

6. Dweck, C.S.: The development of ability conceptions. In: Wigfield, A., Eccles, J.S. (eds.) Development of achievement motivation. A volume in the educational psychology series, vol. xvii, pp. 57–88. Academic, San Diego (2002)
7. Hodgen, J., Pepper, D., Sturman, L., Ruddock, G.: Is the U.K. An outlier? An international comparison of upper secondary mathematics education. Nuffield Foundation, London (2010)
8. Smith, A.: Report of Professor Sir Adrian Smith's review of post-16 Mathematics. Department for Education, London (2017). Retrieved from https://www.gov.uk/government/publications/smith-review-of-post-16-maths-report-and-government-response
9. Watt, H.M.G.: Explaining gendered math enrollments for NSW Australian secondary school students. New Dir. Child Adolesc. Dev. **2005**(110), 15–29 (2005)
10. Nicholas, J., Poladian, L., Mack, J., Wilson, R.: Mathematics preparation for university: entry, pathways and impact on performance in first-year science and mathematics subjects. Int. J. Innov. Sci. Math. Educ. **23**(1), 37–51 (2015)
11. Maltas, D., Prescott, A.: Calculus-based mathematics: an Australian endangered species? Aust Senior Math. J. **28**(2), 39–49 (2014)
12. Mathematical Association of New South Wales (2014). Report on the MANSW 2013secondary mathematics teacher survey. https://www.mansw.nsw.edu.au/documents/item/70
13. Carmichael, C., Callingham, R., Watt, H.M.G.: Classroom motivational environment influences on emotional and cognitive dimensions of student interest in mathematics. ZDM. **49**(3), 449–460 (2017)
14. McPhan, G., Morony, W., Pegg, J., Cooksey, R., Lynch, T.: Maths? why Not? Department of Education, Employment and Workplace Relations, Canberra (2008). https://makeitcount.aamt.edu.au/content/download/33194/469618/version/1/file/MaWhNot_Published.pdf
15. Hine, G.: Reasons why I didn't enrol in a higher-level mathematics course: listening to the voice of Australian senior secondary students. Res. Math. Educ. (2019). https://doi.org/10.1080/14794802.2019.1599998
16. Office of the Chief Scientist: Mathematics, engineering and science in the national interest. In: Austr. Gov. (2012)
17. Lubinski, D., Benbow, C.P.: Study of mathematically precocious youth after 35 years: uncovering antecedents for the development of math-science expertise. Perspect. Psychol. Sci. **1**, 316–345 (2006)
18. Goldin, G.A., Hannula, M.S., Heyd-Metzuyanim, E., Jansen, A., Kaasila, R., Lutovac, S., Di Martino, P., Morselli, F., Middleton, J.A., Pantziara, M., Zhang, Q.: Attitudes, beliefs, motivation and identity in mathematics education: an overview of the field and future directions. Springer Open (2016)
19. Bandura, A.: Self-efficacy: toward a unifying theory of behavioral change. Psychol. Rev. **84**, 191–215 (1977)
20. Shavelson, R.J., Hubner, J.J., Stanton, G.C.: Self-concept: validation of construct interpretations. Rev. Educ. Res. **46**, 407–441 (1976)
21. Marsh, H.W., Shavelson, R.: Self-concept: its multifaceted, hierarchical structure. Educ. Psychol. **20**(3), 107–123 (1985)
22. Marsh, H.W.: The structure of academic self-concept: the Marsh/Shavelson model. J. Educ. Psychol. **82**, 623–636 (1990)
23. Bandura, A.: Self-efficacy: the exercise of control. W.H. Freeman, New York (1997)
24. Pajares, F.: Self-efficacy beliefs and mathematical problem-solving of gifted students. Contemp. Educ. Psychol. **21**, 325–344 (1996)
25. Pajares, F., Graham, L.: Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. Contemp. Educ. Psychol. **24**, 124–139 (1999)
26. Bong, M., Skaalvik, E.: Academic self-concept and self-efficacy: how different are they really? Educ. Psychol. Rev. **15**, 1–40 (2003)
27. Richardson, F.C., Suinn, R.M.: The mathematics anxiety rating scale: psychometric data. J. Couns. Psychol. **19**, 551–554 (1972)
28. Hannula, M.S.: Exploring new dimensions of mathematics-related affect: embodied and social theories. Res. Math. Educ. **14**, 137–161 (2012)

29. Clark, L.A., Watson, D.: Constructing validity: basic issues in objective scale development. Psychol. Assess. **7**, 309–319 (1995)
30. Briggs, S.R., Cheek, J.M.: The role of factor analysis in the evaluation of personality scales. J. Pers. **54**, 106–148 (1986)
31. Plake, B.S., Parker, C.S.: The development and validation of a revised version of the mathematics anxiety rating scale. Educ. Psychol. Meas. **42**, 551–557 (1982)
32. Marsh, H.W., O'Neill, R.: Self description questionnaire III: the construct validity of multidimensional self-concept ratings by late adolescents. J. Educ. Meas. **21**, 153–174 (1984)
33. Byrne, B.M.: Measuring self-concept across the life span: issues and instrumentation. American Psychological Association, New York (1996)
34. Martin, A.J.: The student motivation scale: a tool for measuring and enhancing motivation. Aust. J. Guid. Couns. **11**, 1–20 (2001)
35. Suinn, R.M., Winston, E.H.: The mathematics anxiety rating scale, a brief version: psychometric data. Psychol. Rep. **92**, 167–173 (2003)
36. Marat, D.: Assessing mathematics self-efficacy of diverse students from secondary schools in Auckland: implications for academic achievement. Issues Educ. Res. **15**, 37–68 (2005)
37. Stevens, T., Olivárez Jr., A.: Development and evaluation of the mathematics interest inventory. Meas. Eval. Couns. Dev. **38**, 141–152 (2005)
38. Luttrell, V.R., Callen, B.W., Allen, C.S., Wood, M.D., Deeds, D.G., Richard, D.C.S.: The mathematics value inventory for general education students: development and initial validation. Educ. Psychol. Meas. **70**, 142–160 (2010)
39. Ko, H.K., Yi, H.S.: Development and validation of a mathematics anxiety scale for students. Asia Pac. Educ. Rev. **12**, 509–521 (2011)
40. Butler, K.L.: Motivation for Mathematics: the development and initial validation of an abbreviated instrument. PhD Thesis, University of South Florida (2016)
41. Bandura, A.: Guide for constructing self-efficacy scales. In: Pajares, F., Urdan, T. (eds.) Self-efficacy beliefs of adolescents, vol. 5, pp. 307–337. Information Age Publishing, Greenwich (2006)
42. Streiner, D.: Starting at the beginning: an introduction to coefficient alpha and internal consistency. J. Pers. Assess. **80**, 99–103 (2003)
43. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr. Suppl. **34**(4, Pt. 2), 100 (1969). Supplement, no. 17

Part II
Agricultural Statistics and Policy Analysis

Chapter 6
Modeling for Prospect of Aman Rice Production in Dhaka Division, Bangladesh

Sayed Mohibul Hossen, Md. Takrib Hossain, Aditi Chakraborty, and Mohd Tahir Ismail

Abstract Rice is the dominant food crop of Bangladesh, about 75% of agricultural land is use for production and it contributes 28% of GDP. Aman is one of the main harvest crops and second largest rice crop in the country in respect to the volume of production. The main purpose of this study is to identify the Auto-Regressive Integrated Moving Average (ARIMA) model by Box-Jenkin's approach that could be used to forecast the production of Aman Rice in Dhaka Division Bangladesh. The Secondary data were collected for the year 1972–1973 to 2014–2015 from the Bangladesh Agricultural Research Council (BARC) for the purpose of model identification and forecast up-to the year 2035 of the identified model. Data sets are checked for whether it is stationary or not through graphical method, correlogram and unit root test. Thus Box-Jenkins approach is applied for determination of ARIMA model. The best selected Box-Jenkin's ARIMA model for forecasting the production of Aman Rice is ARIMA (1,1,1). For residual diagnostics correlogram Q-statistic and histogram and normality test were used. The comparison between the original series and forecasted series shows the same manner which indicates the fitted model behaved statistically well and suitable to forecast the production of Aman Rice in Dhaka Division Bangladesh. We have found that the annual production of Aman Rice in Dhaka Division Bangladesh is slightly increasing.

Keywords Aman · Correlogram · Unit root test · ARIMA · Forecasting

S. M. Hossen (✉)
School of Mathematical Science, Universiti Sains Malaysia, George Town, Penang, Malaysia

Department of Statistics, Faculty of Science, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh

Md. T. Hossain · A. Chakraborty
Department of Statistics, Faculty of Science, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh

Md. T. Ismail
School of Mathematical Science, Universiti Sains Malaysia, George Town, Penang, Malaysia

© Springer Nature Singapore Pte Ltd. 2020
A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_6

6.1 Introduction

Bangladesh is the agriculture based developing country and still striving hard for rapid development of its economy. Rice is the staple food of Bangladesh. Bangladesh is not only a rice growing country but also a country of rice eating people [1]. Rice plays the leading role by contributing 91% of total food grain production of all crops. More than 99% of people eat rice as their main food and rice alone provides 76% of calorie and 66% of total protein requirement of daily food intake [2]. The economic development of Bangladesh is fundamentally based on agriculture. The total cultivation area in Bangladesh is about 8.52 million hectares and net cultivated area is 7.45 million hectares and 0.47 million hectares cultivates area are unplanted. Although 78% of total cropped area is devoted to rice production, the country is still suffering from a chronic shortage of food grain [3]. In Bangladesh, the major rice crops namely, Aus, Aman and Boro covering almost 11.0 million hectares of land in Bangladesh which constitute nearly 95% of total food requirements [4]. Aman is one of the main crops in Bangladesh. It is the second largest rice crop in the country in respect to the volume of production while Boro tops the production. The production of Aman depends on the weather condition of the country and farmers usually cultivate Aman in their land. In the year 2016, favorable weather condition prevailed all over the country from sowing to harvesting period of Aman [5]. In Bangladesh, the largest harvest in Aman, occurring in November and December, which occurs more than half of annual production. Among these, transplanted cropping Aman is most important and occupied about 46% of the rice cultivated land in 2009–2010 and 4% of land is occupied by sown Aman respectively. Transplanted Aman is grown throughout Bangladesh and sown, or broadcast Aman is grown mostly in the south and southeastern part of the country.

6.1.1 Objective of the Study

(i) To investigate the pattern of Aman rice production in Bangladesh.
(ii) To find an appropriate model for predicting Aman rice production of Bangladesh.
(iii) To forecast (2015–2035) the production of Aman rice in Bangladesh.

6.1.2 Literature Review

Numerous studies have been made by the researcher to fit an ARIMA model in the agriculture sector in all over the world for different agricultural crops. ARIMA model is used in different agriculture sector to forecast agricultural productions. The relevant work for forecasting by using Box-Jenkins (1970) ARMA model [6], from which we get the idea about forecasting techniques for different types of agricultural productions forecasting such as Goodwin and Ker (1998) added new dimensions to the evolution of this literature [7]. They introduced a Univariate filtering model, an ARIMA (0,1,2) to best represent crop yield series.

Awal Siddique (2011) have worked on the rice production in Bangladesh employing by ARIMA model [8]. Amin et al. (2014) have worked on the time series modeling for forecasting wheat production of Pakistan [9]. They suggest some time series models to forecast the wheat production of Pakistan. Maniha Zakia (2014) has worked on a time series modeling on GDP of Pakistan. ARIMA (1,1,0) has been obtained through the expert modeler by considering best fit model [10]. Maity and Chatterjee (2012) have worked on the forecasting GDP growth rate of India. A very simple tentative ARIMA (1,2,2) model has been fitted on data to estimate the parameters [11]. Yaziz et al. (2011) have worked on a comparative study on Box-Jenkins and Garch models in forecasting crude oil price [12]. Rachana et al. (2010) used of the ARIMA models to forecast pigeon pea production in India [13]. Badmus and Ariyo ARIMA (1,1,1) and ARIMA (2,1,2) for cultivation area and production respectively. Rahman (2010) fitted an ARIMA model for forecasting Boro rice production in Bangladesh [14]. Nasiru Suleman and Solomon Sarpong (2011) have worked on the Milled Rice Production in Ghana Using ARIMA (2, 1, 0) model [15].

Our study has been investigated the behavioral pattern of Aman Rice Production in Dhaka Division, Bangladesh by reviewing the above literatures.

6.2 The Methodology and Model

To test the stationarity of the time series data we use some test such as Graphical analysis, Correlogram and Unit root test. The most frequently used method for the test of a unit root in a parametric framework is the Dickey-Fuller (DF) test [16]. Dickey-Fuller (DF) and Augmented Dickey-Fuller (ADF) test is widely used to check the stationarity. But if the series have no trend and the error terms are auto correlated, then we cannot apply DF test because of time series data have no trend. That’s why we use ADF test to test the stationarity.

If we plot ρ_k against lag, the graph we obtain is known as the population Correlogram. In practical situation we can only compute the Sample Autocorrelation Function (SACF), $\hat{\rho}_K$.To compute this we must first compute the sample covariance at lag k, $\hat{\gamma}_k$ and the sample variance $\hat{\gamma}_0$, which are defined as-.

$$\hat{\gamma}_k = \sum \{(y_t - \overline{y})\,(y_{t+k} - \overline{y})\}\,/n \tag{6.1}$$

$$\hat{\gamma}_0 = \sum (y_t - \overline{y})^2 / n \tag{6.2}$$

Here n is the sample size and $\overline{y}$ is the sample mean. Therefore, the SACF at lag k is-

$$\hat{\rho}_K = \frac{\hat{\gamma}_K}{\gamma_0} \tag{6.3}$$

This is simply the ratio of sample covariance (at lag k) to sample variance. If we plot $\hat{\rho}_K$ against the lag then we get sample correlogram. For any stationary time series the auto correlation of various lags remains around zero. Otherwise the series is non-stationary [17].

The role of ACF is very important at the identification of time series process. It measures the direction and strength of statistical relationship between ordered pairs of observation in a single data series. The ACF at lag k, denoted by ρ_k, defined as -

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{Covariance of k}}{\text{Variance}} \tag{6.4}$$

where k = 0 then ρ =1, because they are measure with the same units of measurement.

ρ_k is unit less, it lies between -1 to $+1$ as any correlation coefficient does. The exponentially declining natures of the ACF plots also observed for taking final decision about the order of auto regression. The number of significant spikes in ACF plots also in taking decision about the degree of moving average [17].

The estimated PACF is broadly like an estimated ACF. An estimated PACF is also a graphical representation of the statistical relationship between sets of ordered pairs drawn from a single time series.

In the Autoregressive (AR) process of order p the current observation y_t is generated by a weighted average of past observations going back p periods, together with a random disturbance in the current period. We denote this process as AR (p) and write the equation as,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{6.5}$$

In the Moving Average (MA) process of order q each observation y_t is generated by a weighted average of random disturbance going back to q periods. We denote this process as MA (q) and write the equation as,

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{6.6}$$

A very popular process in econometric time series is the Autoregressive Integrated Moving Average (ARIMA) process. Autoregressive schemes with moving average error terms of the form are denoted by-.

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{6.7}$$

This process is called Autoregressive Moving Average (ARMA) process of order (p, q) or briefly ARMA (p, q) [18]. Therefore, if we take difference a time series d times and then apply the ARMA(p, q) model to it, then the time series model is ARIMA(p, d, q) where, p is number of autoregressive terms, d the number of times the series has to difference and q the number of moving average terms. The ARIMA(1, 1, 1) process can be written as,

$$\Delta y_t = \phi \Delta y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \tag{6.8}$$

Where, $\Delta y_t = y_t - y_{t-1}$ and $\Delta y_{t-1} = y_{t-1} - y_{t-2}$ are the first differences of y_t. Therefore, ARIMA(p, d, q) is,

$$\Delta^d y_t = \phi_1 \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{6.9}$$

Where, Δ^d indicates the dth difference of y_t.

The forecasting procedure based on ARIMA models, which is usually known as the Box-Jenkins approach [6]. A major contribution of Box-Jenkins has been to provide a general strategy for time-series forecasting, which emphasizes the importance of identifying an appropriate model in an iterative way.

The procedure for Box-Jenkins methods involve four general steps namely; model identification, model estimation, diagnostic checking and use of the fitted model to forecast future values. The first three steps are repeated until an adequate and satisfactory model is formed.

The Mean Square Error (MSE) is another method for evaluating a forecasting technique. This approach penalizes large forecasting errors because the errors are squared, which is important; a technique that produces moderate errors may well be preferable to one that usually has small errors but occasionally yields extremely large ones.

The Mean Absolute Percentage Error (MAPE) provides an indication of how large the forecast errors are in comparison to the actual values of the series. The technique is especially useful when the values are large. MAPE can also be used to compare the accuracy of the same or different techniques on two entirely different series The Jarque-Bera test of normality is an asymptotic or large sample test. It is also based on the OLS residuals. This test first computes the Skewness and Kurtosis measures and uses the following test statistic:

$$JB = n\left[\frac{s^2}{6} + \frac{(k-3)^2}{24}\right] \tag{6.10}$$

Where n = sample size, s = skewness and k = kurtosis. For a normally distributed variables s = 0, k = 3.Therefore, the JB test of normality of the joint hypothesis that s and k are 0 and 3 respectively. In this case the value of the test statistics is expected to be zero. Under the null hypothesis the residuals are normally distributed. Jarque and Bera (1987) showed that this statistic asymptotically the χ^2 distribution with 2 d.f [19].

6.3 Results and Discussion

For any time series analysis, the most important aspect to check is whether the data is stationary or not. If the data is not stationary our main aspect is to change the data set into stationary. Now at first, we do the line graph for annual Aman rice production of Dhaka division in Bangladesh, we observe that the annual Aman rice production has slightly upward trend (Fig. 6.1). To test the stationarity of the data, we use Correlogram with lag length 20 as follows (Table 6.1).

From the ACF and PACF (Table 6.1) we observe that most of the spikes are significant at different lags. Now we will do the unit root test to check the data is stationary or not. The unit root test for annual Aman rice production of Dhaka division given below (Table 6.2):

From the below Table 6.2, we use ADF test and SIC for model selection. Here the null hypothesis states that annual production of Aman rice in Dhaka division has a unit root. This means that the data is not stationary.

If we take an absolute value and ignore the sign, we see the value of t-statistic is smaller than all other test critical values at 1% level, 5% level and 10% level of significance. So, we accept the null hypothesis. Therefore, the data is not stationary. Now we will take first difference $[\nabla y_t = y_t - y_{t-1}]$ to the non-stationary data and to move it to the stationary. After taking first difference to the non-stationary data we will get stationary data (Table 6.3).

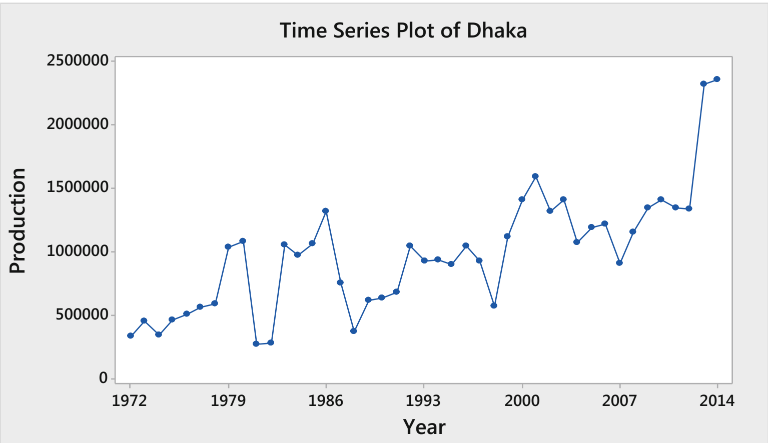
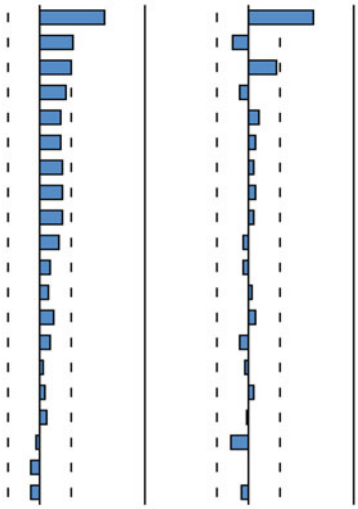


Fig. 6.1 Line graph for annual Aman rice production of Dhaka division

Table 6.1 Correlogram for annual Aman rice production of Dhaka division

Date: 10/07/18 Time: 12:37
Sample: 1972 2014
Included observations: 43



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.649	0.649	19.435	0.000
		2	0.335	–0.150	24.734	0.000
		3	0.310	0.278	29.389	0.000
		4	0.255	–0.097	32.619	0.000
		5	0.199	0.109	34.630	0.000
		6	0.216	0.059	37.072	0.000
		7	0.227	0.056	39.830	0.000
		8	0.229	0.076	42.724	0.000
		9	0.237	0.048	45.927	0.000
		10	0.189	–0.041	48.027	0.000
		11	0.109	–0.047	48.746	0.000
		12	0.096	0.029	49.318	0.000
		13	0.141	0.069	50.597	0.000
		14	0.101	–0.089	51.277	0.000
		15	0.029	–0.038	51.336	0.000
		16	0.055	0.046	51.555	0.000
		17	0.080	–0.008	52.031	0.000
		18	–0.033	–0.175	52.115	0.000
		19	–0.096	0.000	52.858	0.000
		20	–0.095	–0.080	53.612	0.000

Table 6.2 Unit root test for annual Aman rice production of Dhaka division

Null Hypothesis: Production has a Unit root			
		t-statistic	P-value
Augmented Dickey-Fuller test statistic		−1.540017	0.5038
Test critical values:	1% level	−3.596616	
	5% level	−2.933158	
	10% level	−2.604867	

Table 6.3 Unit root test (at first difference) for annual Aman rice production of Dhaka division

Null Hypothesis: D(Production) has a Unit root			
		t-Statistic	P-value
Augmented Dickey-Fuller test statistic		−6.728800	0.0000
Test critical values:	1% level	−3.600987	
	5% level	−2.935001	
	10% level	−2.605836	

First, we choose a tentative model ARIMA (1, 1, 1) which contain one Autoregressive (AR) parameter and one Moving Average (MA) parameter. The estimated parameters are shown in below:

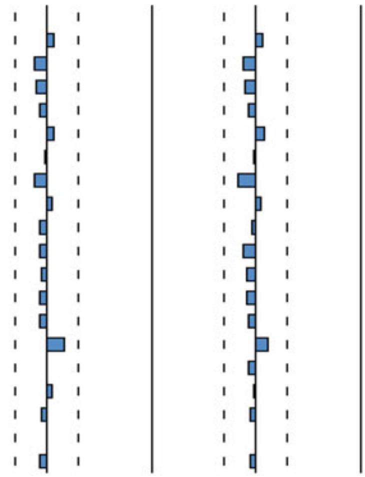
From the below Table 6.4 we use t-Statistic to test the null hypothesis that the parameters are equal to zero. The p value shows that all the coefficients are

Table 6.4 Summary of annual Aman rice production of Dhaka division

Variable	Coefficient	Std. Error	t-Statistic	Probability
C	31958.26	9957.869	3.209347	0.0027
AR(1)	0.490783	0.170181	2.883886	0.0064
MA(1)	−0.952092	0.046130	−20.63946	0.0000
R-squared	0.187506	Adjusted R-squared	0.144743	
Akaike info. criterion	28.11186	Schwarz criterion	28.23724	
Durbin-Watson stat	2.011401	Prob(F-statistic)	0.019346	

Table 6.5 Correlogram for ARIMA (1,1,1) model

Date: 10/07/18 Time: 13:03
Sample: 1974 2014
Included observations: 41
Q-statistic probabilities adjusted for 2 ARMA term(s)



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	−0.003	−0.003	0.0005	
		2	0.079	0.079	0.2801	
		3	−0.123	−0.124	0.9861	0.321
		4	−0.104	−0.112	1.5039	0.471
		5	−0.081	−0.064	1.8230	0.610
		6	0.078	0.082	2.1285	0.712
		7	−0.009	−0.023	2.1327	0.830
		8	−0.126	−0.177	2.9852	0.811
		9	0.043	0.049	3.0854	0.877
		10	−0.064	−0.028	3.3183	0.913
		11	−0.076	−0.122	3.6620	0.932
		12	−0.057	−0.092	3.8571	0.954
		13	−0.075	−0.087	4.2109	0.963
		14	−0.059	−0.059	4.4389	0.974
		15	0.166	0.125	6.3074	0.934
		16	0.003	−0.059	6.3080	0.958
		17	0.041	−0.009	6.4333	0.972
		18	−0.047	−0.046	6.5998	0.980
		19	0.003	0.005	6.6006	0.988
		20	−0.060	−0.046	6.8991	0.991

significant. The R-squared and Adjusted R-squared of *ARIMA* (1, 1, 1) model is 0.187506and 0.144743respectively. The entire coefficient of estimated model is significant at 5% level of significance. The observed R-Squared suggested that the sample regression line fit the data as well. Also, all of the estimated coefficients are satisfied invariability and stability condition. Also, from the Durbin-Watson statistic we see that the value of Durbin-Watson statistic is 2.011401which are greater than 2. This means that the residuals are not autocorrelated. So, we conclude that the model is appropriate (Table 6.5).

Correlogram-Q-statistics states that almost all spikes are significant at different lags support the hypothesis that there is no autocorrelation in the residual. Thus, the model is fully specified. Now we will check the histogram and normality test. This will help us to know about the residuals are normally distributed or not (Fig. 6.2).

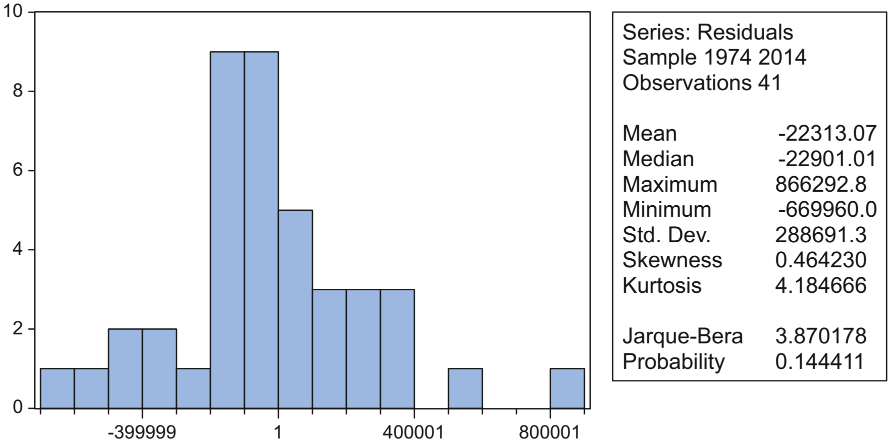


Fig. 6.2 Histogram and Normality Test for ARIMA (1,1,1) Model

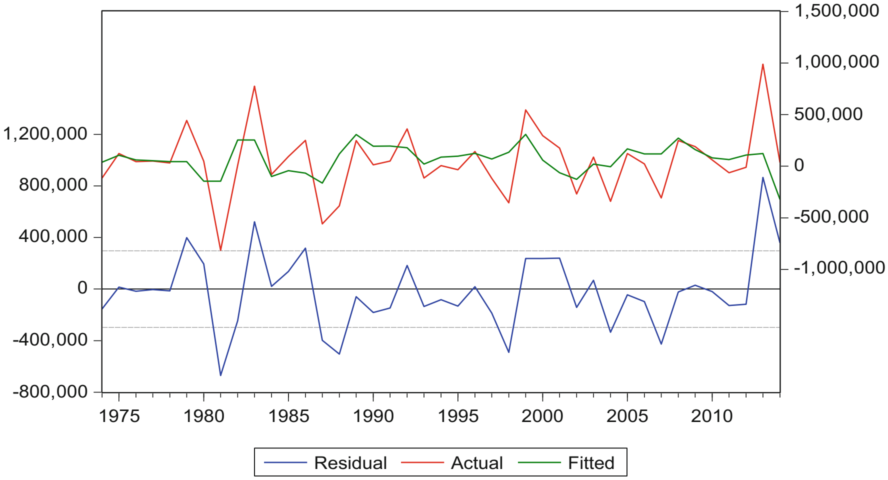


Fig. 6.3 Actual Fitted and Residual Plot for ARIMA (1,1,1) Model

Histogram and Jarque-Bera test indicate that the residuals are normally distributed. Annual Aman rice production of Dhaka division is positively- skewed (Skewness = 0.464230) and leptokurtic (Kurtosis = 4.184666).

We reject H_0 if $\lambda^2_{2,1-\alpha} = 5.99 < JB_{obs}$ or $JB_{obs} > \lambda^2_{2,1-\alpha} = 5.99$ at 5% level of significance, otherwise we accept it. Here the calculated value JB_{obs} of is smaller than the tabulated value $\lambda^2_{2,1-\alpha} = 5.99$ i.e. $JB_{obs} < \lambda^2_{2,1-\alpha} = 5.99$. We may accept the null hypothesis, i.e. residuals are normally distributed. Thus the model is fully specified. Actual fitted and residual plot are shown (Fig. 6.3).

Table 6.6 Chow's tests for ARIMA (1,1,1) model

Chow Breakpoint Test: 2000			
Null hypothesis: No breaks at specified breakpoints			
F-statistic	0.733794	Probability	0.5389
Log likelihood ratio	2.500912	Probability	0.4751
Chow forecast test: Forecast from 2000 to 2014			
F-statistic	7.117661	Probability	0.0001
Log likelihood ratio	29.99619	Probability	0.0000

Stability test are another test for model validation. For stability test the mostly used test is Chow's test. The Chow's test are two categories, they are Chow's Breakpoint test and Chow's Forecast test. They are given (Table 6.6)

Here for Chow's breakpoint test we take breakpoint sample in 2000 and for Chow's forecast test we take forecast sample from 2000 to 2014. From the Table 6.6 we can see that both Chow's tests support the hypothesis that there is no structural break in the model. Thus, the model is fully specified.

6.4 Conclusion

6.4.1 Required Model

According to our above calculation or completing Box-Jenkins process our estimated model *ARIMA* (1, 1, 1). The model is described as below:

$$\nabla y_t = c + \phi \nabla y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t$$

6.4.2 Forecasting

We have estimated the model using the data covered the period from 1972 to 2014. Now we forecast between the observations 2000–2014, which is called in-sample forecast (Fig. 6.4). The out-sample forecast sample of annual Aman rice production of Dhaka division from 2014 to 2035 (Fig. 6.5).

The below figure (Fig. 6.5) shows 20 years (2014–2035) forecasting with 95% confidence interval of annual Aman Rice production of Dhaka division. The upper red and lower red lines on the above and below the blue line shows the 95% confidence upper and lower limits respectively. According to the graph, we can forecast the annual Aman rice production of Dhaka division in Bangladesh

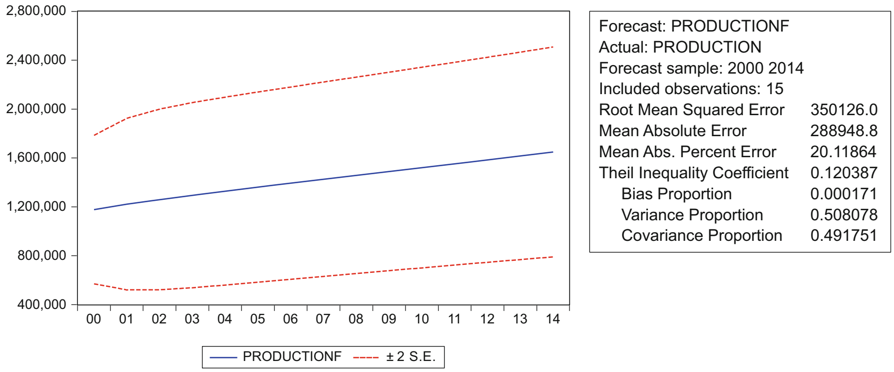


Fig. 6.4 In-sample Forecast of Annual Aman Rice Production from 2000 to 2014

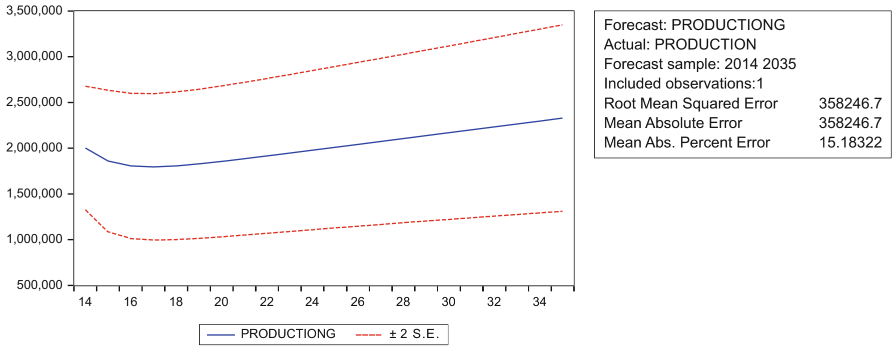


Fig. 6.5 Out-sample Forecast of Annual Aman Rice Production from 2014 to 2035

6.4.3 Summary

Using the appropriate model, we forecast up-to the year 2035. The forecasted values of the annual Aman rice production of Dhaka division in Bangladesh are shown (Table 6.7).

We have shown that the time series ARIMA model can be used to model and forecast the annual Aman rice production of Dhaka division in Bangladesh. The identified *ARIMA* (1, 1, 1) has proved to be adequate in forecasting the annual Aman rice production of Dhaka division in Bangladesh up-to the year 2035. Here we found that the annual Aman rice production of Dhaka division in Bangladesh is slightly increasing.

Table 6.7 Forecasted value up-to the Year 2035 for annual Aman rice production

Year	Production (ton)	Year	Production (ton)
2015	1,872,411	2026	2,060,157
2016	1,820,929	2027	2,092,082
2017	1,811,936	2028	2,124,024
2018	1,823,796	2029	2,155,974
2019	1,845,891	2030	2,187,928
2020	1,873,008	2031	2,219,885
2021	1,902,590	2032	2,251,842
2022	1,933,382	2033	2,283,800
2023	1,964,768	2034	2,315,758
2024	1,996,446	2035	2,347,716
2025	2,028,266		

References

1. Anonymous: Problem and prospects for suitable intensification of rice production in Bangladesh. Project publication no. 12, TCTTI Dhaka p-1 (1998)
2. Bhuiyan, N.I., Paul, D.N.R., Jabber, M.A.: Feeding the extra millions by 2025– Challenges for rice research and extension in Bangladesh. National Workshop on Rice Research and Extension in Bangladesh, Bangladesh Rice Research Institute, Gazipur, 29–3 1 January (2002)
3. Bangladesh Bureau of Statistics: Yearbook of agricultural statistics of Bangladesh. Bangladesh Bureau of Statistics, Ministry of Planning, Dhaka (2012)
4. Rahman, N.M.F.: Forecasting of boro rice production in Bangladesh: an ARIMA approach. J. Bangladesh Agric. Univ. **8**(1), 103–112 (2010)
5. Bangladesh Bureau of Statistics: The statistical yearbook of Bangladesh, pp. 1–2. Bangladesh Bureau of Statistics, Ministry of planning, Dhaka, Agricultural Wing (2015)
6. George, E.P.B., Gwilym, M.J., Gregory, C.R.: Time series analysis: Time series analysis forecasting & control, 4th edn. Wiley/Prentice Hall, Englewood Cliffs (1994)
7. Hamjah, M.A.: Forecasting major fruit crops productions in Bangladesh using Box-Jenkins ARIMA model. J. Econ. Sustain. Dev. **5**(7), (2014)
8. Awal, M.A., Siddique, M.A.B.: Rice production in Bangladesh employing by ARIMA model. Bangladesh J. Agric. Res. **36**, 51–62 (2011)
9. Amin, M., Amanullah, M., Akbar, A.: Time series modeling for forecasting wheat production of Pakistan. J. Anim. Plant Sci. **24**(5), 1444–1451 (2014)
10. Maniha, Z., Time Series, A.: Modeling on GDP of Pakistan. J. Contemp. Issues Bus. Res. **3**(4), 200–210 (2014)
11. Bipasha, M., Bani, C.: Forecasting GDP growth rates of India: an empirical study. Int. J. Econ. Manage. Sci. **1**(9), 52–58 (2012)
12. Siti, R.Y., Maizah, H.A., Lee, C.N., Noryanti, M.: A comparative study on Box-Jenkins and Garch models in forecasting crude oil prices. J. Appl. Sci. **11**(17), 1129–1135 (2011)
13. Rachana, W., Suvarna, M., Sonal, G.: Use of the ARIMA model for forecasting Pigeon Pea production in India. Int. Rev. Bus. Finance. **2**(1), 97–102 (2010)
14. Rahman, N.F.M.: Forecasting of boro rice production in Bangladesh: an ARIMA approach. J. Bangladesh Agric. Univ. **8**(1), 103–112 (2010)
15. Solomon, S.: Forecasting milled rice production in Ghana using Box-Jenkins approach. Int. J. Agric. Manage. Dev. (IJAMAD). (2011)
16. Dickey, D.A., Fuller, W.A.: Distributions of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. **74**(366), 427–481 (1979)

17. Gujarati, D.N., Porter, D.C., Gunasekar, S.: Basic econometrics: econometric modeling specification and diagnostic testing, 4th edn. Mc Graw Hill International (2003)
18. Imon, A.H.M.R.: Introduction to regression Time Series and forecasting: Box-Jenkins ARIMA models. Nanita Prokash. (2017)
19. Jarque, C.M., Bera, A.K.: A test for normality for observations and regression residuals. Int. Stat. Rev. **55**, 163–172 (1987)

Chapter 7
Impacts and Determinants of Adoption in River-Based Tilapia Cage Culture

Mohammad Mizanul Haque Kazal and Md. Sadique Rahman

Abstract Cage culture, an aquaculture technique for tilapia production, expanding quickly in flowing water of rivers and canals in Bangladesh. The present study identifies the determinants of improve cage culture practices adoption and its impact on productivity and profitability. The study employed poisson regression, and inverse probability weighted regression adjustment techniques to achieve the objectives. The findings indicate that the decision to adopt was positively influenced by education (p-value<0.10), extension contact (p-value<0.05), societal membership (p-value<0.05), and number of working person (p-value<0.10) while inversely influenced by farm size (p-value<0.10). Adopters of improve practices received significantly higher productivity and profit compared to non-adopters. More research and investment, as well as modification in extension services and approaches are needed to improve the adoption level and sustain the production. Higher productivity implies higher income, thus may reduce poverty.

Keywords Adoption · Improve practices · Impact evaluation · Tilapia farming

7.1 Introduction

Bangladesh is considered one of the most suitable regions for fisheries in the world, with the world's largest flooded wetland and the third largest aquatic biodiversity in Asia after China and India [1]. The aquaculture plays a particularly crucial role among fish farmers as a main or additional source of employment, livelihood and income in Bangladesh. Aquaculture has increasingly been playing a major role in

M. M. H. Kazal (✉)
Department of Development & Poverty Studies, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh
e-mail: mhkazal@sau.edu.bd

Md. S. Rahman
Department of Management and Finance, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh
e-mail: sadique@sau.edu.bd

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_7

total fish production (3.26 million tons) of the country and presently more than half of the total production (52.92%) comes from aquaculture. The performance of the aquaculture sector of Bangladesh has become the fifth-largest producer of aquaculture products in the world in recent years. It is a clear sign of its increasingly important role in the nation's social and economic development. The steady growth of aquaculture has tremendously contributed in the fish production and livelihood of fish farmers' in Bangladesh.

A new form of aquaculture farming, known as cage culture, is expanding in flowing water of rivers and canals in various parts of the country, raising hopes for an increased production of fish. Fishes, mostly tilapia (Oreochromis niloticus), are now farmed in nearly 6000 cages in rivers where such farming did not exist even a decade ago. Cage culture is a technique to use the inland water-body properly and have significant impacts on aquaculture production, income and employment generation and the nutritional status of people of Bangladesh [2]. Tilapia cage culture has gained momentum in certain parts of Bangladesh such as Dakatia river in Chandpur, different rivers in Laxmipur, Chapainawabganj and Rangamati as evidenced by the growing number of tilapia fish cage farmers in the country. Tilapia in cages is easy to manage, advantageous to rear quality and selective fishes easy to harvest. It is easy to eliminate predation and competition and easier to treat diseases and parasites.

Considering the importance of tilapia cage culture Bangladesh Fisheries Research Institute (BFRI) has identified few 'improve practices' to increase the productivity and return from cage culture and disseminated the improve practices through Department of Fisheries (DoF) in different areas of Bangladesh. A number of technology transfer methods, including training sessions, field days, and demonstrations were conducted by DoF to increase knowledge and adoption of improve practices. Adoption of any new technology depends on the farm-level profitability. But there is dearth of studies regarding productivity and profitability impact of improve practices adoption in Bangladesh. Only a few studies [2, 3] conducted on the profitability aspect of cage farming but used very small sample to achieve the objective which were not representative for the country. Furthermore, previous studies did not consider the selection bias problem which may arise due to non-randomness in sampling. The causal effects of improve cage farming practices adoption on productivity and profitability are yet to be studied empirically in Bangladesh. Identification of the socio-economic determinants of improve practices adoption may contribute to the effective development of future policy. Thus, the present study was conducted to identify the socio-economic determinants of adoption and its impact on the livelihood of the fish farmers.

7.2 Methodology

7.2.1 Data Sources

The study was conducted in three districts (geographical region used for administrative purpose): Chadpur, Laxmipur and Chapainawabganj of Bangladesh where DoF provided training and arranged field days to disseminate the improve practices.

From each district, two major cage farming sub-districts were selected through DoF. A list of cage growers was prepared with the help of DoF officials for each district, which serves as the population for the present study. A total of 85 farmers from each district were selected randomly from that list for face-to face interview. Thus, a total of 255 cage farmers were interviewed. Out of 255 farmers, 8 farmers did not able to provide any information regarding productivity of cage farming. Those farmers were dropped from the analysis, leaving 247 farmers as sample. The cage farmers were divided into adopters and non-adopters based on the four improve practices: i) use of improve fingerlings, ii) maintenance of appropriate stocking density, iii) provide food according to body weight of the tilapia, and iv) apply lime in nursery pond. The distribution of the farmers based on the improve practices is given below (Table 7.1).

7.2.2 Analytical Techniques

Both descriptive statistics and econometric modeling was used to achieve the objectives of the present study.

7.2.2.1 Factors Affecting Adoption

The Poisson regression model, suitable for the estimation of count data [4], was selected for the estimation of the cage farmers' decisions on the number of improve practices to adopt. A random variable Y with parameter γ follows a Poisson distribution if it takes integer values (0,1,2,..) with following probability;

$$\Pr(Y = y) = \frac{e^{-\gamma}\gamma^{y}}{y!}$$

The mean and variance of the distribution are as follows;

$$E(Y) = \gamma \text{ and } Var\ (Y) = \gamma$$

The cage farmers made a series of discrete household decisions that was computed across an aggregation of choices to a Poisson distribution. The key criteria

Table 7.1 Sample distribution of the tilapia farmers

Farmers category	No. of farmers ($n = 247$)
(a) Adopters	
All the four practices	97
Any three practices	119
Any two practices	17
Any one practice	0
(b) Non-adopters	14

of using a Poisson regression is the mean must equal the variance. If overdispersion is present, an alternative model can be used. In this study we have tested the overdispersion by loglikelihood ratio (LR) test. Based on the test result, following Poisson regression model was used to assess the effect of independent variables, x_i, on a scalar dependent variable y.

$$E\ (y/x_i) = \gamma = \exp\ (x_i\ \beta_\mathrm{i})\ \ and\ y = 0\text{–}4$$

Where, x_i is the explanatory variables, β_i unknown parameters to be estimated, and y is the number of improve practices adopted by a farmer.

7.2.2.2 Impact Assessment

Most of the treatment effect models represent the treatment status by a binary (0,1) variable, but in our study adoption of improve practices occurs at different levels (2,3 and 4). Regression and weighting techniques can be employed to estimate multivalued treatment [5, 6]. The present study used inverse probability weighted regression adjustment (IPWRA) estimator due to its double-robust property that ensures consistent results as it allows the outcome and the treatment model to account for mis-specification. Average treatment effect on the treated (ATT) in the IPWRA model was estimated in two steps. In the first step, we estimate the propensity scores using multinomial logit regression and in second step, linear regression was used to estimate the ATT [7]. ATT was computed as follows;

$$\Pr\left(T_i = 2, 3, 4\right) = g\ (J_i\alpha) + v_i$$

$$Y_i = f\ (x_\mathrm{i}\beta) + \mathrm{u_i}$$

Where, J_i is a set of covariates explaining treatment assignment T_i, x_i is a vector of covariates that influence the outcome Y_i. DoF introduced the improve practices with a view to increase the productivity and thus, improve the livelihood of the tilapia farmers. The present study measured the farm-level impact of improve practices adoption in terms of productivity (kg/cage) and profitability (Tk/cage). Description of the independent variables used in the models is given below (Table 7.2).

7.3 Results and Discussion

7.3.1 Descriptive Statistics of the Variables Used in the Models

Differences in selected characteristics of adopters and non-adopters are presented in Table 7.3. The mean difference suggested that there are some differences between adopters and non-adopters in terms of selected household characteristics. The characteristics of adopters and non-adopters were similar in terms of training,

Table 7.2 Definition of the variables used in the different models

Variable	Notation	Description
Family member (No.)	z_1	Total number of family members.
Age (yrs)	z_2	Age in years of the primary farmer, a proxy for experience.
Education (yrs)	z_3	Total years of schooling of the primary farmer, representing the knowledge of the farmer.
Spouse education (yrs)	z_4	Total years of schooling of the primary farmer's spouse.
Training (days)	z_5	Total number of days spends on training during last 12 months.
Farm size (ha)	z_6	Farm size of the primary farmer in hectare.
Societal membership (yes/no)	z_7	One if the primary farmer involved in any societal organization, otherwise 0.
Extension contact (yes/no)	z_8	One if the primary farmer received information from local extension officer regarding cage farming, otherwise 0.
FMWPF (No.)	z_9	Number of family members working in cage farming.

Table 7.3 Descriptive statistics of the socio-economic characteristics of the respondents

Characteristics	All adopters	Non-adopters	Mean diff.
Family member (No.)	4.46	3.85	0.61*
Age (yrs)	38.78	43.71	−4.92*
Education (yrs)	7.63	4	3.63***
Spouse education (yrs)	7.20	6.40	0.80^{ns}
Training (days)	4.40	2.42	1.97^{ns}
Farm size (ha.)	0.33	0.36	-0.03^{ns}
Societal membership (%)	44.21	7.14	37***
Extension contact (%)	89.27	50.00	39***
FMWCF (No.)	1.45	1.14	0.31**

Note: * and *** indicates significant at 10% and 1% level respectively and ns indicates not significant; FMWCF indicates Family members working in cage farm; t-test was used

and farm size. But significant differences exist between adopters and non-adopters with respect to number of family members, age, education, societal membership, extension contact and number of family members working in cage farm, which indicates that the two groups are not directly comparable and justifies the use of treatment effect model.

7.3.2 Adoption Status of Improve Practices

It is evident from the Table 7.4 that most of the farmers (48.18%) adopted 3 practices followed by 39.27% who adopted all the 4 selected improve practices, respectively.

Table 7.4 Adoption status of different improve practices ($n = 247$)

Items	Number of practices				
	1	2	3	4	0
No. of farmers	0	17	119	97	14
Percent of total	0	6.88	48.18	39.27	5.67

Table 7.5 Factors affecting adoption decision: Poisson estimates

Variables	Unit	Coefficient	Robust SE	z
Family member	Number	0.007	0.019	0.370
Age	Years	0.001	0.003	0.470
Education	Years	0.015*	0.008	1.840
Spouse education	Years	−0.003	0.009	−0.280
Training	Days	0.011	0.007	1.600
Farm size	Hectare	−0.032*	0.020	−1.620
Societal membership	Dummy (yes/no)	0.086**	0.038	2.230
Extension contact	Dummy (yes/no)	0.227**	0.094	2.430
FMWCF	Number	0.050*	0.030	1.660
Constant		0.614	0.166	3.700
Log likelihood	−411.08			
LR chi square	30.68***			
Pseudo R^2	0.15			
Goodness of fit	104^{ns}			
No. of observations	247			

Note: FMWCF indicates Family members working in cage farm; *, ** and *** indicates significant at 10%, 5% and 1% level
ns not significant

No farmers adopted 1 practice and around 6% of the farmers did not adopt any of the selected improve cage culture practices.

7.3.3 Factors Affecting Adoption

The LR test of overdispersion parameter ($\alpha = 0$) indicated that alpha was not significantly different from zero and thus, Poisson regression is suitable. It is revealed from Table 7.5 that the estimated pseudo R-squared value is fairly low (0.15), but the overall significance of the Poisson model, reported by the Wald chi-squared value, is satisfactory. Non-significant value of goodness of fit also indicating the good fit of the Poisson model.

Findings indicates that farmers' education, societal membership status, contact with extension and number of family members working person in the cage farm were positive and significantly influenced the adoption of improve practices while farm size of the respondents has negative effect on adoption. The negative association of farm size ($p < 0.10$) may implies that the farmers who have larger amount of land for crop cultivation may not get enough time to involved in cage

Table 7.6 Cost and return from cage cultivation

Items	Adopters			All adopters	Non adopters	Mean diff.
	2	3	4			
Production (kg/cage)	334	455	542	480	268	217***
Ave. price (Tk/kg)	132	122	122	123	125	-2^{ns}
Gross return (Tk/cage)	44,088	55,510	66,124	59,040	33,500	25,540***
TVC (Tk/cage)	35,247	43,181	52,255	45,382	30,399	14,983***
TFC (Tk/cage)	1843	1427	1255	1472	1592	-120^{ns}
Total cost (Tk/cage)	37,090	44,608	53,511	46,854	31,991	12,863***
Net return (Tk/cage)	6998	10,902	12,613	12,186	1509	10,677***

Note: *** indicates significant at 1% level; 1 US$ = Tk. 80 (approx.), t-test was used
ns not significant

farming. The positive and significant association of the extension contact ($p < 0.01$) suggested that exposing farmers to agricultural extension advice could help to increase the adoption of improve practices confirms the findings of [8, 9]. Adoption of new technologies requires some level of technical knowledge, direct contact with extension services increases the acquisition of relevant knowledge. Efforts are necessary to increase the number of extension personnel's in the rural areas to increase the adoption level. Societal membership ($p < 0.05$) also positively influenced the adoption may be due to the fact that farmers who are engaged with different societal membership get the opportunity to meet with different peoples which may influence them to adopt new technologies.

7.3.4 *Cost and Return of Cage Cultivation*

The findings revealed that on an average, the total cost of cage cultivation was higher for adopters compared to non-adopters. These differences were statistically significant at 1% level (Table 7.6). Table 7.6 indicates production per cage was found significantly ($p < 0.01$) higher for adopters (479 kg/cage) than that of non-adopters (263 kg/cage). Due to higher productivity, gross and net return was also significantly ($p < 0.01$) higher for adopters compared to non-adopters. The findings indicate that although the adoption of improve practices need higher capital investment but at the same time it provided significantly higher income which may be useful in reducing poverty and malnutrition to some extent.

7.3.5 *Impact of Improve Cage Culture Practices Adoption*

7.3.5.1 Impact on Productivity

It is evident from the Table 7.7 that the adoption of improve practices significantly affected the productivity of cage farming. Farmers who adopted the improve

Table 7.7 Impact of improve practices on productivity

Number of practices adopted				
Adopters	Non-adopters	ATT	SE	z
2	0	86***	17	5.06
3	0	210***	20	10.50
4	0	266***	34	7.82

Note: *** indicates significant at 1% level

Table 7.8 Impact of improve practices on profitability of cage culture

Number of practices adopted				
Adopters	Non-adopters	ATT	SE	z
2	0	11,922**	4727	2.52
3	0	8655*	4878	1.77
4	0	8811*	4721	1.87

Note: ** and * indicates significant at 5% and 10% level

practices received significantly higher production (86–266 kg/cage) compared to non-adopters which is similar to the findings of [10] indicated that technology adoption significantly affected the productivity. The ATT values indicates that farmers who adopted more practices received higher productivity compared to the farmers who adopted none or less number of improve practices.

7.3.5.2 Impact on Profitability

Findings of the Table 7.8 indicates that all category of adopters received significantly higher profit compared to non-adopters due to higher productivity. According to [11] improve technology adoption increases the income of the adopters. Farmers who adopted more number of improve practices received higher income compared to the farmers who adopted none or less number of practices (Table 7.8). The values of ATT were ranged from Tk. 8811–Tk. 11,922 based on number of adopted improve practices. More awareness building programs along with extension services are warranted to augment the adoption process since adoption enhanced the productivity and income.

7.4 Conclusions and Policy Implications

The present study measured the determinants of improve practices adoption, productivity and profitability impacts of tilapia cage farming in Bangladesh. Findings indicate that most of the farmers adopted three practices out of four recommended practices. Adoption analysis suggested that education, extension contact, and societal membership played a significant role in adoption decision. Results of impact evaluation revealed that farmers who adopted improve practices received a significantly higher productivity and profit compared to non-adopters. Since farmers were reluctant to adopt all the recommended practices, more investment in research and development is needed to develop and optimize the improve practices as a package. Lower adoption may also indicate lack of awareness building programmes. Investment in education and extension services like training sessions, field visit and demonstrations of improve practices is needed to increase the awareness and thus, level of adoption. Extension approaches should also be modified since our findings suggested that the farm size negatively influenced adoption decision. Small and medium farmers should bring under extension coverage by providing more training, and advice for better result. Our findings also suggested that cage culture is cost incentive, so concern authority may arrange credit with low interest for small farmers to augment the adoption process. Since cage culture is profitable, Government may provide incentives to institutions involved in transferring the improve practices at farm level to sustain the production.

Acknowledgments The authors greatly indebted to CRG-BARC-NATP 2 project for funding the research.

References

1. Shamsuzzaman, M.M., Islam, M.M., Tania, N.J., Abdullah Al-Mamun, M., Barman, P.P., Xu, X.: Fisheries resources of Bangladesh: present status and future direction. Aquac. Fish. **2**, 145–156 (2017)
2. Hossain, M.R.A., Rahman, M.A., Akter, S., Hosain, M.M.E., Naser, M.N.: Intervention of tilapia cage culture in the River Dakatia: threaten or blessed to local fish diversity. Int. J. Fish. Aquat. Stud. **5**(1), 228–232 (2016)
3. Baqui, M.A., Bhujel, R.C.: A hands of training helped proliferation of Tilapia culture in Bangladesh. In: 9th international symposium on Tilapia in Aquafish collaborative research program, pp. 311–322. Shanghai, China (2011)
4. Greene, W.: FIML estimation of sample selection models for count data. Working paper EC-97-02. Department of Economics, Stern School of Business, New York University (1997)
5. Uysal, S.D.: Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. J. Appl. Econ. **30**, 763–786 (2015)
6. Frolich, M.: Programme evaluation with multiple treatments. J. Econ. Surv. **18**(2), 181–224 (2004)
7. Imbens, G., Woolridge, J.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**(1), 5–86 (2009)
8. DeGraft-Johnson, M., Suzuki, A., Takeshi, S.T., Otsuka, K.: On the transferability of the Asian rice green revolution to rainfed areas in sub-saharan Africa: an assessment of technology intervention in northern Ghana. Agric. Econ. **45**, 555–570 (2014)
9. Mensah-Bonsu, A., Daniel, B.S., Ramatu, H., Samuel, A.B., Irene, S.E., John, K.M.K., Yaw, B.O.: Intensity of and factors affecting land and water management practices among smallholder maize farmers in Ghana. Afr. J. Agric. Res. Econ. **12**(2), 142–157 (2017)
10. Rand, J., Tarp, F.: Impact of an aquaculture extension project in Bangladesh. J. Dev. Effect. **1**(2), 130–146 (2009)
11. Amankwah, A., Quagrainie, K.K.: Aquaculture feed technology adoption and smallholder household welfare in Ghana. J. World Aquacult. Soc. 1–15 (2018)

Chapter 8
Policy into Practice; Statistics the Forgotten Gatekeeper

Iain H. Hume

Abstract The Australian Federal Government Carbon Farming Initiative (CFI) was a voluntary carbon offsets scheme. It allowed land managers to earn carbon credits by changing land use or management practices to store carbon. The Carbon Farming Futures (CCF) program ran from 2012 to 2017 to identify where farmers can boost productivity and profitability; improve soil and reduce greenhouse gas emissions. The Action on The Ground program was one component of the CCF which aimed to assist landholders in trialing new technologies and practices. Laboratory and field plot research indicated that ploughing nutrients into the soil with crop residues/stubble increased the amount of soil carbon stored. Our project aimed to test if this was possible using farm equipment. The challenge was to design experiments with plots large enough to be managed with commercial farm machinery that deliver enough precision to test different treatments.

We tested carbon sequestration methods in 14 fields in different bioregions. Soil variability was estimated using two electromagnetic surveying instruments. Randomized block experiments were established in fields after areas of similar soil were identified by geostatistics and finite mixture models.

The field experiments failed to reproduce the high sequestration rates of the earlier, more controlled, research. We describe one experiment in this chapter and propose reasons for this failure to sequester carbon. The hybrid demonstration/research approach is applicable in many other situations where agricultural policy or practice change are to be tested at farm scale.

Keywords Soil carbon · Experimental design · Large plots · Policy

I. H. Hume (✉)
NSW Department of Primary Industries, Wagga Wagga, NSW, Australia
e-mail: Iain.Hume@dpi.nsw.gov.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_8

8.1 Background

8.1.1 The Carbon Farming Initiative

The Carbon Farming Initiative (CFI) was a voluntary carbon offsets scheme managed by the Australian Federal Government. It allowed land managers to managers to earn carbon credits by changing land use or management practices to store carbon or reduce greenhouse gas emissions [1]. The action on the ground component of the CFI provided grant funding to farmers and land managers over 6 years 2001–2017.

Action on the Ground funding was to assist farmers and land managers to undertake on-farm trials of abatement technologies, practices and management strategies to measure and demonstrate opportunities to reduce agricultural greenhouse gas emissions of methane and nitrous oxide or increase the sequestration of carbon in soil while maintaining or improving farm productivity [2].

8.1.2 Sequestering Soil Carbon

It has been recognised that crop residues are the primary source of soil organic carbon. This can be as course material >0.4 mm which is less stable than the fine (<0.4 mm) material [3]. This fine carbon fraction has a higher concentration of nutrients than the crop residues, and has a nearly constant ratio of C:N:P:S. Stoichiometry is the process of calculating the quantity of reactants and products needed to balance chemical equations. Balancing the stoichiometric differences between the crop residues and the fine soil organic matter by adding inorganic nutrients to crop residues when they were incorporated increased the sequestration rate of soil carbon by 10% in 5 years [4]. This is the basis of this project, that is, to evaluate the improvement of C sequestration that can achieve through incorporating soil nutrients with crop residues.

8.1.3 Testing Policy

Although the purpose of the Action on the Ground program of the carbon farming initiative was to fund the demonstration of new practices to increase the carbon content of the soil, it had a minor focus on measuring the results of the actions of farmers. Our aim was to achieve both demonstration of practice and robust measurement of the consequences of these new practices through a hybrid design using geostatistics and expectation maximization to guide the design of field experiments.

8.2 Methods

The experiment was carried out over 3 years that encompassed two cropping cycles. Treatments were applied to two crops at each experimental location and the net effect of two treatment cycles assessed by measuring changes in soil carbon over this 2 year period.

8.2.1 Partnerships

The wider experiment was conducted in partnership with five farmer groups, a farm consultant and a university, in 14 fields. This paper focuses on the methods developed and applied, rather than the wider results and considers only one experimental field on a commercial farm.

Experimental Treatments Three treatments were applied:

1. I+: the incorporation of crop residues with additional nutrients required to balance the chemical breakdown of those crop residues;
2. I-: the incorporation of crop residues without additional nutrients; and
3. C: the farmers' usual practice, which in this case was to retain all crop residues with no additional nutrients and not to cultivate or disturb the soil.

Experimental Design A replicated block trial was conducted in the field; each treatment was repeated twice in each of the two blocks. The field 'plots' had to be large enough to have treatments applied and the crop harvested with commercial farm equipment; this was a compromise between scientific rigor and practical field operations. The experimental design rationale was to identify areas of the field that have similar soil properties which are large enough to apply the treatments. Electromagnetic (EM) surveying was used to measure the field's homogeneity.

Electromagnetic Surveys Electromagnetic techniques measure the electrical conductivity of the ground a property which is an integrated function of the soil's texture, water content and mineralogy [5]. Electromagnetic instruments measure the soil's electrical conductivity by sensing the strength of an electromagnetic field they propagate in the soil [6]. These instruments provide a fast and accurate way to measure the heterogeneity of agricultural fields [7] with the EM38 and EM 31 (Geonics. Ltd., Mississauga, ON, Canada) instruments being widely used in agricultural applications [8]. The EM38 is most sensitive to the properties of the top 1 m of soil while the EM31 is responsive to the properties of the deeper (top 3 m) soil. The combination of mobile equipment with GPS technology enables the rapid survey of agricultural fields [9]. The fields in this study were surveyed with both the EM38 and EM31 instruments by a commercial contractor, Terrabyte. The EM38 instrument was towed behind a vehicle while the EM31 was mounted on the vehicle (Fig. 8.1a). The instruments recorded at a rate of one measurement per second and

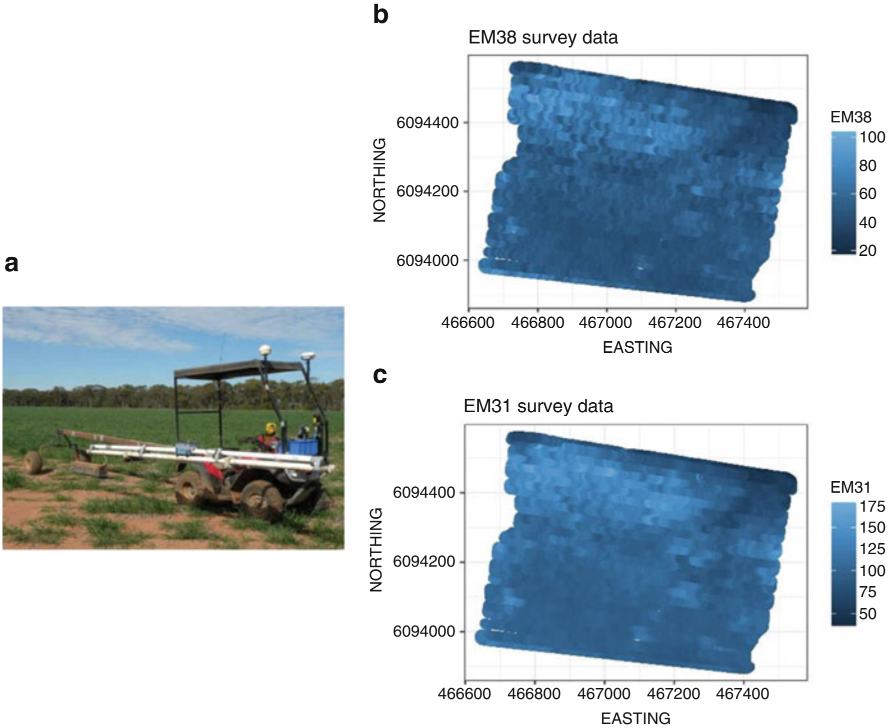


Fig. 8.1 The electromagnetic survey equipment mounted on a vehicle (**a**), and the raw EM38 survey (**b**) and EM31survey (**c**)

data collected by driving along a series of linear parallel transects, about 30 m apart, within fields. Data density within transects varied with vehicle speed and no data were collected between transects. Another source of variation in the density of data was due to the measurement being made around the field boundary which was often closer to the parallel transects than the inter-transect spacing of 30 m. EM survey provided as the raw data as maps showing EM38 and EM31surfaces drawn using 'least squared distance' smoothing and the location of each measurement (Fig. 8.1b, c). The raw GPS Easting and Northing, EM31 and EM38 data were also provided.

8.2.2 Development of Spatial Classes

This section describes how geostatistical and expectation maximization techniques were applied to the electromagnetic surveys to classify soil into areas of similarity.

Geostatistics A filtering process was used to remove the effects of the field boundary and trees within the field. All measurement made within 20 m of the

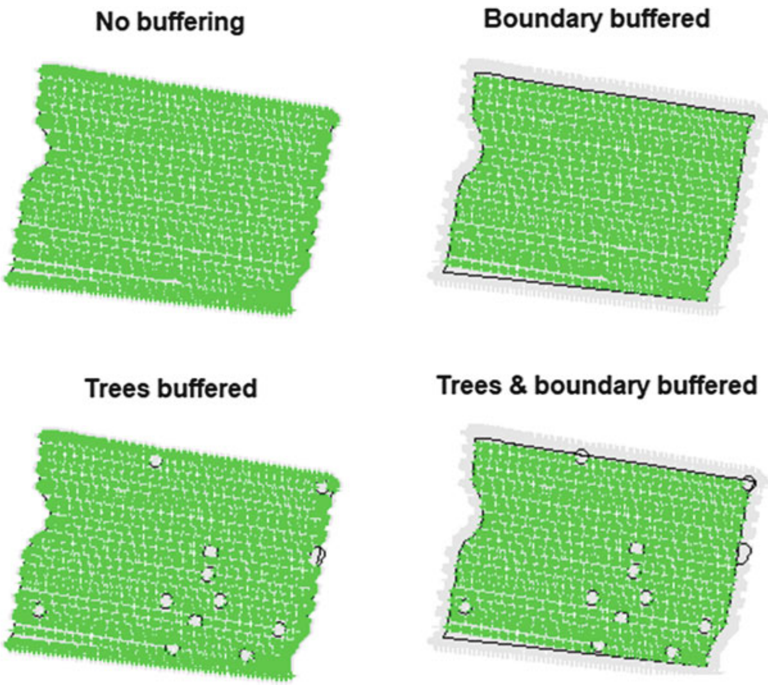


Fig. 8.2 Four levels of buffering applied to the electromagnetic surveys

field boundary (2295 data points) or a tree (133 data points) were removed. This buffering shrank the area surveyed and the data removed because of their proximity to trees were concentrated towards the south eastern quadrant of the survey (Fig. 8.2).

Buffering also changed the distribution of data (Fig. 8.3); with the un-buffered data suggesting a strongly bimodal distribution. This was still present in the data buffered around the trees but was absent from the data buffered to remove the boundary; the data buffered to remove both trees and the boundary were used for all subsequent analysis. The raw data are skewed (Fig. 8.3) and were normalized by taking logs before analysis.

Sample variograms of EM38 and EM31 data were constructed in four directions (0, 45, 90 and 132° from north). Based on the similarity of these variograms the data were assumed to be isotropic up to a distance of 200 m (Fig. 8.4). All subsequent variography and kriging was conducted using isotropic models.

The isotropic sample variograms of EM38 and EM31 data were modelled using spherical, exponential and gausian variogram models fitted using the gstat package [10] of the R statistical package [11]. The best variogram models were judged to be those with the least weighted mean squared error of fitting (Table 8.1). The best models were exponential with a nugget; the nugget, sill and range these parameters differed for the EM38 and EM31 variogram models (Table 8.2).

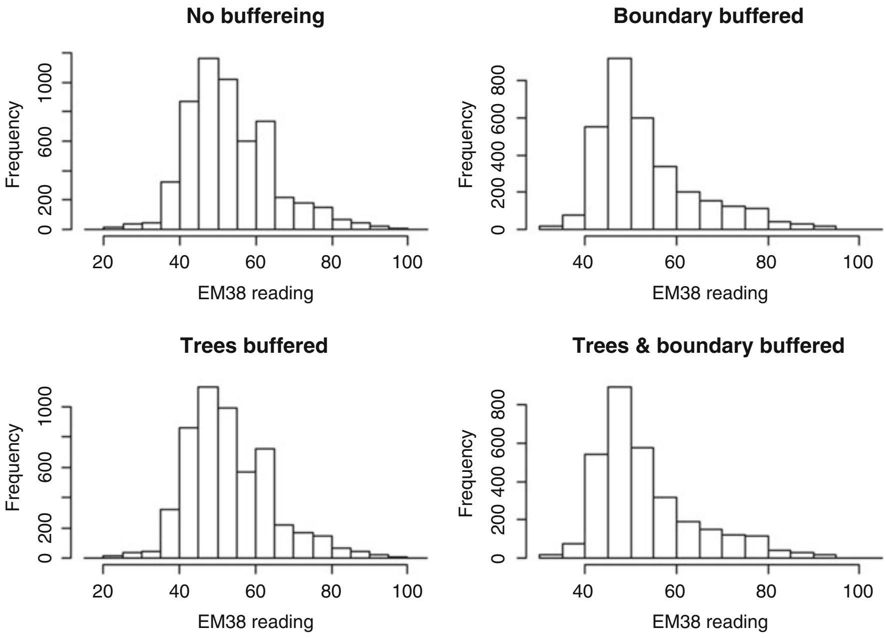


Fig. 8.3 Frequency distributions of the EM38 surveys subjected to four levels of buffering. The distributions of EM31 data (not shown) exhibited the same patterns

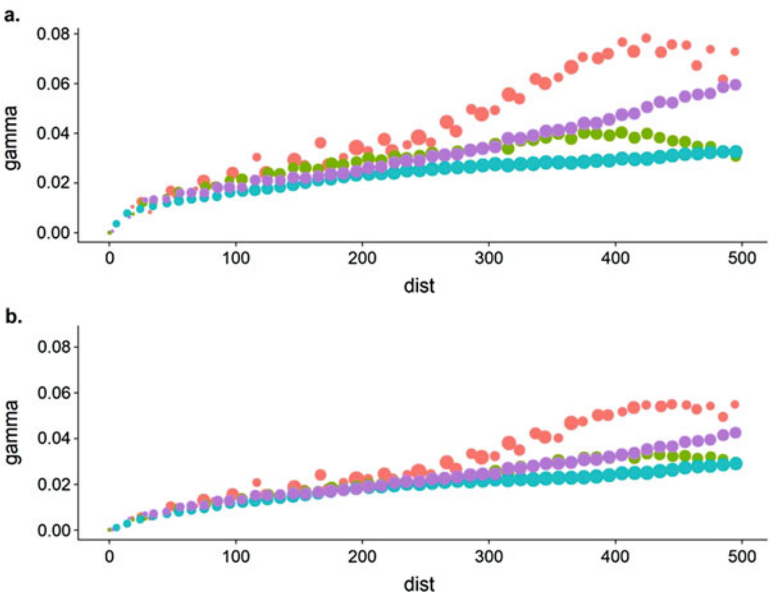


Fig. 8.4 The directional sample variograms of the EM31 (**a**) and EM31 surveys (**b**). The directions of sampling are 0° (●), 45° (●), 90° (●) and 120° (●) from North

Table 8.1 The mean squared errors of the exponential isometric variogram models fitted to the EM38 and EM31 electrical conductivity data buffered to different levels

Instrument	No buffering	Level of Buffering		
		Trees	Boundary	Trees & boundary
EM38	0.0003113	0.0002742	0.0028535	0.0002297
EM31	$1.138e^{-1}$	$1.809e^{-4}$	$1.072e^{-1}$	$8.368e^{-5}$

Table 8.2 The parameters of the isometric exponential variogram models fitted to EM38 and EM31 data buffered to exclude the effects of the field boundary & trees

Instrument	Nugget	Sill	Range
EM38	0.0049	0.3639	175.4
EM31	0.0009	0.3181	189.1

These variograms were used in ordinary Kriging [12] to produce smooth surfaces of electromagnetic data on a 25 m grid. This grid of kriged data was indexed according to the 'East West' columns and 'North South' rows they occupied. This indexing was retained throughout the Expectation Maximization process to allow translation between data and geographic space.

Expectation Maximization Traditional techniques that classify data space like Jenks natural breaks [13] or K means [14] had drawbacks. They produced in discontinuous spatial classes in geographic space. An alternative method was to identify a number of joint finite populations [15, 16] of EM31 and EM38 data. This was applied in the mixtools package [17] of R and could result in discontinuous classes in data space but always delivered contiguous spatial classes. The final, fitted model delivered the following information for each multivariate population or class:

- The proportion of the total population of data in each population (the final mixing proportions);
- Vectors of the mean value of each population in the EM38 and EM31 data;
- The variance covariance matrices of the mixtures; and
- Vectors of posterior probabilities of an observation being in a population.

Classes of spatial data were formed from the vectors of posterior probabilities by assuming that an observation belonged to a class if its posterior probability was greater than $(n - 1)/n$, where n is the number of classes. The initial classification was into two classes or multivariate distributions. Extra classes or distributions were added one at a time and the statistical validity of adding these extra populations was tested by calculating the log likelihood ratio of the n and n + 1 EM models. The extra class is adopted if it passes all the following criteria:

- The log likelihood ratio test returns $p > 0.05$;
- The last class added contains more than 5% of the total population: and
- The new class is large enough and of a shape such that it is possible to apply the treatments using farm equipment.

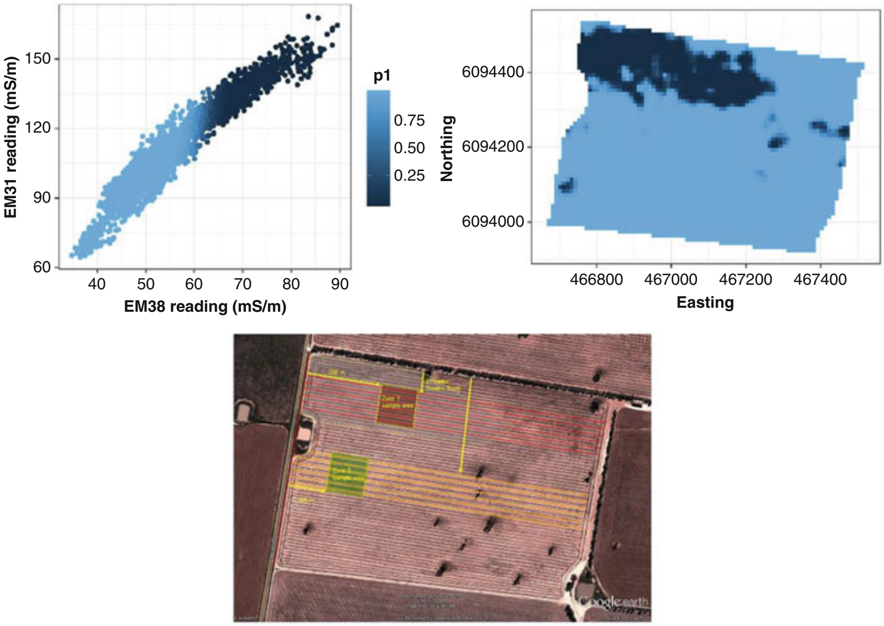


Fig. 8.5 The proportional occupancy of class1 in data space (top left panel) geographic space (top right panel) and the final layout of the experiment (bottom panel)

The classes were plotted in geographic space according to their row and column indexing (Fig. 8.5).

Layout of Experiment The experiment for each field was designed by examining the spatial classes and the direction of field operation. The aim was to locate one block of the experiment wholly within a class identified by electromagnetic surveying. Treatments were randomized within each block and a zone selected in the center of each class where soil carbon and yield were to be measured (Fig. 8.5).

8.2.3 Treatment Application

Treatments were applied as soon after harvest as was practically possible. The total crop residue load present in the strips to have extra nutrients applied was assessed by cutting standing stubble and collecting any residue from the soil surface from three $1 m^2$ quadrats in each treatment strip. These samples were dried at 40°C until they attained a constant weight and the dry weight of crop residue (t ha^{-1}) was calculated.

To satisfy the stoichiometric ratio, each tonne of wheat stubble required 5.8 kg of Nitrogen, 2.2 kg of Phosphorous, and 0.9 kg of Sulfur. These nutrients were supplied as commercially available fertiliser, spread on the soil surface using the farmer's fertliser spreader. The crop residue of the I- treatment and the crop residues and nutrients of the I+ treatment were incorporated into the soil using a disc tillage implement (Speed Tiller™). The control treatment was managed using the farmer's current practice, which in this case was not to cultivate or disturb the soil.

8.2.4 Soil Sampling and Carbon Analysis

At the start of the experiment soil samples were taken immediately before the treatments were applied. These were from three locations in each test strip or plot; at each location soil was sampled in two depth increments (0–10 cm and 10–30 cm).

Soil samples were air dried and ground to pass a size of less than 2 mm. Each sample was analyzed using mid infra-red (MIR) spectroscopy [18] to determine the total amount of organic carbon in the soil and the particulate, humic and recalcitrant fractions of carbon.

8.2.5 Crop Yield

The yield of each treatment strip was estimated as the average value of the yield measured by the yield monitor mounted on the harvester.

8.2.6 Analysis of Treatment Effects

At each location the effect of treatments on total soil carbon and its different fractions and crop yield were analysed using a linear mixed model asreml [19] using the R package "asreml" [20]. Treatments and soil depth were considered as fixed effects and blocking design as a random effect.

8.3 Results

8.3.1 Soil Carbon

The treatments had little effect on the concentration of total organic carbon or any of its three fractions soil C fraction in any given year in the experiment at either 0–10 cm or 10–30 cm depths (Fig. 8.6). The concentration of soil carbon was lower in the deeper soil. There was, however, a trend for each of the fractions of carbon

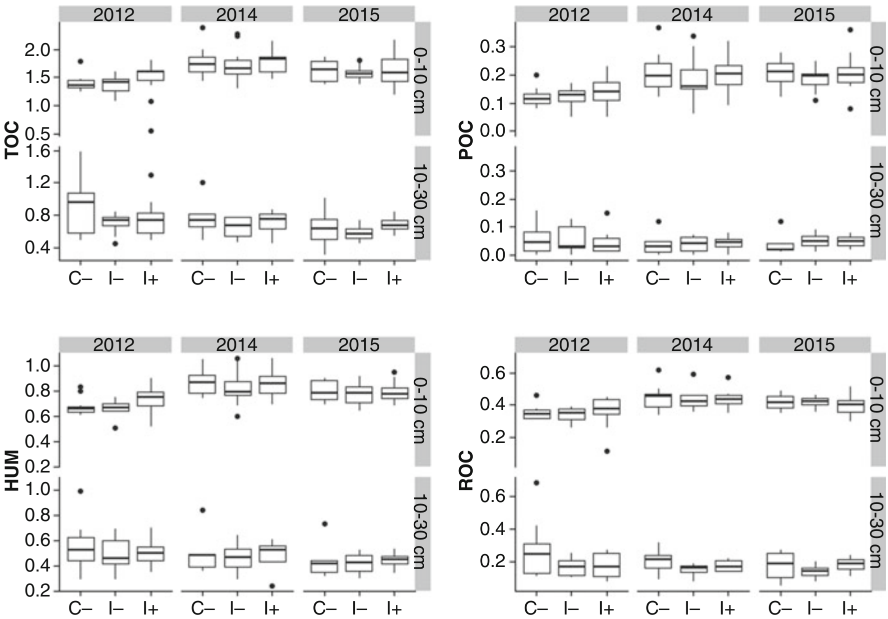


Fig. 8.6 The total organic (TOC), particulate organic (POC), humic (HUM) and recalcitrant (ROC) fractions of soil carbon (%) measured on three occasions at two depths (0–10 cm and 10–30 cm)

to increase in 2013 and 2014. These differences were tested formally by applying linear mixed models to each year with treatment and depth as fixed effects and block as a random effect. These models showed that the incorporation treatments had no significant effect on the stock of total organic carbon (TOC) or any of it components (POC, HUM or ROC) in either 2013 or 2014. However, in all cases TOC and its' fractions decreased significantly ($p < 0.001$) with depth. There is a substantial body of evidence that suggests carbon stocks decrease in the soil used to grow crops and that the most successful way to increase soil carbon is to grow pasture. The values we measured here are typical of the cropping phase of mixed farming rotations [21, 22].

8.3.2 Yield

Crop yield was much lower in 2014 with the two incorporation treatments (I- and I+) yielding less than the control treatment in 2014 (Fig. 8.7). This was confirmed by a linear mixed model fitted to the yield data for each year, with treatment as the fixed term and block as the random effect. This modelling showed no significant effects of treatment in 2013 but a highly significant difference (p = <

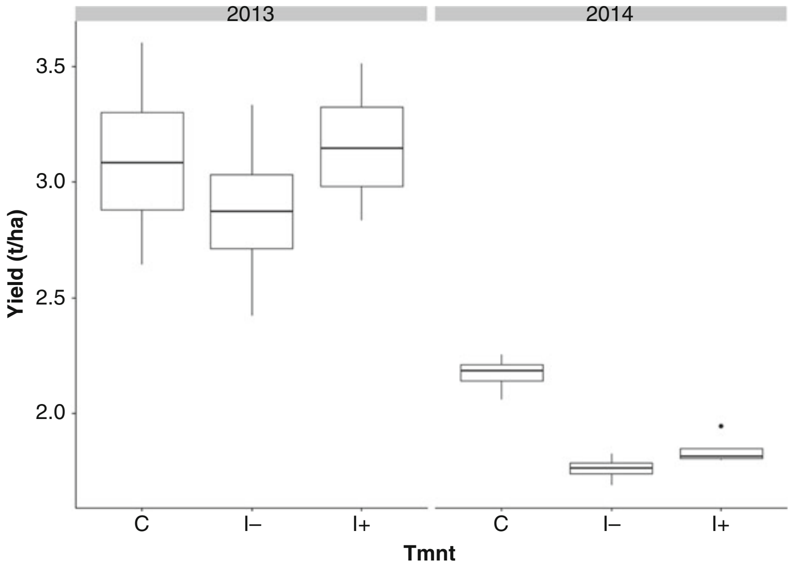


Fig. 8.7 Crop yield (t/ha) measured in 2014 and 2015

0.001) between the control and the I- and the I+ treatments, the was no significant difference between the I- and I+ treatments in 2014. This suggests that here may be a deleterious effect of incorporation on yield.

8.4 Discussion

Close examination of the Kirkby et al. [4] experiment reveals that carbon was lost in the top 10 cm of soil and that sequestration was only achieved when it was calculated over a profile depth of 1.5 m. This is much deeper than the 30 cm suggested for carbon trading [1]. The degree of carbon sequestrated in our 'farmer practice' experiments was unlikely to approach that of Kirkby et al. [4] as our field practice differed from the original the experimental protocol which had the following extra steps to maximize the sequestration of carbon:

1. The fertiliser providing the nutrients was ground to a fine powder before application;
2. Crop residues were pulverized with a flail mower;
3. Fertiliser (nutrients) was spread immediately before 10 mm rain or more was forecast; and
4. A rotary hoe was used to incorporate the fertilizer and crop residues into the soil.

Three of these practices (1,2 and 4) are energy intensive and if subjected to lifecycle analysis [23] this process is unlikely to be carbon neutral and may even consume more carbon than is sequestered.

Geostatisical methods have been applied in farmer-led experiments to analyse the yield response to rates of fertiliser or fungicide [24]. Here geostatistical methods were combined with EM techniques to improve the design of such experiments, which were analysed using traditional statistical methods. Applying geostatisitcal methods to the analysis [for example, 25] would further improve the precision of similar experiments by accounting for spatial trend.

8.5 Conclusions

Our experiment was established as part of an initiative under the Australian Federal Government's climate mitigation scheme. The aim of this Action on the Ground initiative was to demonstrate to farmers the possibility of sequestering soil carbon. Our approach not only allowed the demonstration of practice, but also the rigorous measurement of the success of the farmers' actions in increasing the amount of carbon in the soil. We showed that the suggested practice did not increase soil carbon and that this aspect of the policy was unlikely to achieve its stated goals.

The practice we trialed failed to either sequester soil carbon or increase crop yield. The costs of additional nutrients and incorporation would have a negative impact on farm returns and profitability.

References

1. Australian Ferderal Department of Agriculture: The carbon farming initiative (2019). Available from: http://www.agriculture.gov.au/water/policy/carbon-farming-initiative
2. Australian Government – Department of Agriculture. Action on the ground. 2017.; Available from: http://www.agriculture.gov.au/ag-farm-food/climatechange/carbonfarmingfutures/action-on-the-ground
3. Baldock, J.A., et al.: Quantifying the allocation of soil organic carbon to biologically significant fractions. Soil Res. **51**(8), 561 (2013)
4. Kirkby, C.A., et al.: Inorganic nutrients increase humification efficiency and C- sequestration in an annually cropped soil. PLoS One. **11**(5), e0153698 (2016)
5. McNeill, J.D.: TN 5; Electrical conductivity of soils and rocks, p. 20. Geonics Ltd, Mississauga (1980)
6. McNeill, J.: TN-6 Electromagnetic terrain conductivity measurement at low induction numbers. Geonics Ltd, Mississauga (1980)
7. Heil, K., Schmidhalter, U.: The application of EM38: determination of soil parameters, selection of soil sampling points and use in agriculture and archaeology. Sensors (Basel). **17**(11), (2017)
8. Doolittle, J.A., Brevik, E.C.: The use of electromagnetic induction techniques in soils studies. Geoderma. **223**, 33–45 (2014)

9. Corwin, D., Lesch, S.: Apparent soil electrical conductivity measurements in agriculture. Comput. Electron. Agric. **46**(1), 11–43 (2005)
10. Pebesma, E.J.: gstat: spatial and spatio-temporal geostatistical modelling, prediction and simulation. R package version 1.1.6 (2018)
11. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2014)
12. Cressie, N.C.A.: Statistics for spatial data. Wiley (1993)
13. Jenks, G.F., Caspall, F.C.: Error on choroplethic maps: definition, measurement, and A.o.A.G. reduction. Annals. **61**(2), 217–244 (1971). Error on choroplethic maps: definition, measurement, reduction. Annals, Association of American Geographers, 1971. 61: p. 217–244
14. Hartigan, J.A.: Clustering algorithms. Wiley, New York (1975)
15. McLachlan, G.J., Peel, D.: Finite mixture models. Wiley, New York (2000)
16. Meng, X.-L., Rubin, D.B.: Maximim liklehood estimation via the ECM algorithm: a general framework. Biometrika. **80**(2), 267–278 (1993)
17. Benaglia, T., et al.: mixtools: An R package for analyzing finite mixture models. J. Stat. (2009)
18. Baldock, J.A., et al.: Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. Soil Res. **51**(8), 577 (2013)
19. Gilmour, A.R., Thompson, R., Cullis, B.R.: AI, an efficient algorithm for REML estimation in linear mixed models. Biometrics. **51**, 1440–1450 (1995)
20. Butler, D., asreml: fits the linear mixed model. 2019.
21. Chan, K.Y., et al.: Soil carbon dynamics under different cropping and pasture management in temperate Australia: results of three long-term experiments. Soil Res. **49**(4), 320–328 (2011)
22. Viscarra Rossel, R.A., et al.: Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang. Biol. **20**(9), 2953–2970 (2014)
23. Hauschild, M., et al.: Life cycle assessment theory and practice. Springer (2018)
24. Rudolph, S., et al.: Spatial discontinuity analysis, a novel geostatistical algorithm for on-farm experimentation. In: 13th international conference on precision agriculture. International Society for Precision Agriculture, St Lois (2016)
25. Marchant, B., et al.: Establishing the precision and robustness of farmers' crop experiments. Field Crop. Res. **230**, 31–45 (2019)

Chapter 9
Nutrient Loading in the River Systems Around Major Cities in Bangladesh: A Quantitative Estimate with Consequences and Potential Recycling Options

Shamim Mia, Md. Rushna Alam, Md. Abdus Sattar, and Feike A. Dijkstra

Abstract Biological organisms including humans require mineral nutrients for their growth and development. A significant amount of these nutrients remain unused in the left over materials, known as waste, causing environmental degradation. These nutrients could potentially be a resource for agriculture if recycled and reused. Therefore, a critical examination of nutrient loading from waste such as from human excretion and biowaste it required. Here, we estimated the nutrient loading from municipal organic waste (MOW) and human excreta using linear modelling, explored the potential consequences to ecosystem services and proposed several management strategies. Waste and human excreta generation were calculated from the per capita generation rate while nutrient concentrations were considered as the average of literature values. The daily carbon (C), nitrogen (N), phosphorus (P) and potassium (K) loading from MOW, urine and faeces to water bodies around the major cities in Bangladesh were estimated at 2158, 112, 58 and 91 t, respectively. The nutrient loading to water bodies reduces several ecosystem service

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-981-15-1735-8_9) contains supplementary material, which is available to authorized users.

S. Mia (✉)
Department of Agronomy, Patuakhali Science and Technology University, Patuakhali, Bangladesh

Centre for Carbon, Water and Food, The University of Sydney, Camperdown, NSW, Australia
e-mail: shamim.mia@sydney.edu.au

Md. R. Alam
Department of Aquaculture, Patuakhali Science and Technology University, Patuakhali, Bangladesh

Md. A. Sattar
Department of Disaster Risk Management, Patuakhali Science and Technology University, Patuakhali, Bangladesh

F. A. Dijkstra
Centre for Carbon, Water and Soil, The University of Sydney, Camperdown, NSW, Australia

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_9

with estimated cost of BDT. 42.93 million for reduction in fish production and half of property rent. Combined pyrolysis and composting of municipal organic waste can reduce the volume of organic waste. Similarly, pyrolysis of faeces and then sorption of nutrients using the produced biochar could recycle 440 t C, 71 t N, 113 t P and 73 t K day^{-1}. Combing the nutrient recycling through pyrolysis-composting of MOW and faeces pyrolysis-sorption of nutrients from urine, the estimated nutrient recycling was 39% C, 11% N, 68% P and 11% K of the intrinsic nutrients in waste. Altogether, our study provides a comprehensive understanding of nutrient loading, its consequences and potential recycling options that may help to attain environmental sustainability.

Keywords Biowaste · Human excreta · Ecosystem services · Environmental sustainability

9.1 Introduction

Biological organisms including humans acquire mineral nutrients from food for their growth and development. A significant amount of these nutrients remain unused in the left over materials, known as waste and also in metabolic residues, *i.e.*, urine and faeces. The bio-waste is usually combined with other waste and disposed of while the urine and faeces are fermented in safety tanks that are connected to municipal drainage channels [1, 2]. A sustainable management of waste is a prerequisite for healthy living in a city, which however, often is a challenging task to perform.

Every day, residents in the municipal areas are producing waste with a large share (~75%) of bio-waste/organic waste [3]. When these wastes from our daily household, commercial, industrial and agricultural activities are aggregated, they often each a size beyond the handling capacity of municipalities [4]. Similar to global trends, the waste generation in Bangladesh was projected to increase in the future due to an increase of urban population, shifting life styles, technological developments and growing consumerisms. The metabolic residues (urine and faeces) are also increasing with an increase of urban population. An increase in national income or GDP might accelerate the waste generation [3] suggesting that developing countries will face additional challenges in handling larger volumes of waste. Consequently, sustainable municipal waste (MW) and metabolic residue management are getting and will remain tougher for municipalities of developing countries like in Bangladesh [5]. Reasons include, but not limited to, a large volume of waste generation, lack of proper technologies and policies, non-compliance of citizens to waste sorting and disposal, and limited capacities of the municipalities [4].

Biowaste and metabolic residue carry nutrients including carbon (C), nitrogen (N), phosphorus (P), potassium (K) and other micronutrients, that are basically mined from agricultural soils. These nutrients are released by decomposition and drained to the river system through sewerage channels. Little is known how much nutrients are lost from biowaste and human excretion from municipal areas to around water bodies, although previous studies provided a range of concentrations. For example, municipal wastewater may contain 15–90 mg L^{-1} total N and 5–20 mg L^{-1} total P, respectively.

Many previous studies have reported that excessive nutrients in aquatic ecosystems can lead to increased eutrophication in water bodies ranging from small ponds to coastal waters [6–10]. It is evident that excessive nutrient loading from different point and non-point sources to the small lakes, rivers and the sea have serious adverse impacts on ecosystems services since these nutrient loadings cause imbalances in the nutrient stoichiometry. As a result, the growth of harmful algal blooms increases rapidly, which ultimately affects aquatic ecosystem services [11–14]. Moreover, the decomposition of labile organic C released to water bodies often reduces the dissolved oxygen concentration in water bodies. Such deterioration of aquatic environmental health and quality around major cities is a common phenomenon in south Asia [15, 16]. For example, water quality of rivers surrounding major cities of Bangladesh, as in Dhaka, has declined for various reasons including high concentration of nutrients, creating hypoxic and anoxic conditions, turbidity increment, and bad water smell [16–20]. Consequently, death or limited presence of aquatic flora and fauna are often reported. As far as we know, there is no study that estimates the consequence of nutrient loading from biowaste and human excreta to water bodies around major cities in Bangladesh.

Three R (reduce, recycle and reuse) is considered to be one of the means for attaining sustainability [21]. Recycling of nutrients from bio-waste and metabolic residues can be a potential source of plant nutrients, if they are recycled. Studies suggested that recycling of organic waste could help to compensate future P depletion as human faeces can meet 28% of its total global demand [22]. A number of recycling strategies including composting, nutrient sorption and precipitation, phyto-recycling can be used, either singly or in combination, for sustainable recycling of these nutrients [23–30]. However, in Bangladesh, there is no known study that looked into recycling options. Considering altogether, we estimated nutrient loading from biowaste and human excreta to water bodies and suggested recycling options with quantitative estimated of recyclable nutrients.

9.2 Materials and Methods

9.2.1 Data Collection

We made a rigorous literature collection using Scopus and Google scholar. The collected literature was then synthesized for calculating nutrient loading, determination of its consequences and finding potential recycling options (see the detailed list in SI Table 1). For exploring the waste disposal in Bangladesh we relied on literature data while human excreta disposal was explored by personal communication to waste management personnel, known as Key Informant Interview (KII).

9.2.2 Calculation of Nutrient Loading

Nutrient loading to water bodies was calculated separately from organic waste and human excreta detailed as follows.

A. **Nutrient Loading from Organic Waste**

Nutrient loading from organic waste was calculated following linear modelling.

$$\text{Nutrient loading}\ \left(\text{kg day}^{-1}\right)$$
$$= \text{P} \times \text{PWGR}\ \left(\text{kg day}^{-1}\right) \times \text{FOW} \times \text{FWCR} \times \text{FMA} \times \text{NC}$$

Where

P = Population of the city (World population review, 2018)
PWGR = Per capita waste generation rate (kg day^{-1}) (adjusted from [3])
FOW = Fraction of organic waste (average of literature data [3])
FWCR = Fraction of waste considered to release in the water stream (10%, based on expert judgement)
FMA = Moisture adjustment factor (70% moisture [3])
NC = Nutrient content in the biomass (average of literature data [3])

B. **Nutrient loading from urine and faeces**

We calculated potential nutrient loading from human excreta using the following equation-

$$\text{Nutrient loading}\ \left(\text{kg day}^{-1}\right) = \text{P} \times \text{PWGR}\ \left(\text{kg day}^{-1}\right) \times \text{FWCR} \times \text{NC}$$

Where

P = Population of the city (World population review, 2018)
PWGR = Per capita waste (urine/faeces) generation rate (kg day^{-1}) in dry weight basis (average of relevant literature data)
NC = Nutrient content in the urine/faeces (average of relevant literature data)

The average value of literature data was calculated by Gaussian model fitting to the literature data (Sigma Plot 11.0, Eq. 1) while the simple average was used when the number of observations was small. The Gaussian equation is as follows-

$$\text{y} = \text{a}e^{\left\{\frac{0.5(x-xo)^2}{b}\right\}} \tag{1}$$

9.2.3 Determination of Environmental Consequences of Nutrient Loading

In this study, we analysed the potential consequences of nutrient loadings on the ecosystem services in the river Buriganga near the capital city of Bangladesh. Based on literature data and expert opinion, we identified major ecosystem services of the river, which are under threat. The ecosystem services were then, categorized under three headings, i.e., direct service, indirect service, and optional service. Next, the consequences of nutrient loading to these services were listed and prioritized. Impact of nutrient loading on two of the ecosystem services, i.e., fish production and property rent were estimated quantitatively, discussed in the next sections.

9.2.3.1 Fish Production Calculation

For estimating economic value of potential fish production per year in the Buriganga river, we applied a benefit transfer (BT) approach, a widely used method in situations where time and financial resources are limiting [31, 32]. By this method, valuation of primary non-market services from one location can be transferred to another location [33]. In this case, average fish production in others rivers (318 kg ha^{-1}) in Bangladesh was used to estimate total fish production potential of the river Buriganga while price of fish was grossly considered as BDT. 300, irrespective of fish types [34].

9.2.3.2 Property Rent Calculation

We applied willing to pay (WTP) method to estimate monetary value of property rent. To estimate the impacts of nutrient loading on the rent of a flat/house, we conducted a short online survey using social media, Facebook. A total of 72 respondents were participated on this survey. This survey consisted of two questions, i.e., (a) whether the pollution of the Buriganga river had any impacts on flat/house rent, and (b) how much did a respondent want to pay for a property currently renting at ten thousand BDT, if the river would have been nice and clean.

9.2.4 Determination of Potential Recycling Options

Literature data were synthesized to find suitable recycling options from biowaste and human excreta. Emphasis was paid to harvest synergies through combining multiple options together. We, then, estimated the recyclable nutrients for one of the feasible options, *i.e.*, simultaneous pyrolysis and composting of biowaste while pyrolysis of faeces and use of the produced biochar as sorbent for harvesting nutrients from urine.

9.2.5 Calculation of Nutrient Recycling Potential Through Pyrolysis and Co-composting of Organic Waste

The nutrient recycling potentiality of municipal organic was calculated using the proposed method of Mia et al. [3] that summarized the literature data for calculating nutrient recycling.

9.2.6 Calculation of Nutrient Recycling Potential Through Pyrolysis of Faeces and Use of Biochar as Sorbent to Urine

The potentiality of nutrient recycling from faeces through pyrolysis was calculated based on literature data [35–38]. The average biochar production from faeces and other waste bio-solid were considered for calculating biochar production rate while the average concentration of C, N, P and K in the biochar was used for calculating nutrient content in the faeces biochar. The potentiality of faeces biochar to adsorb nutrients from urine was calculated using the average value of Gaussian model fitting to the Langmuir adsorption maxima of biochar (Sigma Plot, Fig. 9.1). Since there were limited studies on K sorption onto biochar, we considered that K sorption would be similar to ammonium sorption.

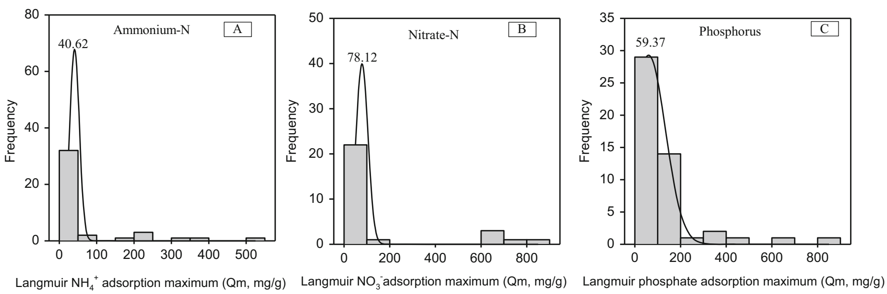


Fig. 9.1 Gaussian model fitting to the literature data. Panel A, B and C represent ammonium-N, nitrate-N and phosphate adsorption maxima, respectively

9.3 Results and Discussion

9.3.1 Waste Disposal Systems in Bangladesh

In Bangladesh, half of waste generated in the urban areas is collected while the rest remains uncollected (Fig. 9.1). The waste collection effort was intensified by collecting waste directly from home by waste collectors, who are paid by the residents. These wastes are then sorted. The biowaste is stored in the temporary waste disposal stations while the recyclable waste (specifically plastic, paper and irons etc.) are sold. The collected biowaste is generally used for land filling (open dumping). Therefore, a fraction of the land filled waste decomposes in the open places and the released nutrients are then mixed with drainage water, which are ultimately loaded in the river water. We estimated this fraction as ~10% of the organic waste. The uncollected waste decomposes in the roads, drainage channels and ultimately goes to water bodies (Fig. 9.2).

9.3.2 Human Excreta Disposal Systems in Bangladesh

According to KII, expert judgement and secondary information, there are two basic human excreta disposal systems in cities of Bangladesh (Fig. 9.3). On-site system in which toilets are connected to a septic tank, and off-site system where toilets are directly connected to a sewerage network. In off-site systems, sludge and wastewater are not separated from the sources but directly drain to the sewerage system. A part of the sewerage waste is being treated in few cities (as in case of Dhaka at Pagla sewage treatment plant). However, little effort is usually made to harvest nutrients from wastewater. Therefore, the nutrients ultimately drain to nearby river channels. Meanwhile, in the on-site system, initially waste is kept in

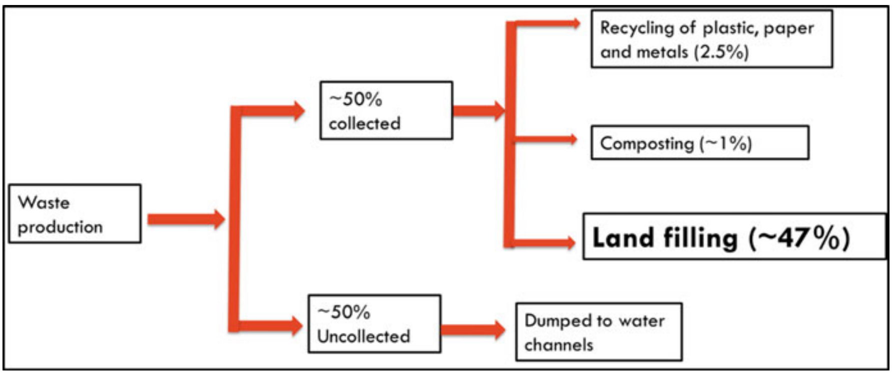


Fig. 9.2 Municipal organic waste management systems in Bangladesh. (Adapted from [3])

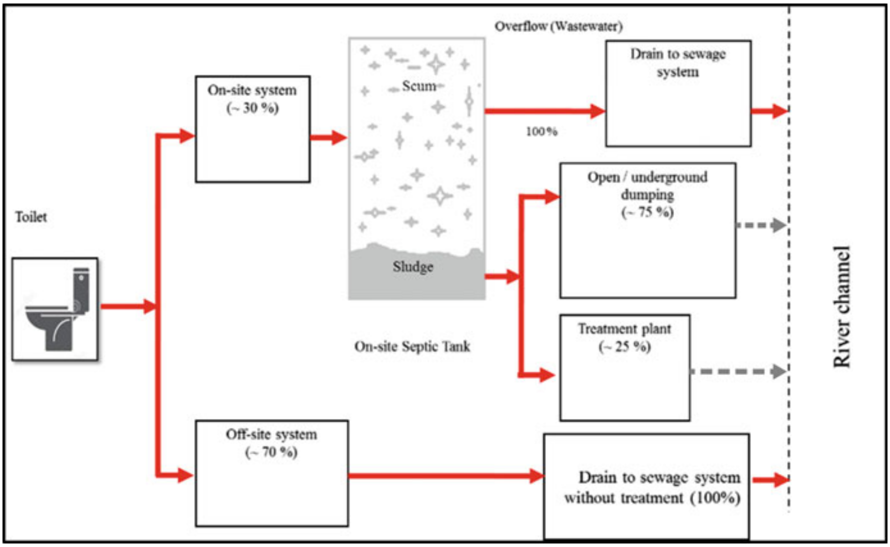


Fig. 9.3 Human excreta management systems in cities of Bangladesh (Based on KII and Expert judgement). The arrows and boxes indicate the current disposal system while filled and dotted arrows represent direct and indirect excreta transportation route

septic tanks where it undergoes anaerobic fermentation. The solid remains inside the tank while liquid often over flows to the drain carrying nutrients. A fraction of sludge is transported via trucks to the treatment plant while the majority is directly dumped into open/underground places. One good example can be found in Faridpur municipality where fecal waste is composted to recycle nutrients.

A large fraction of city dwellers, as high as 55%, specifically living in the urban slums and streets, do not have access to sanitation facilities [39–41]. Therefore, they use open places for excretion. These wastes directly drain to water channels. Therefore, these nutrients are also released to water bodies. Altogether, nutrients from most of human excretion are drained into surrounding water bodies unless lost through gaseous processes or sorbed to soil and fauna along the path of their journey.

9.3.3 Nutrient Loading from Organic Waste Disposal

The nutrient loading from different cities varied widely due to large variation associated with size of population, per capita waste generation rate, fraction of organic waste in the total generated waste. For example, the largest amount of nutrients *i.e.*, C, N, P and K was 314,808, 12,382, 4757 and 12,802 kg day^{-1} respectively in Dhaka while these contributions were in the order of Chittagong, Khulna, Rajshahi, Gazipur, Narayanganj, Comilla, Rangpur, Sylhet, Barishal, and

Mymensingh. However, the contributions of all municipalities combined was higher than Dhaka City Corporation. Combining all cities and municipalities, the daily total nutrient loading from biowaste to water bodies around major cities in Bangladesh was estimate at 882,340 kg C, 34,705 kg N, 13,333 kg P and 35,882 kg K (Table 9.1).

9.3.4 *Nutrient Loading from Human Excreta Disposal*

The human excreta, i.e., urine and faeces generation was different for different cities with the largest production in the Dhaka City Corporation estimating at 569,608 kg dry urine day^{-1} and 517,825 kg dry faeces day^{-1} (Table 9.2). Consequently, C, N, P and K loading from human excreta also varied for different cities. The largest amount of nutrient loading was estimated from Dhaka, which was followed by Chittagong, Khulan, Rajshahi, Comilla, Rangpur, Gazipur, Sylhet, Mymensingh, Narayanganj and Barishal. The variation is primarily due to the size of the urban population. Combining the nutrient loading of all urban areas together, the daily C, N, P and K loading from human excreta to water bodies were estimated at 1,275,862 kg C, 76,985 kg N, 44,827 P and 55,546 kg, respectively. While nutrient loading from organic waste and human excreta combined, C, N, P and K loading to water bodies around cities were estimated at 2,158,202, 112,452, 58,160 and 91,428 kg day^{-1}.

9.3.5 *Impacts of Nutrient Loading on Aquatic Ecosystems*

Water bodies around cities provide ecosystem functions although these service can often be threatened due to pollution including nutrient loading [42, 43]. The threatened services are identified based on published reports and the authors' judgment (Fig. 9.4). Among direct services of an aquatic ecosystem, reduction of property price and rent is one of the dominant loss of services. As opinion received by online survey, 100% of the respondents believe that river pollution could negatively impact nearby house or property rent. Moreover, the majority of the respondents (86%) wanted to pay an additional rent between 4% and 316% (on average, 54%) of current rent if the river would have not been polluted. Specifically, 23 out of 72 respondents were intended to pay 11–25% extra, while 14 of them wished to pay more than 70% of the current rent, if the river would have been pollution free (Fig. 9.5). Not surprisingly, people always prefer a place for living that is not polluted. Nevertheless, the payment rates may depend on various socio-economic factors, for instance, income, education level, awareness of the environment.

Table 9.1 Nutrient loading from municipal organic waste at different cities in Bangladesh

Name of the City	Population of city ("000")	Waste generation Rate (Kg day^{-1} $person^{-1}$)	Fraction of organic waste (%)	Organic waste generation (kg day^{-1})	Nutrient loading (kg day^{-1})			
					C	N	P	K
Dhaka	10,357	0.56	73.14	4,239,841	314,808	12,382	4757	12,802
Gazipur	338	0.56	73.14	138,202	10,261	404	155	417
Narayanganj	224	0.56	73.14	91,548	6797	267	103	276
Rajshahi	700	0.36	78.65	199,975	14,848	584	224	604
Rangpur	343	0.27	82.62	75,923	5637	222	85	229
Barisal	202	0.31	80.77	50,310	3736	147	56	152
Chittagong	3920	0.45	71.92	1,268,489	94,185	3705	1423	3830
Khulna	1342	0.33	79.42	351,773	26,119	1027	395	1062
Sylhet	237	0.39	73.49	68,102	5057	199	76	206
Comilla	389	0.27	80.77	84,236	6254	246	95	254
Mymensingh	225	0.27	82.62	49,815	3699	145	56	150
Other cities	30,448	0.21	80.72	5,265,702	390,978	15,378	5908	15,900
Country total	**48,725**	**0.32**	**76.63**	**11,883,370**	**882,340**	**34,705**	**13,333**	**35,882**

Carbon (C), nitrogen (N), phosphorus (P) and potassium (P) concentration in dry organic waste was considered as 45%, 1.77%, 0.68% and 1.83% respectively. These figure were adopted from a recent review article that used average value of literature data [3]

Table 9.2 Nutrient loading from human faeces and urine at different cities in Bangladesh

	Human excreta generation (kg day^{-1})[a]		Nutrient loading (kg day^{-1})[b]			
Name of the City	Urine	Faeces	Carbon	Nitrogen	Phosphorus	Potassium
Dhaka	569,608	517,825	271,185	313,300	29,749	173,518
Gazipur	18,567	16,879	8840	533	311	385
Narayanganj	12,299	11,181	5855	353	206	255
Rajshahi	38,507	35,007	18,333	1106	644	798
Rangpur	18,872	17,156	8985	542	316	391
Barishal	11,123	10,112	5296	320	186	231
Chittagong	215,612	196,011	102,651	6194	3607	4469
Khulna	73,829	67,117	35,149	2121	1235	1530
Sylhet	13,035	11,850	6206	374	218	270
Comilla	21,418	19,471	10,197	615	358	444
Mymensingh	12,382	11,257	5895	356	207	257
Other cities	1,674,619	1,522,381	797,271	48,107	28,012	34,710
Country total	**2,679,870**	**2,436,246**	**1,275,862**	**76,985**	**44,827**	**55,546**

[a]Faeces and urine generation rates were considered as 0.050 kg dry faeces day^{-1} and 0.055 moisture free urine kg day^{-1}
[b]Carbon, nitrogen, phosphorus and potassium concentration in faeces and urine were considered as 45%, 3.16%, 1.84%, 2.28% and 6.7%, 52.13%, 3.55%, 28.39%, respectively. The nutrient concentration in urine and faeces was calculated as average of relevant literature data (see SI for detailed calculations)

Fish production can be considered as one of the dominant services of an aquatic ecosystem in Bangladesh. However, river water pollution including nutrient loading from the dumping of both biowaste and human excreta resulted in a drastic fall of fish production. In extreme cases during winter season, fish can be hardly found. Similar to fish species, other species were reported to decline. This study estimated that the total annual fish production potential is 143.1 tons pricing at BDT. 42.93 million, which was estimated to be completely destroyed due to pollution [44]. With a study on the river Buriganga it was shown that the annual return from fish production could be BDT. 468 million if the restoration program would have taken place. In addition to these, pollution including the dumping of biowaste (specifically while decomposing) causes serious problems to boat and ship navigation.

9.3.6 Proposed Recycling Options

Considering the context of organic waste and human excreta disposal systems, we propose that the organic waste can potentially be recycled and reused through sorting into different fractions such as (a) pyrolysable- consisting low moisture and high lignin bio-waste, (b) compostable-easily decomposable biowaste with high moisture content (Fig. 9.6). These two methods can be run simultaneously for

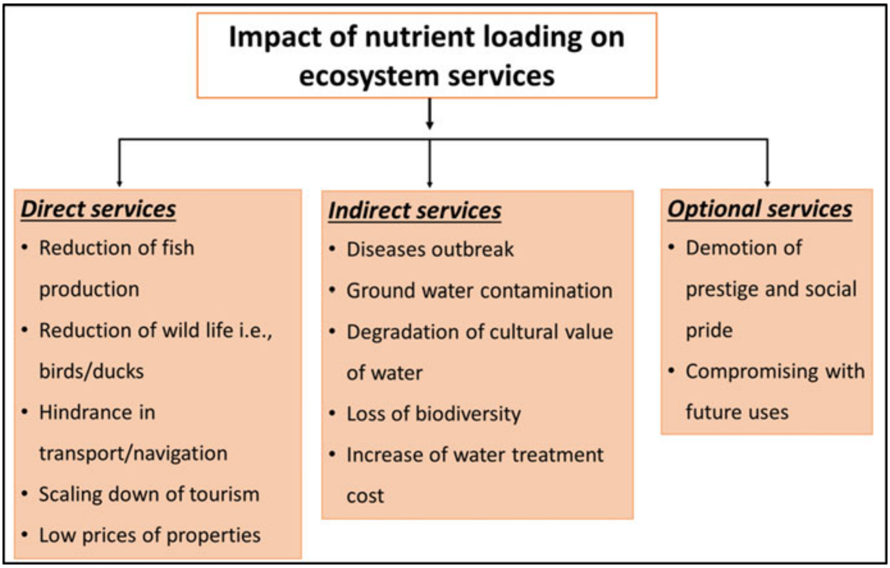


Fig. 9.4 Effect of nutrient loading on ecosystem services of the Buriganga River in Bangladesh. The scheme was proposed based in literature survey and expert judgement. The list of literature used for this graph can be found in Table S1

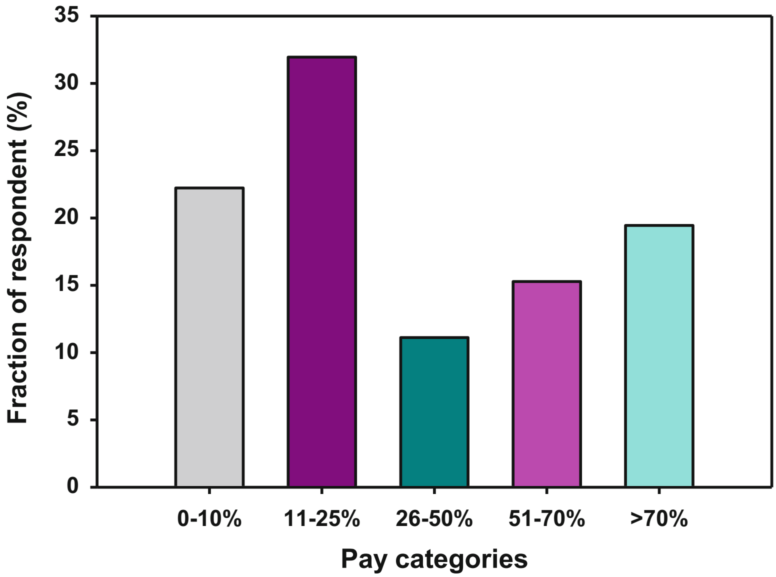


Fig. 9.5 Respondents categories based on fraction (%) of additional rent willing to pay. Data are based on author's online survey

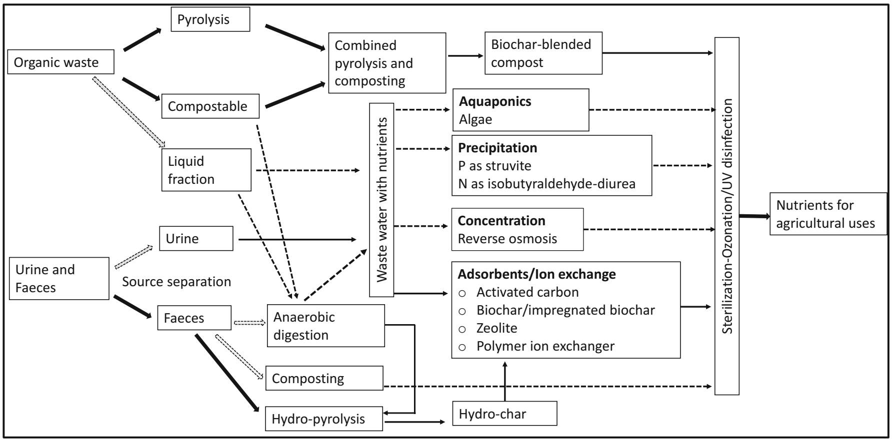


Fig. 9.6 Proposed methods of nutrient recycling from municipal organic waste and human faeces and urine. The solid lines indicate potential methods for Bangladesh while the dotted lines indicate possible alternative options. The scheme is proposed based on literature data and expert judgement. Literature used in this graph can be found in the Table S1

harvesting synergistic benefits [3]. For harvesting nutrients from human excreta, we propose to pyrolysize the bio-solids of faeces after removal of excess moisture and to harvest nutrients from wastewater including urine using different techniques including phyto-accumulation, precipitation, and adsorption. Pyrolysis is relatively a shorter process compared to other available options, e.g. composting and anaerobic digestion while it removes pathogens and anti-biotics. Adsorption could be one of the simple techniques to use in Bangladesh. The produced biochar from fecal pyrolysis can be used as a sorbent for nutrient harvesting from urine. Sorbents can directly be used in the sources of nutrient release. For example, use of sorbents in the outlet of public toilets in cities could harvest a significant amount N from urine since the N concentration is relatively higher in public toilets than wastewater. The nutrient harvesting efficiency will depend on the concentration of nutrients, efficacy of sorbent, sorbent to solution ratio and chemical environment. It is also possible to tailor sorbents that have the capacity to adsorb both cations and anions. Moreover, the liquid and labile fraction can be digested anaerobically to generate methane for getting energy. Nutrient harvested through different means such as biochar-compost and nutrient loaded biochar/hydrochar/zeolite need to be sterilized for removing any harmful organisms including *Escherichia coli*.

9.3.7 Quantitative Estimate of Nutrient Recycling Using Pyrolysis of Faeces and Sorption of Nutrients from Wastewater with Biochar

Nutrient recycling potential using pyrolysis of faeces is presented in Fig. 9.7. Combining all the municipalities and cities together, the daily carbon recycling potentiality was estimated at ~394 t while the N, P and K recycling potential was estimated at 33, 41 and 19 t day$^{-1.}$

When the produced biochar would have been used in the wastewater containing urine, the daily nutrient cycling potential was estimated at either of 46 t C, 39 t NH_4^+-N, 71 t NO_3^--N, or 54 t P (Fig. 9.8). Considering a similar percentage of K recovery to NH_4^+, 39 t K could be recycled (Fig. 9.8). However, it is unknown how much nutrients could be recycled when all these nutrients are aimed to harvest simultaneously. The recycling potential of the produced biochar was lower than the nutrient loading from urine, estimated at only 8%, 57% and 5% in case of N, P and K, respectively. Combing the nutrient recycling through pyrolysis-composting and pyrolysis-sorption based recycling, 39% C, 11% N, 68% P and 11% K of the nutrients loaded to water bodies around the cities could have been recovered for agricultural uses (Fig. 9.9).

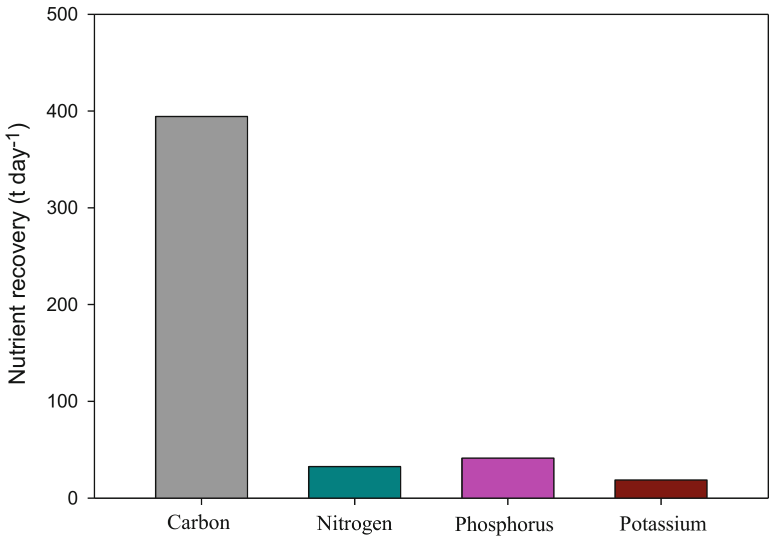


Fig. 9.7 Nutrient recycling potential through pyrolysis of faeces. The graph was made based on the average production of biochar from bio-solids including faeces and nutrient concentration in produced biochar. (Data were collected from [35–38])

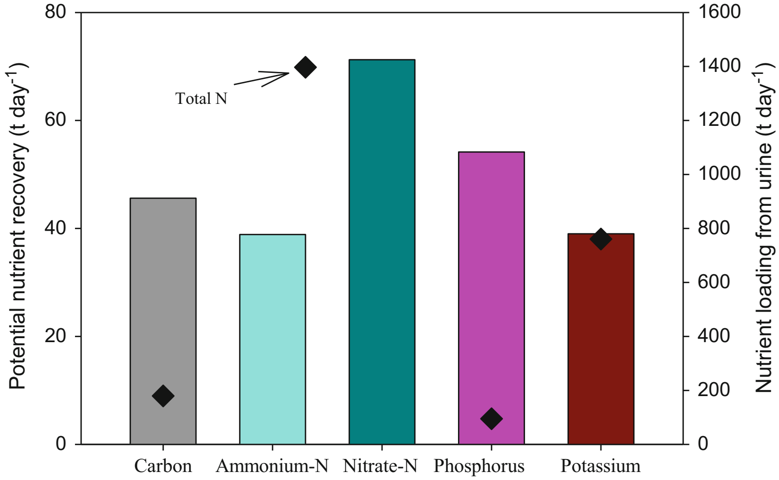


Fig. 9.8 Nutrient recycling potentiality of faeces biochar when applied to urine. The literature used to produce this graph in listed in Fig. 9.1

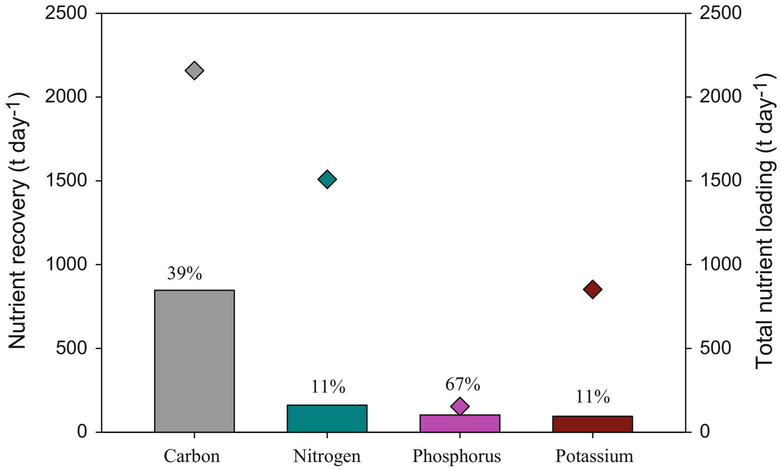


Fig. 9.9 Total nutrient loading and potential recovery with the proposed methods

9.3.8 Limitation of the Study

We acknowledge that the reduction in ecosystem services was possibly caused by combined pollution of heavy metals, household and industrial waste disposal and also petroleum discharge from ships, launches, cargoes, boats, etc. Since there were no studies on the effect of nutrient loading on ecosystem services in Bangladesh, it was not possible to separate individual effects. Albeit efforts were made to use best possible estimates, our model based estimate of nutrient loading and recycling may carry some uncertainties.

9.4 Conclusions and Future Direction of Research

Nutrient recycling from waste materials including human excreta remain one of the prime potential alternatives to chemical fertilizers. However, challenges associated with recycling of nutrients from waste are diversified including suitable technologies and pathogenic and heavy metal removal. In this study, we provided combined pyrolysis-composting and pyrolysis-sorption based recycling options with quantitative estimates of recyclable nutrients. Our proposed model could recycle 39% C, 11%N, 68%P and 11%K of the nutrients loaded to water bodies around the cities. Therefore, the proposed method could have significant impact in recycling nutrients to agriculture while improving environmental quality.

References

1. Mahmud, A.H.: City waste management: only 51% collected for disposal, p. 5. Dhaka Tribune, Dhaka, Bangladesh (2014)
2. Yousuf, T.B., Rahman, M.: Monitoring quantity and characteristics of municipal solid waste in Dhaka City. Environ. Monit. Assess. **135**(1–3), 3–11 (2007)
3. Mia, S., Uddin, M.E., Kader, M.A., Ahsan, A., Mannan, M.A., Hossain, M.M., Solaiman, Z.M.: Pyrolysis and co-composting of municipal organic waste in Bangladesh: A quantitative estimate of recyclable nutrients, greenhouse gas emissions, and economic benefits. Waste Manag. **75**, 503–513 (2018)
4. Bhuiyan, S.H.: A crisis in governance: urban solid waste management in Bangladesh. Habitat Int. **34**(1), 125–133 (2010)
5. Omari, A., Said, M., Njau, K., John, G., Mtui, P.: Energy recovery routes from municipal solid waste, A case study of Arusha-Tanzania. J. Energy Techno Pol. **4**(5), 1–8 (2014)
6. Seitzinger, S.P., Mayorga, E., Bouwman, A.F., Kroeze, C., Beusen, A.H.W., Billen, G., Van Drecht, G., Dumont, E., Fekete, B.M., Garnier, J., et al.: Global river nutrient export: a scenario analysis of past and future trends. Global Biogeochem. Cycles. **24**(2), (2010)
7. Nixon, S.W.: Coastal marine eutrophication: a definition, social causes, and future concerns. Ophelia. **41**(1), 199–219 (1995)
8. Bouwman, A.F., Beusen, A.H.W., Billen, G.: Human alteration of the global nitrogen and phosphorus soil balances for the period 1970–2050. Global Biogeochem. Cycles. **23**(4), (2009)
9. Barile, P.J.: Evidence of anthropogenic nitrogen enrichment of the littoral waters of East Central Florida. J. Coast. Res. **204**, 1237–1245 (2004)
10. Sattar, M.A., Kroeze, C., Strokal, M.: The increasing impact of food production on nutrient export by rivers to the Bay of Bengal 1970–2050. Mar. Pollut. Bull. **80**(1–2), 168–178 (2014)
11. Anderson, D.M., Glibert, P.M., Burkholder, J.M.: Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. Estuaries. **25**(4), 704–726 (2002)
12. Cugier, P., Billen, G., Guillaud, J.F., Garnier, J., Ménesguen, A.: Modelling the eutrophication of the Seine Bight (France) under historical, present and future riverine nutrient loading. J. Hydrol. **304**(1), 381–396 (2005)
13. Billen, G., Garnier, J.: River basin nutrient delivery to the coastal sea: Assessing its potential to sustain new production of non-siliceous algae. Mar. Chem. **106**(1), 148–160 (2007)
14. Howarth, R.W.: Coastal nitrogen pollution: a review of sources and trends globally and regionally. Harmful Algae. **8**(1), 14–20 (2008)
15. Pote, S.E., Singal, S.K., Srivastava, D.K.: Assessment of surface water quality of Godavari River at Aurangabad. Asian J. Water Environ. Pollut. **9**(1), 117–122 (2012)

16. Islam, M., Akhtar, M., Masud, S.: Prediction of environmental flow to improve the water quality in the river Buriganga. In: *Proceedings of the 17th IASTED international conference on Modelling and simulation*, pp. 62–67. ACTA Press, Montreal, Canada (2006)
17. Ali, M.Y., Amin, M.N., Alam, K.: Ecological health risk of Buriganga River, Dhaka, Bangladesh. Hydro Nepal J. Water Energy Environ. **3**(3), 25–28 (2009)
18. Mowla, Q.A., Mozumder, M.A.K.: Deteriorating Buriganga river: it' s impact on Dhaka's urban life. PSC J. **2**(2), 1–10 (2015)
19. Hossain, A.M.M.M., Rahman, S.: Impact of land use and urbanization activities on water quality of the Mega City, Dhaka. Asian J. Water, Environ. Pollut. **9**(2), 1–9 (2012)
20. Kibria, M.G., Kadir, M.N., Alam, S.: Buriganga river pollution: its causes and impacts. In: International conference on recent innovation in civil engineering for sustainable development, pp. 323–328 (2015)
21. Yousuf, T.B., Reza, A.: 3R (Reduce, Reuse and Recycle) action plan for the city corporations in Bangladesh: paradigm shift of waste management to resource management. In: WasteSafe 2013 – 3rd international conference on solid waste management in the developing countries (2013)
22. Foix-Cablé, M., Darmawan, R.A., Sahnoun, M., Hindersin, S., Kerner, M., Kraume, M.: Water Sci. Technol.**78**(7), 1556–1565 (2018)
23. Ban, Z.S., Dave, G.: Laboratory studies on recovery of n and p from human urine through struvite crystallisation and zeolite adsorption. Environ. Technol. (United Kingdom). **25**(1), 111–121 (2004)
24. Ganrot, Z., Dave, G., Nilsson, E.: Recovery of N and P from human urine by freezing, struvite precipitation and adsorption to zeolite and active carbon. Bioresour. Technol. **98**(16), 3112–3121 (2007)
25. Jorgensen, T.C., Weatherley, L.R.: Ammonia removal from wastewater by ion exchange in the presence of organic contaminants. Water Res. **37**(8), 1723–1728 (2003)
26. Mnkeni, P.N.S., Kutu, F.R., Muchaonyerwa, P., Austin, L.M.: Evaluation of human urine as a source of nutrients for selected vegetables and maize under tunnel house conditions in the Eastern Cape, South Africa. Waste Manag. Res. **26**(2), 132–139 (2008)
27. Posadas, E., García-Encina, P.A., Soltau, A., Domínguez, A., Díaz, I., Muñoz, R.: Carbon and nutrient removal from centrates and domestic wastewater using algal-bacterial biofilm bioreactors. Bioresour. Technol. **139**, 50–58 (2013)
28. Spångberg, J., Tidåker, P., Jönsson, H.: Environmental impact of recycling nutrients in human excreta to agriculture compared with enhanced wastewater treatment. Sci. Total Environ. **493**, 209–219 (2014)
29. Van Voorthuizen, E.M., Zwijnenburg, A., Wessling, M.: Nutrient removal by NF and RO membranes in a decentralized sanitation system. Water Res. **39**(15), 3657–3667 (2005)
30. Vinnerås, B., Jönsson, H.: The performance and potential of faecal separation and urine diversion to recycle plant nutrients in household wastewater. Bioresour. Technol. **84**(3), 275–282 (2002)
31. Desvousges, W.H., Naughton, M.C., Parsons, G.R.: Benefit transfer: conceptual problems in estimating water quality benefits using existing studies. Water Resour. Res. **28**(3), 675–683 (1992)
32. Brouwer, R.: Environmental value transfer: state of the art and future prospects. Ecol. Econ. **32**(1), 137–152 (2000)
33. Brookshire, D.S., Neill, H.R.: Benefit transfers: conceptual and empirical issues. Water Resour. Res. **28**(3), 651–655 (1992)
34. Fisheries, D. of. Year book of Fisheries Statistics of Bangladesh 2016–17. Fisheries Resources Survey System (FRSS) (2017)
35. Liu, X., Li, Z., Zhang, Y., Feng, R., Mahmood, I.B.: Characterization of human manure-derived biochar and energy-balance analysis of slow pyrolysis process. Waste Manag. **34**(9), 1619–1626 (2014)
36. Krounbi, L., Enders, A., van Es, H., Woolf, D., van Herzen, B., Lehmann, J.: Biological and thermochemical conversion of human solid waste to soil amendments. Waste Manag. **89**, 366–378 (2019)

37. Ward, B.J., Yacob, T.W., Montoya, L.D.: Evaluation of solid fuel char briquettes from human waste. Environ. Sci. Technol. **48**(16), 9852–9858 (2014)
38. Yacob, T.W., (Chip) Fisher, R, Linden, K.G., Weimer, A.W.: Pyrolysis of human feces: Gas yield analysis and kinetic modeling. Waste Manag. **79**, 214–222 (2018)
39. Alam, M.S., Mondal, M.: Assessment of sanitation service quality in urban slums of Khulna city based on SERVQUAL and AHP model: a case study of railway slum, Khulna, Bangladesh. J. Urban Manag. **8**(1), 20–27 (2019)
40. UN-Habitat: The challenge of slums: Global report on human settlements 2003. Manag. Environ. Qual. Int. J. **15**(3), 337–338 (2004)
41. Angeles, G., Lance, P., Barden-O'Fallon, J., Islam, N., Mahbub, A.Q.M., Nazem, N.I.: The 2005 census and mapping of slums in Bangladesh: Design, select results and application. Int. J. Health Geogr. **8**(1), 1–19 (2009)
42. Costanza, R., Arge, R., DeGroot, R., Farberk, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., Neill, R.V.O., Paruelo, J., et al.: The value of the world' s ecosystem services and natural capital. Nature. **387**(May), 253–260 (1997)
43. de Groot, R.S.: Functions of Nature: evaluation of nature in environmental planning, management and decision making. Wolters, Nordhoff BV, Groningen (1993)
44. Alam, K.: Cost-benefit analysis of restoring buriganga river, Bangladesh. Int. J. Water Resour. Dev. **24**(4), 593–607 (2008)

Chapter 10
Analysing Informal Milk Supply Chains Data to Identify Seasonal Occurrences of Antibiotic Residues

Naveed Aslam, Sosheel S. Godfrey, Mateen Abbas, Muhammad Y. Tipu, Muhammad Ishaq, David M. McGill, Hassan M. Warriach, Muhammad Husnain, and Peter C. Wynn

Abstract Informal milk marketing chains provide the major milk supply link from smallholder dairy production systems to consumers in developing countries. Prevalence risk of antibiotic residues was investigated in milk samples (n = 528) collected from different levels of informal supply chains in Pakistan from 2012–2013. After screening, all positive samples were further analyzed by High Performance Liquid Chromatography to quantify individual β-lactam residues. Fifteen percent of the total samples were found to be positive for at least one of the antibiotics. All positive samples (81/528) were positive for amoxicillin. Percentage of positive samples for ampicillin and penicillin was 12.1 and 6.4% respectively. Percentages of positive samples collected from farmers, small collectors, large collectors and retailers were 17.5, 15.1, 8.3 and 13.5 respectively. When relating to season of collection 11.3, 10.2, 19.1, 17.9 and 16.3% of samples in autumn, monsoon,

N. Aslam (✉)
Graham Centre for Agricultural Innovation, Charles Sturt University, Wagga Wagga, NSW, Australia

Faculty of Veterinary and Agricultural, University of Melbourne, Werribee, VIC, Australia

S. S. Godfrey · P. C. Wynn
Graham Centre for Agricultural Innovation, Charles Sturt University, Wagga Wagga, NSW, Australia

M. Abbas · M. Y. Tipu
Quality Operation Laboratory, University of Veterinary and Animal Sciences, Lahore, Pakistan

M. Ishaq · H. M. Warriach
ASLP Dairy Project, University of Veterinary and Animal Sciences, Lahore, Pakistan

D. M. McGill
ASLP Dairy Project, University of Veterinary and Animal Sciences, Lahore, Pakistan

Faculty of Veterinary and Agricultural, University of Melbourne, Werribee, VIC, Australia

M. Husnain
Department of Clinical Medicine and Surgery, University of Veterinary and Animal Sciences, Lahore, Pakistan

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_10

spring, summer and winter were found positive. Concentrations of amoxicillin, ampicillin and penicillin in positive samples were 79.5 ± 12.15, 106.6 ± 12.49 and 13.7 ± 4.2μgkg^{-1} respectively. Occurrence of these high levels of drug residues demand mass awareness programs associated with the implementation of legislation and its enforcement designed to protect the consumer.

Keywords Beta lactams · Traditional milk marketing chains · Smallholder dairy production · Antimicrobial residues

10.1 Introduction

Food safety is an essential part of every dairy and meat industry because of the need to retain consumer confidence in the products reaching them [1, 2]. One of the most important safety issues is the use of antibiotics for prophylactic and/or treatment of animals for bacterial infections which then find their way into milk and meat reaching the commercial marketplace due to the failure of unscrupulous or uneducated farmers to adhere to recommended withholding periods [3, 4]. β-lactams (penicillin, amoxicillin and ampicillin) are one of the most commonly used antibiotic classes throughout the world [3, 5].

Development of antibiotic resistant microorganisms [6], hypersensitivity reactions [3], destruction of gastrointestinal flora in children [2] and economic losses in the manufacturing industry for cheese and yogurt [7] are some of the major problems that antibiotic residues can cause within the human health and dairy industries. A significant part of the population (5–10%) is sensitive to penicillin at a level as low as 1 μgkg^{-1} which may cause asthma, anaphylactic shock, hives and skin rashes [5]. Maximum residue limits (MRL) for each of penicillin, amoxicillin and ampicillin in milk is 4 ugL^{-1} according to the EU standards (EU regulation 2377/90) [8]. On the other hand, USFDA has set these limits as 10, 10 and 5 μgL^{-1} for ampicillin, amoxicillin and penicillin respectively [9].

The dairy industry has an important impact in the socio-economic lives of people living in Southeast Asia [10]. Dairy production has engaged around 150 million farm households in developing countries where production mainly comes from smallholder dairy production systems [11]. Thirty seven percent (8.8 million) of the households in Pakistan keep livestock for milk production [12] making it the third largest milk producing country in the world [13]. An average household spends more than 20% of its income on milk and milk products [14]. Informal milk supply chains in Pakistan handle 97% of the total milk production [15]. In these chains, milk passes from small-holder dairy farmers in villages through a series of milk collectors handling increasing volumes to distribute to vendors in urban areas within limited infrastructure and without any cool chain system in place. Farmers, small collectors, large collectors, retailers, bakers and confectioners and processing plants are some of the most important contributors to milk value chains [16, 17]. This structure of the dairy industry not only has the potential to provide good quality food but also offers employment opportunities to the poor population [11]. Misuse of oxytocin by farmers to enhance milk let down and penicillin by different marketing agents as a bacterial inhibitor has been reported [18].

The potential for harmful effects of antibiotic residues in milk for end-consumers dictates the need to monitor this contamination closely. The aims of the present study were therefore; (a) to observe the prevalence of β-lactam residues at various levels of informal milk supply chains in Pakistan and (b) to monitor the seasonal variation in occurrence of these antibiotic residues. This will further help in assessing the risk of exposure of antibiotic residues for consumers in Pakistan which will lead to legislation designed to prevent this health risk to the Pakistani population.

10.2 Materials and Methods

10.2.1 Experimental Site

Three informal milk marketing chains, one from each Kasur, Okara and Pakpattan districts of province Punjab were selected for this survey. Distances of these chains from their end point in Lahore were 82, 145 and 182 km respectively. All three chains started from small villages in the different districts and ended in the supply of milk to metropolitan Lahore.

10.2.2 Sample Collection

Samples from three informal milk marketing chains were collected from bulk tank milk of farmers, small collectors, large collectors and retailers on a monthly basis. Milk samples were collected from October, 2012 to September, 2013. Table 10.1 identifies the major participants and their position in milk value chains in all three districts selected for sample collection.

A total of 528 milk samples were collected from the chains described in one complete year. A total of 245 milk samples from smallholder dairy farmers, 106 samples from small collectors, 36 samples from large collectors and 141 samples

Table 10.1 The number of different participants in milk value chains selected for sample collection

Districts/marketing chain	Kasur	Okara	Pakpattan
Farmer bulk milk	6–8	6–8	6–8
Small collector	3	3	3
Large collector	1	1	1
Retailer	4	4	4

from retailers were collected during the 12 month survey period. Furthermore, the year was divided into five seasons: winter, spring, summer, monsoon and autumn to determine the seasonal effect. Winter season was classified as the months from December to February; spring from March to April; summer from May to July; the monsoon season from August to September and autumn from October to November. Monthly data for presence or absence of antibiotic residues in milk was pooled and analysed according to these seasons. Samples were then transported on ice to the laboratory and stored at −18 °C pending analysis.

10.2.3 Antibiotic Residues Analysis

Milk samples collected were analysed for β-lactams (penicillin G, ampicillin, amoxicillin) as follows. Initially all 528 milk samples were qualitatively analysed for positive and negative samples and a subset of these samples was assessed by high performance liquid chromatography (HPLC) to quantify antibiotic residues.

10.2.4 Growth, Identification and Purification of Bacillus subtilis and Preparation of Plates

The primary bacterial culture of *Bacillus subtilis* was obtained from the department of Microbiology, University of Veterinary and Animal Sciences, Lahore. It was inoculated into Tryptic Soy Broth (TSB) and incubated for 24 h. Slides were prepared and gram staining was performed for identification of the bacterium. Plates were prepared by pouring Muller Hinton agar and cooling. Sterilized cotton swabs were used for swabbing *Bacillus subtilis* on prepared plates.

10.2.5 Screening of Milk Samples for Antibiotic Residues

All milk samples were screened for the presence and absence of antibiotic residues (penicillin G, ampicillin, amoxicillin) using *Bacillus subtilis* (Qualitative Field Disc Assay) as described by the Association of Official Analytical Chemists [5, 19]. The round blank discs were prepared using filter paper (Whatman 1®) obtained with a punch machine. Discs were dipped into milk samples that had been thoroughly mixed. They were then dried and placed on pre-prepared agar plates containing *Bacillus subtilis*. A total of six discs were placed on each plate. Plates were incubated at 37 °C for 24 h and examined for zones of inhibition. Standard test discs were also used for comparison of zone of inhibition.

10.2.6 High Performance Liquid Chromatography (HPLC) Analyses

Preparation of Standard Ampicillin, amoxicillin and penicillin supplied by Sigma Aldrich (Madrid, Spain) were weighed (100 mg each). Individual antibiotics were then dissolved in 100 mL of HPLC grade methanol (Sigma, Aldrich) in different volumetric flasks. One mL of each antibiotic solution was mixed with 100 mL of phase A (formic acid 0.1% (v/v) in water). Standard solutions were finally prepared at a concentration of 10 μgmL^{-1}.

Sample Preparation and Extraction Acetonitrile (10 mL) was added to milk samples (2 mL) in 20 mL centrifuge tubes. This mixture was mixed using a vortex mixer (Barstead International, USA) and kept in the dark for 10 min. Protein precipitation resulted in extraction of antibiotic residues by shaking for 20 min. The mixture was then centrifuged for 15 min at 3000 g and the supernatant was collected and dried under a mild stream of nitrogen. The residue obtained was dissolved with mobile phase A (50 mL), filtered through polyamide filter paper (0.45 μm) and was transferred to vials. Final filtrate (20 μL) was then injected into the HPLC [20].

HPLC HPLC analysis was performed using a system (Model # SP-20A) provided by Shimadzu Scientific Instruments (Sydney, Australia). Chromatographic separation of antibiotic residues was conducted using an analytical reverse phase column RP-C18 (250 mm × 4.6 mm, 5 μm) provided by Waters Corporation (Milford, USA). Formic acid 0.1% (v/v) in water (phase A) and formic acid 0.1% (v/v) in acetonitrile (phase B), were the two mobile phases used for these analyses. A gradient program was used at a flow rate of 0.45 $mLmin^{-1}$ that is 97% A to 40% A (5 min), 40% A to 0% A (5 min), 0% A to 97% A (10 min) and for 12 min at 97% A. Temperature of the column was maintained at 40 °C.

10.3 Results

10.3.1 Screening of Milk Samples Through Qualitative Field Disc Assay

All milk samples collected were screened for the presence or absence of antibiotic residues with the Qualitative Field Disc Assay using *Bacillus subtilis*. All positive milk samples had a clear zone of inhibition around them. Samples having a zone of inhibition of more than 1 mm were considered positive. Out of 528 milk samples, 81 samples were found to have a clear zone of inhibition of more than 1 mm (Fig. 10.1). The average length of zone of inhibition was 5 ± 0 mm for all positive samples.

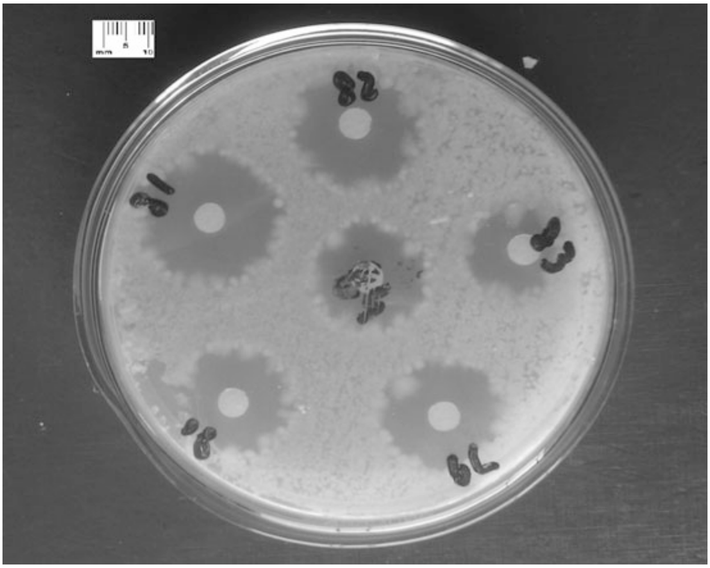


Fig. 10.1 Clear zone of inhibitions in antibiotic residue positive samples

10.3.2 Occurrence of β-Lactam at Different Levels of Milk Marketing Chains in Different Seasons of the Year

Fifteen percent of the samples (81) were found positive for at least one of the three antibiotics they were analysed for. All positive samples (81) were positive for amoxicillin. On the other hand 64 (12.1%) and 34 (6.4%) samples were observed as positive for ampicillin and penicillin respectively. Percentages of positive samples collected from farmers, small collectors, large collector and retailers were 17.5 (43/245), 15.1 (16/106), 8.3 (3/36) and 13.5 (19/141) respectively. Details of samples for each of the antibiotic residues exceeding EU standards (4 μgkg^{-1}) at different stages of the milk supply chains are provided in Table 10.2. All samples which exceeded EU standards were also above the USFDA standards (10 μgkg^{-1}) except for one collected from a farm.

Contamination with amoxicillin and ampicillin was found to be at higher levels than both EU (4 μgkg^{-1}) and USFDA (10 μgkg^{-1}) standards at all levels of the informal supply chains. Penicillin residues were found marginally higher than EU (4 μgkg^{-1}) standards at the farmer, small collector and retailer levels. The highest contamination of amoxicillin was found at the retailer level while highest contamination for ampicillin was found in samples collected from the farmer. Mean values of all positive samples for the concentration of amoxicillin, ampicillin and penicillin at different levels of supply chains are provided in Table 10.3.

Large variation was also seen between seasons of the year. For example 11.3% (10/88) of samples collected in the autumn, 10.2% (9/88) in the monsoon season,

Table 10.2 Number of positive samples below and exceeding EU standards (4 μgkg^{-1}) for amoxicillin, ampicillin and penicillin at different levels of milk supply chains

Level of marketing chain	No. of positive samples	Samples exceeded EU standards	Amoxicillin		Ampicillin		Penicillin	
			$>4 \mu gkg^{-1}$	$< 4 \mu gkg^{-1}$	$>4 \mu gkg^{-1}$	$< 4 \mu gkg^{-1}$	$>4 \mu gkg^{-1}$	$< 4 \mu gkg^{-1}$
Farmers	43	41	41	2	34	2	10	11
Small collectors	16	15	15	1	10	1	2	5
Large collectors	3	3	3	0	3	0	0	1
Retailers	19	17	17	2	12	2	3	2
Total	81	76	76	5	59	5	15	19

Table 10.3 Mean concentrations (μgkg^{-1}; mean ± SEM) for β-lactams (amoxicillin, ampicillin and penicillin) in milk collected from different stages of informal value chains

Antibiotic	Farmers	Small collector	Large collector	Retailer
Amoxicillin (μgkg^{-1})	80.11 ± 15.65	49.0 ± 6.29	93.83 ± 62.35	101.6 ± 36.45
Ampicillin (μgkg^{-1})	121.8 ± 34.12	65.09 ± 16.26	76.62 ± 19.15	84.01 ± 26.08
Penicillin (μgkg^{-1})	6.89 ± 3.0	5.56 ± 4.57	0.56 ± 0.56	4.15 ± 2.45

19.1% (17/89) in spring, 17.9% (23/128) in summer and 16.3% (22/135) in winter were positive for at least one β-lactam antibiotic residue.

Table 10.4 explains in detail the number of samples which exceeded EU standards (4 μgkg^{-1}) in different seasons of the year. All samples that exceeded EU standards for ampicillin and amoxicillin were also above the minimum allowable USFDA level. Only one sample in winter which exceeded the EU standard also exceeded the USFSA limit. On the other hand, one sample that did not exceed the USFDA standard (5 μgkg^{-1}) for penicillin was higher than the EU standard.

The highest contamination of milk samples with amoxicillin was observed during the monsoon season. However ampicillin contamination was found more in summer and slightly less in the monsoon season. The highest concentration of penicillin was also found during the monsoon season. Mean concentrations (μgkg^{-1}) of amoxicillin, ampicillin and penicillin in different seasons (autumn, monsoon, spring, summer and winter) of the year are presented in Table 10.5.

10.3.3 Amoxicillin

Out of 81 positive samples analysed for amoxicillin, 76 samples exceeded EU standards (4 μgkg^{-1}). The mean concentration for all positive samples for amoxicillin was 79.5 ± 12.15μgkg^{-1}. The number of positive samples that exceeded the EU standard along the marketing chains in various seasons is given in Fig. 10.2. The mean concentrations of all positive samples for amoxicillin in the districts of Kasur, Okara and Pakpattan were 62.8 ± 16.54, 62.8 ± 16.54 and 95.9 ± 22.92 μgkg^{-1} respectively. The corresponding proportion of positive samples that exceeded EU standards was 90.3, 92.6 and 100% respectively.

10.3.4 Ampicillin

Contamination of milk samples with ampicillin that exceeded both EU (4 μgkg^{-1}) was 92.2%. All samples that exceeded EU standards were also above USFDA (10 μgkg^{-1}) standards. The concentration of all positive samples for ampicillin was 106.6 ± 12.49 μgkg^{-1} (mean ± SE). Concentrations of ampicillin were 94.12 ± 19.42 μgkg^{-1}, 121.7 ± 51.42 μgkg^{-1} and 82.55 ± 20.83 μgkg^{-1} in

Table 10.4 The effect of season on the prevalence of samples that exceeded the EU standards (4 $\mu g kg^{-1}$) for amoxicillin, ampicillin and penicillin

Level of marketing chain	No. of positive samples	Samples exceeded EU standards	Amoxicillin		Ampicillin		Penicillin	
			$>4 \mu g kg^{-1}$	$< 4 \mu g kg^{-1}$	$>4 \mu g kg^{-1}$	$< 4 \mu g kg^{-1}$	$>4 \mu g kg^{-1}$	$< 4 \mu g kg^{-1}$
Autumn	10	10	10	0	7	0	0	2
Monsoon	9	9	9	0	6	2	4	2
Spring	17	16	16	1	14	1	4	6
Summer	23	21	21	2	14	2	3	3
Winter	22	20	20	2	18	0	4	6
Total	81	76	76	5	59	5	15	19

Table 10.5 Mean concentrations (μgkg^{-1}) of β-lactams (amoxicillin, ampicillin and penicillin) in milk samples collected in the various seasons of the year

Antibiotic	Autumn	Monsoon	Spring	Summer	Winter
Amoxicillin (μgkg^{-1})	44.02 ± 7.31	105.6 ± 47.91	87.97 ± 24.93	69.17 ± 17.26	89.67 ± 31.93
Ampicillin (μgkg^{-1})	53.79 ± 18.37	117 ± 40.1	89.14 ± 26.82	120.6 ± 59.90	101 ± 23.45
Penicillin (μgkg^{-1})	0.11 ± 0.08	15.52 ± 12.45	8.2 ± 4.52	2.03 ± 1.01	6.32 ± 3.2

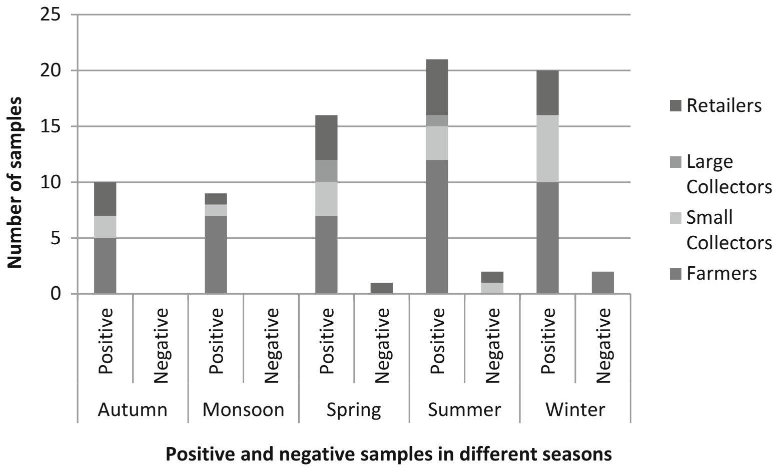


Fig. 10.2 Number of samples collected from different participants of milk supply chains with amoxicillin concentrations above (positive) and below (negative) the EU standard (4 μgkg^{-1}) in different seasons of the year

the districts of Kasur, Okara and Pakpattan respectively. Detail of the distribution of positive samples that exceeded the the EU standard at different levels of the marketing chains in different seasons is given in Fig. 10.3. The percentage of the positive samples exceeding EU standards for different districts was 88.9% (Kasur), 90.5% (Okara) and 100% (Pakpattan).

10.3.5 Penicillin

The percentage of samples positive for penicillin was 13.7 ± 4.2. The prevalence of these higher concentrations varied along the marketing chain (Fig. 10.4). The average concentration of penicillin in the district of Kasur (2.86 ± 1.66 μgkg^{-1}) was found to be lower than the EU standard (4 μgkg^{-1}). However concentrations were found to be higher in the districts of Okara (4.18 ± 1.7) and Pakpattan

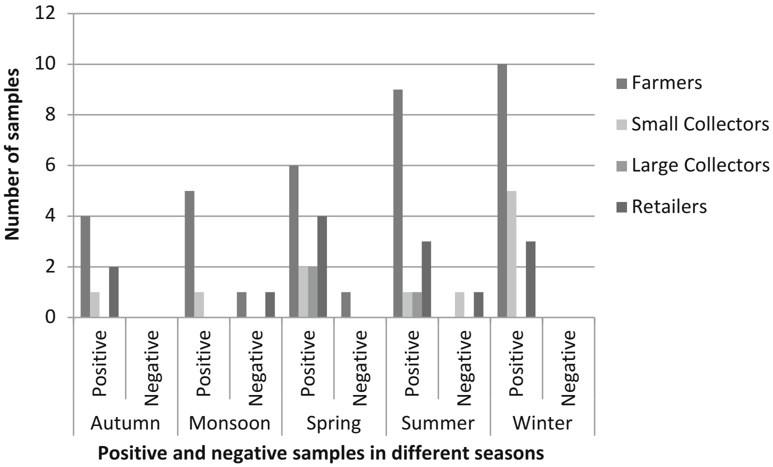


Fig. 10.3 The number of samples for different participants in milk supply chains with ampicillin concentrations above (positive) and below (negative) the EU standard (4 μgkg-1) in different seasons of the year

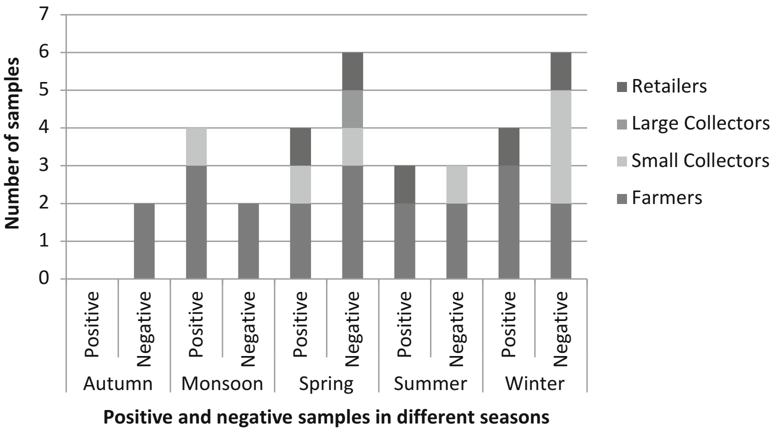


Fig. 10.4 The number of samples collected from different participants of milk supply chains with penicillin above (positive) and below (negative) the EU standards (4 μgkg^{-1}) in different seasons of the year

(11.5 ± 5.94). Only fifteen (44.1%) samples out of 34 positive samples across all 3 chains exhibited concentrations greater than the EU standard (4 μgkg^{-1}). These corresponded to 33.3, 42.9 and 54.5% of positive samples found in the districts of Kasur, Okara and Pakpattan, respectively.

10.4 Discussion

The presence of antibiotic residues in market milk and meat has been reported previously but levels of contamination of milk obtained from informal supply chains in particular have not been identified. In a survey conducted by M Khaskheli et al. [5] in three cities (Hyderabad, Latifabad and Qasimabad) in province Sindh of Pakistan, 36.5% of the samples were found to be positive for β-lactams. Mean concentrations of amoxicillin, ampicillin and penicillin were 36.11, 46.91 and 59.53 μgkg^{-1} respectively in their study. The percentage of positive samples found in our study was less than half of what was observed by M Khaskheli et al. [5]. Despite the lower percentage of positive samples, mean concentrations of amoxicillin (79.5 μgkg^{-1}) and ampicillin (106.6 μgkg^{-1}) were found to be higher showing the increased potential risk associated with the use of this antibiotic in mastitis prophylaxis. In contrast the concentration of penicillin (13.7 μgkg^{-1}) was found to be lower than reported previously (59.53 μgkg^{-1}), showing a lower risk of exposure to penicillin in Punjab as compared to Sindh province. N Noori et al. [3] evaluated the quality of milk powder produced in dairy factories in Iran for contamination with β-lactams and tetracyclines. Thirty percent of the total samples were found to be positive for β-lactams in a total of 240 milk samples collected over a period of 12 months [3]. The percentage of contaminated samples in their report (37%) was higher than observed (15%) in our present study. The presence of antibiotic residues indicates negligence on the part of the farmers in not adhering to the appropriate withholding periods for each of the antibiotics tested for in this study. This may be associated with the over dosing of antibiotics to treat animals. Of course this may also be due to the ignorance on the part of farmers with a limited education.

KG Aning et al. [21], LR Kurwijila et al. [22] and EK Kang'ethe et al. [23] observed antimicrobial drug residues at different points in marketing chains (producer-sellers, processors, wholesalers and retailers) in Ghana, Tanzania and Kenya. Their results showed 31.1%, 36% and 16% of the total samples to be positive in each of these countries. There was no difference in the prevalence of antibiotic residues observed during the dry and wet seasons at any level of milk marketing chains except in the city of Dar es Salaam in Tanzania, where the incidence of antibiotic residues was higher in the wet season in milk purchased from vendors/hawkers. Although a significant number of samples was tested (n = 228, 986 and 854 for the studies of Aning et al., Kujwijila et al. and Kang'ethe et al. respectively), no quantitative analysis was conducted. In addition samples were qualitatively analysed using the CharmAIM-96 test kit which is relatively non-specific for particular antimicrobials (Sulfamethazine, Gentamicin, Tylosin, Tetracyclines and β-lactams). In contrast our survey has detected contamination with only β-lactams but over all seasons (summer, winter, spring, autumn and monsoon) of the year.

The higher concentration of amoxicillin and penicillin in the monsoon season is an indication of greater prevalence of diseases induced through higher humidity specifically in this season. Comparable levels were also reported for the summer season (120.6 vs 117 $\mu g kg^{-1}$). Haemoprotozoan and associated diseases (babesiosis, theileriosis and anaplasmosis) have been reported in both summer and monsoon seasons in Pakistan (A Durrani et al. [24]) and India (R Velusamy et al. [25] MA Alim et al. [26]). As a result uneducated farmers resort to the use of antibiotics as soon as they observe any physical abnormality. β-lactams are normally the first choice for these farmers as they are readily available and do not require veterinary prescription. The prevalence of contamination reported in the present study was half of that observed by KG Aning et al. [21] and LR Kurwijila et al. [22], and similar to that obtained by EK Kang'ethe et al. [23].

Higher contamination of milk with amoxicillin and ampicillin across the districts studies are indicative of their prevalence of use in disease prophylaxis. Although our study has highlighted the problem of antibiotic residues entering the human food chain, the incidence of tetracyclines and other antibiotics is not known and requires investigation. In addition to the problem, residues history tells us that the risk of developing drug resistant bacteria causes major issues for the animal and veterinary pharmaceutical industries.

10.5 Conclusion

Given the socio-economic status of the population of Southeast Asia particularly Pakistan, informal supply chains are the most practical systems for collection and distribution of the essential dietary animal protein source, milk. Due to the large population dependent on this product, there is a need to monitor more closely its quality. Thus the presence of herbicides, pesticides and other antibiotics used in agriculture needs to be monitored. Legislation for minimum acceptable levels should be established and then enforced.

Acknowledgement Authors really acknowledge the help of Biomin® Singapore for its support by providing funds to purchase kits for ELISA analysis of milk samples. Moreover, we would like to thank Mr. Sajid and Mr. Faisal in Quality Operation Lab in UVAS Lahore for their support during analysis. Moreover, Charles Sturt University's writing support for timely submission was exceptional.

References

1. Irum, A., Saeed, H., Ali, S., et al.: Estimation of amoxicillin residues in commercial meat and milk samples. J. Microbiol. Biotechnol. Food Sci. **4**, 48–50 (2014)
2. Jabbar, A., Rehman, S.U.: Microbiological evaluation of antibiotic residues in meat, milk and eggs. J. Microbiol. Biotechnol. Food Sci. **2**, 2349–2354 (2013)
3. Noori, N., Karim, G., Raeesian, M., et al.: Antibiotic residues and aflatoxin M1 contamination in milk powder used in Tehran dairy factories, Iran. Iran. J. Vet. Med. **7**, 221–226 (2013)

4. Sats, A., Mootse, H., Lepasalu, L., et al.: Use of *Delvotest T* for quantitative estimation of beta lactam antibiotic residues in waste milk and for evaluation of thermal treatment efficiency- a methodical pilot study. Agron. Res. **12**, 807–812 (2014)
5. Khaskheli, M., Malik, R.S., Arain, M.A., et al.: Detection of beta lactam antibiotic residues in market milk. Pak. J. Nutr. **7**, 682–685 (2008)
6. Anderson, A.D., Nelson, J.M., Rossiter, S., et al.: Public health consequences of use of antimicrobial agents in food animals in the United States. Microb. Drug Resist. **9**, 373–379 (2003)
7. Allison, J.R.D.: Antibiotic residues in milk. Br. Vet. J. **141**, 9–16 (1985)
8. Ghidini, S., Zanardi, E., Varisco, G., et al.: Residues of B-lactam antibiotics in bovine milk: confirmatory analysis by liquid chromatography tandem mass spectrometry after microbial assay screening. Food Addit. Contam. **20**, 528–534 (2003)
9. Graham, T., Hennes, R.F..: Beta lactam test methods for use under appendix N and section 6 of the grade "A" pasteurized milk ordinance (M-a-85, revision # 13). In: USFDA (ed.), MD, USA (2010)
10. Fao.: Status of and prospects for smallholder milk production- a global perspective. In: Hemme, T., Otte, J. (eds.). Rome, Italy (2010)
11. Afzal, M.: Re-designing smallholder dairy production in Pakistan. Pak. Vet. J. **30**, 187–190 (2010)
12. Government of Pakistan.: Agricultural census. In: statistics PBo (ed.), Agricultural census organization, Lahore (2010)
13. Hemme, T.: Dairy sector and chain profile. Dairy report: for a better understanding of milk production world-wide. In: International Farm Comparison Network (IFCN), Kiel, Germany, p. 62 (2012)
14. Government of Pakistan.: Household integrated economic survey (HIES) 2011–12. In: statistics Pbo (ed.), Islamabad (2013)
15. Zia, U.: Analysis of milk marketing chain-Pakistan. Ital. J. Anim. Sci. **6**, 1384–1386 (2007)
16. Muhammad, Z.U., Akhter, S.N., Ullah, M.K.: Dairy supply chain management and critical investigations on dairy informal channel partners in Pakistan. IOSR J. Bus. Manag. **16**, 81–87 (2014)
17. Tariq, M., Mustafa, M.I., Iqbal, A., et al.: Milk marketing and value chain contraints. Pak. J. Agric. Sci. **45**, 195–200 (2008)
18. Fao.: Dairy development in Pakistan. In: Zia U, Mahmood T, Ali MR (eds). Rome, Italy (2011)
19. Mangsi, A.S., Khaskheli, M., Soomro, A.H., et al.: Detection of antimicrobial drug residues in milk marketed at different areas of Sindh province. IOSR J. Agric. Vet. Sci. **7**, 65–69 (2014)
20. Freitas, A., Barbosa, J., Ramos, F.: Development and validation of a multi-residue and multiclass ultra-high-pressure liquid chromatography-tandem mass spectrometry screening of antibiotics in milk. Int. Dairy J. **33**, 38–43 (2013)
21. Aning, K.G., Donkor, E.S., Omore, A., et al.: Risk of exposure to marketed milk with antimicrobial drug residues in Ghana. Open Food Sci. J. **1**, 1–5 (2007)
22. Kurwijila, L.R., Omore, A., Stall, S., et al.: Investigation of the risk of exposure to antimicrobial residues present in marketed milk in Tanzania. J. Food Prot. **69**, 2487–2492 (2006)
23. Kang'ethe, E.K., Aboge, G.O., Arimi, S.M., et al.: Investigation of the risk of consuming marketed milk with antimicrobial residues in Kenya. Food Control. **16**, 349–355 (2005)
24. Durrani, A., Mehmood, N., Shakoori, A.: Comparison of three diagnostic methods for Theileria annulata in Sahiwal and Friesian cattle in Pakistan. Pak. J. Zool. **42**, 467–471 (2010)
25. Velusamy, R., Rani, N., Ponnudurai, G., et al.: Influence of season, age and breed on prevalence of haemoprotozoan diseases in cattle of Tamil Nadu, India. Vet. World. **7**, 574–578 (2014)
26. Alim, M.A., Das, S., Roy, K., et al.: Prevalence of hemoprotozoan diseases in cattle population of Chittagong division, Bangladesh. Pak. Vet. J. **32**, 221–224 (2012)

Part III
Data Science and Image Processing Statistics

Chapter 11
Detection of Vegetation in Environmental Repeat Photography: A New Algorithmic Approach in Data Science

Asim Khan, Anwaar Ulhaq, Randall Robinson, and Mobeen Ur Rehman

Abstract Environment change being one of the major issues in today's world needs special attention of the researchers. With the advancement in computer vision researchers are equipped enough to come up with algorithms accomplishing automated system for environment monitoring. This paper proposes an algorithm which can be used to observe the change in vegetation utilizing the images of a particular site. This would help the environment experts to put on their efforts in a right direction and right place to improve the environment situation. The proposed algorithm registers the image so that comparison can be carried out in an accurate manner using single framework for all the images. Registration algorithm aligns the new images with the existing images available in the record of the same particular site by performing transformation. Registration process is followed by segmentation process which segments out the vegetation region from the image. A novel approach towards segmentation is proposed which works on the machine learning based algorithm. The algorithm performs classification between vegetation patches and non-vegetation patches which equips us to perform segmentation. The proposed algorithm showed promising results with F-measure of 85.36%. The segmentation result leads us to easy going calculation of vegetation index. Which can be used to make a vegetation record regarding particular site.

A. Khan (✉) · R. Robinson
College of Engineering and Science, Victoria University, Melbourne, VIC, Australia
e-mail: asim.khan@live.vu.edu.au

A. Ulhaq
School of Computing and Mathematics, Charles Sturt University, Port Macquarie, NSW, Australia

Centre of Applied Informatics, Victoria University, Melbourne, VIC, Australia
e-mail: aulhaq@csu.edu.au

M. Ur Rehman
Avionics Department, Air University, Islamabad, Pakistan

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_11

Keywords Vegetation index · Image registration · Image segmentation · SVM · Flucker post dataset · F-measure

11.1 Introduction

In today's world population have increased rapidly with the advancement of technology which have created a need and importance of Environment monitoring. Humans have effected the environment to such an extent that the time has come to pay our concentration to this issue. The change in climate and the change in land are interconnected powers of global change prompted by human beings [1, 2]. Environment monitoring is now an important task to do so that an information can be processed to assess environmental effectiveness. It can proved to be helpful in monitoring humidity and temperature. One of the most important components of environmental monitoring is to observe the information regarding vegetation which is essential to predict at an early stage about ongoing trends.

Now a day's satellite and aerial podiums are utilized to get remote sensing images which can be used for environment monitoring. But from a very long time photographs taken from ground are used to record the change in environment and ecosystem [3]. Repeat photography is one the most often method used for monitoring any change, where images are taken repeatedly of a particular location in different times. So by comparing the repeated photographs analysis can be carried out regarding the change. While in the case of analyzing the environmental change time difference between image repetition is in months and years. These repeated photography communicates about the change in environment with respect to time [4].

The advancement in remote sensing images have made these repeat photography methods to lose its importance. But if environment analysis needs to be carried out in a transparent manner, revitalization of this technology is necessary so that researchers can utilize these images in their research. As repeat photography gives a detailed information regarding the transformation in vegetation regarding some particular site.

Repeated photography is used since decades for different monitoring tasks like geomorphological development [5–7], change in trees arrangement [8, 9], costal lacales [10], plant phenology [11–14] and many others. Literature shows that researchers have immensely utilized the repeat photography for vegetation cover [15–19]. Repeat photography involves many issues like background clutter, high interclass variation and illumination variance [20, 21]. However with the advancement of technology and introduction to new approaches, these limitation are encountered and repeated photography is used to carry out different analysis.

In [22, 23] authors have proposed an algorithm which divides image into rectangular grid and than using those rectangular grid percentage of vegetation cover is calculated. Point sampling is also used to quantify the vegetation change [24], where classification is performed to classify each image in different cover types, which are treated as different classes. Moreover different cover types represents different quantitative measures.

In literature, papers can be found which utilizes automated segmentation for extracting the vegetation region for calculating vegetation index. But the approaches proposed utilizes remote sensing imagery for this purpose [25, 26]. However the advancement in computer vision field enables us to use approaches to get the automated system for the segmentation of vegetation so that qualitative measure regarding vegetation can be calculated. Fortin et al. in [27] have proposed an algorithm for estimation of landscape composition from repeat photography but like other researchers they have used manual segmentation scheme.

This research is carried out to come up with an approach of automated system for segmentation and than vegetation measure in repeat photography with the help of advanced computer vision techniques. The proposed algorithm showed promising results for segmentation as well as calculation of quantitative measure for vegetation approximation. As in the field of ecology the vegetation index of some particular site give a lot of information regarding its environment.

Rest of the paper is organized in following manner: Sect. 11.2 discusses about the dataset used in this paper, Sect. 11.3 discusses about the main methodology where discussion about each step is described in detail, Sect. 11.4 carries details regarding the results of the proposed algorithm on the dataset discussed in Sect. 11.2, Sect. 11.5 is the part where conclusion is carried out regarding the proposed algorithm on the research topic.

11.2 Dataset

The paper uses Fluckerpost dataset [28] which have different scene images taken in different time spans. Figure 11.1 shows some of the examples images of two different scenes and its 3 different time span images. The dataset contains images of 22 locations in total and in each location there are images of multiple scenes. The dataset is the initiative towards monitoring of more than 150 precious environment.

11.3 Proposed Methodology

The proposed methodology involves multiple steps which leads to the calculation of vegetation index present in an image. Figure 11.2 shows the general flow diagram followed for the extraction of vegetation index for any input image with regard to previous images of the particular scene input image belongs to. This section further discusses about all the building blocks of the proposed methodology.

Fig. 11.1 Figure shows dataset images where row figures shows different time spans and columns represent different scenes

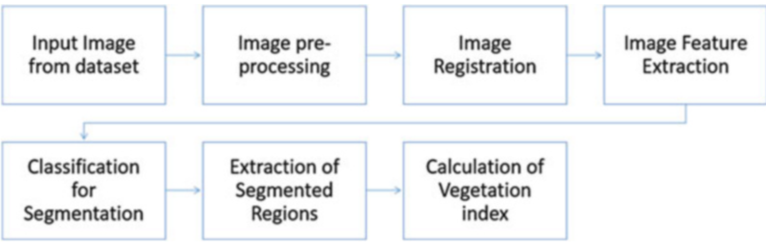


Fig. 11.2 General flow diagram

11.3.1 Image Pre-processing

As the images in the dataset are taken at different time spans with different cameras so the dataset images have varying light intensities. Therefore the dataset images needs to be pre-processed before their registration process. For this process Image histogram equalization is adopted to encounter the said problem. The image was divided into R, G and B channels after which image histogram equalization was applied on the individual channels. The resultant channels were combined together to get the RGB image.

11.3.2 Image Registration

Acquired images are of the same scene but they are not on a single coordinate system because they are captured from different viewpoints as can be seen in Fig. 11.1 as well. Images registration is the process that can be used to tackle this problem. Image Registration transforms one image with respect to other in such a manner that they both can be compared. In [29] authors have described the general steps which needs to be followed for image registration in any research problem. Those steps are,

- Feature Detection
- Feature Matching
- Model Estimation for Transformation
- Image Transformation

11.3.2.1 Feature Detection

Gradients in an image holds great importance as they represents the information regarding edges. Similarly in case of mitotic and non-mitotic cells gradients can play a vital role. As discussed earlier as well that mitotic and non-mitotic cells differentiates from each other on the bases of texture/shape. So gradients give us thorough information about edges and edges represents the texture/shape of the image. Therefore using HOG can be significant for the good performance of cell detection. HOG generates the histogram of orientation of a window/patch selected. The working steps of HOG are discussed below,

- In first step image is divided into small patches and each patch is known as detection window. The size of the window NxN is dependent on the input image size and the number of gradients we want to extract.
- Color and Gamma normalization is performed for the extraction of strong gradients of the image.

- Gaussian derivative is used to compute the gradients fx and fy in horizontal and vertical direction respectively. For which following filter kernels are used.

$$[-1, 0, 1] \quad and \quad [-1, 0, 1]^T \tag{11.1}$$

- Now orientation binning is performed in which cell histograms are developed. In spatial cell each pixel hold a specific weight which is derived from the gradient calculated. Histogram channels are spread evenly with a difference of 20 degree.
- Contrast Normalization is performed on the overlapping block in cells. The normalization factor used can be seen in below equation,

$$Norm : L = \frac{B}{\sqrt{||B||_2^2 + t^2}} \tag{11.2}$$

  Where 'B' represents the un-normalized vector and 't' represents threshold value which will remain constant for all windows.
- The last step is collection of all HOGs in each block. This prepares our feature vector which can be now used for classification.

11.3.2.2 Feature Matching

In this step features of two images are matched together to calculate correspondence between them. We have used Brute-Force Matcher method for calculating similarity between detected features. Brute-Force Matcher works on a simple phenomenon where distance between K1 (key point features extracted from image 1) and K2 (key point features extracted from image 2) is calculated. The minimum Euclidean distance between K1 and K2 will be considered as matched points.

11.3.2.3 Model Estimation for Transformation

After feature matching task most important task is to remove the outliers. Outliers are those features points which do not fit in our model. Removing them is necessary so that an accurate transformation parameters can be calculated. Proposed architecture utilizes RANSAC for outlier removal task. RANSAC algorithm works on 4 main steps which are,

- From the dataset containing outliers select a random subset to start the process.
- Model fitting is performed on the designated subset.
- Number of outliers are calculated for the fitted model.
- Repeat all above three steps for the selected number of iterations. Desired model with the check on number of outlier against it, is considered to be best fit model.

For model estimation we have used similarity transform as global mapping model. This model tackles three main transformation constraints which are translation, rotation and scaling. This model is also known as shape preserving mapping model because it conserves the angle as well as curvature present in an image that needs to be transformed. So in our case it is really important to conserve the angles because we need to segment out the output image of registration block.

Equations 11.3 and 11.4 shows formulation used to calculate u and v which transformation coordinates. Where i and j are the coordinates of original image. s, p and t shows scale factor, rotation and translation factor respectively which are calculated using global mapping model

$$u = s * (i * cos(p) - j * sin(p)) + t_i \tag{11.3}$$

$$v = s * (i * sin(p) - j * cos(p)) + t_j \tag{11.4}$$

11.3.2.4 Image Transformation

Mapping function is constructed in previous step which would be used to transform the image so that two image can have same scale, rotation and translation. Which is necessary so that we can know the change in vegetation index between two images in an accurate manner. Each pixel is individually transformed using the equations 11.3 and 11.4 discussed in previous step.

11.3.3 Image Feature Extraction

After Image registration we need to extract the features from the registered image. For which image was divided into small chunks of size 4×4. From each chunk/patch color and texture features were extracted. Figure 11.3 shows the algorithm adopted after image registration for the process of segmentation and then finally achieving the vegetation index. The 8 features extracted for the proposed are discussed below.

11.3.3.1 Color Features

Three color features were extracted. R, G and B values of all 16 pixels from block of 4×4 was extracted. The three features extracted were the mean values of all three channels in a block.

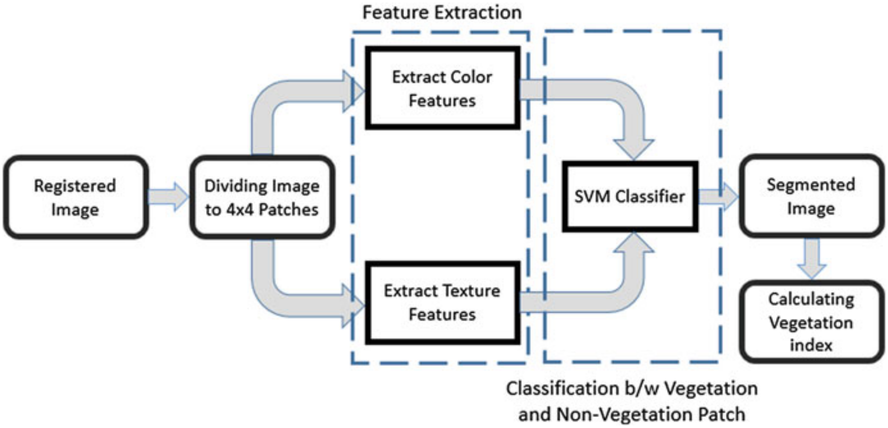


Fig. 11.3 Proposed segmentation algorithm for vegetation index calculation

11.3.3.2 Texture Features

Five texture features were extracted which were Energy, Entropy, Contrast, Homogeneity and Correlation. Equations 11.5, 11.6, 11.7, 11.8, and 11.9 are used for the extraction of these features after converting the patch from RGB to gray scale. Where 'I' is the pixel value of gray scale patch, l and m are the coordinates of the patch.

$$Energy = \sum_{l} \sum_{m} I^2(l, m) \tag{11.5}$$

$$Entropy = -\sum_{l} \sum_{m} I(l, m) log(I(l, m)) \tag{11.6}$$

$$Contrast = \sum_{l} \sum_{m} (l - m)^2 I(l, m) \tag{11.7}$$

$$Homogeneity = \sum_{l} \sum_{m} \frac{I(l, m)}{1 + |l - m|} \tag{11.8}$$

$$Correlation = \frac{\sum_{l} \sum_{m} (l - \mu_x)(m - \mu_y) I(l, m)}{\sigma_x \sigma_y} \tag{11.9}$$

11.3.4 Classification for Segmentation

After feature extraction of each patch of the image we need to classify the patches to vegetation and non-vegetation classes. Each chunk was labelled with vegetation or non-vegetation using Image labeler App of matlab. Support Vector Machine was used as a classifier. SVM was trained using 100 different scene images taken from flucker post project dataset [28].

SVM is a linear classifier which generates a boundary between the two classes for classification purpose. The boundary is created using the optimal hyperlane which the best fit model for classification of class P and Q. The optimal hyperlane equation is shown below,

$$p\ .\ x + q = 0 \tag{11.10}$$

and the above equation is required to fulfill two conditions which are,

– Hyperlane should be such that it satisfies below two equations,

$$f(x) = p\ .\ x + q \qquad \textit{should only be positive if } x \in P \tag{11.11}$$

$$f(x) \leq 0 \quad if\ x \in Q \tag{11.12}$$

– The hyperlane should be at maximum possible distance from all the observations which would depict robustness of the system. The hyperlane distance from the observation is defined as,

$$\frac{|p\ .\ x + q|}{||p||} \tag{11.13}$$

After classification of all patches vegetation classified patches are segmented out from the image and combine together in a single image.

11.3.5 Calculating Vegetation Index

After segmentation process an image is achieved which only contains vegetation content in the image. So we can now calculate the vegetation index with the help of pixels being segmented out. The equation used for calculating the vegetation index is shown below,

$$Vegetation\ Index = \frac{Number\ of\ pixels\ segmented}{Total\ number\ of\ pixels\ in\ original\ image} \tag{11.14}$$

11.4 Results

As discussed earlier, 100 images from fluckerpost dataset were taken into account for training purpose. For testing purpose we took other 50 images from the same dataset. On those images registration was applied on the basis of the model image from each site. After which the developed algorithm was applied to segment the vegetation region in the image. Example images can be seen in Fig. 11.4. Figure 11.4

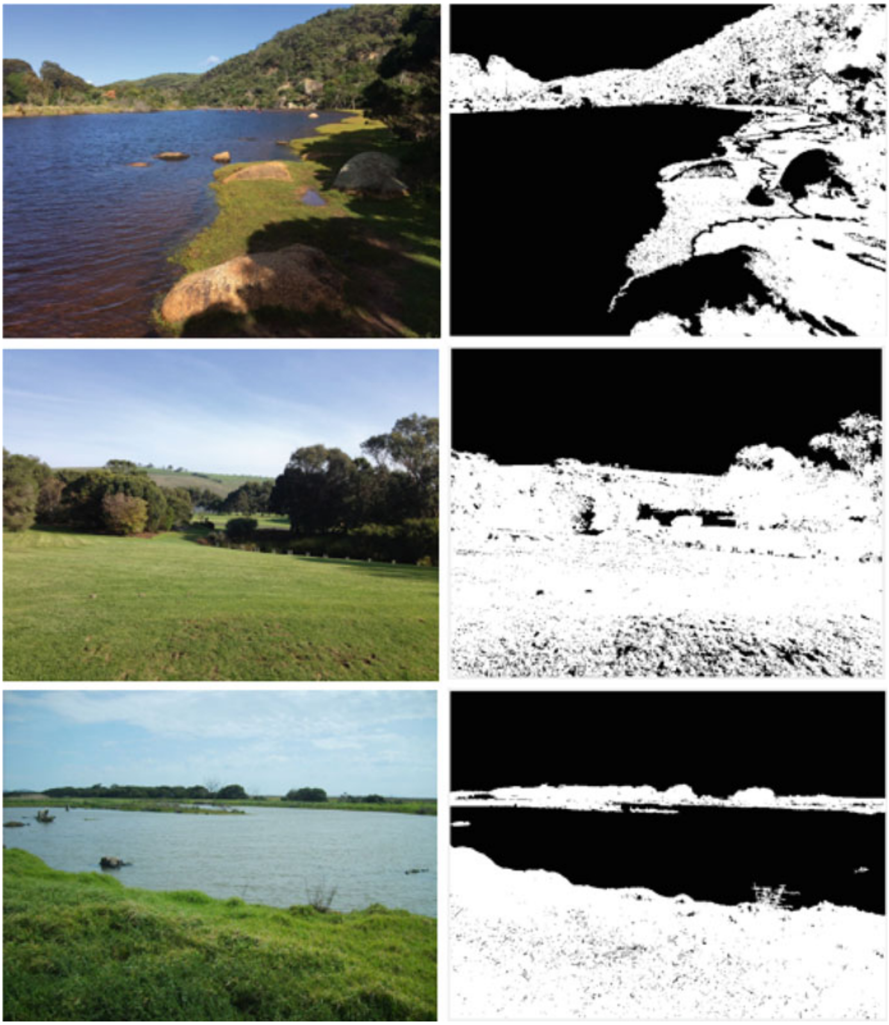


Fig. 11.4 Image A shows the example registered imaged followed by the segmentation mask for vegetation region extraction in Image B

Table 11.1 Image number and calculated vegetation index

Image number	Calculated vegetation index
1	0.43
2	0.62
3	0.39

clearly depicts that high performance segmentation is performed for the extraction of vegetation region.

For further evaluation of the proposed segmentation algorithm few numerical approaches were adopted. F-measure of the classification process carried out by SVM based on given features was evaluated on 50 test images. The equation of F-measure is shown below,

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11.15}$$

Where precision is the ratio between number of correctly classified vegetation patches and total number of patches classified by SVM as vegetation region. And recall is the ratio between number of correctly classified vegetation and the number of vegetation patches present in the whole labeled image.

F-Measure achieved for the 50 test images from fluckerpost dataset is 85.36%. Moreover we have calculated the vegetation index of the images as well. The Table 11.1 shows vegetation index of the segmented images.

11.5 Conclusion

Authors have proposed a novel approach towards automatic environment analysis. Where the algorithm will carry out processing to calculate the vegetation index from the images of same site acquired in different phases of a year as well as in different years. The proposed algorithm for segmentation of vegetation region have shown promising results. A machine learning algorithm is appointed for this job which was trained using color and texture features extract from the patches of image. Segmentation algorithm is analyzed using precision and recall values, where calculated f-measure is 85.36%.

References

1. Dale, V.H., Efroymson, R.A., Kline, K.L.: The land use–climate change–energy nexus. Landsc. Ecol. **26**, 755–773 (2011)
2. De Baan, L., Alkemade, R., Koellner, T.: Land use impacts on biodiversity in LCA: a global approach. Int. J. Life Cycle Assess. **18**, 1216–1230 (2013)

3. Pickard, J.: Assessing vegetation change over a century using repeat photography. Aust. J. Bot. **50**, 409–414 (2002)
4. Klett, M.: Three methods of presenting repeat photographs. In: Webb, R.H., Boyer, D.E., Turner, R.M. (eds.) Repeat Photography: Methods and Applications in the Natural Sciences, pp. 32–45. Island Press, Washington (2010)
5. Khan, S.F., Kamp, U., Owen, L.A.: Documenting five years of landsliding after the 2005 Kashmir earthquake, using repeat photography. Geomorphology **197**, 45–55 (2013)
6. Frankl, A., Nyssen, J., De Dapper, M., Haile, M., Billi, P., Munro, R.N., Deckers, J., Poesen, J.: Linking long-term gully and river channel dynamics to environmental change using repeat photography (Northern Ethiopia). Geomorphology **129**, 238–251 (2011)
7. Conedera, M., Bozzini, C., Scapozza, C., Rè, L., Ryter, U., Krebs, P.: Anwendungspotenzial des WSL-Monoplotting-Tools im Naturgefahrenmanagement. Schweiz. Z. Forstwes. **164**, 173–180 (2013)
8. Roush, W., Munroe, J.S., Fagre, D.B.: Development of a spatial analysis method using ground-based repeat photography to detect changes in the alpine treeline ecotone, Glacier National Park, Montana, U. S. A. Arctic Antarct. Alp. Res. **39**(2), 297–308 (2007)
9. Van Bogaert, R., Haneca, K., Hoogesteger, J., Jonasson, C., De Dapper, M., Callaghan, T.V.: A century of tree line changes in sub-Arctic Sweden shows local and regional variability and only a minor influence of 20th century climate warming: twentieth century tree line changes in the Swedish sub-Arctic. J. Biogeogr. **38**, 907–921 (2011)
10. Reimers, B., Griffiths, C., Hoffman, M.: Repeat photography as a tool for detecting and monitoring historical changes in South African coastal habitats. Afr. J. Mar. Sci. **36**, 387–398 (2014)
11. Julitta, T., Cremonese, E., Migliavacca, M., Colombo, R., Galvagno, M., Siniscalco, C., Rossini, M., Fava, F., Cogliati, S., Morra di Cella, U., Menzel, A.: Using digital camera images to analyse snowmelt and phenology of a subalpine grassland. Agric. For. Meteorol. **198–199**, 116–125 (2014)
12. Luo, Y., El-Madany, T.S., Filippa, G., Ma, X., Ahrens, B., Carrara, A., Gonzalez-Cascon, R., Cremonese, E., Galvagno, M., Hammer, T.W., Pacheco-Labrador, J., Martín, M.P., Moreno, G., Perez-Priego, O., Reichstein, M., Richardson, A.D., Römermann, C., Migliavacca, M.: Using near-infrared-enabled digital repeat photography to track structural and physiological phenology in Mediterranean tree–grass ecosystems. Remote Sens. **10**, 1293 (2018)
13. Moore, C.E., Brown, T., Keenan, T.F., Duursma, R.A., van Dijk, A.I.J.M., Beringer, J., Culvenor, D., Evans, B., Huete, A., Hutley, L.B., Maier, S., Restrepo-Coupe, N., Sonnentag, O., Specht, A., Taylor, J.R., van Gorsel, E., Liddell, M.J.: Reviews and syntheses: Australian vegetation phenology: new insights from satellite remote sensing and digital repeat photography. Biogeosciences **13**, 5085–5102 (2016)
14. Snyder, K., Wehan, B., Filippa, G., Huntington, J., Stringham, T., Snyder, D.: Extracting plant phenology metrics in a great basin watershed: methods and considerations for quantifying phenophases in a cold desert. Sensors **16**, 1948 (2016)
15. Manier, D.J., Laven, R.D.: Changes in landscape patterns associated with the persistence of aspen (Populus tremuloides Michx.) on the western slope of the Rocky Mountains, Colorado. For. Ecol. Manag. **167**, 263–284 (2002)
16. Rhemtulla, J.M., Hall, R.J., Higgs, E.S., Macdonald, S.E.: Eighty years of change: vegetation in the montane ecoregion of Jasper National Park, Alberta, Canada. Can. J. For. Res. **32**, 2010–2021 (2002)
17. Hendrick, L.E., Copenheaver, C.A.: Using repeat landscape photography to assess vegetation changes in rural communities of the southern Appalachian mountains in Virginia. 29. Mountain Research and Development, USA, pp. 21–29. https://doi.org/10.1659/mrd.1028 (2009)
18. Herrero, H.V., Southworth, J., Bunting, E., Child, B.: Using repeat photography to observe vegetation change over time in Gorongosa national park. Afr. Stud. Quart. **17**, 65–82 (2017)
19. Masubelele, M.L., Hoffman, M.T., Bond, W.J.: A repeat photograph analysis of longterm vegetation change in semi-arid South Africa in response to land use and climate. J. Veg. Sci. **26**, 1013–1023 (2015)

20. Clark, P.E., Hardegree, S.P.: Quantifying vegetation change by point sampling landscape photography time series. Rangel. Ecol. Manag. **58**, 588–597 (2005)
21. Kull, C.A.: Historical landscape repeat photography as a tool for land use change research. Norsk Geografisk Tidsskrift – Norw. J. Geogr. **59**, 253–268 (2005)
22. Hall, F.C.: Ground-based photographic monitoring. General Technical Report PNW-GTR- 503. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland (2001)
23. Roush, W., Munroe, J.S., Fagre, D.B.: Development of a spatial analysis method using ground-based repeat photography to detect changes in the alpine treeline ecotone, Glacier National Park, Montana, U. S. A. Arctic Antarct. Alp. Res. **39**(2), 297–308 (2007)
24. Clark, P.E., Hardegree, S.P.: Quantifying vegetation change by point sampling landscape photography time series. Rangel. Ecol. Manag. **58**, 588–597 (2005)
25. Blaschke, T.: Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. **65**, 2–16 (2010)
26. Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A., Fisher, P.F.: A critical synthesis of remotely sensed optical image change detection techniques. Remote Sens. Environ. **160**, 1–14 (2015)
27. Fortin, J.A., Fisher, J.T., Rhemtulla, J.M., Higgs, E.S.: Estimates of landscape composition from terrestrial oblique photographs suggest homogenization of Rocky Mountain landscapes over the last century. Remote Sens. Ecol. Conserv. **5**(3), 224–236 (2018)
28. Fluckerpost Homepage: http://www.flukerpost.com/. Last Accessed 30 July 2019
29. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**, 977–1000 331 (2003)

Chapter 12
Can Data Fusion Increase the Performance of Action Detection in the Dark?

Anwaar Ulhaq

Abstract Automated human action detection and recognition is a challenging research problem due to the complexity of its data. Contextual data provides additional cues about the actions like if we know car and man, we can short-list actions involving car and man, i.e., driving, opening the car door etc. Therefore, such data can play a pivotal role in modelling and recognizing human actions. However, the visual context during night is often badly disrupted due to clutter and adverse lighting conditions especially in outdoor environments. This situation requires the visual contextual data fusion of captured video sequences. In this paper, we have explored the significance of contextual data fusion for automated human action recognition in video sequences captured at night-time. For this purpose, we have proposed an action recognition framework based on contextual data fusion, spatio-temporal feature fusion and correlation filtering. We have performed experimentation on multi-sensor night vision video streams from infra-red (IR) and visible (VIS) sensors. Experimental results show that contextual data fusion based on the fused contextual information and its colourization significantly enhances the performance of automated action recognition.

Keywords Data fusion · Action detection · Night vision

12.1 Introduction

Recognizing human actions in different visual conditions is still a challenging computer vision problem. Different challenging scenarios are considered in literature like action in large collections [1], group actions [2], crowd based actions [3],

A. Ulhaq (✉)
School of Computing and Mathematics, Charles Sturt University, Port Macquarie, NSW, Australia

Centre of Applied Informatics, Victoria University, Melbourne, VIC, Australia
e-mail: aulhaq@csu.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_12

actions inside movies [4], single and two person action recognition [5], actions recognition from side of a video [6], actions across different viewpoints [7] occluded actions [8]. However, these approaches consider action recognition in bright and good quality videos and do not discuss the case of adverse lighting conditions or night-time scenarios where clutter and disrupted context makes automated action recognition a daunting task.

Action contexts is a valuable priori knowledge in terms of modelling action instances. Contextual action recognition is discussed by various researchers. The context of scenes is utilized for recognizing events by [9] however, it uses only the static images. Contextual action recognition is presented by [10] which is based on bag-of-features framework. This work considered the annotated actions in movies and with script mining for visual learning. A similar technique [11] extracts overall object based context by detectors and their descriptors with supervised learning. Modelling of scene and object context is designed by [12] for Hollywood2 action dataset. These approaches aim action recognition in high-resolution videos. Therefore, the main advantage available to these approaches is the ease in finding visual interest points and relevant detectors because of their high quality visual context. However, achieving the same objectives in night-time imagery is cumbersome due to clutter and low lighting conditions resulting in low accuracy of action recognition approaches.

Human action activity recognition in single spectrum is discussed in [13, 14] which perform recognition in infra-red spectrum. However, these approaches ignore action contexts which is poorly captured by infra-red sensors (IR). These approaches, therefore, cannot be classified as contextual action recognition approaches. In this paper, we build upon the idea of our previous work [15–19] and further explore the use of contextual information in recognizing human actions.

Moreover, night-time imagery lacks colour information that provides great help to our visual perception. Due to unnatural appearance and IR imagery limitations, multi-sensor systems are employed for better contextual awareness [23]. Another approach to optimize these systems is to introduce pseudo-color information. A recent study about perceptual evaluation [20] of such colour transformed multi spectral systems has concluded that pseudo-colourization better illustrates the gist of a night scene by improving the fixation behaviour of human eye compared to large-scale imagery.

We address the following research questions: (i) Can accuracy of action recognition be increased by context enhancement? And (ii) Does psedu-colourization of night-time imagery contributes to the increase of recognition and detection accuracy?

This paper claims the following contributions: (1) It considers a diverse scenario of action recognition in clutter and adverse lighting conditions at night-time and answers how colourization and contexts can be utilized to enhance automated action recognition at night, (ii) it proposes to integrate motion, colour and context information in a single action recognition framework by space-time feature fusion and correlation filtering, (iii) it proposes a new spatio-temporal frequency domain filter for action recognition named (Action-02MCF).

The paper is organized as follows: Sect. 12.2 describes the prior related work. Section 12.3 illustrates the context data fusion of multi-sensor videos. The proposed framework is presented in Sect. 12.4. Experimental results are discussed in Sect. 12.5. The conclusion and references are provided at the end.

12.2 Prior Work

There are various techniques for human action recognition which can be categorized on the basis of action dataset used, as the performance of these approaches vary in different circumstances. Action-scene context is acquired through movie-script mining by [1] for realistic action recognition in movies. Spatio-temporal action context was utilized by [11] based on space-time features. Similarly, [21] employs convolution neural network for contextual action recognition. However, these approaches use high resolution action datasets for which the extraction of spatio-temporal interest points is straight-forward.han2005human

However, recognition of human actions in low-quality night-time videos is not well explored area of research and very few approaches can be cited in this category. The utility of thermal imagery is analysed by [22] for human action recognition. This approach is built upon the histogram of oriented gradients and nearest neighbour classification. A similar work [13] uses gait energy images. However, it is limited to walking activity which is easier to recognized. Contextual action recognition based on 3D-FFT and contextual cues is proposed in [16]. This approach uses context of night vision into consideration. However, its recognition performance is low. Our approach can be categorized in similar domain.

In this paper, we present robust action recognition which can deal with low quality night-time video sequences. In case of night-time videos, we consider the registered videos collected from low-light visible and IR spectrum. We enhance the context through video fusion [19]. Our action recognition approach is based on space-time interest point detection and frequency domain correlation analysis and can detect and classify human actions in a robust manner.

12.3 The Context Data Fusion in Night-Time Videos

In this section, we discuss the motivation behind contextual data fusion for night-time video sequences, video fusion and colourization for context enhancement.

12.3.1 The Motivation

The aim of context data fusion is a pre-processing step to give day-like appearance to night-time videos. It involves video fusion applied on registered video streams collected from infra-red (IR) and visible (VIS) spectrum Context data fusion helps to reveal a camouflaged target and to assist target localization [23]. Here we present and discuss context data fusion briefly.

12.3.2 Contextual Data Fusion

The objective of employing video fusion is to generate a single enhanced video from complementary videos, that is more suitable for the purpose of human visual perception, action and context recognition. If we denote A as IR video sequence and B as visible video sequence, we intend to generate another video sequence C by fusing visual information from A and B. Figures 12.1 and 12.2 give illustrations of video fusion results.

For simultaneous fusion and colourization, we used automatic colour transfer based video fusion (FACE) [17] which enhances video context by colour transfer from a source image. The illustration of this approach is given in Fig. 12.2.

12.4 Spatio-Temporal Feature Extraction and Fusion

In this section, we discuss feature extraction from video sequences containing valuable information regarding motion, colour and context.

Fig. 12.1 An example scenario of video fusion of registered video streams: (**a**) an IR video sequence (**b**) a low-light visible video stream and (**c**) a fused video generated from (**a**) and (**b**)

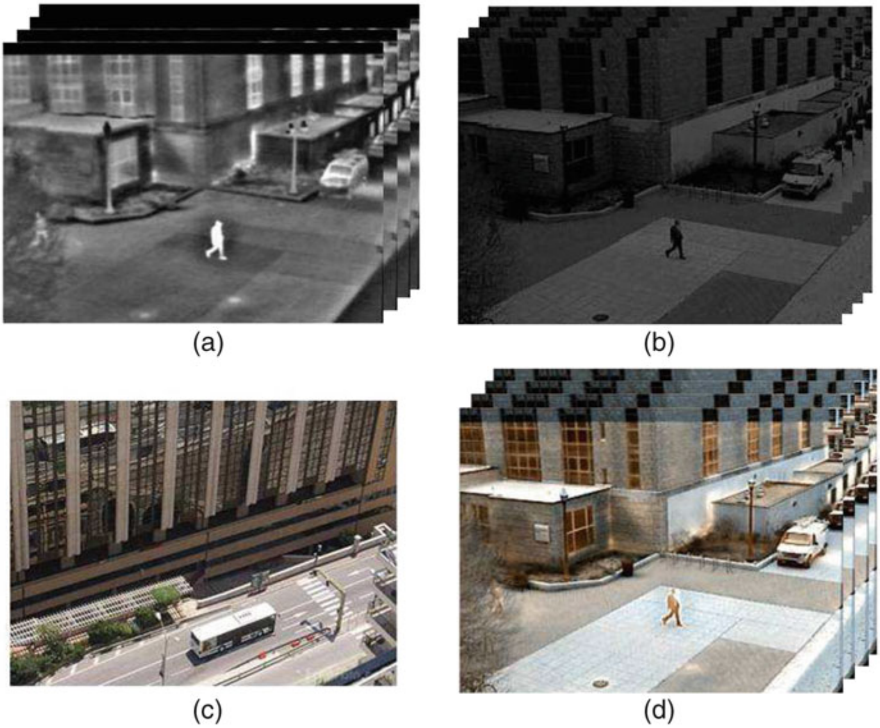


Fig. 12.2 Colour Transfer based Video Fusion method: (**a**) an IR video sequence (**b**) low light visible domain video sequence (**c**) a source colour image for the purpose of colour transfer and (**d**) a colour fused video generated from (**a**) and (**b**) and (**c**)

12.4.1 Motion and Colour Information Cues

To capture motion and colour information in video sequences, we use multi-channel 3D colour cuboid as spatio-temporal features with HOG3D descriptor [24]. To construct multi-channel framework, an opponent colour space conversion is applied to RGB frames as described below:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} (R-G)/\sqrt{2} \\ (R+G-2B)/\sqrt{6} \\ (R+G+B)/\sqrt{3} \end{pmatrix} \tag{12.1}$$

Above transformation de-correlates the colour channels and new photometric image representations, $\mathcal{N} = [\frac{O_1}{O_3}, \frac{O_2}{O_3}]$, $\mathcal{C} = [O_1, O_3]$ can be achieved. Chromatic combination with intensity channel $\mathcal{I}$ can produce $\mathcal{IN}$ and $\mathcal{IC}$ representations. We prefer to use $\mathcal{IN}$ based on best recognition performance in [24].

Let volume V having n channels is given by $V = (V^1, V^2, \ldots\ldots., V^n)^T$ with individual channels are presented in scale-space by $V^j = G(.; \sigma_o; \tau_o) * I^j(.)$ with G as a 3D Gaussian kernel with equal scale along spatial and temporal dimension i.e. $\sigma_o; \tau_o$ and I^j is image function of channel j. Multi-channel spatio-temporal cuboid features Γ are then extracted as:

$$\Gamma = \sum_{i=1}^{n} (G(.; \sigma_o * a_{av}(.; \tau_i);) * V^i)^2 + (G(.; \sigma_o * a_{od}(.; \tau_i);) * V^i)^2 \tag{12.2}$$

where Gaussian kernel $G(.; .)$ is applied spatially and Gabor filter pair a_{av}, a_{od} are applied along temporal dimension.

Multi-channel HOG3D descriptor, α encapsulates the histogram of 3D gradient directions and can be described as:

$$\alpha = (\alpha^1, \alpha^2, \ldots\ldots., \alpha^n)^T = \frac{L.\nabla'}{||\nabla||^2} \tag{12.3}$$

where L is n matrix holding the polyhedron face centre locations as the unit sphere centred at the gradient location is approximated by a regular n-sided polyhedron, f is the histogram of 3D gradient directions and For a 3D gradient ∇, and a multi-channel formation is given using concatenation approach where $\nabla' = \{\nabla^j\}$, $j = \{1, 2, \ldots, n\}$. We use gradient direction with channel concatenation and the dimensionality of the final descriptor is $n \times d$.

12.4.2 Contextual Information Cues

To extract contextual information, we propose multi-channel Gist features as an extension of Gist features [25] that are triggered at temporal locations specified by multi-channel cuboid features. The Gist feature is a vector $\beta = \{\beta^j\}$, where $j = \{1, 2, \ldots., n\}$ represents channel j, n is the number of channels and each individual feature, cuboid Gist β is averaged Gist feature for all frames described by a triggered spatio-temporal cuboid and is computed as:

$$\beta^j = \sum_{i=1}^{N_f} \sum_{x,y} w(x, y) \times |I_i(x, y) \otimes h_i(x, y)|^2 \tag{12.4}$$

Where $\otimes$ represents the image convolution, $\times$ denotes pixel wise multiplication, $I(x, y)$ stands for the frame of video sequence, and N_f is number of total frames specified by a given spatio-temporal cuboid. $h_i(x, y)$ is a filter from multi-scale oriented Gabor filters and w denotes a spatial window to compute the average output energy of each filter. The impulse response of a Gabor filter is considered a Gaussian modulated by a harmonic function:

$$h(x, y) = \cos(2\pi \frac{x'}{\lambda} + \phi) exp(\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \tag{12.5}$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, θ is rotation angle, ϕ is the phase offset, γ is the spatial aspect ratio and λ is the harmonic wavelength, σ is the standard deviation of the Gaussian function. The window $w(x, y)$ divides an image into a grid of 4×4 non-overlapping windows, producing a descriptor of size $4 \times 4 \times 6 \times 4 = 384$.

12.4.3 Spatio-Temporal Feature Fusion

Feature fusion combines the different complementary features into a single discriminative feature vector. We take spatio-temporal feature as $\alpha \in \mathbb{R}^{p\times n}$ and contextual features as $\beta \in \mathbb{R}^{q\times n}$. Therefore for n video sequences, we get $p + q$ features. Let $S_{\alpha\alpha} \in \mathbb{R}^{p\times p}$ and $S_{\beta\beta} \in \mathbb{R}^{q\times q}$ denote the within -sets covariance matrices. The overall covariance matrix $(p + q) \times (p + q)$ S contains all the information on the associations between pairs of features:

$$S = \begin{pmatrix} Cov(\alpha) & Cov(\alpha, \beta) \\ Cov(\beta, \alpha) & Cov(\beta) \end{pmatrix} = \begin{pmatrix} S_{\alpha\alpha} & S_{\alpha\beta} \\ S_{\beta\alpha} & S_{\beta\beta} \end{pmatrix} \tag{12.6}$$

CCA aims to find the linear combinations $\alpha^* = W_\alpha^T Y$ that maximize the pairwise correlations across the two datasets,

$$Corr(\alpha^*, \beta^*) = \frac{Cov(\alpha^*, \beta^*)}{var(\alpha^*).var(\beta^*)} \tag{12.7}$$

where $Cov(\alpha^*, \beta^*) = W_\alpha^T S_{\alpha\beta} W_\beta$, $var(\alpha^*) = W_\alpha^T S_{\alpha\alpha} W_\alpha$, and $var(\alpha^*) = W_\beta^T S_{\beta\beta} W_\beta$. Maximization is performed using Lagrange multipliers by maximizing the covariance between α^* and β^* subject to constraints $var(\alpha^*) = var(\beta^*) = 1$. The transformation matrices W_α and W_β are produced by solving the eigenvalue equations

$$\begin{aligned} A^2 \bar{W_\alpha} &= S_{\alpha\alpha}^{-1} S_{\alpha\beta} S_{\beta\beta}^{-1} S_{\beta\alpha} \bar{W_\alpha}, \\ A^2 \bar{W_\beta} &= S_{\beta\beta}^{-1} S_{\beta\alpha} S_{\alpha\alpha}^{-1} S_{\alpha\beta} \bar{W_\beta} \end{aligned} \tag{12.8}$$

where $\bar{W_\alpha}$ and $\bar{W_\beta}$ are the eigenvectors and A^2 are the diagonal matrix of eigenvalues of square of the canonical correlations. A feature level fusion is then performed as:

$$f = \begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} W_\alpha^T \alpha \\ W_y^T \beta \end{pmatrix} = \begin{pmatrix} W_\alpha & 0 \\ 0 & W_\beta \end{pmatrix}^T \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{12.9}$$

Where f is called fused discriminant feature and represents fused motion, colour and contextual information based feature.

12.5 Action-02MCF: 3D Feature-Based Zero-Aliasing Maximum-Margin Correlation Filter

Notation: Vectors are represented by the lower case letters f, matrices by upper case letters F, $\mathcal{F}^{-1}(\hat{f})$ denotes inverse Fourier transform of $\hat{f}$, where $\hat{}$ represents variables in frequency domain, $'$ denotes transpose operation. $*$ is complex conjugate and $\dagger$ represents its transpose. For implementation detail, please refer to [26].

12.6 Experimental Results and Discussion

This section describes our experimental data, set-up, results and performance comparison with discussion.

12.6.1 Actions Dataset and Experimental Set-Up

In absence of any benchmark night-vision action dataset, we have recorded Night Vision Action Dataset (NV) using two different cameras. One of them is IR camera, Raytheon Thermal IR-2000B and the other is low-light visual camera, Panasonic WV-CP470. The thermal and visual videos are registered before fusion process by selecting corresponding points in corresponding views and following a computation of a least-squared error fitting homography. In addition to these videos, this dataset includes 20 video sequences collected from TNO image fusion dataset [20], Eden project dataset [27] and Ohio-state University thermal dataset. This dataset comprises eight action categories including: walking, wave1, wave2, stand-up, sit-down, clapping, pick-up and running performed by different actors.

12.6.2 Experiment No. 1: The Role of Contextual Data Fusion in Terms of Recognition Accuracy

Firstly, the training of Action-02MCF filters is performed for each action category. During the testing phase, test action video is correlated with the synthesized filter to find correlation peak. For validation, leave one out cross validation strategy is used. The confusion matrices are shown in Fig. 12.3.

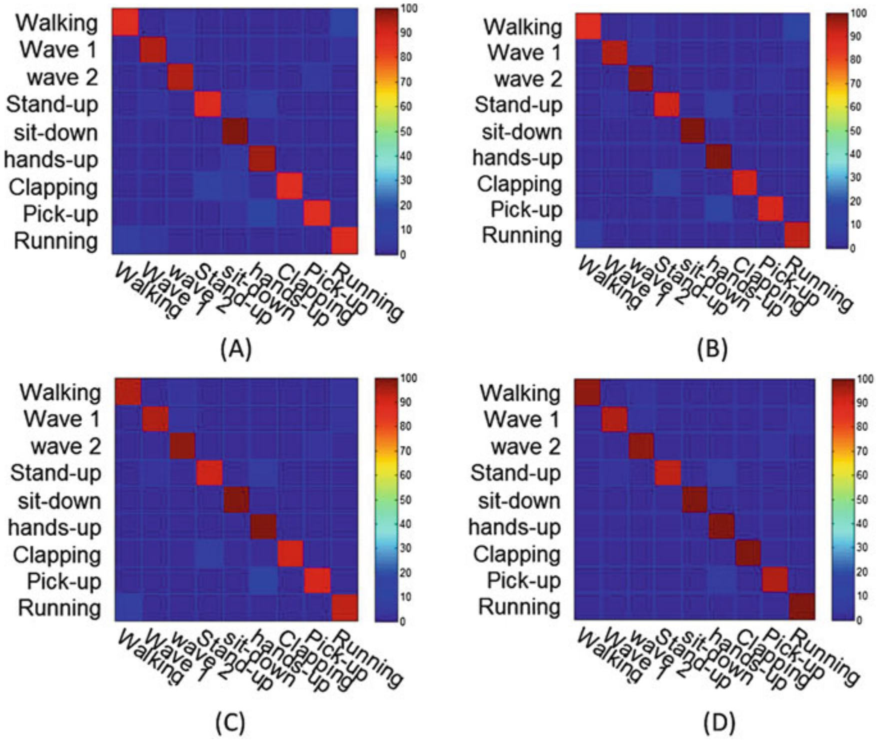


Fig. 12.3 Confusion matrices illustrate action recognition performance against NV action dataset. (**a**) Baseline (without colour and contextual information) 91.33%, (**b**) Without Colour information,92.88% (**c**) Without contextual information 94.88%, (**d**) With fused colour and contextual information 97.44%

We experimented with different versions of our recognition framework to know the impact of each visual cue (colour and context) on recognition performance. At first, a baseline is developed using space-time cuboid features and excluding context and colour information in feature extraction process, and we call it Action-baseline. At second, we fuse contextual information through Gist features but no colour channel is used, and we call it Action-02MCF At third, we use multi-channel space-time cuboid features without Gist features and we call it Action-02MCF At last, we utilized fused discriminant features and we call it Action-02MCF.

We calculated confusion matrices for each case against NV dataset and displayed in Fig. 12.4. We found that Action-02MCF (Fig. 12.3d) that use both colour and context information fusion alongside motion cues outperforms others (Action-02MCF (Fig. 12.3a–c)) It demonstrated that colour information is significantly important alongside motion information in action recognition process. Contextual information also plays an important role especially in actions which involve full-body motion. Therefore, a information fusion of motion, colour and contextual cues can enhance action recognition performance.

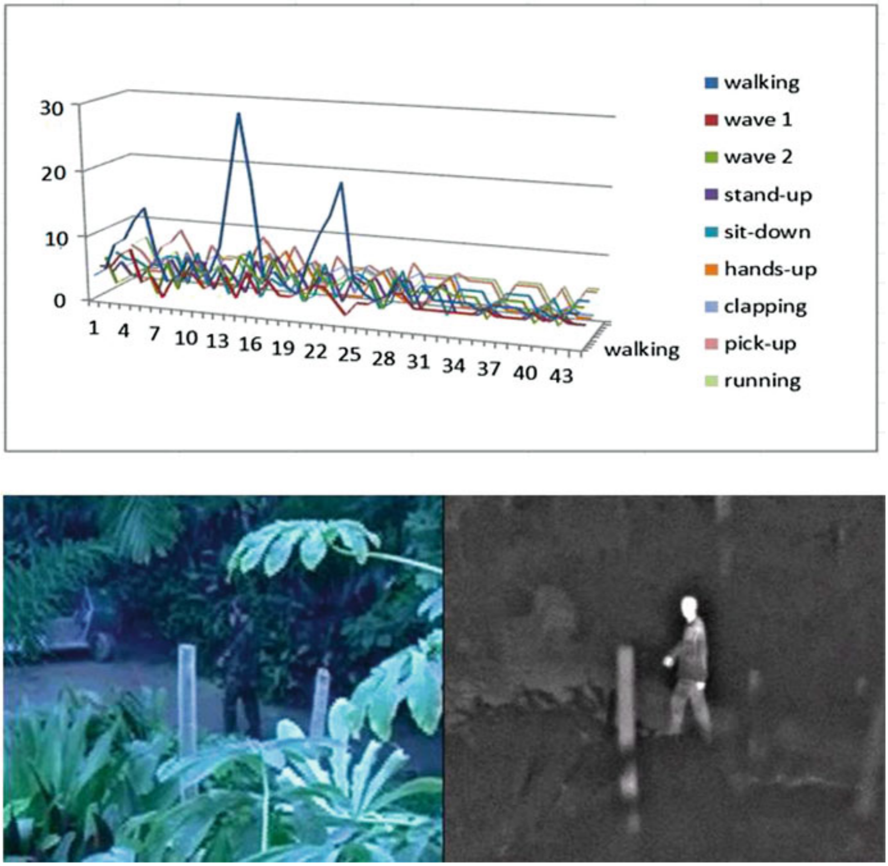


Fig. 12.4 The plot of PSR (Peak-to-Sidelobe-Ratio) by correlating the trained walking-Action-02MCF with a night-time video sequence. As visible from the plot, PSRs produced by walking-Action-02MCF are comparatively much higher than the responses by action filters for other actions. A representative frame from respective videos (both domains) is shown below

12.6.3 Experiment No. 2: Filter Performance for Action Detection and Localization

To measure the detection and localization performance of the proposed filter, we utilize the probability of detection vs false alarms per second (FA/s). A performance metric denoted as P is utilized which is equal to the integration of an ROC curve from 0 to 5 FA/s. An ideal ROC curve must have, P = 5. To evaluate this performance, we applied the proposed filter to each test video and varied the threshold of the detection to generate ROC curves. The detection is labelled a true positive detection if the ground truth and the centre of the bounding box lie within 3 frames of each other and $\leq$ 8 pixels Euclidean distance in spatial domain to keep greater than 50% bounding box overlap for each action. We then plot the values of performance metric, P against all actions and do comparison with similar approach, Action-DCCF filter [16]. The corresponding filter is selected due to similarity and code availability. Figure 12.4 displays the sample detections with the original and estimated bounding boxes while Fig. 12.5 illustrates the plot of recognition rate vs normalized correlation error (E).

12.6.4 Experiment No.3: Quantitative Evaluation of Filter Robustness

We use another quantitative metric, named Peak-to-Sidelobe-Ratio (PSR) described in [28] which calculated the ratio of peak response to the local surrounding response. Figure 12.6 plots PSRs for walking action present in test video sequence (sample frame displayed) using Action-02MCF for walking action trained on NV dataset.

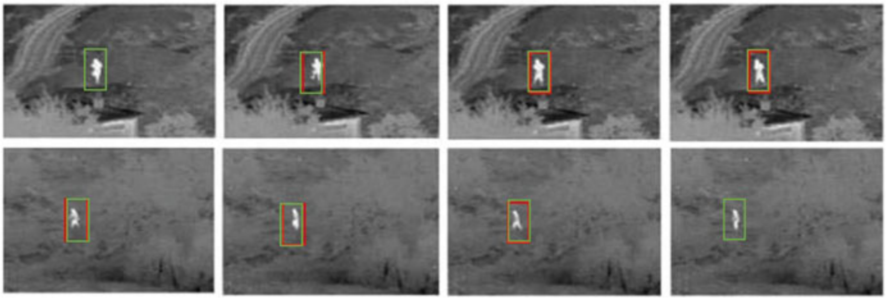


Fig. 12.5 Action instance detection in three NV Actions instances where red bounding box is the actual ground truth while green bounding box shows detection by 3D-SDCF

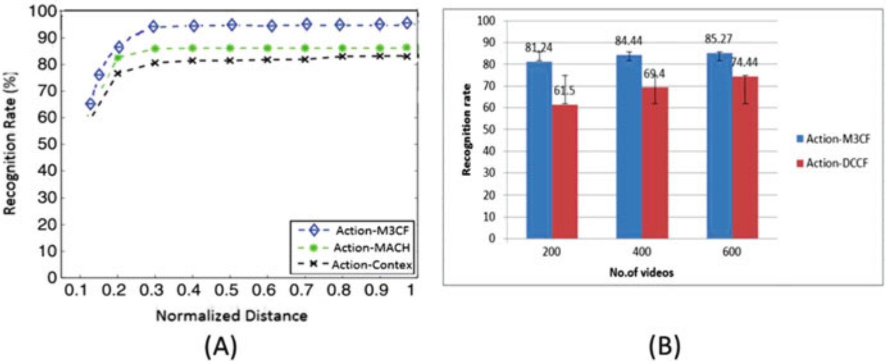


Fig. 12.6 (**a**) The plot of recognition rate vs normalized correlation error (E) for different approaches. For every value of E, Action-02MCF outperforms other classifiers, (**b**) Recognition rate vs no. of training videos

12.7 Conclusion

In this paper, we explored and discussed the role of diverse data like colour, context and motion information for automated action recognition. We use Canonical Correlation Analysis for feature fusion of colour, contextual and motion data. In addition, we introduced a robust feature based space-time action recognition called Action-02MCF. The proposed filter combines the maximum margin property of SVMs, localization criteria of correlation filters and sparsity of spatio-temporal features in a single framework. Furthermore, we discovered that colourization clue is an important data to increase action recognition performance alongside contextual data.

References

1. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos “in the wild", in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003, (2009).
2. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks, in: IEEE International Conference on Computer Vision, pp. 742–749, (2003).
3. Siva, P., Xiang, T.: Action detection in crowd., in: British Machine Vision Conference, pp. 1–11, (2010).
4. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, (2008).
5. Islam, S., Qasim, T., Yasir, M., Bhatti, N., Mahmood, H., Zia, M.: Single-and two-person action recognition based on silhouette shape and optical point descriptors, Signal, Image and Video Processing **12** (5) (2018) 853–860.
6. Pei, L., Ye, M., Zhao, X., Xiang, T., Li, T.: Learning spatio-temporal features for action recognition from the side of the video, Signal, Image and Video Processing **10** (1) (2016) 199–206.
7. Ulhaq, A., Yin, X.S., He, J., Zhang, Y.: On space-time filtering framework for matching human actions across different viewpoints, IEEE Transactions on Image Processing **27** (3) (2018) 1230–1242.
8. Weinland, D., Özuysal, M., Fua, P.: Making action recognition robust to occlusions and viewpoint changes, in: Computer Vision–ECCV, Springer, pp. 635–648, (2010).
9. Li, L.-J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition, in: IEEE 11th International Conference on Computer Vision, pp. 1–8, (2007).
10. Marszalek, M., Laptev, I., Schmid, C.: Actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936, (2009).
11. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition., in: ICCV, Vol. 9, pp. 1933–1940, (2009).
12. Jiang, Y.-G., Li, Z., Chang, S.-F.: Modeling scene and object contexts for human action retrieval with few examples, IEEE Transactions on Circuits and Systems for Video Technology, **21** (5) (2011) 674–681.
13. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery, in: CVPR Workshops., IEEE, pp. 1–17, (2005).
14. Li, J.F., Gong, W.G.: Application of thermal infrared imagery in human action recognition, in: Advanced Materials Research, Vol. 121, pp. 368–372, Trans Tech Publ, (2010).
15. Mirza, A., Qamar, S., et al.: An optimized image fusion algorithm for night-time surveillance and navigation, in: Proceedings of the IEEE Symposium on Emerging Technologies, 2005., IEEE, 2005, pp. 138–143.

16. Anwaar, H., Iqbal, G., Murshed, M.: Contextual action recognition in multi-sensor nighttime video sequences, in: Digital Image Computing Techniques and Applications (DICTA), pp. 256–261, (2011).
17. Ulhaq, A., Yin, X., He, J., Zhang, Y.: Face: Fully automated context enhancement for nighttime video sequences, J. Vis. Commun. Image Representation **40** (2016) 682–693.
18. Ulhaq, A.: Action recognition in the dark via deep representation learning, in: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), IEEE, pp. 131–136, (2018).
19. Anwaar, H., Iqbal, G., Murshed, M.: Automated multi-sensor color video fusion for nighttime video surveillance, in: IEEE Symposium on Computers and Communications (ISCC), pp. 529–534, (2010).
20. Toet, A., de Jong, M.J., Hogervorst, M.A., Hooge, I.T.: Perceptual evaluation of colorized nighttime imagery, in: IS&T/SPIE Electronic Imaging, pp. 1–14, (2014).
21. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r∗ cnn, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088, (2015).
22. Li, J.F., Gong, W.G.: Application of thermal infrared imagery in human action recognition, in: Advanced Materials Research, Vol. 121, pp. 368–372, Trans Tech Publ (2010).
23. Shah, P., Reddy, B.C.S., Merchant, S.N., Desai, U.B.: Context enhancement to reveal a camouflaged target and to assist target localization by fusion of multispectral surveillance videos, Signal, Image and Video Processing **7** (3) (2013) 537–552.
24. Everts, I., Gemert, J., Gevers, T.: Evaluation of color stips for human action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2850–2857, (2013).
25. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition, Progress in brain research 155 (2006) 23–36.
26. Ulhaq, A., Yin, X., Zhang, Y., Gondal, I.: Action-02mcf: A robust space-time correlation filter for action recognition in clutter and adverse lighting conditions, in: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 465–476, Springer (2016).
27. Lewis, J., Nikolov, S., Loza, A., Canga, E.F., Cvejic, N., Li, J., Cardinali, A., Canagarajah, C., Bull, D., Riley, T., et al.: The eden project multi-sensor data set, The Online Resource for Research in Image Fusion (ImageFusion. org).
28. Kumar, B.V.K.V., Mahalanobis, A., Juday, R.D.: Correlation Pattern Recognition, Cambridge University Press, New York, NY, USA (2005).

Chapter 13
Data Privacy and Security in the Cloud

Peter Padiet, Md. Rafiqul Islam, and Azizur Rahman

Abstract Cloud computing have slowly made it way and has gained popularity with its services and products and they have repeatedly increased, and many users are attracted to migrate their data into the cloud. But there are still issues which concern users when outsourcing their data and the business application into the cloud, security and privacy are very critical and created trust concern. There are five attributes related to security and privacy as follow; privacy-preservability, availability, integrity, confidentiality and accountability. The aim of this paper is to improve security in the cloud because the uses of cloud services has been more increasing and in the next few years would be likely the cloud services will be a big driving force, better to be prepare and develop majors to encounter and security loophole that may arise in the near future.

Keywords Cloud · Privacy · Security · SaaS

13.1 Introduction

The aim of the research is to enhance the privacy and security in the cloud, the uses of cloud services has been more increasing attracted to uses cloud services by migrating most of their data to the cloud environment.

Cloud computing has become as one of the growing industries by making more services or applications available on the Internet but still there are many challenges that are facing the industry e.g. lack of expertise who are committed to developing the security strategies needed in the twenty-first century.

P. Padiet · Md. R. Islam (✉)
Charles Sturt University, Wagga Wagga, NSW, Australia
e-mail: mislam@csu.edu.au

A. Rahman
School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW, Australia

© Springer Nature Singapore Pte Ltd. 2020
A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_13

Cloud computer is a computer paradigm whereby a larger pool of the systems are being connected either privately or public networks so that to provide the dynamically scalable infrastructure for example; data, application and files storage and with the advent of the technology and also the cost of an application hosting, computation content storage and the delivery is also reduce significantly.

In the last few years, Cyber Security has become as one of the growing industries but still there are many challenges that are facing the industry e.g. lack of expertise who are committed to develop the security strategies needed in the twenty-first century.

Security is one of the main concerns in the cloud services and I am very determined to work hard in collaboration to other researches who shared same ambition to make sure the security should be worrying anyone although there are always shortfalls. In the new few years big Corporate and Governments will be more likely to be using cloud services. Most client are attracted because of cost efficient and resources availability [1].

To make data access more secure all users should have multiple verification layers which allow users to be verify before accessing their data. Users' needs to be authenticated with unique access codes which can be encrypted. Each user should have multiple tokens to go through difference level of authentications.

13.2 Literature Review

As seen, many users including big organizations, private and governments have moved to host their data in the cloud, there is a complete loss of control, trust and multi-tenancy with the self-managed clouds that have been directing to the security. Cyber security is not new it has been existing and rapidly growing due to increase and advancing of technology.

The complete loss of control, data, applications and the resources have been located where the user identity management and the user access rules work on the security policies and the enforcement which has been through the providers of the cloud [3]. The consumer works on the assurance of the resource availability, monitoring and the repairing of the services. There is a confidentiality issue where the fear of the loss of control over the data along with the increased attack surface. The entity outside the organisation works on the stores and the data computation where the attackers are able to target the link of communication between the cloud providers and the client. The identity and the access management and the privacy issues are the biggest challenges.

Cloud computing is considered by many users as being very promising for their IT applications; but there are still some outstanding issues to be solved e.g. data access and privacy concern in cloud environment to create trust between the users and the cloud providers [2].

For private, enterprises and government users can store their data in the cloud computing environment and deploy their applications, but the significant barrier is data security which are accompanied by issues including but not limited to privacy, compliance, legal matters and trust. The security is a very complex and large topic to address. There are needs to expand security due to the amount of data being uses in the cloud by big corporate, governments and individual [4].

There is a need to introduce a data monitoring mechanism in the cloud to make sure the data will not be manipulated to ensure Data Integrity by protecting unauthorized data modification, deletion.

In any system data integrity is one of the elements considered as one of the most critical. It means the general protection of data from unauthorized access including data deletion, modification of data.

Data confidentiality is very significant for users to store their private or confidential data in the cloud [8].

13.2.1 Relevant Technologies & Its Applications

Cloud computing provided services to users on demand at an affordable cost, it is made affordable to Government, private agencies, big businesses and small enterprises and that make it more attractive to many users prompting many users to outsources their data for the cost of storage and infrastructure cost as well [7].

There are three relevant technologies of Cloud computing which are:

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

SaaS it is used to deploy provider's application to run on the cloud infrastructure and it offer access everywhere which also increases the security risk. This service model is very important to the implement policies and for identity management and it provide access control to the applications. Salesforce.com is one of the examples only provide access to authorized salespeople download customers sale information.

PaaS this environment is shared for development used by customers to control application which have been deployed (example) Microsoft™ Windows AZURE but doesn't deal with the fundamental cloud infrastructure. This cloud services involved with very strong authentication used to identify users and audit trail with capability of supporting compliance policy and privacy mandate. Uses for the development, test, deployment, host and maintaining of the applications with the integrated development. This is based on handling the creation of the tools, modification and testing the deployment. This is based on the built-in scalability which includes the load balancing and the failover. The integration has been with the web services and the databases. This has been working on the agile software development and includes the Google App Engine and the Microsoft Azure services.

IaaS works on the delivering of the servers, storage network and the operating systems. This is on demand service which can handle the clients with the resources under the distributed service, allowing the dynamic scaling and have a utility of the pricing model. This includes the users on a single hardware piece. IaaS has been very volatile where there has been rapidly growing organisation to work on the limited capital expenditure. There has been off shoring and the outsourcing of the data storage and the processing is difficult [5].

Data security has constantly been an issue of concern in information technology. In the cloud computing environment, it become mainly an issue when the data been located in different cloud computing infrastructure around the globe. The two main issues of concern to users are; Data security and privacy protection. Data security and privacy protection are become the main important issues of concern in the technology development, many users are focusing on how the technology can be secure and convenience to use in regard to cloud computing.

This study is aiming to review security techniques and challenges in both hardware and software to ensure data protection in the cloud computing [1].

13.2.2 Data Integrity

Generally, Data integrity is the way of protecting data from being access by unauthorized person to avoid deletion, modification, or fabrication.

Data integrity is considered as one of the most critical fundamentals in information system [1].

13.2.3 Data Confidentiality

Data confidentiality is significant for users when they store their private and confidential data the cloud. In order to ensure data confidentiality users, must use Authentication and access control strategies [1].

Increasing cloud reliability and trustworthiness can be used to address issues of authentication, data confidentiality, and access control in cloud computing.

It is still a big challenge to could providers although cloud providers try to create trust between users and their services, it is a big challenge for users to store their sensitive data in the cloud [1].

13.2.4 Data Availability

Data availability means the availability of data to users to access their data on their daily basis to certify the users need, there are many risks that can cause failure for data interruption, like hard disk damage, network failure, power failure or fire and any other risks that can interrupt data availability can providers have to guarantee the availability of data if any avoidable risks occur.

The issue of data being store in many servers either locally or including overseas or foreign countries is one of the serious concern of clients because the cloud providers can be subject to laws of the country they resided. It is very important that cloud providers to provide users with guarantees that their data will be safe in the cloud provider servers and provide explanation to users regarding local jurisdiction or local laws. This paper will focus on the data storage location, safety of the data against the local jurisdiction laws, availability of the data against risks, data security management [1].

13.2.5 Privacy Definitions

Privacy can be define in so many ways across different entities but in the cyber security context, privacy is the personal/individual, government and organizational information or secrecy that should only be known within a certain boundaries without being expose to those who have nothing to do with the information, information can be obtain with consent per ethical needs to know basis but any information being access without a consent from the an entity can be consider as breach of privacy [1].

13.3 Research Significance

There is no mechanism available to trace who had access to the database in the cloud, there is no transparency from the cloud providers as well, users do not have access to the physical infrastructure where their data being stored. Users have no access to information about privacy and data protection policies from the Country their data being stored. Data access can be subject to the Country that the data being stored or where the cloud service provider reside.

Users always in a loop wondering how long the data can be still available in the cloud after their contract finished. If the data remain in the cloud environment after the contract finished, then it can be considered as data breach or data privacy breach, there must be a clear principal regarding data deletion or data destruction after the contract. Because if the data remain in the cloud environment for a certain time without any agreement from the user and the cloud provider then there might be no security grantee and it can be access by unauthorized person.

Users remain wondering to find answers about who are data administrators who look after their data in the cloud environment, are administrators local or international employees, there must be a clear documentation that to educate users more about data access management in the cloud to avoid doubt regarding privacy and data security protection.

Data security had constantly remained a main issue in information technology. In the cloud computing environment, it has become predominantly serious because the data is most stored or placed in different places and this can even be stored all globally. The two main factors of user's concerns are privacy and Data security protection in the cloud environment. Although various methods on the issues of cloud computing have been investigated by both academics and industries, privacy and data security protection have become important for the forthcoming development of the cloud computing technology within the government agencies, industry and in the business.

Data security and privacy protection problems are applicable equally to hardware and the software in the cloud architecture. Many questions remain answer and many users always look for answers to know exactly what happen to their data behind the cloud environment, many users outsource their data to the cloud platforms but sometime there are no clear policies regarding data protection although they sometime read them on a provided contract document. Many users don't fully understand the services they are sign on regarding if their data will be subject to country their data being host because each country have their own data security and privacy protection policy. There always hidden policies which are not disclose to users regarding data access and this remain as big issue of concern regarding data security and privacy protection. Nobody has gone to the bottom of this issue and until then the data security and privacy protection will remain the issue of concern when using cloud services. User lost control of their data control because being manage in the cloud where user only have limited control [1].

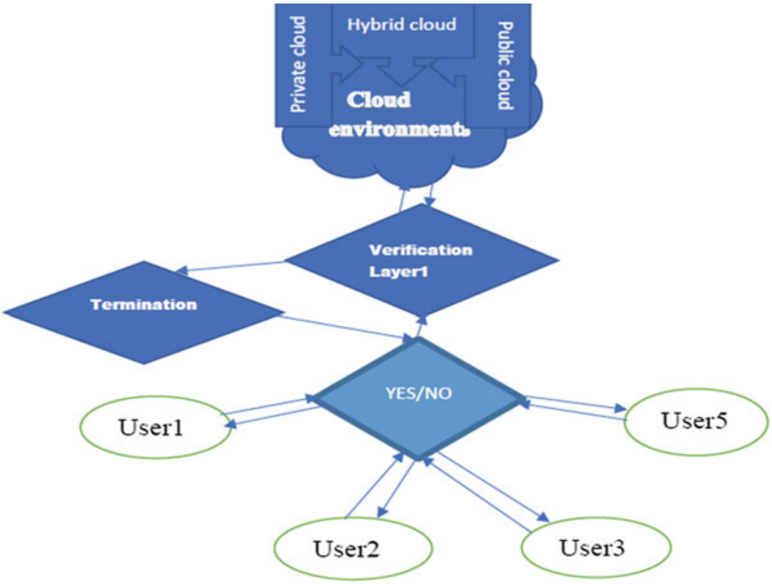
13.3.1 Research Gap and Relevant Issues

Sun et al. [1] This article summarized more on Data Mining, Security, Privacy, models and have address more regarding data mining and its implications and also act as educational steps for cloud providers to adopt the ideas to increase their productions.

Building trust between cloud providers and users is one of the critical or difficult thing to achieve due to the complexity of the data and privacy protection. In order to achieve the trust between cloud providers and users, there is a need to investigate regulations which are already in place and do necessary adjustment to improve the trust needed. Although many researches have already proposed many techniques to deal with data and privacy protection still there is a need to scrutinise data and privacy protection policies to breach the gap. Since the issue of data and privacy protection remain as a big challenge, this survey paper will focus more on where the data being store in the cloud, local government policies regarding data and privacy protection in the hosted country, data retention in the cloud after contract termination [1].

13.4 Proposed Method

The investigation aims to find possible measures in place to protect the privacy of all users either individual, governments, small, medium, big businesses in the cloud environment. To investigate measures that being put in place concerning the cloud environment being use by organisations to make sure the necessary protection needed is being deployed to protect data integrity. To investigate and find out what are most suitable control measures being uses by most organisations and businesses to protect their data not to be exploit.



Proposed methodology

The above diagram represents steps they users should go through before being granted full access to their data in the cloud.

To make data access more secure all users should have multiple verification layers which allow users to be verify before accessing their data. Users needs to be authenticated with unique access codes which can be encrypted. Each user should have multiple tokens to go through difference level of authentications, the first step is for the customer to login using unique username and a password. Second step, user should be verify with their face being scan in the verification layer1 and if the user fail to be fully identify with the image/photo then the section should be terminate and prompt a user to go back to logon screen and if the user is fully verified then he/she should be able to proceed to the next level whereby the user should presented with data sample and data should be verify against the image provided in the verification layer1 and if the data match up the image provided during the verification stage then the user will have a full access to the data base on the verification confirmation [9].

The aim for the multiple security verification layers is to prevent any malware that can get access to the confidential information through individual access point.

Security program to be run as follow:

If the user has a valid login credentials, then access 1 allow,

Else if user doesn't have valid login credential then terminate the session at initial point.

If the user is fully verified at the initial stage, then user will be required to prompt to be verify by their biometric or picture.

If the user is fully verified at the verification layer1 with their biometric identity, then you will be granted access to their data.

Or else if the user failed the verification at the layer1 stage then terminate the session and prompt user with access denied due to invalid information.

If the user failed verification at any stage during the login procedures, then user will always be prompted to start from the begin.

13.5 Investigation Methods and Sample Collection

Privacy literature reviews are non-systematic, many of them cannot be reproduced. Only some studies presented a methodical literature reviews that are related to security and privacy in support of big data. With anticipation the research on big data count on methodical Literature Reviews. This research will focus on data and privacy security of users in the cloud, although many researchers have done a lot of studies regarding Cyber Security generally but there is no specific finding or there is no specific solution regarding data security and privacy protection due its complexity. There are no specific literature reviews related to big data privacy that provide research making investigation and practical effect assessment on computational assets [10].

Considering the cloud computing, there is a need to protect against the different internal and the external threats where the monitoring of the security services will help in the improvement of the effectiveness of the customer [5].

The platform, control and the service monitoring help in handling the platforms which are being monitored. The vulnerability detection and the management help in enabling the automation verification with the management of the handle the unauthorised access to the administrative services. There is a continuous system for the patching and the fortification where the security patching is based on handling the operations and working on the different efforts.

Investigate aimed at cyber security issue and the appropriate method to be use in this research is; data acquisition. This method is to address the objectives and questionnaire for this research. This technique will be used to obtain information using design and provided questionnaires from many different businesses and organizations that have access to cloud computing environment.

The questionnaires will be mainly from data and privacy protection regarding cloud computing usage. The target sample data to be use for this investigation will be about 250–300 users and it could be more than the target number in order to present a quality result.

The sample questionnaires will be flexibility with options of hardcopy and softcopy including online version, this will make very convenience to many users who will take part in this investigation.

The investigation will be limited to governments, businesses and organizations that are involve in the use of private network system and have access to cloud computing environment or infrastructures. Other organization or institutions that uses public networks could be form part of the investigation for the comparison purpose regarding the risk challenge each entity faces.

The findings could benefit many organizations that are involve, with their data, systems and networks protection from cyber-attack. After successfully complete an investigation on this proposal the finding will be provide to organizations which took part in the investigations as form of feedback per the recommending in the finding [1].

13.6 Ethical Issues

To find out ethical issues regarding the survey or work you will do involving data that may be concern with privacy, the best way is to start by questioning yourself and Assessing and potential conclusions regarding ethics issues by showing your trustworthiness that will to development of more cyber security ethics practices.

Does my action cause any effect or harm to those who are involved?

What will be the likely outcome or results and what quality report will it represent?

To be consider as a cyber security professional you should behave ethically to gain trust from others and prove to your supervisors beyond reasonable doubt because of the responsibility you shoulder when gaining access to sensitive and valuable information. Consumers should be involved when it comes to cyber security ethical issues because of their data and privacy protection concern. There are challenges involved with trust from the stakeholders regarding their systems and sensitive data and their privacy.

13.7 Recommendation

User should be required to have multiple verification layers for authentication purpose before accessing the information or data in the cloud. Organizations should educate their employees to make them aware about the risk that will be involved when missing using access level, educating users is very important because it will reduce the risk of insiders attack, insiders attack is concerning and it can only be prevent within the organization with strong policies that restrict data access by unauthorized users.

Cloud providers should have to be very transparency with the services the provided to their clients/users. Users should be presented with risks that may occur or any hidden risks that users may not be aware of, cloud provided should not authorize data access by any third party without consent from their clients/users.

Activities log should be made accessible to users as part of transparency, Activities log should include report on who have had access to the data, the date data was last modified and any data transactions and who performed those activities.

13.8 Discussion and Conclusion

The aim of this paper is to investigate most of the hidden security concern like data administration issues in the cloud, regulations or laws in the countries where the data kept, data integrity and more. Privacy and data security are major issues of concern including the compliance and the trust. There have been issues related to the resource security, management and the monitoring process. The other relevant issues of the data security, governance and management which includes the recognition of the tracing, data confidentiality and avoidance of the malicious insiders with the illegal operations. The other issues of concerns are with the data privacy, protection, availability, location and the secured transmission process [6].

Cloud computing shows as potential and is an emerging technology intended for the subsequently invention of IT applications. Difficulty and hurdles toward the fast development of cloud computing are data security and privacy issues. Building trust between cloud providers and users is one of the critical or difficult thing to achieve due to the complexity of the data and privacy protection. In order to achieve the trust between cloud providers and users, there is a need to investigate regulations which are already in place and do necessary adjustment to improve the trust needed. Although many researches have already proposed many techniques to deal with data and privacy protection still there is a need to scrutinise data and privacy protection policies to breach the gap. Since the issue of data and privacy protection remain as a big challenge, this survey paper will focus more on where the data being store in the cloud, local government policies regarding data and privacy protection in the hosted country, data retention in the cloud after contract termination [1].

To make data access more secure all users should have multiple verification layers which allow users to be verify before accessing their data. Users' needs to be authenticated with unique access codes which can be encrypted. Each user should have multiple tokens to go through difference level of authentications.

References

1. Sun, Y., Zhang, J., Xiong, Y., Zhu, G.: Data security and privacy in cloud computing. Int. J. Distribut. SensorNet. https://doi.org/10.1155/2014/190903 (2014)
2. Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**(5), 662–679 (2012)
3. Tatwani, L.N., Tyagi, R.K.: Security and Privacy Issues in Cloud Computing. Int. Res. J. Comput. Electron. Eng. (IRJCEE, USA), ISSN (2015)
4. Sen, J.: Security and privacy issues in cloud computing. In: Architectures and Protocols for Secure Information Technology Infrastructures, pp. 1–45 (2013)
5. Rahul, S., Sharda, D.J.: Security & privacy issues in cloud computing. Int. J. Eng. Res. Technol. **2**(3) (March-2013)). ESRSA Publications (2013, March)
6. Jansen, W.A.: Cloud hooks: security and privacy issues in cloud computing. In: System Sciences (HICSS), 2011 44th Hawaii International Conference on, pp. 1–10. IEEE (2011, January)
7. Chen, D., Zhao, H.: Data security and privacy protection issues in cloud computing. In: Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on, vol. 1, pp. 647–651. IEEE (2012, March)
8. Xiao, Z., Xiao, Y.: Security and privacy in cloud computing, In: IEEE Communications Surveys & Tutorials, vol. 15, no. 2, pp. 843–859, Second Quarter (2013)
9. James, B.: Security and Privacy Challenges in Cloud Computing Environments (2010)
10. Izang, A., Adebayo, O., Okoro, A., Taiwo, O., Oluwabunmi. Security and Ethical Issues to Cloud Database (2017)

Chapter 14
Evaluating Faster-RCNN and YOLOv3 for Target Detection in Multi-sensor Data

Anwaar Ulhaq, Asim Khan, and Randall Robinson

Abstract Intelligent and autonomous systems like driverless cars are seeking the capability to navigate around at any time of the day and night. Therefore, it is vital to have the capability to reliably detect objects to predict any situation. One way to capture such imagery is through multi-sensor data like FLIR (Forward Looking Infrared) and visible cameras. Contemporary deep object detectors like YOLOv3 (You Look Once Only) (Redmon and Farhadi, Yolov3: an incremental improvement, arXiv) and Faster-RCNN (Faster Region based Convolutional Neural Networks) (Girshick Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1440–1448, 2015) are well-trained for daytime images. However, no performance evaluation is available against multi-sensor data. In this paper, we argue that diverse contextual multi-sensor data and transform learning can optimise the performance of deep object detectors to detect objects around the clock. We explore how contextual multi-sensor data can play a pivotal role in modelling and recognizing objects especially at night. For this purpose, we have proposed the use of contextual data fusion on available training data before training these deep detectors.We show that such enhancement significantly increases the performance of deep learning based object detectors.

Keywords Object detection · YOLOv3 · Faster-RCNN · Deep learning

A. Ulhaq
School of Computing and Mathematics, Charles Sturt University, Port Macquarie, NSW, Australia

Centre of Applied Informatics, Victoria University, Melbourne, VIC, Australia

A. Khan (✉) · R. Robinson
College of Engineering and Science, Victoria University, Melbourne, VIC, Australia
e-mail: asim.khan@live.vu.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_14

14.1 Introduction

Automated object detection is aimed at finding any instances of objects of interest from the known object categories (such as cars, humans etc.) in a captured image or video and, additionally it seeks object's spatial location and the extent (e.g., a bounding box). Object detection also plays a fundamental role to solve other high level vision problems like activity detection. Therefore, it has a wide range of applications. It has been an active area of research over the last few decades. Several survey papers [1–3] are written about the topic. However, current success of deep learning models has revolutionized the field and many state-of-the art object detection networks are known today. The development of this filed has moved from the detection of a single object category towards the more challenging task of building general purpose object detection systems whose object detection accuracy and timing rivals that of humans. The success of Deep CNNs in image classification led to success in object detection, resulting in development of pioneer work of Region based CNN (RCNN) detector of Girshick et al. [4]. This development started a new era of sophisticated deep learning approaches for object detection. The availability of GPU computing resources and abundance of large scale datasets and challenges such as ImageNet [5] and MS COCO [6] caused a so-called revolution in object detection.

Detecting objects in different visual conditions is still a challenging computer vision problem. These variations are generally caused by changes in imaging conditions, leading changes in object appearance in images (like illumination, pose, scale, occlusion, clutter, shading, blur and motion). Particularly, an image can be captured sin different light conditions: sunlight (dawn, day, dusk), moonlight, or darkness. It is also effected by other factors like weather conditions, cameras, backgrounds, illuminations, occlusion conditions,locations, and viewing distances. It can further be amplified or degraded by digitization artefacts, noise corruption, poor resolution, and other distortions (Fig. 14.1).

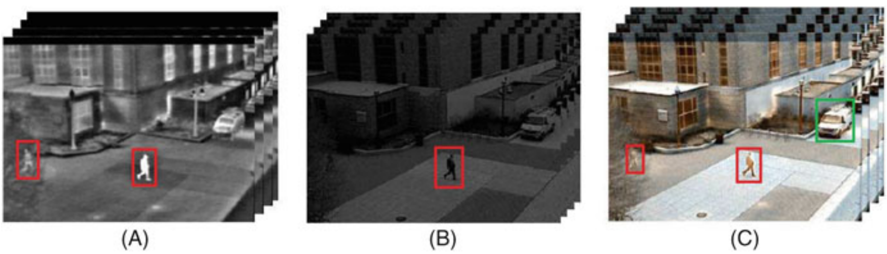


Fig. 14.1 Object detection in different image spectrum: (**a**) an IR video sequence with two detections but car is missing (**b**) low light visible domain video sequence with only one detection, two detections missing, (**c**) three successful detections in a colour fused video generated from (**a**) and (**b**)

Object context is a valuable priori knowledge in terms of modelling their instances. Therefore, the main advantage available to object detection in daylight based imagery is the ease in finding visual interest points and descriptors because of their high quality visual context. However, achieving the same objectives in night-time imagery is cumbersome due to clutter and low lighting conditions resulting in low accuracy of action recognition approaches. Most recently, some methods are developed that can work on dark images. One of the salient work is [7] that provides excellent performance on raw sensed data and its enhancement. However, it does not deal with already acquired images in which raw sensor data is not available. In this paper we try to address this issue by integrating contextual data fusion and deep object detector models.

Moreover, night-time imagery lacks colour information that provides great help to our visual perception. Due to unnatural appearance and IR imagery limitations, multi-sensor systems are employed for better contextual awareness [8–10]. Another approach to optimize these systems is to introduce pseudo-colour information. A recent study about perceptual evaluation [11] of such colour transformed multi spectral systems has concluded that pseudo-colourization better illustrates the gist of a night scene by improving the fixation behaviour of human eye compared to large-scale imagery.

We address the following research questions: (i) Can accuracy of object detection be increased by context enhancement? And (ii) Does psedu-colourization of night-time imagery contributes to the increase of object recognition and detection?

This paper claims the following contributions: (1) It considers a diverse scenario of object recognition in clutter and adverse lighting conditions at night-time and answers how colourization and contexts can be utilized to enhance automated object recognition at night, (ii) it proposes to integrate motion, colour and context information in a single object detection framework by training deep detectors on enhanced imagery. In terms of novelty, it may be the first paper that compares the deep learning based object detectors for night vision imaging data and studies the importance of contextual information.

The paper is organized as follows: Sect. 14.2 illustrates the contextual data fusion of multi-sensor videos. Deep object detectors are presented in Sect. 14.4. Experimental results are discussed in Sect. 14.5. The conclusion and references are provided at the end.

14.2 The Contextual Data Fusion in Night-Time Videos

In this section, we discuss the motivation behind contextual data fusion for night-time video sequences, video fusion and colourization for context enhancement.

Fig. 14.2 An example scenario of video fusion of registered video streams: (**a**) an IR video sequence (**b**) a low-light visible video stream and (**c**) a fused video generated from (**a**) and (**b**)

14.2.1 The Motivation

The aim of context data fusion is a pre-processing step to give day-like appearance to night-time videos. It involves video fusion applied on registered video streams collected from infra-red (IR) and visible (VIS) spectrum Context data fusion helps to reveal a camouflaged target and to assist target localization. Here we present and discuss context data fusion briefly.

14.2.2 Contextual Data Fusion

The objective of employing video fusion is to generate a single enhanced video from complementary videos, that is more suitable for the purpose of human visual perception, action and context recognition. If we denote A as IR video sequence and B as visible video sequence, we intend to generate another video sequence C by fusing visual information from A and B. Figures 14.2 and 14.3 give illustrations of video fusion results.

For simultaneous fusion and colourization, we used automatic colour transfer based video fusion (FACE) [12] which enhances video context by colour transfer from a source image. The illustration of this approach is given in Fig. 14.3.

14.3 Deep Object Detection

Object recognition involves a range of related computer vision tasks like predicting the type or class of an object in an image (classification), indicating their location with a bounding box (localization) or highlighting the specific pixels of the object instead of a coarse bounding box (semantic segmentation or instance segmentation).

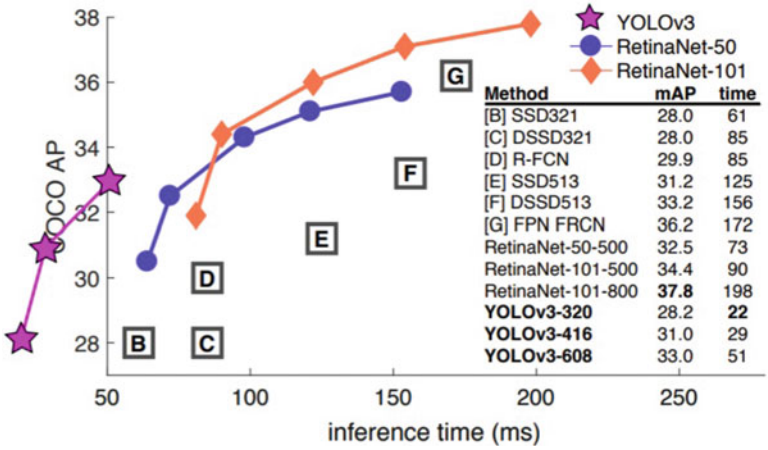


Fig. 14.3 Figure adapted from [18] It shows that the object instances detection by YOLOv3 is much faster than other methods using the same GPU

Here we are presenting the recent top-performing deep learning models for object detection in ILSVRC (an annual academic competition with a separate challenge for object detection). It includes Faster-RCNN [13] and YOYLV3 [14].

Faster Region Based Convolutional Neural Networks The R-CNN was developed by Ross Girshick, et al [15]. The proposed R-CNN model was comprised of three modules called Region Proposal (to generate and extract category independent region proposals, e.g. candidate bounding boxes.), Feature Extractor (to extract features from each candidate region using a deep CNN) followed by a Classifier (to classify features as one of the known class). The feature extractor was based on t AlexNet [16].Region proposals are bounding boxes, designed to accelerate the object class prediction. An extension to address the speed issues of R-CNN was proposed by [4]. It was based on a single model instead of a pipeline to learn and output regions and classifications directly. A further extension of this work is [13] that is improved for both speed of training and detection. The main contribution was the introduction of a Region Proposal Network. Convolutional neural network for proposing regions and the type of object to consider in the region. This small network uses anchor boxes for binary class prediction, indicating the presence or absence of an object, so-called “objectness” of the proposed region. It uses the output of a pre-trained deep CNN, such as VGG-16, and passes a small network over the feature map and outputs multiple region proposals and a class prediction for each object.

You Only Look Once YOLO or “You Only Look Once,” was developed by Joseph Redmon, et al. [17]. It is based on a single neural network trained end to end that takes a image as input and predicts bounding boxes and class labels for each bounding box directly.

It divides the input image into a grid of cells (each cell is responsible for predicting a bounding box if the centre of a bounding box is found within it). Each grid cell predicts a bounding box with the x, y coordinate and the width and height, confidence and the class prediction on each cell. An extension named YOLOv2 employs the anchor boxes like Fater-RCNN (pre-defined bounding boxes). The choice of these bounding boxes is pre-processed using a k-means analysis on the training images. Finally the confidence prediction provides the IOU (intersection of union) between the predicted box and any ground truth box. Each grid cell additionally predicts C conditional class probabilities conditioned on the grid cell containing an object. It was followed by a deeper feature detector network and representational changes into another version called YOLOV3 [14]. It is considered state-of-the art object detector in-terms of speed and accuracy (Please see the analysis study adapted from [14]).

14.4 Experimental Results and Discussion

This section describes our experimental data, set-up, results and performance comparison with discussion.

14.4.1 Multi-sensor Datasets and Experimental Set-Up

Night Vision Dataset (NVD) To develop our benchmark night-vision object detection dataset, we have recorded Night Vision Dataset (NV) using two different cameras. One of them is IR camera, Raytheon Thermal IR-2000B and the other is low-light visual camera, Panasonic WV-CP470. The thermal and visual videos are registered before fusion process by selecting corresponding points in corresponding views and following a computation of a least-squared error fitting homography. In addition to these videos, this dataset includes 20 video sequences collected from TNO image fusion dataset [11], Eden project dataset [19] and Ohio-state University thermal dataset. This dataset comprises four main objects including human, camp, fence and buildings.

Multi-spectral Object Detection dataset (MOD) [20] The multispectral images dataset includes an RGB image, NIR image, MIR image, and FIR image where each spectral image is created by using light of different wavelengths. This dataset contains 7, 512 images. It also includes ground truth with bounding box coordinates and labels. It has five classes of obstacles commonly encountered during driving (bike, car, carstop, colorcone, person) are annotated in this dataset.

Firstly, the training of both detectors is performed for each dataset separately. During the testing phase, test video is presented to the model to find the bounding boxes.

We experimented with different versions of our detection framework to know the impact of each visual cue (colour and context) on recognition performance. At first, a baseline is developed using single channels like RGB and IR excluding context and colour information. At second, we fuse contextual information and get final fused data for training for our detectors.

We calculated average precision (AP) for each object detection against NV dataset, then calculated mean, mAP for all categories and displayed it in Table 14.1. We repeated the same evaluation methodology for MOD dataset and displayed it in Table 14.2. these experiments show that we get better mAP for both datasets in case of fused multi-sensor data. Here, we only show our results for YOLOv3 detector. It validates our initial proposal that multi-sensor contextual information can increase object detection due to more complete nature of the fused information.

In addition to mAP comparisons, we have displayed visualisations of object detections for both datasets and detectors in Figs. 14.4 and 14.5. It shows that both detectors improve their detection performance based on the fused representation.

Table 14.1 Mean avg.precision for detections on four categories of VVD dataset

Data	mAP	Human	Camp	Fence	Building
RGB	0.59	0.72	0.69	0.41	0.56
NIR	0.73	0.85	0.75	0.69	0.65
ColorContextual Fused	0.80	0.92	0.85	0.72	0.72

Table 14.2 Mean avg.precision for detections on five categories of MOD dataset

Data	mAP	Person	Car	Colorcone	Carstop	Bike
RGB	0.70	0.82	0.79	0.61	0.68	0.58
NIR	0.65	0.75	0.85	0.40	0.55	0.72
ColorContextual Fused	0.84	0.94	0.91	0.71	0.80	0.88

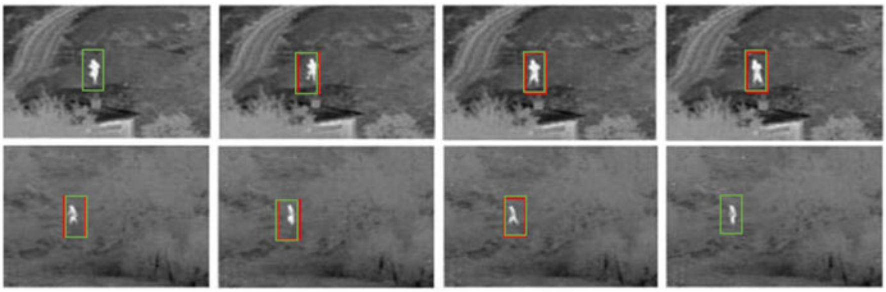


Fig. 14.4 Human instance detection in three NV object instances where red bounding box is the actual ground truth while green bounding box shows detection by YOLOv3

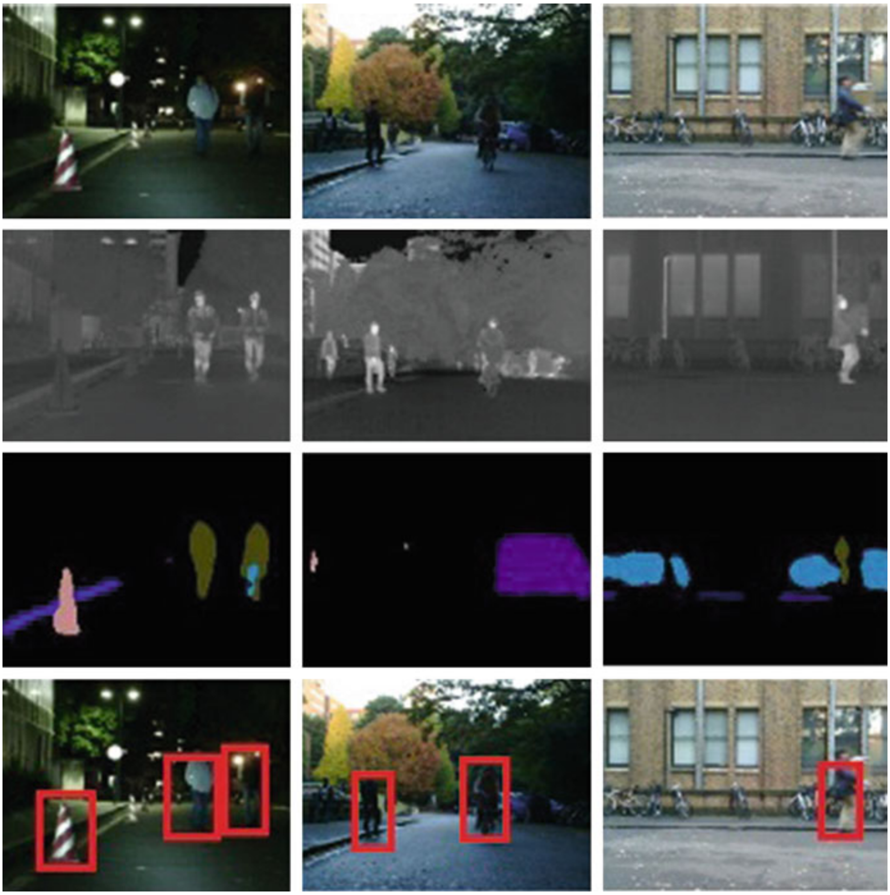


Fig. 14.5 Objects instance detection in MOD dataset: (Top) two rows from RGB and thermal images, (bottom) the object instances where segmented object is by Faster RCNN and bounding box is shows detection by YOLOv3

14.5 Conclusion

In this paper, we explored and discussed the role of diverse data like colour, contextual information for automated object recognition. We used image and video fusion of contextual data with colour transfer approach and used it to train Faster RCNN and YOLOV3 object detectors. Results show that detector performance is considerably increased by using contextual data fusion. In comparison, YOLOv3 outperforms Faster-RCNN in terms of accuracy and speed. Furthermore, we discovered that colourization clue is an important data to increase object recognition performance alongside contextual data.

References

1. Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., Li, J.: Salient object detection: a survey. Comput. Vis. Media **5**, 1–34 (2014)
2. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process. Mag. **35**(1), 84–100 (2018)
3. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: a survey. arXiv preprint arXiv:1809.02165
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
6. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
7. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3291–3300 (2018)
8. Anwaar, H., Iqbal, G., Murshed, M.: Automated multi-sensor color video fusion for nighttime video surveillance. In: IEEE Symposium on Computers and Communications (ISCC), pp. 529–534 (2010)
9. Anwaar, H., Iqbal, G., Murshed, M.: On dynamic scene geometry for view-invariant action matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3305–3312 (2011)
10. Ulhaq, A.: Action recognition in the dark via deep representation learning. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), pp. 131–136. IEEE (2018)
11. Toet, A., de Jong, M.J., Hogervorst, M.A., Hooge, I.T.: Perceptual evaluation of colorized nighttime imagery. In: IS&T/SPIE Electronic Imaging, pp. 901412 (2014)
12. Ulhaq, A., Yin, X., He, J., Zhang, Y.: Face: fully automated context enhancement for night-time video sequences. J. Vis. Commun. Image Represent. **40**, 682–693 (2016)
13. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
14. Redmon, J., Farhadi, A.: Yolov3: An Incremental Improvement (2018). *arXiv preprint arXiv:*1804.02767
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
16. Yuan, Z.-W., Zhang, J.: Feature extraction and image retrieval based on alexnet. In: Eighth International Conference on Digital Image Processing (ICDIP 2016). International Society for Optics and Photonics, vol. 10033, pp. 100330E (2016)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
18. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
19. Lewis, J., Nikolov, S., Loza, A., Canga, E.F., Cvejic, N., Li, J., Cardinali, A., Canagarajah, C., Bull, D., Riley, T., et al.: The Eden project multi-sensor data set. The Online Resource for Research in Image Fusion (ImageFusion.org)
20. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5108–5115. IEEE (2017)

Chapter 15
Wavelet-Based Quantile Density Function Estimation Under Random Censorship

Esmaeil Shirazi and Hassan Doosti

Abstract In this paper, the estimation of a quantile density function in the presence of right censored data is investigated. A new wavelet-based methodology for the estimation of the quantile function will is proposed. In particular, an adaptive hard thresholding wavelet estimator is constructed. Under mild assumptions on the model, we prove that it enjoys powerful mean integrated squared error properties over Besov balls. While existing estimators of the quantile density function are not good at the tails, our proposed estimators perform well at the tails. The comparison of the proposed estimator has been made with estimators given by Jones (1992) Ann Inst Stat Math 44(4):721–727 and Soni et al. (2012) Comput Stat Data Anal 56(12):3876–3886 graphically and in terms of the mean integrated square error (MISE) for the uncensored case.

Keywords Adaptivity · Quantile density function · L^2 risk function · Wavelets · Hard thresholding

15.1 Introduction

The use of quantile function estimation has been around for decades in exploratory data analysis, statistical analysis, reliability and medical studies. (See, for example, [9, 10, 13, 14, 17, 19]). For a unified study of this concept, one can refer to [4, 6, 10, 11]. The concept of quantiles has been used by [5, 12, 15] for modelling competing risk models.

E. Shirazi
Gonbad Kavous University, Khonbad Kavous, Iran

H. Doosti (✉)
Macquarie University, Sydney, NSW, Australia
e-mail: hassan.doosti@mq.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_15

In classical statistics, most of the distributions are defined in terms of their cumulative distribution function (CDF) or probability density function (PDF). There are some distributions which do not have the cdf/pdf in an explicit form but a closed form of the quantile function is available, for example Generalised Lambda distribution (GLD) and Skew logistic distribution [4]. Karian and Dudewicz [7] showed the significance of different Lambda distributions for modeling failure time data. Quantile measures are less influenced by extreme observations. Hence the quantile function can also be looked upon as an alternative to the distribution function in lifetime data for heavy tailed distributions. Sometimes for those distributions whose reliability measures do not have a closed or explicit form, the reliability characteristics can be represented through quantile function.

Formally, let $X_1, X_2, \ldots, X_n$ be independent identically distributed (i.i.d.) survival times with a common distribution function F and a density function f. Also let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. censoring times with a common distribution function G. It is assumed that X_m is independent of Y_m for every m. Rather than observing $X_1, X_2, \ldots, X_n$, the variables of interest, in the randomly right-censored models, one observes $Z_m = min(X_m, Y_m)$ and $\delta_m = I(X_m \leq Y_m)$, $m = 1, 2, \ldots, n$, where $I(A)$ denotes the indicator function of the set A. The quantile function of X is defined as

$$Q(x) = F^{-1}(x) = \inf\{y \in R;\ F(y) > x\} \tag{15.1}$$

It satisfies F (Q(x)) = x. In the same way that the CDF can be differentiated to give the PDF, [6, 11] defined the derivative of $Q(x)$ as the quantile density function. That is, $g(x) = Q'(x)$. Differentiating (15.1), we get

$$g(x) = \frac{1}{f(Q(x))}, \quad x \in [0, 1] \tag{15.2}$$

Note that the sum of two quantile density functions is again a quantile density function. This idea is useful in modeling data. Nair and Sankaran [10] defined the hazard quantile function

$$R(x) = r(Q(x)) = \frac{f(Q(x))}{1 - F(Q(x))} = \frac{1}{(1-x)g(x)}, \quad x \in (0, 1).$$

Hence g(x) appears in the expression for hazard quantile function and it would be useful to study nonparametric estimators of this unknown quantile density function.

The estimation of a quantile density function for right censored data, have been studied elsewhere in some literatures. For example, the kernel method were used by [18, 20] for studying non-parametric estimators of quantile density estimation. They proposed some smooth estimators and investigated their asymptotic properties. Recently, [1, 16] discussed the nonparametric wavelet estimators of the quantile density function and proposed linear and nonlinear wavelet based estimators for the uncensored case.

In this article we take a new and novel approach to quantile density function estimation based on wavelet methods for right censored data. We show that these estimators attain optimal and nearly optimal rates of convergence over a wide range of Besov function classes, and in particular enjoy those rates without the extraneous logarithmic penalties that given in [1].

The paper is structured as follows. Section 15.2 presents additional assumptions on the model. We describe in Sect. 15.3 our wavelet-based framework and strategy. Results are given in Sect. 15.4.

15.2 Wavelet Estimators

We start this section by introducing the concept of Multiresolution Analysis (MRA) on R as described in [8]. Let ϕ be a scale function and ψ its associated wavelet basis of $L^2[0,1]$, and define $\phi_{i_0 j}(x) = 2^{i_0/2}\phi(2^{i_0}x - j)$ and $\psi_{ij}(x) = 2^{i/2}\psi(2^i x - j)$. We assume that the father and mother wavelets, $\phi(x)$ and $\psi(x)$, are bounded and compactly supported over [0, 1], that $\int \phi = 1$ and that the wavelets are *r-regular*. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives. Given a square-integrable g, put $\alpha_{i_0 j} = \int g\phi_{i_0 j}$ and $\beta_{ij} = \int g\psi_{ij}$. An empirical wavelet expansion for all $g \in L^2(R)$ is given by

$$g(x) = \sum_{j\in Z} \alpha_{i_0 j}\phi_{i_0 j}(x) + \sum_{i\geq i_0}\sum_{j\in Z} \beta_{ij}\psi_{ij}(x),$$

As is done in the wavelet literature, we investigate wavelet-based estimators asymptotic convergence rates over a large range of Besov function classes B^s_{pq}, $s > 0$, $1 \leq p, q \leq \infty$. The parameter s measures the number of derivatives, where the existence of derivatives is required in an L^p-sense, whereas the parameter q provides a further finer gradation.

The Besov spaces include, in particular, the well-known Sobolev and Hölder spaces of smooth functions H^m and C^s and (B^m_{22} and $B^s_{\infty,\infty}$ respectively), but in addition less traditional spaces, like the spaces of functions of bounded variation, sandwiched between $B^1_{1,1}$ and $B^1_{1,\infty}$. The latter functions are of statistical interest because they allow for better models of spatial of inhomogeneity (e.g. [3, 8]).

For a given *r-regular* mother wavelet ψ with $r > s$, define the sequence norm of the wavelet coefficients of a regression function $g \in B^s_{pq}$ by

$$|g|_{B^s_{pq}} = (\sum_j |\alpha_{i_0 j}|^p)^{1/p} + \{\sum_{i=i_0}^{\infty}[2^{i\sigma}(\sum_j |\beta_{ij}|^p)^{1/p}]^q\}^{1/q} \tag{15.3}$$

Where $\sigma = s + 1/2 - 1/p$. Meyer [8] shows that the Besov function norm $\|g\|_{B^s_{pq}}$ is equivalent to the sequence norm $|g|_{B^s_{pq}}$ of the wavelet coefficients of g. Therefore we will use the sequence norm to calculate the Besov norm $\|g\|_{B^s_{pq}}$ in the sequel.

We also consider a subset of Besov space B^s_{pq} such that $sp > 1$, $p, q \in [1, \infty]$. The spaces of regression functions that we consider in this paper are defined by

$$F^s_{p,q}(M) = \{g : g \in B^s_{pq}, \|g\|_{B^s_{pq}} \leq M, supp\ g \subseteq [0, 1]\},$$

i.e., $F^s_{p,q}(M)$ is a subset of functions with fixed compact support and bounded in the norm of one of the Besov spaces B^s_{pq}. Moreover, $sp > 1$ implies that $F^s_{p,q}(M)$ is a subset of the space of bounded continuous functions.

In our random censorship model, we observe $Z_m = min(X_m, Y_m)$ and $\delta_m = I(X_m \leq Y_m)$, $m = 1, 2, \ldots, n,$. When data is of the form (Z_i, δ_i), $i = 1, \ldots, n$, The wavelet expansion of quantile density function $g(x)$ (see [1]) is given by

$$g(x) = \sum_{j \in Z} \alpha_{i_0 j} \phi_{i_0 j}(x) + \sum_{i \geq i_0} \sum_{j \in Z} \beta_{ij} \psi_{ij}(x), \tag{15.4}$$

The wavelet coefficients $\alpha_{i_0 j}$ and β_{ij} are unknown and need to be estimated. Our approach is based on the following remark: by the change of variable $x = F(y)$ with $y \in X_1(\Omega) = [0, 1]$, we can rewrite $\alpha_{i_0 j}$ as

$$\alpha_{i_0 j} = \int_{[0,1]} g(x)\phi_{i_0 j}(x)dx = \int_{[0,1]} \frac{1}{f(F^{-1}(x))}\phi_{i_0 j}(x)dx = \int_{[0,1]} \phi_{i_0 j}(F(x))dx \tag{15.5}$$

Similarly,

$$\beta_{ij} = \int_{[0,1]} \psi_{i_0 j}(F(x))dx$$

Since F is unknown, we estimate it by the empirical estimator:

$$\hat{F}_n(x) = 1 - \Pi_{i=1}^n \big[1 - \frac{\delta_{(i)}}{n - i + 1}\big]^{I(Z_{(i)} \leq x)}$$

where, $Z_{(m)}$ is the *m-th* ordered Z-value and $\delta_{(m)}$ is the concomitant of the *m-th* order Z statistic, i.e., $\delta_{(m)} = \delta_j$ if $Z_{(m)} = Z_j$. This leads the following integral estimator for $\alpha_{i_0 j}$ and β_{ij}

$$\hat{\alpha}_{i_0 j} = \int_{[0,1]} \phi_{i_0 j}(\hat{F}(x))dx, \qquad \hat{\beta}_{ij} = \int_{[0,1]} \psi_{i_0 j}(\hat{F}(x))dx \tag{15.6}$$

Accordingly, the proposed linear wavelet estimator of $g(x)$ is

$$\hat{g}_L(x) = \sum_{j \in Z} \hat{\alpha}_{i_0 j} \phi_{i_0 j}(x) + \sum_{i \geq i_0}^{i_1} \sum_{j \in Z} \hat{\beta}_{ij} \psi_{ij}(x), \tag{15.7}$$

Clearly, $\hat{\alpha}_{i_0 j}$ and $\hat{\beta}_{ij}$ are not unbiased estimators for $\alpha_{i_0 j}$ and β_{ij}. However, using the dominated convergence theorem, one can prove that they are asymptotically unbiased. The proposed non-linear wavelet estimator of $g(x)$ is

$$\hat{g}_H(x) = \sum_{j \in Z} \hat{\alpha}_{i_0 j} \phi_{i_0 j}(x) + \sum_{i \geq i_0}^{i_1} \sum_{j \in Z} \hat{\xi}_{ij} \psi_{ij}(x), \tag{15.8}$$

Here, the hard thresholded empirical wavelet coefficients are $\hat{\xi}_{ij} = \hat{\beta}_{ij} I\left(|\hat{\beta}_{ij}| > \lambda\right)$, where $\lambda > 0$ is a threshold parameter and the empirical wavelet coefficients $\hat{\alpha}_{i_0 j}$ and $\hat{\beta}_{ij}$ are defined from equations (15.6). Also, the smoothing parameters i_0 and i_1 satisfying $2^{i_0} \simeq n^{1/1+2s}$ and $2^{j_1} \simeq n(\log n)^{-2}$.

15.3 Asymptotic Results

In following theorems, we consider the rate of convergence of wavelet estimators $\hat{g}_L(x)$ and $\hat{g}_H(x)$ under L^2 risk function. They attain optimal and nearly optimal rates of convergence over a wide range of Besov space classes. Moreover,C denotes any constant that does not depend on i, j and n. It's value may change from one term to another and may depends on ϕ or ψ.

Theorem 1 *Assume $g \in F^s_{\nu,\pi}(M)$ with $s > 1/r$, $r \geq 1$ and $\pi \geq 1$. Let $\hat{g}_L(x)$ be as in (15.7) with i_0 being the integer such that $2^{i_0} \simeq n^{1/1+2s}$. Then there exists a constant $C > 0$ such that*

$$E\left(\int_{[0,1]} \left(\hat{g}_L(x) - g(x)\right)^2\right) \leq C n^{-\frac{2s}{2s+1}}$$

In following, Theorem 2 explores the rates of convergence of $\hat{g}_H(x)$ under the L^p risk over Besov balls.

Theorem 2 *Let $\hat{g}_H(x)$ be as in (15.8) with i_s being the integer satisfying*

$$2^{i_s} \simeq \left(\frac{n}{\log_2 n}\right)^{1/1+2s}$$

and λ_n being the threshold:

$$\lambda_n = \kappa \sqrt{\frac{\log n}{n}}$$

Assume that $g \in F^s_{\nu,\pi}(M)$ with $s > 1/\nu$, $\nu \geq 1$ and $\pi \geq 1$. Then there exists a constant $C > 0$ such that

$$E\left(\int_{[0,1]}\left(\hat{g}_H(x)-g(x)\right)^2\right)\leq C\left(\frac{\log_2 n}{n}\right)^{-\frac{2s}{2s+1}}$$

If we do a global comparison between the results of Theorems 1 and 2, the rates of convergence achieved by $\hat{g}_H(x)$ are better than the one achieved by $\hat{g}_L(x)$. Moreover, let us recall that $\hat{g}_H(x)$ is adaptive while $\hat{g}_L(x)$ is not adaptive due to its dependence on s in its construction.

15.3.1 Auxiliary Results

In the following section we provide some asymptotic results that are of importance in proving the theorem. The proof of Theorem 1 is a consequence of Propositions 1 and 2 of [1] and we describe them below. They show that the estimators $\hat{\beta}_{jk}$ defined by (15.6) satisfy a standard moment inequality and a specific concentration inequality. Before presenting these inequalities, the following lemma determines an upper bound for $|\hat{\beta}_{ij}-\beta_{ij}|$.

Lemma 1 *Suppose that the assumptions of Theorem 1 are satisfied. Then, for any $i\in\{i_0+1,\ldots,R\}$ and any $j\in\{0,\ldots,2^i-1\}$, the estimator $\hat{\beta}_{ij}$ defined by (15.6) satisfies*

$$|\hat{\beta}_{ij}-\beta_{ij}|\leq K2^{3i/2}\int_{[0,1]}|\hat{F}(x)-F(x)|dx\leq K2^{3i/2}\sup_{[0,1]}|\hat{F}(x)-F(x)|. \tag{15.9}$$

with $K=\sup_{[0,1]}|\psi'(x)|$ and $\psi'_{ij}(x)=2^{3i/2}\psi(2^ix-j)$.

Proposition 1 *Let $p\geq 2$. Suppose that the assumptions of Theorem 1 are satisfied, then there exists a constant $C>0$ such that, for any $j\geq j_0$, and n large enough, the estimator $\hat{\beta}_{jk}$, defined by (15.6) satisfies the following:*

$$E\left(|\hat{\beta}_{jk}-\beta_{jk}|^{2p}\right)\leq Cn^{-p} \tag{15.10}$$

The expression in (15.10) holds for $\hat{\alpha}_{jk}$ as well, replacing $\hat{\beta}_{jk}$ by $\hat{\alpha}_{jk}$ and β_{jk} by α_{jk}.

Proposition 2 *Let $p\geq 2$. Under the assumptions of Theorem 1, there exists a constant $c>0$ such that, for any $j\geq j_0$, and large enough n, the estimators $\hat{\xi}_{jk}$ defined by (15.6) satisfy the following concentration inequality:*

$$\mathbb{P}\left(\left(\sum_{(i)}|\hat{\beta}_{jk}-\beta_{jk}|^p\right)^{1/p}\geq cn^{-1/2}(\log(n))^{1/2}\right)\leq Cn^{-p}, \tag{15.11}$$

for some constant $C>0$.

15.4 A Simulation Study

In this section, we consider two distributions $Exp(1)$ and Generalised Lambda $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The corresponding quantile functions are $Q_E(u) = -\log(1-u)$ and $Q_{GL}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}$, where λ_1 and λ_2 are location and inverse scale parameters, respectively, and λ_3 and λ_4 jointly determine the shape (with λ_3 mostly affecting the left tail and λ_4 mostly affecting the right tail). Though the GLD is defined on the entire real line, we consider those choices of parameters that give support as $(0, \infty)$. For $Exp(1)$, the quantile density function is $q_E(u) = \frac{1}{1-u}, 0 < u < 1$ and for $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, the quantile density function is $q_{GL}(u) = \frac{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}}{\lambda_2}, 0 < u < 1$.

We consider performance of linear wavelet estimator $\hat{g}_L(x)$ and the hard thresholding wavelet estimator $\hat{g}_H(x)$ presented in Sect. 15.2 and compare them with a kernel estimator $\hat{g}_{KS}$ proposed by [18]. In the following examples, we simulate a random censored sample from $Exp(1)$ with a censoring distribution is assumed to be $Exp(0.25)$ and Generalised Lambda $GLD(1, -1, -1/8, -1/8)$ with a censoring distribution is assumed to be Uniform(0, 4.1), for different sample sizes and calculate the mean integrated square error (MISE) based on aforementioned competitors, where the MISE criterion of the estimator $\hat{g}$ is defined as

$$MISE = \frac{n-1}{N(n+1)^2} \sum_{b=1}^{N} \sum_{i=1}^{n} \left(\hat{g}_b(x_i) - g(x_i) \right)^2 .$$

with $\hat{f}_b(.)$ being defined as an estimator of $f(.)$ at the bth replication. The results in this simulation study are obtained using Daubechies's compactly supported wavelet Symmlet 4 (see [2], p. 198) and Coiflet 2 (see [2], p. 258), and primary resolution level $j_0 = 6$ based on $N = 100$ replications. The code was written in MATLAB environment using the WaveLab software. Lower values of ANorm are indicative of better performance. We also list the corresponding standard errors.

Figures 15.1 and 15.2 show the results of simulation for the Generalised Lambda Distributions and the exponential distribution, respectively. In each figure, we show the original quantile density function *pdf* with black line along with two versions of the new wavelet estimator of g, namely (i) the hard thresholding estimator ($\hat{g}_H$) with red line, the linear wavelet estimator ($\hat{g}_L$) with blue line, (iii) the smoothed version of the linear wavelet estimator ($\hat{g}_{SL}$) with green line and the yellow line is Soni's kernel estimator $\hat{g}_K$, respectively. We conclude that the smoothed linear wavelet estimators is closer to the unknown quantile density function as compared to the other tow estimators we have studied.

Tables 15.1 and 15.2 show the values of MISE and Standard Deviations for each one of the three different quantile estimators. From Tables 15.1 and 15.2 we conclude that

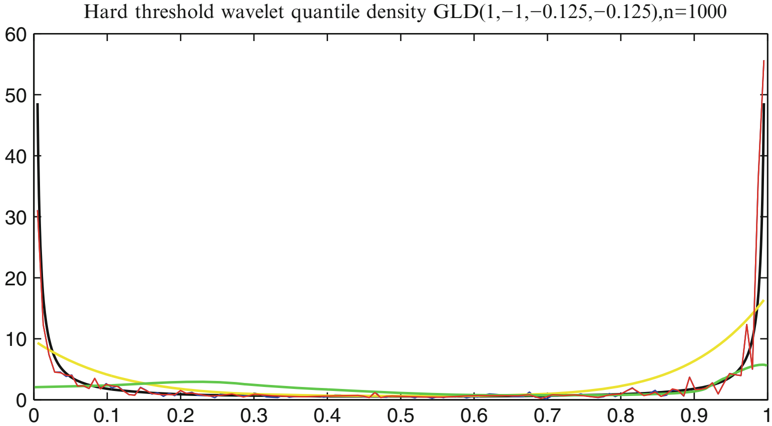


Fig. 15.1 Quantile Generalised Lambda density function estimation; the true density function (*black line*), the smoothed version of linear estimator (*yellow line*), the hard thresholding estimator (*red line*), Soni's kernel estimator (*green line*) and the linear wavelet estimator (*blue line*)

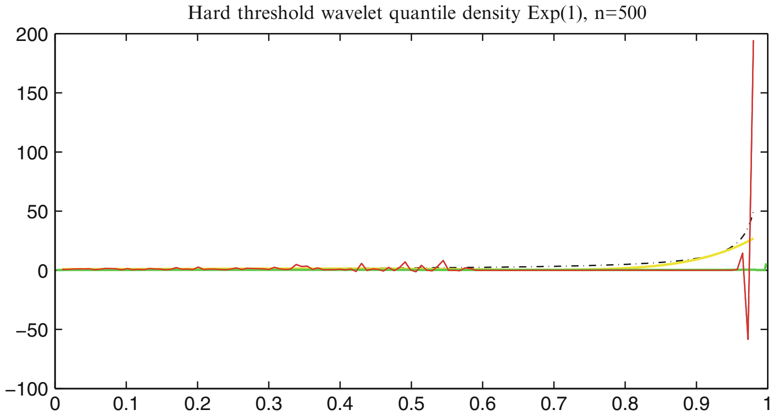


Fig. 15.2 Quantile Exponential density function estimation; the true density function (*black line*), the smoothed version of linear estimator (*yellow line*), the hard thresholding estimator (*red line*), Soni's kernel estimator (*green line*) and the linear wavelet estimator (*blue line*)

– In all cases, the MISEs for the smooth version of the wavelet estimator are smaller than those of the competitors.
– The smoothed linear wavelet estimator has the best performance compared to the other estimators.
– The hard threshold estimator is almost better than the kernel based estimator.

Table 15.1 Computed values of MISE and Standard Deviation in Quantile Generalised Lambda density function estimation; MISEs are located in the first row and Standard Deviations are in the second row

	MISE and Standard Deviation				
Estimation methods	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$
Hard threshold estimator	24.5237	19.7307	20.4130	14.4886	25.6089
	27.1769	17.2885	21.4917	16.8630	26.6924
Smoothed linear estimator	6.4516	5.1365	6.3780	6.1865	7.8016
	3.6516	7.9642	17.3534	17.1474	29.2123
Kernel estimator	15.4660	39.3248	99.5636	246.5960	602.7817
	0.8587	1.3256	1.4007	1.8740	2.3804

Table 15.2 Computed values of MISE and Standard Deviation in Quantile Exponential density function estimation; MISEs are located in the first row and Standard Deviations are in the second row

	MISE and Standard Deviation				
Estimation methods	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$
Hard threshold estimator	75.2404	62.1327	59.4890	38.1547	26.9764
	12.5016	10.3790	8.0070	11.6341	10.2135
Smoothed linear estimator	34.8566	38.4708	32.0460	13.0490	12.8654
	11.7462	10.8183	18.0920	27.4975	10.6689
Kernel estimator	199.0126	406.7701	826.2722	1663.9	3344.23
	0.9596	0.4948	0.9697	1.0860	0.8665

15.5 Conclusion

In this paper, the estimation of a quantile density function in the presence of right censored data is investigated. Two wavelet based estimators are proposed and the corresponding properties of the estimators are investigated analytically and through numerical studies. Asymptotic results prove that the estimators enjoy powerful mean integrated squared error properties over Besov balls. In addition, a simulation study illustrates that new proposed estimators perform better than existing kernel based estimators proposed at the tails.

References

1. Chesneau, C., Dewan, I., Doosti, H.: Nonparametric estimation of a quantile density function by wavelet methods. Comput. Stat. Data Anal. **94**, 161–174 (2016)
2. Daubechies, I.: Ten lectures on wavelets. In: CBMS-NSF Regional Conferences Series in Applied Mathematics. SIAM, Philadelphia (1992)
3. Donoho, D.L., Jonestone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. **90**, 1200–1224 (1995)

4. Gilchrist, W.: Statistical Modeling with Quantile Functions. Chapman and Hall, New York (2000)
5. Jeong, J.H., Fine, J.P.: Parametric regression on cumulative incidence function. Biostatistics **8**, 184–196 (2009)
6. Jones, M.C.: Estimating densities, quantiles, quantile densities and density quantiles. Ann. Inst. Stat. Math. **44**(4), 721–727 (1992)
7. Karian, Z.A., Dudewicz, E.J.: Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods. CRC, London (2000)
8. Meyer, Y.: Wavelets and Operators. Cambridge University Press, Cambridge (1992)
9. Nair, N.U., Sankaran, P.G., Kumar, B.V.: Total time on test transforms of order n and its implications in reliability analysis. J. Appl. Probab. **45**, 1126–1139 (2008)
10. Nair, N.U., Sankaran, P.G.: Quantile based reliability analysis. Commun. Stat. Theory Methods **38**, 222–232 (2009)
11. Parzen, E.: Non parametric statistical data modeling. J. Am. Stat. Assoc. **74**, 105–122 (1979)
12. Peng, L., Fine, J.P.: Nonparametric quantile inference with competing risks data. Biometrika **94**, 735–744 (2007)
13. Reid, N.: Estimating the median survival time. Biometrika **68**, 601–608 (1981)
14. Sankaran, P.G., Nair, N.U.: Nonparametric estimation of hazard quantile function. J. Nonparametr. Stat. **21**, 757–767 (2009)
15. Sankaran, P.G., Nair, N.U., Sreedevi, E.P.: A quantile based test for comparing cumulative incidence functions of competing risks models. Stat. Probab. Lett. **80**, 886–891 (2010)
16. Shirazi, E., Doosti, H.: Nonparametric estimation of a quantile density function under L^2 risk via block thresholding method, Comm. Stat. - Simul. Comput. (2019) http://dx.doi.org/10.1080/03610918.2019.1656250
17. Slud, E.V., Byar, D.P., Green, S.B.: A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data. Biometrics **40**, 587–600 (1984)
18. Soni, P., Dewan, I., Jain, K.: Nonparametric estimation of quantile density function. Comput. Stat. Data Anal. **56**(12), 3876–3886 (2012)
19. Su, J.Q., Wei, L.J.: Nonparametric estimation for the difference or ratio of median failure times. Biometrics **49**, 603–607. 365 (1993)
20. Zhou, Y., Yip, P.S.F.: Nonparametric estimation of quantile density function for truncated and censored data. J. Nonparametric Stat. **12**, 17–39 (1999)

Part IV
Health Statistics and Social Policy

Chapter 16
Factors Associated with Coronary Heart Disease among Elderly People in Different Communities

Kanis Fatama Ferdushi, Anton Abdulbasah Kamil, Mohammad Nayeem Hasan, and Tanjila Islam

Abstract Coronary Heart Disease (CHD) is one of the major causes of morbidity and mortality in many developing countries, including Bangladesh. A stratified random sampling with proportional allocation technique was used to collect data from elderly people from urban, rural, and ethnic areas of the Sylhet region. A total of 230 (110 women and 120 men) people aged 60 years or above from the above mentioned areas were included in this study. A multiple logistic regression model was used to evaluate major risk factors associated with CHD for this sample group. The prevalence of CHD was higher for males than females, at 47.80% and 52.20% for females and males, respectively. Elderly people in urban areas were significantly (AOR = 4.03; 95% CI: 1.22–3.29) more likely to have CHD as compared to elderly persons living in rural areas. Elderly persons of ethnic origin were found to be less likely to suffer from CHD (AOR = 0.04, 95% CI: 0.01–0.17514.69) in comparison to urban elderly. The risk factors smoking (44.8%) and hypertension (51.3%) were also positively associated with CHD. Elderly persons who reported to exercise regularly were found to have 89% (AOR = 0.11, 95% CI: 0.03–0.50) less risk of suffering from CHD as compared to those who did not exercise regularly. The findings of this study further indicated that factors such as a high BMI (overweight/obese), high sugar intake, high soft drink consumption, diabetes, and mental stress have a significant influence on CHD.

Keywords Coronary heart disease · Elderly · Risk factor · Adjusted odds ratio

K. F. Ferdushi (✉) · M. N. Hasan · T. Islam
Department of Statistics, Shahjalal University of Science and Technology, Sylhet, Bangladesh

A. A. Kamil
Faculty of Economics, Administrative and Social Sciences, Istanbul Gelisim University, Istanbul, Turkey

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_16

16.1 Background

Coronary heart disease (CHD) is one of the most prevalent of chronic diseases of our modern era [1, 2]. Known as a ‘silent killer’ due to its lack of easily observable symptoms, CHD has become one of the leading causes of death globally [3]. Over 80% of all reported cases of CHD have been registered in low- or middle-income countries in 2005 [4]. Further, the prevalence of CHD is predicted to continue to increase in low- and middle-income countries due to lifestyle conversions associated with increasing urbanization, economic development, and globalization. Significant risk factors associated with CHD include age, family history, gender, hypertension, high cholesterol, diabetes, smoking, obesity, excessive alcohol intake, excessive stress, and physical inactivity [5]. Obesity, among the various risk factors associated with the onset of CHD [6], has nearly tripled since 1975. It is estimated that approximately 2 billion people worldwide are overweight or obese; of these, over 650 million are obese [7, 8]. Of the above factors significantly associated with CHD, obesity, intake of processed foods (associated with increased prevalence of high blood pressure), and physical in activity are known to increase in populations undergoing urbanization. It is well established that the risk of dying from CHD increases substantially with age [9]. As a chronic disease, CHD is the largest contributor to mortality in individuals >60 years [10], being responsible for 17% of all deaths [11]. The average percentage of CHD is 73.3% between age 60 and 79 [12].

Although Bangladesh has the highest rates of coronary artery disease among all Asian countries, it is an issue that remains severely understudied [13]. Bangladesh is a densely populated country with population of over 160 million [14]. In 1981, the percentage of the elderly population was 11.35%, and in 2011, 16.57%. While no official census has been carried out since 2011, population estimates by the United Nations (UN) predict this number to have further increased. In Bangladesh, approximately 99.6% of males and 97.9% of females are endangered to at least one of the risks of cardiovascular disease at a younger age [15]. By 2020, 85% of the global CH disease burden is expected to be borne by developing nations, and the increase in coronary artery disease (CAD) mortality in developing countries between 1990 and 2020 is projected to be 120% in women and 137% in men [16, 17]. In recent years, factors such as rapid urbanization, increased life expectancy, unhealthy dietary habits, and lifestyle changes have led to an increase in the rate of CHD in Bangladesh [18]. A recent report has highlighted CHD as a significant, urgent medical and public health concern in Bangladesh due to its alarmingly increased impact on national mortality rates. In the advent of the new millennium, factors such as poor dietary habits, excess saturated and trans-fat consumption, high salt intake, and low levels of physical activity have been identified as major contributors towards the increased onset of this chronic disease in Bangladesh [5]. Bangladesh has been experiencing a rough epidemiological transition due to the limited availability of resources and strategies for prevention of these chronic diseases in low-income areas [19]. Moreover, better healthcare accessibility for elderly people (aged over 60 years) is a significant issue of concern in the context of Bangladesh, as studies have revealed that lack of access to healthcare significantly increases the risk of death among elderly people [20] and speeds up the longevity of health to access health care service [21]. CHD has become a major burden of disease for people of this country. Day by day, the mortality rate of CHD increases; today, it bears responsibility for approximately 75% of all deaths reported in the past few years in Bangladesh [22]. showed that weighted pooled prevalence of overall CHD in the Bangladeshi population was higher in urban areas (8%) compared to rural areas (2%) but there is no study has been conducted based on ethnic people and sylhet region. In this study, we have estimated the effect of factors such as socioeconomic status, dietary habits, and traditionally CHD-associated risk factors (hypertension, diabetes, smoking etc.) on the prevalence rate of this disease in Bangladesh.

16.2 Methods

16.2.1 Study Design

A stratified random sampling with proportional allocation technique was used to collect survey locations, number of respondents and the data. In order to capture the socioeconomic and demographic conditions as well as the ethnicity distribution of the Sylhet region, the total study area was divided into three strata, namely an urban area, a rural area, and an ethnic minority area. From the above strata, a ward in an urban area, a union in a rural area, and an ethnic community were randomly selected and surveyed. Data from elderly people (age 60 years or above) at the household level was collected with the use of a structured questionnaire. Three randomly selected areas were considered from three strata. The recognized sample size determination formula for these strata was $n = z^2[p(1 - p)/d^2] * D_{eff}$; where p is the indicator percentage, Z is the value of normal variate with a 95% confidence interval, and D_{eff} is the design effect. Attained values were calculated on the basis of 50% indicator percentage (proportion of households having elderly person), 95% confidence interval, 0.10p relative precision, and highest response distribution with an assumed design effect of 2.00. Using this design, a total of 230 households were calculated as the substantial number sample for strata. A total 230 elderly people were successfully interviewed. The sylhet region was selected purposely so as to capture data from elderly persons from an ethnic community.

In stratified random sampling, the allocation of a given sample of size n to different strata is done in proportion to their sizes:

$$n_i = \frac{nN_i}{N}$$

Where N_i is the total number of units in the ith stratum. The sampling fraction is equal to the sample size divided by the population size. Using this formula, a total of

90, 100, and 40 elderly persons belonging to urban, rural, and ethnic communities, respectively, were interviewed.

16.2.2 Response Variable

A prior CHD diagnosis was selected as the dependent variable in this study. An elderly person was considered to suffer from heart disease if they had been diagnosed with CHD prior to data collection. The CHD variable was coded as 1 if the elderly suffered from heart disease, and as 0 if otherwise.

16.2.3 Predictor Variables

Predictor variables considered for this study included socioeconomic factors such as income, occupation, wealth index, and education, as well as variables such as age, gender, residence (strata), and marital status. Risk factors such as hypertension, diabetes, respiratory, obesity (measured as body mass index, BMI), mental stress and smoking were also considered as predictor variables. Lastly, factors related to dietary and fitness habits, such as food consumption, level of exercise, sugar intake, red meat consumption, use of cooking oil, use of salt, soft drink consumption, and fast food consumption, were also considered as possible predictor variables on the basis of previous studies.

16.2.4 Statistical Analyses

A complex multiple logistic regression model was used to identify risk factors associated with CHD in the elderly population of the Sylhet region. Variables were selected in two stages so as to assure the final model was correctly specified. In the first stage of variable selection, a chi-square test was carried out to assess the statistical significance of all variables considered as possible predictor variables. Based on the chi-square test, only 18 variables were statistically significant with respect to the CHD variable at a 95% level of significance. Variables with a p-value >0.05 were excluded from the model. A stepwise (forward) logistic regression was next carried out, further narrowing down the variables to 11. Lastly, a multiple logistic regression was carried out with the selected predictor variables. The final adjusted model included the independent variables: sex, residence, education, occupation, BMI, exercise, sugar intake, cooking oil, smoking habit, mental stress, and hypertension. The hosmer and lemeshow goodness-of-fit test was used to assess the overall fit of the final model. Statistical analysis and data management for this study were carried out using R and SPSS (IBM SPSS 25).

16.3 Results

A total of 230 elderly generated as a sample. Among them, 43.50%, 39.10% and 17.00% elderly were included from rural, urban and ethnic, respectively (Fig. 16.1). Among 230 elderly, 52.20% were male and 47.80% were female (Fig. 16.2). From (Fig. 16.3), it was depicted that the percentage of CHD in the elderly were high in urban and it was gradual decreases in urban and ethnic.

Fig. 16.1 Percentage of elderly by residence/strata

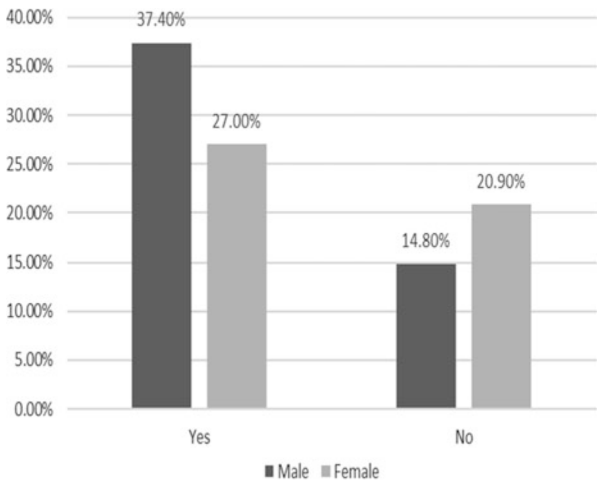
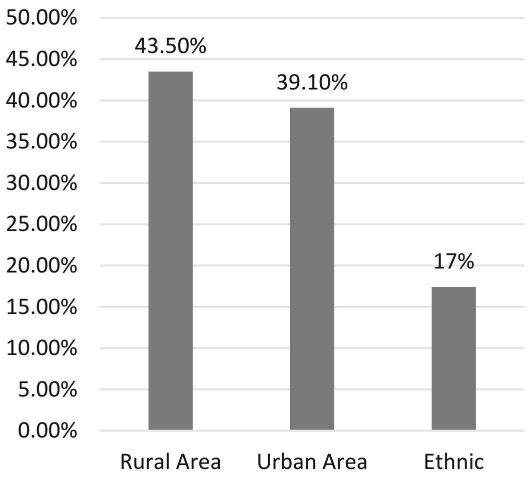


Fig. 16.2 Percentage of elderly by sex



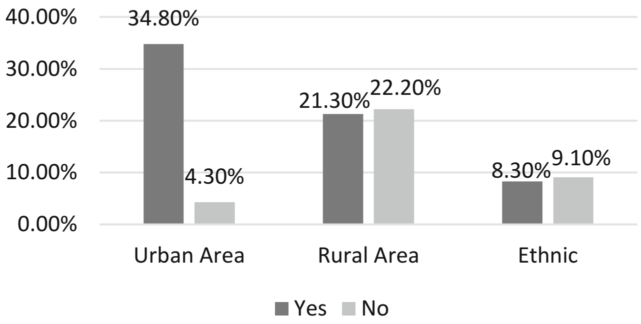


Fig. 16.3 Percentage of elderly by residence and coronary heart disease

16.3.1 Socio-economic Status

From the Table 16.1 it can be said that 27.0% female elderly and 37.4% male elderly had CD. Among all of the elderly people 72.2% were illiterate and 27.8% were literate. Most of the elderly (43.5%) lived in rural areas. Among all, 42.6% were employed others were unemployed. About all of the elderly, 83.9% reported their yearly income was less than 50,000 TK. From the BMI status who had CHD, 35.7% of them were overweight, 5.2% were normal weight and 23.5% were underweight.

16.3.2 Food Consumption

Table 16.2 represents the association of food consumption of elderly and their CHD. More than one third, 34.8% elderly had reported to take sugar and had CHD. Most of the elderly (32.8%) who had CHD and used to eat red meat. Again, 43.9% elderly people had CHD and used to have soft drinks and above half of the elderly (50.4%) had CHD who used to eat fast food. From this food consumption table, it can be assured that sugar intake, meal intake (three or two times a day), cooking oil, taking extra salt in meal, fast food, drinking water have some correlation with CHD whether taking red meat and soft drinks do not have that correlation with heart disease.

16.3.3 Risk Factors

Table 16.3 indicates the risks of other diseases related with CHD. In this table, 47.0% people had CHD and also had diabetes, 17.4% people had CHD but they were free from diabetes. Elderly people who had smoking habit (44.8%) suffering from CHD where 16.1% elderly were not suffering from CHD who did not use to

Table 16.1 Chi-square test for identifying socio-economic status associate with CHD

Socio-economic variables		Coronary heart disease Yes, *n* (%)	No, *n* (%)	Total, *n* (%)	*P*-value
Age, mean (SD)		65.84(8.60)	68.28(9.69)	66.71(9.06)	0.027
Gender	Male	86(37.4)	34(14.8)	120(52.2)	0.019
	Female	62(27.0)	48(20.9)	110(47.8)	
Marital status	Married	122(53.0)	62(27.0)	184(80.0)	0.231
	Others	26(11.3)	20(8.7)	46(20.0)	
Education level	Illiterate	116(50.4)	50(21.7)	166(72.2)	0.006
	Literate	32(13.9)	32(13.9)	73(27.8)	
Residence	Urban area	80(34.8)	10(4.3)	90(39.1)	0.000
	Ethnic	19(8.3)	21(9.1)	40(17.4)	
	Rural area	49(21.3)	51(22.2)	100(43.5)	
Occupation	Employed	73(31.7)	25(10.9)	98(42.6)	0.008
	Unemployed	75(32.6)	57(24.8)	132(57.4)	
Wealth index (yearly income)	Less than 50,000 taka	64(27.8)	129(56.1)	193(83.9)	0.158*
	Between 50,000-1,50,000	13(5.7)	14(6.1)	27(11.7)	
	More than 1,50,000	6(2.6)	4(1.7)	10(4.3)	
BMI	Underweight	54(23.5)	33(14.3)	87(37.8)	0.000
	Normal	12(5.2)	21(9.1)	33(14.3)	
	Overweight	82(35.7)	28(12.2)	110(47.8)	
Exercise	Yes	83(36.1)	74(32.2)	157(68.3)	0.000
	No	65(28.3)	8(3.5)	73(31.7)	
Total		148(64.3)	82(35.7)	230(100)	

Note: *indicates not significant

in habit of smoking. 48.7% elderly were suffering from mental stress with CHD. Elderly who were suffering from hypertension (51.3%) also had CHD. The results also indicate that diabetes, smoking, chest pain, mental stress and hypertension have clear relation with heart disease. Only respiratory did not show any relation with CHD.

16.3.4 Multiple Logistic Regression Model

Table 16.4 shows that there were significant differences in CHD and residence of elderly when the socio-economic factors, food consumption and other risk factors was adjusted. For instance, female elderly was 93% (AOR = 0.07; 95% CI: 0.02–0.26) less likely to have CHD as compared with male elderly. Elderly in ethnic were

Table 16.2 Chi-square test for identifying food consumption associate with CHD

Food consumption		Yes, *n* (%)	No, *n* (%)	Total, *n* (%)	*P*-value
Sugar intake	Yes	80(34.8)	12(5.2)	92(40.0)	0.000
	No	68(29.6)	70(30.4)	138(60.0)	
Red meat	Yes	75(32.8)	39(17.0)	114(49.8)	0.680*
	No	72(31.4)	43(18.8)	115(50.2)	
Meal intake	Two times	18(7.9)	19(8.3)	37(16.2)	0.040
	Three times	128(56.1)	63(27.6)	191(83.8)	
Cooking oil	Palm oil	36(15.7)	52(22.6)	88(38.3)	0.000
	Soya bean oil	107(46.5)	28(12.2)	135(58.7)	
	Others	5(2.2)	2(0.9)	7(3.0)	
Extra salt	Low	10(4.3)	16(7.0)	26(11.3)	0.001
	Medium	47(20.4)	12(5.2)	59(25.7)	
	High	91(39.6)	54(23.5)	145(63.0)	
Soft drinks	Yes	101(43.9)	65(28.3)	166(72.2)	0.091*
	No	47(20.4)	17(7.4)	64(27.8)	
Fast food	Yes	116(50.4)	50(21.7)	166(72.2)	0.005
	No	32(13.9)	32(13.9)	64(27.8)	
Total		148(64.3)	82(35.7)	230(100)	

Note: *indicates not significant

Table 16.3 Chi-square test for identifying risks factors associate with CHD

Risk factors		Yes, *n* (%)	No, *n* (%)	Total, *n* (%)	*P*-Value
Hypertension	Yes	118(51.3)	28(12.2)	146(63.5)	0.000
	No	30(13.0)	54(23.5)	84(36.5)	
Diabetes	Yes	108(47.0)	35(15.2)	143(62.2)	0.000
	No	40(17.4)	47(20.4)	87(37.8)	
Smoking	Yes	103(44.8)	45(19.6)	148(64.3)	0.031
	No	45(19.6)	37(16.1)	82(35.7)	
Chest pain	Yes	29(12.6)	32(13.9)	61(26.5)	0.029
	No	116(50.4)	53(23.0)	169(73.5)	
Depression	Yes	112(48.7)	28(12.2)	140(60.9)	0.000
	No	36(15.7)	54(23.5)	90(39.1)	
Respiratory	Yes	20(8.7)	13(5.7)	33(14.3)	0.696*
	No	128(55.7)	69(30.0)	197(85.7)	
Total		148(64.3)	82(35.7)	230(100)	

Note: *indicates not significant

a 96% (AOR = 0.04; 95% CI: 0.01–0.175) lower chance to have CHD compared to rural elderly people. It is worthwhile to mention that people of the elderly of urban areas were 4.03 times more likely (AOR = 4.03, 95% CI: 1.22–13.29) to hold CHD compared to rural elderly people when other factors were adjusted. The odds ratio for the employed elderly had 98% (AOR = 0.02, 95% CI: 0.01–0.12) lower chance compare to who had not involvement in job. The odds of underweight

Table 16.4 Multiple logistic regression of CHD among elderly people

Variables		Adjusted OR	95% CI	*P*-value
Gender	Female	0.07	[0.02–0.26]	0.000
	Male	Ref	–	–
Residence	Urban area	4.03	[1.22–13.29]	0.022
	Ethnic	0.04	[0.01–0.175]	0.000
	Rural area	Ref	–	–
Education	Literate	0.21	[0.06–0.79]	0.021
	Illiterate	Ref	–	–
Occupation	Employed	0.023	[0.01–0.12]	0.000
	Unemployed	Ref	–	–
BMI	Underweight	0.77	[0.02–0.33]	0.000
	Normal	0.71	[0.02–0.31]	0.000
	Overweight	Ref	–	–
Exercise	Yes	0.11	[0.03–0.50]	0.004
	No	Ref	–	–
Sugar intake	Yes	4.32	[1.35–13.81]	0.014
	No	Ref	–	–
Cooking oil	Good	0.07	[0.01–4.15]	0.066
	Bad	Ref	–	–
Smoke	Yes	11.68	[3.06–44.57]	0.000
	No	Ref	–	–
Mental stress	Yes	5.50	[1.69–17.90]	0.005
	No	Ref	–	–
Hypertension	Yes	4.59	[1.36–15.55]	0.014
	No	Ref	–	–
Diabetes	Yes	3.95	[1.43–10.94]	0.008
	No	Ref	–	–

elderly and normal-weight elderly affected by CHD were 33% (AOR = 0.77, 95% CI: 0.02–0.33) and 39% (AOR = 0.71, 95% CI: 0.02–0.31) less likely than those elderly who were overweight. Elderly people who were exercise regularly had 89% (AOR = 0.11, 95% CI: 0.03–0.50) less risk to develop CHD compared to those elderly who were not involve in regular exercise.

Elderly who reported take sugar were (AOR = 4.32, 95% CI: 1.35–13.81) more likely to have CHD compared to the people who were not take soft drinks. After adjusting for the risk factors, we found that smoke had 11.68 times (AOR = 11.68, 95% CI: 3.06–44.57) higher odds of CHD compared to people who had not smoked. On the other hand, the likelihood of mental stress and blood pressure had increased 5.50 times higher the risk [95% CI: 1.69–17.90] and 4.59 times higher the risk [95% CI: 1.36–15.55] of having CHD in comparison to the elderly had no mental stress and no hypertension.

Table 16.5 Test for goodness of fit of the final model

Hosmer and Lemeshow goodness of fit test		
Value	df	P-value
8.792	8	.360

16.3.5 Hosmer and Lemeshow Goodness of Fit Test

Hosmer and Lemeshow goodness of fit test has been given in Table 16.5.

So, the Hosmer-Lemeshow test does not gives us significant evidence of good fit on 36% of occasions.

16.4 Discussion

This study shows that rural communities have lower risk factors of heart diseases than urban westernized populations. Another study reveals that this difference may be attributable to a low fat (15–20 g per day) diet based on whole grain (400 g per day) combined with physically demanding occupations [23] which is similar to our results. A recent report published that the prevalence of stroke of elderly people (aged >30 years) in rural population has been found to be 0.94% in general, 1.45% in male and 0.45% in female [24]. In Bangladesh, year of 2014, non-communicable diseases (NCDs) represented 59% of the total deaths; CHD was the single-most important contributor, being responsible for 17% of total death [25]. Recent studies suggest that socio-economic statuses, as assessed by occupation, education and income level [26] is closely related with the quality of diet though differences in the amount of food or in nutrient intake among social classes are often small [27, 28]. Occupation is considered as an important marker of socio-economic status which gives the economic condition of people and income ranges. Income provides access to goods and services, including quality education and medical care, which may protect against disease [29]. This study concludes that unemployed people have higher chance of heart diseases compared with the employed people and normally people of urban areas are more sufferer than the rural people [30]. Sitting for prolonged periods would also cause the loss of opportunity for cumulative energy expenditure resulting from the thousands of intermittent muscular contractions throughout the 16-h period that people are awake. This may have chronic effects on the propensity to become overweight [31]. And this study also reveals that people who are inactive are being attacked by heart diseases more often than the people who're physically active. Male are suffering from heart diseases than the female and BMI Index is another most important risk factor for heart disease. A study shows that Body mass index (BMI), waist to hip ratio, obesity, independent fat distribution and weight gain since age 21 are associated with an increased risk of coronary heart disease and is getting stronger day by day [32]. Along with the previous studies, this study also reveals that people who are overweight they are in higher risk of attacked by heart diseases than the people who have normal weight and underweight. There is an inverse relation between education and long-term risk of coronary heart disease, this study indicates that heart diseases rarely vary from literate to illiterate people. But practically literate people are more conscious about the nutrition value of food [33, 34]. In an absolutely theoretical basis, the higher educational level could be related with more stressed occupation, less available time for cooking and consequently with a bigger trend for consumption of ready-to eat food or fast-food habits [35]. Food consumption is most important risk factors for heart diseases. People who are used to fast foods, intake extra salt and sugar, soft drinks, polyunsaturated fat foods are in higher risk of heart diseases. Heart disease is significantly associated with a lower intake of saturated fat and higher intakes of polyunsaturated fat, alcohol, folate, vitamin C, and vitamin E [36]. Healthy nutrient food consumption, proper meals every day and quitting smoking can reduce the risk of heart disease. In Bangladesh people are used to tobacco consumption i.e. smoking which is a risk factor for heart disease. From this heart disease people are attacked with many diseases like diabetes, hypertension, chest pain, back pain etc. which make people more inactive. Diabetes has long been recognized as one of the most important major cardiovascular risk factors, and in particular from cardiovascular disease diabetic subjects have a risk of early death [37].

16.5 Conclusion

In this study, we found that there is a wide range of factors, which are significantly associated with CHD among elderly (age > 60 years). The finding demonstrates that the primary risk factors for cardiovascular disease are smoking, hypertension, sedentary lifestyle, and diabetes. Moreover, a significant proportion of the elderly men and women are not aware of the consequence of this disease, and a small proportion of them are taking healthy dietary. Exercise is vital to reduce risk of heart disease. Health professionals have a crucial role in influencing lifestyle decisions. However, most health care delivery in south Asia is through formally and informally trained and traditional health professionals, who should become part of prevention strategies for CHD to obtain the maximum effect. CHD can be included cost effectively in existing training programs for these health professionals.

Declaration
Conception, Design, Writing, Review and Edit: KFF
Review and Edit: AAK
Analysis: MNH
Writing: TI

Acknowledgement We would like to thank SUST research center for funding this research.

References

1. Cassar, A., et al.: Chronic coronary artery disease: diagnosis and management. Mayo Clin. Proc. **84**(12), 1130–1146 (2009)
2. Iyngkaran, P., et al.: Risk stratification for coronary artery disease in multi-ethnic populations: are there broader considerations for cost efficiency? World J. Methodol. **9**(1), 1–19 (2019)
3. Benjamin, E.J., et al.: Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation **137**(12), e67–e492 (2018)
4. WHO (World Health Organization).: Cardiovascular Diseases (CVDs). http://www.whoint/mediacentre/factsheets/fs317/en/index.html. Last accessed 2015/01
5. Saquib, N., et al.: Cardiovascular diseases and type 2 diabetes in Bangladesh: a systematic review and meta-analysis of studies between 1995 and 2010. BMC Public Health. **12**(1), 434 (2012)
6. Lim, S.S., et al.: A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. **380**(9859), 2224–2260 (2012)
7. The, G.B.D.O.C, et al.: Global, regional and national prevalence of overweight and obesity in children and adults 1980-2013: a systematic analysis. Lancet. **384**(9945), 766–781 (2014)
8. WHO (World Health Organization).: Obesity and Overweight (2014)
9. Castelli, W.P.: Epidemiology of coronary heart disease: the Framingham study. Am. J. Med. **76**, 4–12 (1984)
10. WHO.: The Atlas of Heart Disease and Stroke. WHO, Geneva, (2004)
11. WHO.: Noncommunicable Diseases (NCD) Country Profiles. Bangladesh, http://apps.who.int/iris/bitstream/10665/128038/1/9789241507509_eng.pdf?ua=1 (2014). Last accessed 14 Dec 2014
12. Nag, T., Ghosh, A.: Cardiovascular disease risk factors in Asian Indian population: a systematic review. J. Cardiovasc. Dis. Res. **4**(4), 222–228 (2013)
13. Islam, A.K.M.M., Majumder, A.A.S.: Coronary artery disease in Bangladesh: a review. Indian Heart J. **65**(4), 424–435 (2013)
14. Bangladesh Bureau of Statistics (BBS).: http://www.bbs.gov.bd/Home.aspx (2016). Last accessed 16 Nov 2016
15. Sameh E. S., Zunaid, A.K., Tarcey, Michel. M. E.: Tackling Noncommunicable Diseases in Bangladesh: Now Is the Time. Directions in Development – Human Development. The World Bank 136 (2013)
16. Bulato, R.A., Stephens, P.W.: Global Estimates and Projections of Mortality by Cause. Population, Health and Nutrition Department, Washington, DC. World Bank **1007** (1992)
17. Reddy, K.S., Yusuf, S.: Emerging epidemic of cardiovascular disease in developing countries. Circulation. **97**, 596–601 (1998)
18. Joshi, P., et al.: Risk factors for early myocardial infarction in South Asians compared with individuals in other countries. JAMA. **297**(3), 286–294 (2007)
19. Gaziano, T.A., et al.: Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I follow-up study cohort. Lancet. **371**(9616), 923–931 (2008)
20. Abdullah, J.M., Ahmad, M.M., Saqib, S.E.: Understanding accessibility to healthcare for elderly people in Bangladesh. Dev. Pract. **28**(4), 552–561 (2018)
21. Gu, D., Zhang, Z., Zeng, Y.: Access to healthcare services makes a difference in healthy longevity among older Chinese adults. Soc. Sci. Med. **68**(2), 210–219 (2009)
22. Chowdhury, M.Z.I., et al.: Prevalence of cardiovascular disease among Bangladeshi adult population: a systematic review and meta-analysis of the studies. Vasc. Health Risk Manag. **14**, 165–181 (2018)
23. Singh, R.B., et al.: Prevalence of coronary artery disease and coronary risk factors in rural and urban populations of North India. Eur. Heart J. **18**(11), 1728–1735 (1997)

24. Zaman, M.M.A.C., et al.: Prevalence of stroke in a rural population of Bangladesh. Glob. Heart. **10**, 333–334 (2015)
25. World Health Organization.: Noncommunicable Diseases (NCD) Country Profiles, 2014,Bangladesh. Available from: http://apps.who.int/iris/bitstream/10665/128038/1/9789241507509_eng.pdf?ua=1 (2014)
26. Krieger, N., Williams, D.R., Moss, N.E.: Measuring social class in US Public Health Research: concepts, methodologies, and guidelines. Annu. Rev. Public Health. **18**(1), 341–378 (1997)
27. Galobardes, B., Morabia, A., Bernstein, M.S.: Diet and socioeconomic position: does the use of different indicators matter? Int. J. Epidemiol. **30**(2), 334–340 (2001)
28. Groth, M.V., Fagt, S., Brøndsted, L.: Social determinants of dietary habits in Denmark. Eur. J. Clin. Nutr. **55**, 959–966 (2001)
29. Psaltopoulou, T., et al.: Socioeconomic status and risk factors for cardiovascular disease: impact of dietary mediators. Hell. J. Cardiol. **58**(1), 32–42 (2017)
30. Fatema, K., et al.: Prevalence of risk factors for cardiovascular Diseases in Bangladesh: a systematic review and meta-analysis. PLoS One. **11**(8), e0160180 (2016)
31. Hamilton, M.T., Hamilton, D.G., Zderic, T.W.: Role of low energy expenditure and sitting in obesity, metabolic syndrome, type 2 diabetes, and cardiovascular disease. Diabetes. **56**(11), 2655–2667 (2007)
32. Rimm, E.B., et al.: Body size and fat distribution as predictors of coronary heart disease among middle-aged and older US men. Am. J. Epidemiol. **141**(12), 1117–1127 (1995)
33. Patterson, R.E., et al.: Is there a consumer backlash against the diet and health message? J. Am. Diet. Assoc. **101**(1), 37–41 (2001)
34. Wardle, J., Parmenter, K., Waller, J.: Nutrition knowledge and food intake. Appetite. **34**(3), 269–275 (2000)
35. Kirkpatrick, S.T.V.: The relationship between low income and household food expenditure patterns in Canada. Public Health Nutr. **6**, 589–597 (2003)
36. van Dam, R.M., et al.: Patterns of food consumption and risk factors for cardiovascular disease in the general Dutch population. Am. J. Clin. Nutr. **77**(5), 1156–1163 (2003)
37. The, D.S.G: Glucose tolerance and mortality: comparison of WHO and American Diabetes Association diagnostic criteria. Diabetes epidemiology: collaborative analysis of diagnostic criteria in Europe. Lancet. **617-621**, 354 (1999)

Chapter 17
Finite Mixture Modelling Approach to Identify Factors Affecting Children Ever Born for 15–49 Year Old Women in Asian Country

Md. Karimuzzaman, Md. Moyazzem Hossain, and Azizur Rahman

Abstract The number of ever born children is one of the main components of population dynamics that determine the size, structure, as well as the composition of a countries' population. Children ever born refer to the number of children born alive to the person up to a specified reference date and served as a response variable here. A secondary dataset is used in this paper that is obtained from a countrywide representative survey entitled Bangladesh Demographic and Health Survey (BDHS) 2014. This study aims to identify the socioeconomic and demographic factors influencing children ever born to the women of 15–49 years old in Bangladesh. The first attempt of this paper is to identify the best-fitted model among generalized Poisson, Negative Binomial, truncated, COM and finite mixture regression model form. The results suggest that among the model considered in this study Finite Mixture Negative Binomial Regression with three components gives the best-fitted model to estimate the number of ever born children in Bangladesh. It reveals that respondents age, residential status, family size and intention of using contraception have shown positive impact and respondents education, drinking water, toilet facility, religious status, household head age, wealth index, age at first birth, and husband education shows a negative impact on ever born children.

Keywords Children ever born · Poisson · negative binomial · finite mixture regression · Bangladesh

Md. Karimuzzaman
Jahangirnagar University, Savar, Dhaka, Bangladesh

Md. Moyazzem Hossain (✉)
School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK

Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh

A. Rahman
School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW, Australia

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_17

17.1 Introduction

Bangladesh had achieved the target of reduction of infant mortality as there was 151 in the year 1991 and 41 in the year 2017 per 1000 alive birth but still, the infant mortality is high compared to the developed countries [1]. Moreover, high infant mortality, as well as economic security at future life, has influenced the decision about the family size [2]. Bangladesh is a country of 160 million people with 40% Children of its population, in recent time tendency of having more children is decreased significantly as in 2015 the fertility rate was 2.1533 and 2017 that was 2.076 [3]. Due to the modernization of societies, the number of children per woman has decreased substantially throughout the world and reached just below 2.5 children per woman globally [4]. The number of children born alive among the married women of age 15 years or more or the total summary of experienced live birth of women in her life can be considered as ever born children of a mother [5] which is acted as a count variable.

Many researchers measure the consequence of ever born children of a mother to accompanying factors as Satyavada and Aravinda [6] displays education has a substantial positive as well as negative influence on contraception and ever instinctive children of a mother respectively. Interestingly in Japan, younger women have lower fertility intentions but people who live in rural areas with larger family members shows the inverse [7]. A recent study showed that the number of children ever born has significant influence's on respondent's age, religion, and both respondents as well as husband education [5], Where another study indicate education, employment as well as food security as key responsible childbearing factors among the women [8]. Moreover, in a society couples with more daughters and a strong fondness for sons continue childbearing longer than those with more sons [9].

Bangladesh is passing through the demographic dividend from the 1980s and will continue until 2040 and if the country fails to take potential economic benefits from it then Bangladesh has to pay the huge cost with unemployment, unbearable stain on education, health and old age security [10]. Moreover, it is mentionable that the projected population of Bangladesh will be 200 million within 2053 [11]. To handle the created problem by population pressure, the government of Bangladesh takes several actions to ensure the basic needs for these populations in future. The government also trying to influence its people to control the live birth by several social and national campaigning, but due to the increasing speed of population growth as well as with the problem of overpopulation, these steps may not enough. Probably most common regression approach for handling count data (the number of kids ever born considered as a count variable in nature) is Poisson regression as linear regression fails to allow for the limited number of possible values of the response variable with a problem of under-dispersion and over-dispersion. To find out the factors affecting Children ever born in Botswana [12], Ethiopia [13] and as well as Nigeria [14] was studied by application of the Poisson regression model. Another more formal way for handling count data is Negative binomial

(NB) Regression which can also handle over-dispersion in data. To determine the influencing factors for the fertility of Sudanese [15] women as well as analysis of regional differentials in under-five mortality in Kenya [12] were done through Negative Binomial Regression.

Although these models can handle over-dispersion there is another problem of having zeros in data. Thus zero-augmented models were addressed for fixing this issue by capturing zero counts [16, 17]. In addition, Zero-truncated Count Regression Models were used in the estimation of the fertility of US women [18] and statistical modeling of fertility experience among women of reproductive age in Nigeria [14] as well as estimation of the number of ever born children of Bangladesh [5]. Moreover, logistic regression [8], multiple classification [19] and Support Vector Machine [20] also used by several researchers to estimate the number of ever born children. The Conway-Maxwell-Poisson (COM-Poisson) regression model became popular in recent study's which permits modeling with both over and under-dispersion distribution. It allows the modeling of mean and variance distinctly with the same covariate that affect average levels and variability in a different way [21, 22]. Sellers et al. [21], demonstrated a lifted COM-Poisson regression model for a dataset having no zero and they also discoursed several extensions like zero-inflated and zero-deflated COM-Poisson distributions by including an extra parameter.

On the other hand, Finite Mixture Model (FMM) makes a new way of controlling the count data. However, in order to identify the unknown distributional shapes of a dataset mixture models are being used as more convenient [23]. Due to the flexibility, researchers are more interested in Finite Mixture model and the application of this model has expanded enormously in the statistical as well as general scientific literature. Moreover, Finite mixture models (FMMs) can contain any number of observed and unobserved subpopulations that can help to classify the observations and adjust for clustering. The most interesting advantage of FMMs models is to model unobserved heterogeneity and the specific models from sub-population need not be limited to a mixture of normal densities. Another great advantage of FMMs models includes modeling the binary, ordinal, nominal, and count responses with the mixtures of linear and generalized linear regression models which also allows the inclusion of covariates with subpopulation-specific effects and the vector of regression coefficients to vary from component to component [24, 25]. Because of the flexibility of finite mixture models, many researchers have found many new and interesting fields of application. The finite mixture model for count data was applied in the study of examining the incidence of sudden infant death syndrome [26], motor vehicle crashes [23], criminal careers [27], insurance rate making [28], highway safety [29] and use of E-cigarette among teenage cigarette smokers [30].

From the above discussion, it is clear that there are several models employed by several researchers to analyze the count data. In previous literature, researchers are taken their decision mostly by using standard Poisson and negative binomial count regression approaches. However, very few researchers select appropriate count regression models for predicting the ever born children. Furthermore, no existing literature found by the author of this manuscript of using a finite mixture model to predict the ever born children of Bangladesh. Therefore, this study is going to find out the influential factors for the number of children ever born and make a prediction by best-fitted count regression approaches.

17.2 Methods and Materials

17.2.1 Data Sources and Variables

A secondary dataset is used in this paper that is obtained from a countrywide representative survey entitled Bangladesh Demographic and Health Survey (BDHS) 2014 published by the Bangladesh Bureau of Statistics (BBS). The survey was conducted through two-stage stratified sampling wherein the first stage considers total 600 Enumeration Areas were selected with probability proportional to the Enumeration Areas size and second stage sampling was done through a systematic sampling of average 30 households per Enumeration Areas for both urban and rural distinctly for entire Bangladesh. For the 2014 BDHS survey, 17,989 households were selected. The details of the sampling procedure are available in the report of Bangladesh Demographic and Health Survey 2014 [31]. The explanatory variables are selected on the available literatures as well as the existence relationships with the number of children ever born. Women were chosen as the respondents in order to study about childbearing, prenatal and postnatal care mostly in demographic point of view [5, 32].

The factors and variables being considered for analysis to includes:

- *The number of children*: It defines the total children ever born which is considered as the depended variable in the study.
- *Age*: Its define Respondent's current age in years.
- *Division:* It referred to the seven administrative Division of Bangladesh as Dhaka, Barisal, Khulna, Chittagong, Rajshahi, Rangpur and Sylhet.
- *Residence:* It's referred to where the respondent's live with two categories Urban and Rural.
- *Education:* Its represent the highest education level of respondent's with four categories No education; Primary; Secondary; and Higher Secondary.
- *Drinking Water:* It define, from which source the respondent's drink the water with two categories Safe sources and Unsafe sources.
- *Toilet:* It's referred that whether the respondents has a better toilet facility or not.
- *Religion:* From which religion the respondents belongs among Muslim; Hinduism; Christianism; Buddhism; & Others.
- *Family Size*: This variable indicates the total number of family member in the respondent's family.

- *Household Head Age:* This variable indicates the age of respondent's household head.
- *Wealth index:* This referred to the health conditions of respondents with five categories as poorest; poor; middle; richer and richest.
- *First Birth-Age:* This variable describes the age of respondent at first birth in years.
- *Contraceptive:* This referred to the current situation of respondents whether they use contraceptive or not.
- *Body Mass:* This variable referred Body Mass Index of the respondents.
- *Husband Education:* Its referred Husband/partner's education level with five categories as No education; Primary; Secondary; Higher Secondary; and don't know.
- *Husband Profession:* Its define the Husband/partner's profession labeled as No income, Lower Income, Medium income and High income.

17.2.2 Methods

All the mathematical illustration of Count Regression approaches that are mentioned in the previous section are included in the following section.

Zero Truncated Poisson Regression Model
Since this study includes the observation having the response value larger than zero, so the zero-truncated models were applied for predicting the ever born children [5]. As Bonat et al. [22], Cameron and Trivedi [33] and Hilbe [35] illustrate the zero truncated model for at least one must occurred event Y_i so the zero truncated Poisson distribution with mean μ_i can be defined as,

$$f\left(Y_i, \mu_i\right) = P\left(Y_i = y_i \,|y_i > 0\right) = \frac{\exp\left(-\mu_i\right) {\mu_i}^{y_i}}{\left[1 - \exp\left(-\mu_i\right)\right] y_i!} \tag{17.1}$$

where, $\mu_i = \exp\left({X_i}^T \beta_i\right)$ with k-dimensional vector covariates of $X_i = [x_{1i}, x_{2i}, \ldots, x_{ki}]$.

The estimates can be illustrated through maximum likelihood estimation (MLE) from the following log-likelihood function,

$$l\left(Y_i; \mu_i \,|y_i > 0\right) = \sum \left\{ \begin{array}{l} y_i\left[{X_i}^T \beta_i\right] - \exp\left({X_i}^T \beta_i\right) - \ln\left(y_i!\right) \\ \qquad\qquad - \ln\left[1 - \exp\left\{-\exp\left({X_i}^T \beta_i\right)\right\}\right] \end{array} \right\} \tag{17.2}$$

Differentiating (17.5) with respect to β_i yields MLE of β_i.

Zero Truncated Negative Binomial Regression Model
There is no presence of zeros in the response variable in this study, hence the zero truncated negative binomial regression were applied to determine the number of

children ever born. Cameron and Trivedi [33] and Hilbe [35] mentioned the applied model for zero truncated negative binomial distribution of Y_i as below,

$$f\left(Y_i,\mu_i,k\right)=P\left(Y_i=y_i\left|y_i\right.>0\right)=\frac{\left|\overline{\left(k+y_i\right)\left(\frac{k}{k+\mu_i}\right)^k\left(1-\frac{k}{k+\mu_i}\right)^{y_i}}\right.}{\left[1-\left(\frac{k}{k+\mu_i}\right)^k\right]y_i\left|\overline{(k)}\right.} \tag{17.3}$$

where, $\mu_i = \exp\left(X_i{}^T\beta_i + \varepsilon_i\right)$ with combine effect of unobserved variables ε with k-dimensional vector covariates of $X_i = [x_{1i}, x_{2i}, \ldots, x_{ki}]$. The Log-likelihood function of zero-truncated negative binomial regression can be obtained similarly as zero-truncated Poisson regression with the help of Eq. (17.3).

Finite Mixture Poisson and Negative Binomial Regression Model

To examine the influences of different factors on the total number of children ever born, Finite Mixture Models (FMM) were applied where the FMM captures deviations in the study factor with over varied and mutually exclusive subpopulations ([33, 37, 38]). Distinct from one-population estimates through a single component distribution, a FMM can account an environment for observed heterogeneity that clusters around a restricted set of sub populations [33, 38]. The FMM suggests that a population is organized of two or more distinct unobserved subpopulations with anonymous mixing weights or proportions or class which estimated beside through the additional factors of model [33, 38]. Even if the underlying mixing distribution is continuous therefore finite mixture produces a valid estimate [33].

The Finite Mixture Poisson-K (FMP-*K)* regression model assumes the mixture of Poisson distributions with marginal distribution of y_i as,

$$P\left(y_i\left|X_i,\Theta\right.\right)=\sum_{k=1}^{K}w_k\left(\frac{e^{-\mu_{k,i}}\left(\mu_{k,i}\right)^{y_i}}{y_i!}\right) \tag{17.4}$$

with $E\left(y_i\left|X_i,\Theta\right.\right) = \sum_{k=1}^{K} w_{k,i}\mu_{k,i}$ and $Var\left(y_i\left|X_i,\Theta\right.\right) = E\left(y_i\left|\Theta\right.\right) + \sum_{k=1}^{K} w_{k,i}{\mu^2}_{k,i} - E(y_i\left|\Theta\right.)^2$, where, $\mu_{k,\,i} = \exp\left(X_i\beta_k\right)$ and $\Theta = \{(\beta_1, \ldots, \beta_K), w\}$. Unless all the component's means are the same ($\mu_{1,\,i} = \ldots = \mu_{k,\,i}$), the variance is constantly greater than the mean.

For the Finite Mixture Negative Binomial-K (FMNB-K) regression model, it is assumed that the mixture of negative binomial distributions with marginal distribution of y_i. The model can be defined as below,

$$P\left(y_i\,|X_i,\Theta\right)=\sum_{k=1}^{K}w_k\left[\frac{\left|\overline{(y_i+\phi_k)}\right.}{\left|\overline{(y_i+1)}\right.\left|\overline{(\phi_k)}\right.}\left(\frac{\mu_{k,i}}{\mu_{k,i}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{k,i}+\phi_k}\right)^{\phi_k}\right] \tag{17.5}$$

with $E\left(y_i\,|X_i,\Theta\right)=\sum_{k=1}^{K}w_{k,i}\mu_{k,i}$ and $Var\left(y_i\,|X_i,\Theta\right)=E\left(y_i\,|X_i,\Theta\right)+\left(\sum_{k=1}^{K}w_{k,i}{\mu^2}_{k,i}\left(1+\frac{1}{\phi_k}\right)-E(y_i\,|X_i,\Theta)^2\right)$, where, $\mu_{k,\,i}=\exp\,(X_i\beta_k)$ and $\Theta=\{(\beta_1,\ldots,\beta_K),(\phi_1,\ldots,\phi_k),w\}$. In this case, even if all the component's means are the same, the variance is must be greater than the mean. The FMNB-K model is converted to the FMP-K model when β_K drives to infinity in separate component.

COM-Poisson Regression

The Conway-Maxwell Poisson (COM-Poisson or CMP) regression can define by random variable of weighted Poisson distribution as its belongs to the family of weighted Poisson distribution mentioned in [21] as,

$$f\left(y;\lambda,v\right)=\frac{\exp\left(-\lambda\right)\lambda^y w_y}{w_y!},\ y=0,1,\ldots, \tag{17.6}$$

where, $W=\sum_{i=0}^{\infty}\exp\left(-\lambda\right)\lambda^i w_i/i!$ is a normalizing constant [21, 22]. The COM-Poisson is obtained when $w_y=(y!)^{1\,-\,v}$ for $v\geq 0$. The series of $W$for COM-Poisson distribution is denoted by $Z(\lambda,v)$ and can be written as $\sum_{i=0}^{\infty}\frac{\lambda^i}{(i!)^v}$.

Sellers and Shmueli [39] suggested a regression model with the help of COM-Poisson distribution that can be defined as,

$$Y_i\sim CP\left(\lambda_i,v\right),\text{with }\lambda_i=g^{-1}\left({x_i}^T\beta\right)$$

where λ is covariates and v is the dispersion parameter with the condition $v>1$ and $v<1$ for under-dispersion and over-dispersion respectively.

In Bangladesh, there is a significant proportion of reproductive-age women who did not conceive any baby. As a consequence, this paper includes directly the statistical models which are based on a zero-truncated probability distribution for comparison instead of considering the standard Poisson and Negative binomial distributions. Though all the explained regression models are the specific model of the generalized linear model, however, in this study, generalized linear model-based Poisson and Negative binomial regression were also applied to make the comparison among them. One of the main reason behind the inclusion of Generalized linear model in this study is its well-known popularity among researcher, however, for a better explanation about standard Poisson, Negative binomial and Generalized models, readers are suggested to see [22, 33–36].

17.3 Results and Discussion

In this study, the first attempt was to get the underlying influence of factors which is related to having children by descriptive analysis. Generalized Poisson and Negative Binomial Regression, COM-Poisson Regression, Zero Truncated Poisson and Negative Binomial regression and Finite Mixture Poisson and Negative Binomial Regression with two components were fitted at the second step and make a comparison to identify the best-fitted model. The fitted model was assessed by using the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) and log-likelihood values where smaller values are chosen for taking the decisions [40]. After getting Finite Mixture Model as the best-fitted model then the third step was used to determine the number of components in the FMM analysis by estimating isolated models with two to four components without covariates and finally the best model is assessed by mentioned model selection criteria. The final step was to estimate the number of ever born children by the selected model and point toward the influencing factors from the study.

The range of the ever born children is 1–15 and noticed that there were 43,772 respondents with a mean and standard deviation of the response variable were 3.70 and 1.948 respectively. The minimum number of children was one whereas the maximum number of children was 15. Table 17.1 represents the number of ever born children and their frequencies along with the percentage. The evidence shows that 22.82% of the women had 3 children and 21.54% of the women had 2 children which were the second-highest. The rest of the frequencies given in Table 17.1 depicts that there is a very tiny percentage of having more than 5 children.

As mentioned before, the first step of this study was to check the influencing factors by descriptive analysis. The box-plot of different explanatory variables along with the number of ever born children is illustrated in Fig. 17.1. Under the variable division, Chittagong and Sylhet have more children than the rest of the division as 17.34% and 14.6% of the total children respectively (Fig. 17.1). Moreover, respondents who live in rural areas, have no education, drink safe water, follow

Table 17.1 Frequency distribution of number of children ever born

No. of children	1	2	3	4	5	6	7	
Frequency	3975	9424	9990	7520	5490	3618	1834	
Percent	9.08	21.52	22.82	17.17	12.54	8.26	4.18	
No. of children	8	9	10	11	12	13	14	15
Frequency	944	549	250	99	24	26	14	15
Percent	2.15	1.25	0.5	0.2	0.05	0.05	0.03	0.03

Islam as religion and also use contraception as well as having better toilet facilities have more children than others with the percentage of 69.1, 34, 92, 91.1, 61, 66.4 (Fig. 17.1).

On the other hand, respondents having a husband or partner who does not have any education as well as belongs to the lower-income group got more children than others (Fig. 17.1). Moreover, respondents with middle wealth condition have 21.3%

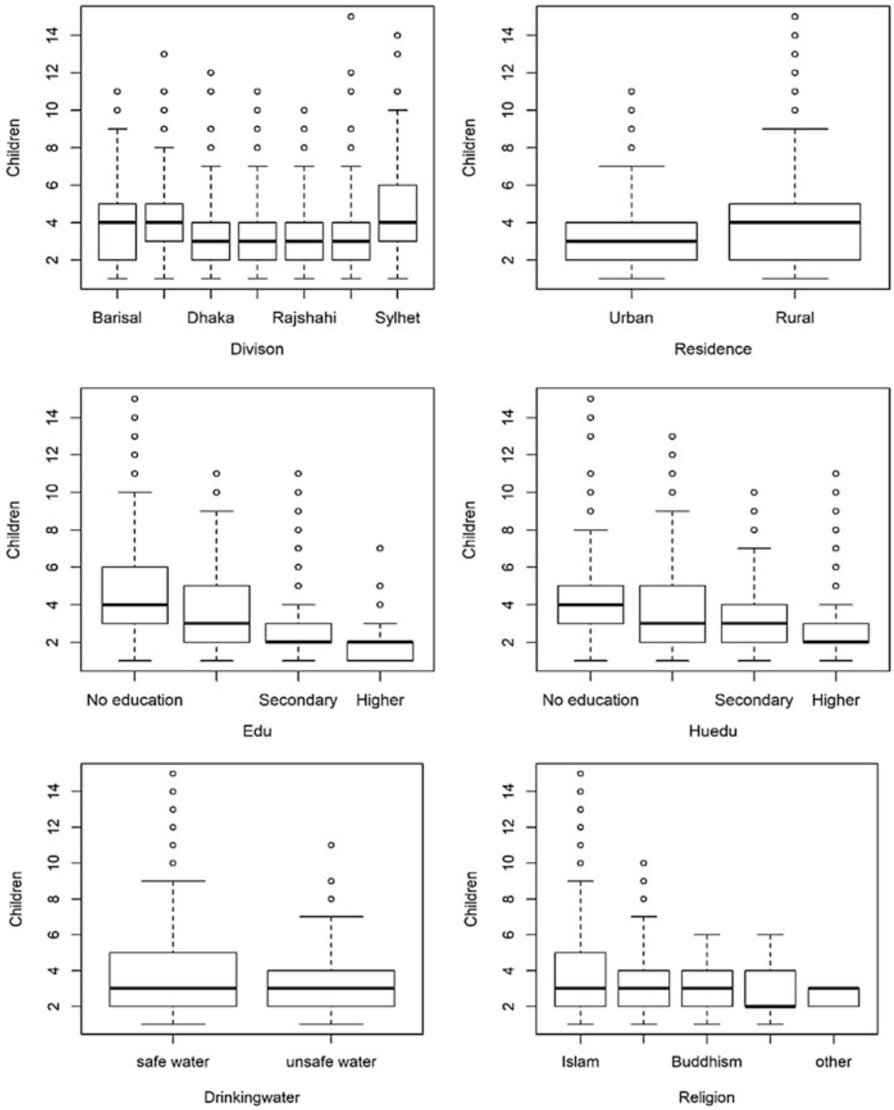


Fig. 17.1 Predictor against different variables

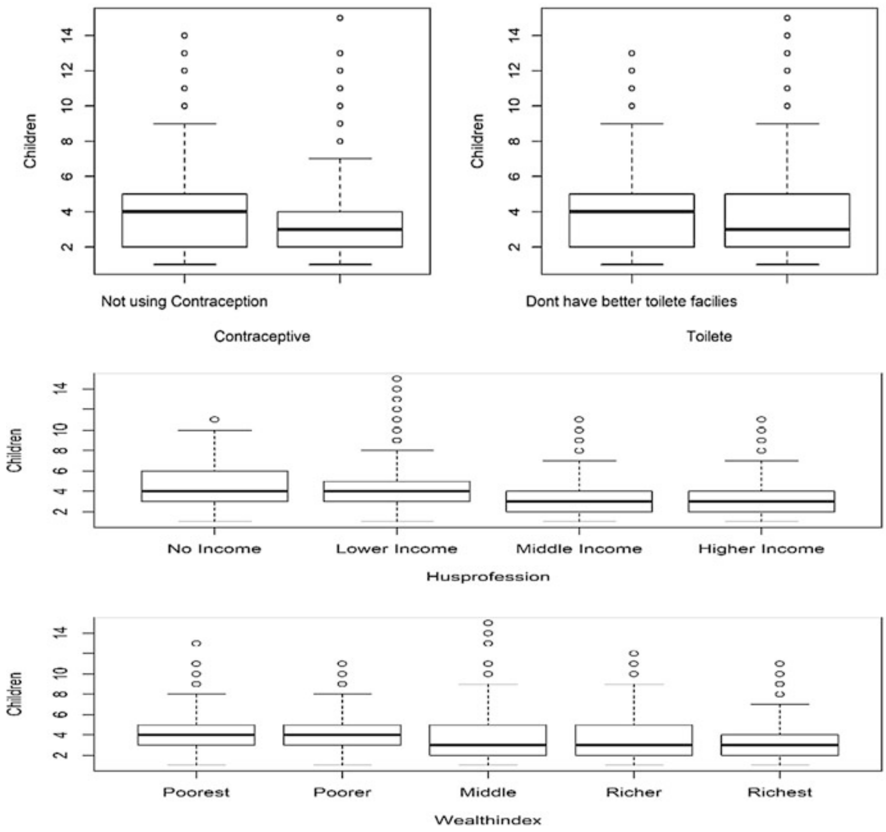


Fig. 17.1 (continued)

of the total children. Besides, most of the respondents have 3 children in the family with the average age about 34 years, family size is more than 5, husband age just above 44 years, age of the first birth is more than 17 years and on an average of body mass index approximately 2330 (Table 17.2).

As mentioned before the second step is used to find out the most appropriate model from the different count regression models considered in this study. The most appropriate model is selected through some well-known model selection criteria like AIC, BIC and Log-Likelihood and the results reveal that the Finite Mixture Negative Binomial Regression is the best-fitted model. However, the FMM suggests that a population is organized of two or more distinct unobserved subpopulations with anonymous mixing weights or proportions or class which estimated along with the other model is the best-fitted model (Table 17.3).

To identify the best-fitted component is the subsequent step of this study. After fitting up to 4-component Negative Binomial Finite Mixture Regression Model, the best-fitted model is selected with the help of AIC and BIC. The results depict that the

Table 17.2 Summary statistic of some selected variables against response variable

No. of Children	Statistic	Age	Family size	Household-head age	Age at first birth	Body mass index
1	Mean	24.37	5.32	43.32	19.09	2274.59
	n	3975	3975	3975	3975	3975
2	Mean	30.52	5.03	42.27	18.14	2367.88
	n	9424	9424	9424	9424	9424
3	**Mean**	**34.83**	**5.266**	**44.26**	**17.39**	**2330.52**
	n	**9990**	**9990**	**9990**	**9990**	**9990**
4	Mean	37.51	5.63	46.44	17.03	2302.85
	n	7520	7520	7520	7520	7520
5	Mean	39.85	5.83	48.14	16.96	2247.97
	n	5490	5490	5490	5490	5490
6	Mean	40.87	6.45	48.93	16.82	2191.24
	n	3618	3618	3618	3618	3618
7	Mean	42.16	6.80	50.73	16.67	2226.34
	n	1834	1834	1834	1834	1834
8	Mean	42.36	7.19	49.87	16.59	2129.05
	n	944	944	944	944	944
9	Mean	43.29	7.60	55.36	15.57	2377.60
	n	549	549	549	549	549
10	Mean	45.2	9.28	53.76	17	2535.28
	n	250	250	250	250	250
11	Mean	44.11	10.22	52	15.33	2969
	n	99	99	99	99	99
12	Mean	44	10	40	14	2280.5
	n	24	24	24	24	24
13	Mean	43.5	11	56.5	15	2056.5
	n	26	26	26	26	26
14	Mean	49	13	59	16	2553
	n	14	14	14	14	14
15	Mean	42	4	56	14	2068
	n	15	15	15	15	15

Note: *n* represent the number of respondents

3-component Negative Binomial Finite Mixture Regression Model is more suitable for predicting the number of ever born children in Bangladesh (Table 17.4).

From the evidence of Table 17.5, it may be concluding that among the components, the respondents age, residence, family size and using contraception have a significant positive impact on ever born children. Hence a conclusion about residence can be drawn as if a person lived in rural areas having more children than its counterpart. As respondents age and family size increases the likelihood of getting more children also increases.

On the other hand, respondent's education, drinking water, toilet facility, religion, household head age, wealth index, age at first birth, and husband education

Table 17.3 Results of model selection criteria

Model	AIC	BIC	Log-likelihood
Generalized Poisson Regression	151041.5	151328.2	−75487.75
Generalized Negative Binomial	151044.1	151339.4	−75488.03
Zero Truncated Poisson Regression	147712.5	147851.5	−73840.7
Zero Truncated Negative Binomial Regression	147714.5	147862.2	−73840.2
COM-Poisson Regression	151043.5	151338.8	−75487.74
Finite Mixture Poisson Regression	153168.8	153455.4	−76551.3
Finite Mixture Negative Binomial Regression	**30691.13**	**30977.8**	**−15312.57**

Note: Akaike information criteria (AIC) $= -2\ \ln(L) + 2k$, where k=number of parameters and Bayesian information criteria (BIC) $= -2\ \ln(L) + k\ \ln(n)$, where n=number of observations

Table 17.4 Results of number of components selection of finite mixture model

Model selection criteria	2 component	3 component	4 component
AIC	30691.13	**15471.5**	15534.07
BIC	30977.8	**15888.46**	16081.34

Table 17.5 Fitted values of finite mixture negative binomial regression model

Variables	Coefficient	Std. Err.	Z	p-value	95% CI	
Component1						
Age	0.03819	0.00081	47.39	<0.001	0.03661	0.03977
Division	−0.00033	0.003	−0.11	0.912	−0.00622	0.00556
Residence	0.04497	0.01011	4.45	<0.001	0.02515	0.06479
Education	−0.0626	0.00975	−6.42	<0.001	−0.08171	−0.04349
Drinking water	−0.08862	0.02207	−4.02	<0.001	−0.13188	−0.04537
Toilet	−0.03184	0.0097	−3.28	0.001	−0.05085	−0.01282
Religion	−0.02157	0.03449	−0.63	0.532	−0.08916	0.04602
Family size	0.0706	0.00924	7.64	<0.001	0.05249	0.08872
Household head age	−0.0024	0.00043	−5.64	<0.001	−0.00323	−0.00157
Wealth index	−0.04215	0.00434	−9.72	<0.001	−0.05064	−0.03365
Age at first birth	−0.03777	0.00146	−25.79	<0.001	−0.04064	−0.0349
Contraceptive	0.00148	0.00057	2.58	0.01	0.00036	0.00261
Body mass index	0.00029	0.001	−0.44	0.66	−0.00002	0.00001
Husband education	−0.02471	0.00627	−3.94	<0.001	−0.037	−0.01242
Husband profession	−0.0001	0.00031	−0.32	0.748	−0.0007	0.0005
Constant	0.53148	0.06658	7.98	<0.001	0.40098	0.66198
Component2						
Age	0.03833	0.00026	149.11	<0.001	0.03783	0.03884
Division	0.00089	0.00087	1.03	0.304	−0.00081	0.00259
Residence	0.048	0.00395	12.14	<0.001	0.04026	0.05575
Education	−0.05994	0.00262	−22.91	<0.001	−0.06507	−0.05481
Drinking water	−0.08069	0.00727	−11.1	<0.001	−0.09493	−0.06644

(continued)

Table 17.5 (continued)

Variables	Coefficient	Std. Err.	Z	p-value	95% CI	
Toilet	−0.02702	0.004	−6.75	<0.001	−0.03486	−0.01917
Religion	−0.09087	0.00559	−16.27	<0.001	−0.10182	−0.07993
Family size	0.05591	0.00097	57.89	<0.001	0.05402	0.0578
Household head age	−0.00229	0.00016	−14.18	<0.001	−0.0026	−0.00197
Wealth index	−0.04076	0.00167	−24.44	<0.001	−0.04403	−0.03749
Age at first birth	−0.03746	0.0006	−62.86	<0.001	−0.03862	−0.03629
Contraceptive	0.00175	0.00018	9.57	<0.001	0.00139	0.00211
Body mass index	−0.00079	0.00222	−0.36	0.722	−0.00001	0.00036
Husband education	−0.02217	0.00227	−9.77	<0.001	−0.02662	−0.01772
Husband profession	0.00002	0.0001	0.19	0.849	−0.00017	0.00021
Constant	0.587	0.01816	32.32	<0.001	0.5514	0.6226
Component3						
Age	0.03833	0.00067	57.53	<0.001	0.03702	0.03964
Division	0.00088	0.00222	0.4	0.692	−0.00347	0.00523
Residence	0.04797	0.009	5.33	<0.001	0.03033	0.06562
Education	−0.05997	0.00586	−10.24	<0.001	−0.07146	−0.04849
Drinking water	−0.08079	0.01489	−5.43	<0.001	−0.10997	−0.05161
Toilet	−0.02707	0.00961	−2.82	0.005	−0.04591	−0.00823
Religion	−0.09092	0.01074	−8.47	<0.001	−0.11197	−0.06987
Family size	0.05324	0.00124	42.82	<0.001	0.0508	0.05567
Household head age	−0.00229	0.00037	−6.26	<0.001	−0.00301	−0.00157
Wealth index	−0.04077	0.00431	−9.46	<0.001	−0.04922	−0.03232
Age at first birth	−0.03746	0.00134	−27.92	<0.001	−0.04009	−0.03483
Contraceptive	0.00175	0.00053	3.31	0.001	0.00071	0.00278
Body mass index	−0.00081	0.00403	−0.2	0.84	−0.00001	0.00001
Husband education	−0.0222	0.00539	−4.12	<0.001	−0.03277	−0.01163
Husband profession	0.00002	0.00027	0.06	0.95	−0.00051	0.00054
Constant	0.6163	0.03481	17.7	<0.001	0.54807	0.68453

are negatively associated with the number of ever born children. Respondent's education and their partner's education labels categorize as no education; primary; secondary; and higher secondary, the results depict that as the education level increase the number of children declines. Similarly, if the respondents drink unsafe water and do not have better toilet facilities can influence to have more children. Moreover, the inverse relationship is observed between the number of children of respondents and the age of the first birth. Another conclusion can be drawn through the respondent's religion if respondents follow Islam may have declined the number of children.

Finally, Fig. 17.2 illustrates the histogram of the predicted and actual number of children. From the figure, it can be observed that there is a similarity between the actual and predicted number of ever born children however, there is an existing gap between the predicted and actual number of children.

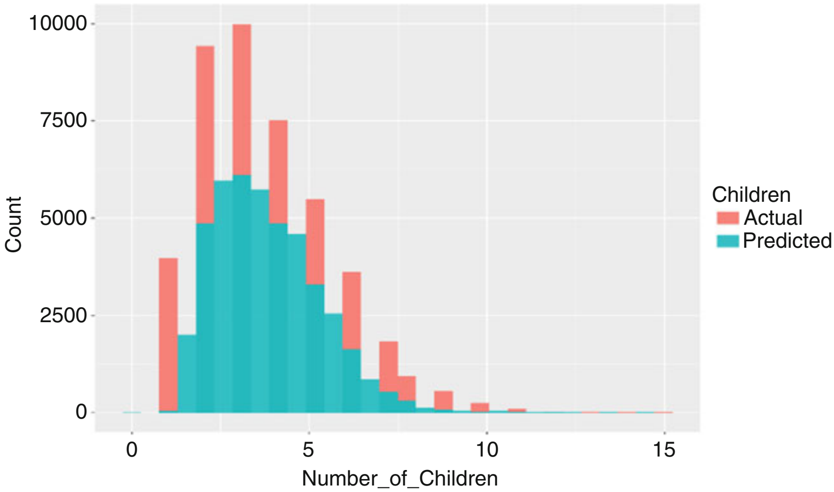


Fig. 17.2 Comparison of the actual and predicted number of children by FMNB-3 fitted model

17.4 Conclusion

The previous literature is taking as the basis of considering the influential factors of the number of ever born children in Bangladesh. Most of the researchers take their decision by using standard Poisson and negative binomial count regression approaches. However, very few researchers comparing the best-fitted models for predicting the ever born children. Furthermore, there is no literature exist of using a finite mixture model to predict the ever born children of Bangladesh. From the descriptive statistics like summary statistics and box-plots are used as an implication to have a better idea about the regressor. Finally, GLM Poisson and negative binomial regression; the COM-Poisson regression; Finite Mixture (FM) Poisson and negative binomial regression were compared with the help of AIC, BIC, and Log-likelihood values and the results suggest that the FM Negative Binomial as the most suitable model.

Again, the authors make a comparison among different component base Finite Mixture Negative Binomial Model and it is observed that the 3- component base Finite Mixture Negative Binomial Regression as the best-fitted model. After that, the ever born children was predicted through the best-fitted regression model. The results depict that respondents age, residential status, family size and intention of using contraception have a significant positive impact, however, respondent's education, drinking water, toilet facility, religious status, household head age, wealth index, age at first birth, and husband's education shows negative influence on the number of ever born children. The author of this paper thinks that the outcome of this paper will be helpful to the researchers who are concern about demographic research.

References

1. United Nations Development Programme.: Human Development Report 2015: work for human development. UN. http://hdr.undp.org/en/content/human-development-report-2015-work-human-development (2016). Accessed on 15 Mar 2019
2. Sayem, A.M., Nury, A.T.M.S., Hossain, M.D.: Achieving the millennium development goal for under-five mortality in Bangladesh: current status and lessons for issues and challenges for further improvements. J. Health Popul. Nutr. **29**(2), 92–102 (2011)
3. WB Homepage.: Fertility rate, total (births per woman) | Data. https://data.worldbank.org/indicator/SP.DYN.TFRT.IN?locations=BD. Accessed on 4 May 2019
4. Our World in Data Homepage.: Fertility rate. https://ourworldindata.org/fertility-rate. Accessed on 16 May 2019
5. Kiser, H., Hossain, M.A.: Estimation of number of ever born children using zero truncated count model: evidence from Bangladesh demographic and health survey. Health Inform. Sci. Syst. **7**(1), 3 (2019)
6. Satyavada, A., Adamchak, D.J.: Determinants of current use of contraception and children ever born in Nepal. Soc. Biol. **47**(1–2), 51–60 (2000)
7. Matsumoto, Y., Yamabe, S.: Family size preference and factors affecting the fertility rate in Hyogo, Japan. Reprod. Health. **10**, 6 (2013)
8. Haque, A., Hossain, T., Nasser, M.: Predicting the number of children ever born using logistic Regression model. Biometr. Biostat. Int. J. **2**(4), 00034 (2015)
9. Uddin, I., Bhuyan, K.C., Islam, S.S.: Determinants of desired family size and children ever born in Bangladesh. J. Fam. Welf. **57**(2), 39–47 (2011)
10. Matin, K.A..: The demographic dividend in Bangladesh: an illustrative study. In: 18th Biennial Conference of the Bangladesh Economic Association Proceedings, pp. 1–20. Dhaka, Bangladesh (2012)
11. Population Pyramids Homepage.: Population Pyramids of the World from 1950 to 2100. https://populationpyramid.net/bangladesh/2050/. Accessed 4 May 2019
12. Dwivedi, V.K., Sediadie, T., Ama, N.O.: Factors affecting children ever born (CEB) in Botswana: application of Poisson Regression model. Res. J. Math. Stat. Sci. **4**(10), 1–9 (2016)
13. Mekonnen, W., Worku, A.: Determinants of fertility in rural Ethiopia: the case of Butajira Demographic Surveillance System (DSS). BMC Publ. Health. **11**, 782 (2011)
14. Kareem, Y.O., Yusuf, O.B.: Statistical modeling of fertility experience among women of reproductive age in Nigeria. Int. J. Stat. Appl. **8**(1), 23–33 (2018)
15. Ahmed, H.M.M., ALI, H.M.H.: Using count Regression models to determinate factors influencing fertility of Sudanese women. J. Glob. Econ. Manag. Bus. Res. **6**(4), 277–285 (2015)
16. Mullahy, J.: Specification and testing of some modified count data models. J. Econ. **33**(3), 341–365 (1986)
17. Lambert, D.: Zero-inflated Poisson Regression, with an application to defects in manufacturing. Technometrics. **34**(1), 1–14 (1992)
18. Poston, D.L., McKibben, S.L.: Using zero-inflated count Regression models to estimate the fertility of U. S. Women. J. Mod. Appl. Stat. Methods. **2**(2), 371–379 (2003)
19. Islam, M.R., Islam, M.R., Alam, M.R., Hossain, M.M.: Affecting socio-demographic factors on children ever born for women who have experienced domestic violence and women who have not experienced domestic violence in Bangladesh. Am. J. Sociolog. Res. **2**(5), 113–119 (2012)
20. Ahmmed, F., Naser, M.: Modeling and predicting of children ever born in Bangladesh. In: International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment Proceedings, pp. 1–10. Rajshahi, Bangladesh (2012)
21. Sellers, K.F., Borle, S., Shmueli, G.: The COM-Poisson model for count data: a survey of methods and applications. Appl. Stoch. Model. Bus. Ind. **28**(2), 104–116 (2012)

22. Bonat, W.H., Zeviani, W.M., Ribeiro Jr, E.E.: Regression Models for Count Data: Beyond the Poisson model. Statistics and Geoinformation Laboratory Federal University of Parana, Brazil, http://cursos.leg.ufpr.br/rmcd/rmcdbook.pdf (2017), Accessed on 12 Apr 2019
23. Park, B.-J., Lord, D.: Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. **41**(4), 683–691 (2009)
24. McLachlan, G.J., Lee, S.X., Rathnayake, S.I.: Finite mixture models. Ann. Rev. Stat. Appl. **6**(1), 355–378 (2019)
25. Stata Finite Mixture Models Reference Manual Release 16. Stata Press, Texas, 2019
26. Dalrymple, M.L., Hudson, I.L., Ford, R.P.K.: Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. Comput. Stat. Data Anal. **41**(3–4), 491–504 (2003)
27. Land, K.C., McCall, P.L., Nagin, D.S.: A comparison of Poisson, negative binomial, and semiparametric mixed Poisson Regression models: with empirical applications to criminal careers data. Sociol. Method. Res. **24**(4), 387–442 (1996)
28. Bermúdez, L., Karlis, D.: A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. Comput. Stat. Data Anal. **56**(12), 3988–3999 (2012)
29. Park, B.-J., Lord, D., Wu, L.: Finite mixture modeling approach for developing crash modification factors in highway safety analysis. Accid. Anal. Prev. **97**(74–287), 274–287 (2016)
30. Azagba, S., Wolfson, M.: E-cigarette use and quantity of cigarette smoking among adolescent cigarette smokers: a finite mixture model analysis. Drug Alcohol Depend. **185**, 33–39 (2018)
31. National Institute of Population Research and Training (NIPORT).: Bangladesh Demographic and Health Survey 2014. NIPORT, Mitra and Associates, and ICF International, Bangladesh, Dhaka. https://dhsprogram.com/pubs/pdf/FR311/FR311.pdf accessed on 2018/06/25 (2016)
32. Sahle, G.: Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia. Health Inform. Sci. Syst. **4**(1), 4 (2016)
33. Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press, Cambridge (2013)
34. Poisson Regression, Chapter 325, NCSS Statistical Software. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf. Accessed 10 Mar 2019
35. Hilbe, J.M.: Modeling Count Data. Cambridge University Press, Cambridge (2014)
36. Negative Binomial Regression, Chapter 326, NCSS Statistical Software. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf. Accessed 11 Mar 2019
37. Dávila, V.H.L., Cabral, C.R.B., Zeller, C.B.: Finite Mixture of Skewed Distributions. Springer, Cham (2018)
38. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
39. Sellers, K.F., Shmueli, G.: Predicting Censored Count Data with COM-Poisson Regression. Robert H. Smith School Research Paper No. RHS-06-129, (2010). https://ssrn.com/abstract=1702845. Accessed on 17 Mar 2019
40. Cameron, A., Trivedi, P.: Generalized Count Regression. In Regression Analysis of Count Data (Econometric Society Monographs, pp. 111–176). Cambridge University Press, UK, Cambridge (2013)

Chapter 18
An Assessment of Influencing Factors for Motherhood During Childhood in Bangladesh Using Factor Analysis and Logistic Regression Methods

Mohammad Salim Zahangir and Mosammat Zamilun Nahar

Abstract Though Bangladesh has achieved great success in family planning as well as maternal and child sector nowadays, it still needs further improvement. This study deals with popular phenomenon motherhood in childhood and its influential factors in Bangladesh. Data are obtained from the 2014 Bangladesh Demographic and Health Survey (BDHS). It can be seen that 62.1% of women age 18 or below become a mother or pregnant in childhood. The relationship between factors obtained by factor analysis and motherhood in childhood is assessed by both linear discriminant and logistic regression analyses. The covariates that are found to be significant by the χ^2-test are also analysed by binary logistic regression technique for examining their effects on childbearing in childhood. The analysis reveals that respondent's education, husband's education and age at first marriage are significantly negatively associated and respondent's current age is significantly positively associated with the chance of becoming a mother in childhood. Region, wealth index, husband's occupation and husband's age are also significant to some extent. In conclusion, the prevalence of motherhood in childhood can be reduced by educating women upto secondary or higher levels, alleviating poverty and limiting the provisions of early marriage.

Keywords Motherhood in childhood · Factor analysis · Binary logistic regression

18.1 Introduction

The United Nations convention on the rights of the child defines a "child" as a person under the age of 18 unless the laws of a specific country set the legal age of adulthood at a younger age. Childhood or adolescence (10–19 years) is a period

M. S. Zahangir (✉) · M. Z. Nahar
Department of Statistics, University of Chittagong, Chattogram, Bangladesh
e-mail: salim.zahangir@cu.ac.bd

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_18

that has special importance in the course of an individual's life [1]. In fact, it is the time to gain knowledge and build up career to finding life projects. However, motherhood in childhood is a huge global problem, especially in developing countries. There is a concern that forming a family at younger ages limits women's future opportunities [2]. Due to early uptake of responsibilities of motherhood and childcare, women do not get the opportunity to develop themselves; as a result, the quality of women's life is reduced. Moreover, teenagers are not psychologically and physically mature to bear the burden of childbearing and rearing. Complications from pregnancy or childbirth are the second cause of death for women of ages 15–19 years globally [3].

The State of World Population 2013 reports that most (95%) of the world's births to adolescents occur in developing countries, and nine in ten of these births occur within marriage or a union [4]. One-fifth of women (about 19%) in developing countries becomes pregnant before reaching the age of 18. Of the 7.3 million adolescent women below 18 years who give birth every year in developing countries, 2 million are under the age of 15. Child marriage, an inferior category of age at first marriage—a demographic factor, is the main reason of early pregnancy/motherhood. Various socio-cultural and economic (e.g., illiteracy or lower level of education, poverty and lack of access to reproductive health), and security (e.g., sexual violence) factors are also responsible for early pregnancy.

Early entry into motherhood can have various immediate and long-lasting consequences, especially on maternal and child health. For instance, early pregnant women are at increased risk of toxaemia, anaemia, excessive bleeding, prolonged labor-pain, premature delivery and higher levels of blood pressure [5], higher rate of low birth weight, stillbirth and perinatal mortality [6, 7]. The ultimate consequences are morbidity and mortality of women during pregnancy and delivery [5]. In developing countries, the risk of maternal death for women under 15 is double of women age 15 or older. This younger group also suffers notably higher rates of obstetric fistulae than their older peers. Beside health-related consequences, early pregnancy may have negative social and economic impacts on women's individual, familial and communal levels [3]. Researchers found a positive relationship between lower educational attainment and early childbearing [8–10]. Teenagers who become pregnant have drop out of school. Girls with little or no education have fewer skills and opportunities to get a job. This can also have an economic value with a country losing out on the annual income a teen girl would have earned during her lifetime, if she had not had an early pregnancy [3].

Adolescent births are declining globally but are increasing in three regions such as South Asia, sub-Saharan Africa, and Latin America. Bangladesh is often characterized by early marriage [11–13] and early motherhood [14]. In fact, early marriage and early motherhood are traditions in the patriarchal Bangladeshi culture [15]. Over 82% of women become mothers by age 19 [16]. Apparently, age at first birth is much lower in Bangladesh [16, 17]. According to the State of World Population 2013 report, about 17% of Bangladeshi women get married before reaching 15, while most of them give birth to two children before reaching 18 [4]. Considering the rate of adolescent pregnancy, Bangladesh ranks third in the world and highest in the Asia. The British Medical Journal reports that one in three women aged 15–19 becomes a mother or pregnant in Bangladesh today, with adolescent mothers are more likely to suffer birth complications than adult women [18]. Save the Children [14] reported that women who become mothers by age 14 face the greatest risks. According to Early Motherhood Risk Ranking, Bangladesh is the first in South Asia and 13th in the world. With respect to fertility, early childbearing is related to a higher completed fertility [19]. Suchitta [20] claimed that early motherhood is a major constraint to economic development and social progress of Bangladesh.

There is a large body of research on adolescent motherhood in the global context [4, 21–24]. All those researches showed that lower socioeconomic attainments, early age of marriage, lower contraceptive use and higher spousal age difference are the key determinants of adolescent motherhood. There also have several studies that investigated factors affecting adolescent fertility in Bangladesh [15, 25–29]. Among recent studies, Islam et al. [29] observed trends and determinants of adolescent motherhood in Bangladesh using all seven DHS datasets (1993/94 to 2014). The authors found that current age, respondent's education, husband's education, place of residence, region of residence, exposure to mass media and spousal age gap are important determinants of adolescent fertility. Kamal [15] analyzed the 2007 BDHS data. The author observed with almost all the above-mentioned factors that wealth index and ever used any contraceptive are significantly associated with adolescent fertility in Bangladesh. Moreover, using the four different BDHS datasets (1993–2004), Nahar and Min [28] showed that women's higher education, delayed age at first marriage and media contact exerted strong influence in delay of having babies among adolescent women in Bangladesh. They also noticed significant variations across divisions in having the first birth of women age 19 or earlier.

It should be mentioned that in almost everywhere, especially in developing regions, teenage pregnancy has been a source of worry for policy makers, social workers and other human service providers due to its negative consequences to the girl-child [30]. Taking the context "understanding girl-child childbearing" seriously into account, little attention has been paid to study the context. The aim of this study is to provide some data and ideas for the design of specific interventions that are essentials of these girls. Thus, this is an attempt to investigate the socio-cultural, economic and demographic factors and their effects on motherhood in childhood among Bangladeshi women who are 18 years of younger or below 18.

18.2 Data and Methodology

18.2.1 The Data

This study uses the latest Bangladesh Demographic and Health Survey (BDHS) data conducted in 2014. The survey is based on a two-stage stratified sampling technique. In the first stage, probability proportional to size sampling and in the second stage, systematic sampling techniques are applied. Among a total of 17,989 selected households throughout the country, 17,565 were occupied. Interviews were successfully completed in 17,300 households, in which a total of 18,245 ever-married women aged 15–49 years were identified and finally 17,863 were interviewed. The survey was designed to provide information on the basic national indicators of social progress, including fertility, fertility preferences and regulation, childhood mortality and causes of death, maternal and child health, nutrition status of mothers and children, awareness and attitudes towards HIV/AIDS, and prevalence of non-communicable diseases. Information on the timing of the first birth was collected from women in the standard DHS individual module. The detailed descriptions of the survey are given in the report book (see, [31]).

18.2.2 Unit of Analysis

While 18 is the legal age at marriage of women in Bangladesh, a large majority (58.1%; source: BDHS 2014) of women aged 15–49 become mothers/pregnant by age 18. In the case of considering adolescent women, there may have a chance to include some of those who marry within 18–19 years of age. According to the United Nations' definition of child, "motherhood in childhood" means a woman becomes a mother by her 18th year birthday. Since births are confined to marriage in Bangladeshi society, this study considers the ever-married women who are 18 years of younger or below 18. It should be noticed that "age" is computed in completed years by the BDHS surveys. For instance, age 18 means completed 18 years. Upto age 18, there are total 1460 ever-married women in the BDHS 2014 data. Thus, the unit of analysis in this study is 1460 respondents.

18.2.3 Variables

18.2.3.1 Dependent Variable

"Motherhood in childhood"—a dichotomous variable, is considered the dependent variable in this study.

18.2.3.2 Independent Variables

In choosing independent variables or covariates, this study relies on previous literature on early motherhood in the global context, mainly in Bangladesh. Moreover, according to availability of data, total 13 covariates on socio-cultural, economic and demographic issues are chosen here. The list of covariates and their suitable categories are presented in Table 18.1.

Table 18.1 Percentage of childbearing status among ever-married women aged 15–18 by socioeconomic and demographic characteristics, BDHS 2014

Covariates	Category	No. of respondents	% of starting childbearing	Chi-square test	
				Value	P-value
Type of place of residence	Urban	440 (30.1)	61.1	0.226	0.635
	Rural	1020 (69.9)	62.5		
Current age	15 years	291 (19.9)	40.0	75.360	0.000
	16 years	431 (29.5)	52.9		
	17 years	548 (37.5)	65.0		
	18 years	291 (19.9)	72.3		
Respondent's education	Illiterate & primary[a]	447 (30.6)	71.4	23.710	0.000
	Secondary & above	1013 (69.4)	57.9		
Husband's education	Illiterate	209 (14.3)	76.1	52.070	0.000
	Primary	479 (32.8)	69.5		
	Secondary & above	772 (52.9)	53.6		
Region	Barisal	210 (14.4)	59.0	13.670	0.034
	Chittagong	262 (17.9)	64.9		
	Dhaka	234 (16.0)	58.1		
	Khulna	209 (14.3)	57.9		
	Rajshahi	184 (12.6)	62.5		
	Rangpur	218 (14.9)	61.5		
	Sylhet	143 (9.8)	74.1		
Access to mass media[b]	No	543 (37.2)	67.6	11.240	0.001
	Yes	917 (62.8)	58.8		
Religion	Muslims	1360 (93.2)	61.6	1.610	0.204
	Others	100 (6.8)	68.0		
Wealth index[c]	Poor	934 (64.0)	65.4	14.220	0.001
	Middle	335 (22.9)	58.2		
	Rich	191 (13.1)	52.4		
Respondent's working status	Not working	1252 (85.8)	62.1	0.027	0.868
	Working	208 (14.2)	61.5		
Husband's occupation	Agriculture	156 (10.7)	62.8	6.800	0.078
	Business	301 (20.6)	67.1		
	Service	39 (2.7)	48.7		
	Others	964 (66.0)	60.9		
Age at first marriage	10–14 years	579 (39.7)	72.9	47.790	0.000
	15–18 years	881 (60.3)	54.9		
Ever used any contraception	No	433 (29.7)	58.9	2.616	0.106
	Yes	1027 (70.3)	63.4		

(continued)

Table 18.1 (continued)

Covariates	Category	No. of respondents	% of starting childbearing	Chi-square test	
				Value	P-value
Husband's age[d]	15–22 years	384 (26.3)	52.6	33.040	0.000
	23–25 years	447 (30.6)	60.4		
	26–55 years	596 (40.8)	70.5		
	Missing	33 (2.3)			
Total		1460 (100)	62.1		

Figures in parentheses indicate percentage of respondents
[a] Illiterate women and women with primary education are added together as individually they are rather inadequate in size
[b] Access to mass media is obtained by combining three variables: frequencies of listening to the radio, watching TV and reading newspapers at least once in a week
[c] Wealth index is a composite measure of a household's cumulative living standard. The index is constructed and provided by DHS, Macro International
[d] Husband's age is used to investigate whether it has the similar relationship like respondent's age with childhood motherhood

18.2.4 Methods

This study deals with ever-married Bangladeshi women to observe the recent and the most recent past situations of motherhood status. This section would help to understand the reproductive behavior among women in childhood. Both bivariate and multivariate statistical methods are employed to examine the relationship between motherhood in childhood and its underlying factors. In order to evaluate individual factors and to extract significant factors of motherhood in childhood, factor analysis based on principal components is used as an alternative approach. Linear discriminant analysis is employed using the factor scores derived from the factor analysis to determine as to how correctly women are classified into different motherhood status. Binary logistic regression method is also applied, uses the predicted probabilities to assign cases into the categories of the dependent variable and then compares the results with their actual categories. The method can also be used to explain the effect of covariates on the dependent variable.

Based on the dependent variable, the logistic regression model is briefly described here. Let Y represents childbearing status in childhood. It is categorized as "1" for ever-married women age 18 or below who become mothers/pregnant by age 18 and "0" for those belonging in the opposite group. Let $x_1, x_2, \ldots, x_k$ be the explanatory variables. If $P(Y = 1) = p$ is the probability of becoming a mother or pregnant by age 18, then logistic regression model can be expressed as:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

$$or, \quad p = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}.$$

The above expression is used to predict the childhood status using a cut value or threshold probability of 0.5.

18.3 Results and Discussions

Figure 18.1 represents the percentage of ever-married women age 18 or below, giving first birth (in single years) by age 18. It shows that childbearing starts very early (at age 12) in Bangladesh. Of all childhood mothers, the highest 25.3% become a mother at age 16. Table 18.1 shows that total 62.1% of women age 18 or below become a mother/pregnant in childhood. However, about 73% of them are intended to give birth in childhood and the remaining are mistimed (see Table 18.2).

The proportion of women who give birth by age 18 is varied across categories of each covariate. Here, it is assumed that each of selected covariate is associated with

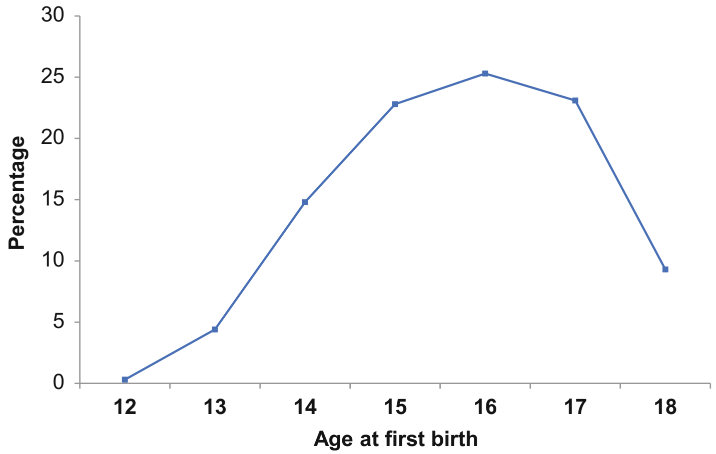


Fig. 18.1 Percentage of ever-married women age 18 or below by their age at first birth, BDHS 2014

Table 18.2 Percentage of ever-married women age 18 or below giving at least one birth in the last five years preceding the survey by their intention status, BDHS 2014

Birth intention[e]	Frequency	Percent
Intended	491	72.6
Mistimed	185	27.4
Total	676	100

[e] All eligible women were asked whether each birth in the 5-year preceding the survey was planned (wanted then), mistimed (wanted but later), or unwanted. The pregnancies reported as wanted later or wanted no more is termed as "unintended" or "mistimed"

motherhood in childhood. The χ^2-test for independence of attributes is employed to verify the assumption; findings are shown in Table 18.1. Of all 13 covariates, type of place of residence, religion, respondent's working status and ever used any contraception has no significant association with motherhood in childhood.

Current age is found to be positively associated with motherhood in childhood in Bangladesh. Nguyen [24] has found the similar relationship in the study of adolescent pregnancy in Vietnam. In Table 18.1, the percentage of becoming a mother in childhood is the highest for women of age 18 (72.3%) and the lowest for women of age 15 (40.0%). Female education is inversely associated with childbearing in childhood. More than seven-tenths (71.4%) of illiterates or primary school graduates start childbearing by age 18, and that of secondary or higher school graduates is about 58%. The percentage of women with illiterate husbands becoming a mother in childhood is the highest (76.1%), followed by those with primary school graduated (about 70%) and secondary or higher school graduated (53.6%) husbands, respectively. Kamal [15] has also found the similar findings in studying adolescent fertility in Bangladesh.

Age at first marriage is also negatively associated with childbearing in childhood. Women age 18 or below marrying within ages 10–14, 72.9% of them become a mother by age 18, and that of those marrying within ages 15–18 is only 54.9%. Haque [26] has also found a decreasing pattern in adolescent childbearing with age at first marriage. Husband's age is positively associated with motherhood in childhood. Of women age 18 or below whose husbands are 26 years or older, 70.5% become a mother in childhood, whereas those whose husbands are 22 years or younger, 52.6% of them become a mother in childhood.

Now, to check the strength of covariates in relation to motherhood in childhood factor analysis based on principal components is applied. In the principal component analysis, orthogonal rotation (varimax option) is used to obtain non-correlated factors [32] and the Kaiser eigenvalue 1 criterion is used for choosing the number of factors. Adequacy of factors and the homogeneity of variance across variables for motherhood in childhood are checked by Kaiser-Meyer-Olkin's (KMO) and Bartlett's tests. Factor loadings greater than 0.5 and cross-loading (loadings with a negative value) less than -0.4 are considered to explain the relationship between the covariates and factors [33].

The scree plot of factor analysis for motherhood in childhood among women upto age 18 is displayed in Fig. 18.2. The plot shows that every eigenvalue after the sixth component is less than 1 and therefore, the first six factors will be extracted for motherhood in childhood among women age 18 or below. These six factors are computed by factor analysis that provides two more outputs: total variance explained by factors associated with motherhood in childhood, presented in Table 18.3, and rotated component matrix for motherhood in childhood, shown in Table 18.4. Among 13 covariates, six extracted factors explain the variation of motherhood in childhood by 16.55%, 11.06%, 10.43%, 8.88%, 8.34% and 7.77%, respectively and the cumulative percent of the variance of those six components is 63.02% (see Table 18.3).

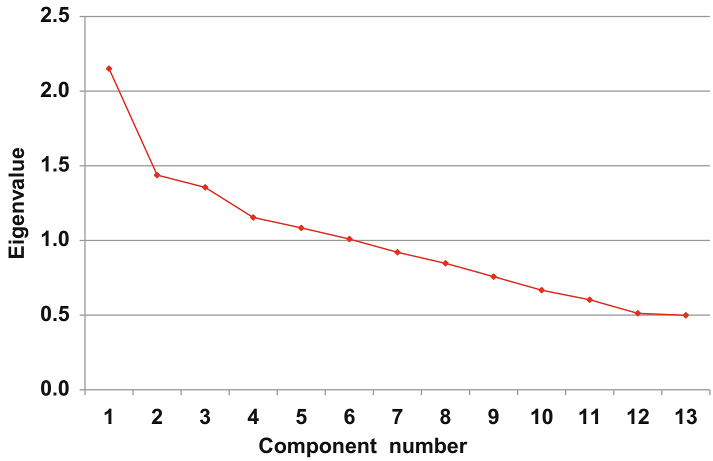


Fig. 18.2 Scree plot: Eigenvalues by factors influencing motherhood in childhood, BDHS 2014

Table 18.3 Total variation explained by factors related to starting childbearing in childhood, BDHS 2014

	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
Comp.	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.15	16.55	16.55	2.15	16.55	16.55	1.75	13.47	13.47
2	1.44	11.06	27.61	1.44	11.06	27.61	1.59	12.22	25.69
3	1.36	10.43	38.04	1.36	10.43	38.04	1.43	10.99	36.67
4	1.15	8.88	46.92	1.15	8.88	46.92	1.17	9.02	45.69
5	1.08	8.34	55.25	1.08	8.34	55.25	1.17	9.00	54.69
6	1.01	7.77	63.02	1.01	7.77	63.02	1.08	8.34	63.02

Extraction method: Principal component analysis. Comp. means component

The rotated component matrix represents that type of place of residence, access to mass media and wealth index are highly correlated with the first principal component. The second principal component is found to have a high correlation with respondent's education and husband's education. The third, fourth, fifth and sixth principal components are highly correlated with respondent's current age and age at first marriage, husband's occupation and husband's age, respondent's working status and ever used any contraception, and region and religion, respectively. Beyond the significance of covariates, respondent's education and type of place of residence are strong determinants of childhood in motherhood as they are highly correlated with only one component and less correlated with the other two components.

Linear discriminant analysis is used to check the validity of findings from factor analysis. To compare the predictive power of this model, binary logistic regression is also used. The relationship between factors obtained by factor analysis

Table 18.4 Rotated component matrix for starting childbearing in childhood, BDHS 2014

Variables	Component 1	2	3	4	5	6
Respondent's current age	0.081	−0.034	0.779	0.257	0.154	0.053
Type of place of residence	−0.760	0.113	−0.012	0.088	−0.091	−0.005
Division	0.101	−0.401	−0.157	0.139	−0.397	0.561
Respondent's education	0.048	0.746	−0.003	0.094	−0.029	−0.035
Religion	−0.085	0.054	0.150	−0.080	0.183	0.825
Access to mass media	0.674	0.289	−0.011	−0.033	0.027	0.066
Wealth index	0.774	0.180	0.069	0.102	−0.005	−0.126
Respondent's occupation	0.074	−0.241	0.058	−0.062	0.666	0.008
Husband's education	0.263	0.730	0.090	0.013	−0.07	−0.004
Husband's occupation	0.114	−0.211	0.171	−0.656	0.229	−0.128
Age at first marriage	−0.005	0.165	0.762	−0.181	−0.264	0.061
Husband's age	0.060	−0.053	0.214	0.755	0.171	−0.143
Ever used any contraception	0.059	0.268	−0.319	0.119	0.588	0.152

Extraction method: Principal component analysis
Rotation method: Varimax with Kaiser normalization
Rotation converged in 14 iterations

Table 18.5 Classification of individual woman using scores derived from factor analysis for motherhood in childhood

Starting childbearing: Actual group	Predicted group by Discriminant analysis: No	Yes	Classification rate (%)	Logistic regression: No	Yes	Classification rate (%)
No	66.8 (370)	33.2 (184)	63.6	36.1 (200)	63.9 (354)	66.6
Yes	38.4 (348)	61.6 (558)		14.7 (133)	85.3 (773)	

Figures in parentheses indicate number of respondents

and childbearing status in childhood is assessed by both linear discriminant and binary logistic regression analyses. In both models, childbearing in childhood is the dependent variable and factor scores obtained by the factor analysis are the independent variables. The results of the two models are shown in Table 18.5. It shows that the discriminant analysis classifies 66.8% for women who are not mothers and 61.6% for those who are currently mothers. The overall correct classification rate is 63.6%. About 33.2% of women who do not start childbearing are predicted as a mother, while 38.4% who are not mothers or pregnant, are predicted as a mother/pregnant. These results imply that there is a significant relationship between 13 explanatory variables and motherhood status during childhood. In the logistic regression analysis, 36.1% of non-mothers are correctly classified and 85.3% among mothers. The overall correct classification rate is 66.6%. It also shows that logistic regression provides somewhat better results than discriminant analysis. In this analysis, discriminant analysis can classify 370 respondents as a good group out of 554 good respondents and 558 respondents as a bad group out of 906 bad respondents. Table 18.6 represents that discriminant analysis obtains a large amount of costs through "bad accepted (348)" rather than logistic regression (133). So, logistic regression analysis achieves less cost of misclassification.

The binary logistic regression technique is applied to examine the effect of covariates on childbearing in childhood. All four covariates that are found to be insignificant by the χ^2-test are excluded from the logistic regression analysis. The regression coefficients, odds ratios and their respective 95% confidence intervals (CI) are presented in Table 18.7. The odds ratio indicates the relative chance of the other groups in relation to specific baseline group by the exponent of the regression coefficient, Exp(β). Values greater than 1 indicate that the chance of becoming a mother or pregnant in childhood is higher for this group compared to the reference group, and vice versa.

There is a strong positive association between the current age and the onset of childbearing in childhood. Women age 16, 17 and 18 are 2.59 (CI = 1.69–3.96), 6.80 (CI = 4.36–10.59) and 11.86 (CI = 7.51–18.72) times more likely to be a mother in childhood than the reference group (e.g., women age 15). This may be because women of age 15 do not have physical maturity like their older peers. Female education is inversely associated with motherhood in childhood. Illiterate or primary educated women are 1.37 (CI = 1.02–1.85) times more likely to give birth in childhood than those who are secondary or higher educated. Moreover, women with illiterate and primary school graduated husbands are 2.28 (CI = 1.46–3.57) and 1.76 (CI = 1.31–2.36) times more likely to start childbearing in childhood with respect to those with secondary or higher school graduated husbands. Focusing on mass media, it shows that women who have no exposure to any media are 1.14 (CI = 0.85–1.52) times more likely to be a mother in childhood than those who have any exposure to that. It is partly due to fact that women can be conscious about the adverse effects of early motherhood through mass media. According to region of residence, women who live in Sylhet and Chittagong divisions are 1.78 (CI = 1.05–3.01) and 1.56 (CI = 1.03–2.36) times more likely to be a mother in childhood than those residing in Dhaka division. A significantly higher chance of becoming a mother in childhood is found among poor women (OR: 1.69; CI = 1.13–2.51) than those who are economically rich. In case of husband's occupation, the chance of becoming a mother in childhood is significantly higher among women with businessmen husbands (OR: 2.19; CI = 1.02–4.68) than those with service-holder husbands.

Table 18.6 Predictive models comparison

Models	Good news		Bad news		Success rate
	Accepted	Rejected	Accepted	Rejected	
Discriminant analysis	370	184	558	348	63.6%
Logistic regression	200	354	773	133	66.6%

Table 18.7 Binary logistic regression coefficients and odds ratios of starting childbearing by selected covariates among ever-married women aged 15–18, BDHS-2014

Covariates	Category	β	P-value	Exp(β)	95% CI for Exp(β)
Current age	15 years			RC	
	16 years	0.950	0.000	2.585	(1.686, 3.962)
	17 years	1.917	0.000	6.797	(4.363, 10.589)
	18 years	2.473	0.000	11.855	(7.508, 18.717)
Respondent's education	Illiterate & primary	0.317	0.036	1.373	(1.021, 1.845)
	Secondary & above			RC	
Husband's education	Illiterate	0.826	0.000	2.284	(1.461, 3.571)
	Primary	0.565	0.000	1.759	(1.313, 2.357)
	Secondary & above			RC	
Region	Barisal	0.030	0.896	1.031	(0.655, 1.622)
	Chittagong	0.443	0.037	1.557	(1.028, 2.357)
	Dhaka			RC	
	Khulna	0.166	0.461	1.181	(0.759, 1839)
	Rajshahi	0.029	0.900	1.030	(0.650, 1.632)
	Rangpur	0.117	0.612	1.124	(0.715, 1.769)
	Sylhet	0.574	0.033	1.775	(1.048, 3.007)
Wealth index	Poor	0.522	0.010	1.686	(1.132, 2.511)
	Middle	0.262	0.204	1.300	(0.868, 1.946)
	Rich			RC	
Access to mass media	No	0.132	0.372	1.141	(0.854, 1.524)
	Yes			RC	
Husband's occupation	Agriculture	0.457	0.273	1.580	(0.698, 3.578)
	Business	0.783	0.043	2.188	(1.024, 4.677)
	Service			RC	
	Others	0.424	0.258	1.528	(0.733, 3.189)
Age at first marriage	≤14 years	1.636	0.000	5.133	(3.732, 7.059)
	15–18 years			RC	
Husband's age	15–22 years			RC	
	23–25 years	0.016	0.919	1.016	(0.743, 1.390)
	26–55 years	0.475	0.003	1.608	(1.180, 2.191)
Constant		−3.387	0.000	0.034	

RC means reference category (A category of each variable, which has the lowest proportion of starting childbearing in childhood, is considered as a reference category). CI means confidence interval

The chance of becoming a mother by age 18 has sharply declined with increasing age at first marriage. For instance, among women age 18 or lower marrying before age 15, are 5.13 (CI = 3.73–7.06) times more likely to be a mother in childhood than those marrying within ages 15–17. Regarding husband's age, the chance of giving a birth by age 18 is significantly higher for the women whose husbands are 26 or higher ages (OR: 1.61; CI = 1.18–2.19) with respect to those belonging in the reference category.

18.4 Conclusions and Policy Implications

This study notices that motherhood starts very early in Bangladesh. Slightly more than three-fifths (62.1%) of women age 18 or below become a mother or pregnant in childhood. The prevalence of motherhood in childhood is also perceived among all groups of women age 18 or below, controlled by covariates, especially among women who own and their husbands are illiterate and marry very early.

Of all selected covariates, age at first marriage plays the most influential role in motherhood in childhood. This means that the earlier a woman gets married, the earlier she gives birth. Since childbearing outside of marriage is illegal or unacceptable in Bangladeshi society, an increase in the consented age at first marriage should be the first and foremost agenda for reducing the number of cases of motherhood in childhood. To reduce early marriage, the legal age of marriage law should be properly implemented across regions and in rural and urban areas of Bangladesh. Female education is not strong enough like male education in influencing motherhood in childhood. This may be because males are not controlled by their age, whereas females upto age 18 are limited here. However, secondary or higher educated women are not more prone to become a mother in childhood like those who are illiterates or primarily school graduates. The similar but rather stronger inference can be drawn for women with secondary or higher educated husbands. Apparently, upto a certain level of education, especially female education is very crucial in order to raise their consciousness about adverse consequences of motherhood in childhood. In light of this fact, the government should be more concerned to increase the level with increasing the rate of education.

It is important to notice that there is much more than one-third (40.8%) of women age 18 or below whose husbands are 26 years or older. This indicates that when a woman marries very early, she has a higher chance to have a more aged husband. Due to a larger age gap between couples, child brides are often inept at negotiating safe sex with their husbands. This might be the main reason for making them more vulnerable to sexually transmitted infections, including HIV, and put them at higher risk of early pregnancy even if they have a poor BMI [4]. The reality is that they are typically ignored by or beyond the reach of national health, education, and development institutions. It is, therefore, crucial to provide user-friendly reproductive health services to young women throughout the country, especially in poor communities to enable them in order to avoid early sexual activity and ultimately early childbearing. Indeed, the government, civil society, national and international communities should do more to protect them and support their safe and healthy life. Above all, focusing on adverse consequences of motherhood in childhood, integrated social awareness programs, including campaigns from the parts of social, electronic and print media can take notable contributions in order to reduce the prevalence of motherhood in childhood in Bangladesh.

18.5 Limitations of the Study

There are some limitations in the study. Firstly, in the BDHS 2014 survey, each woman was asked a wide range of retrospective questions on birth histories of women. There were no questions about the most background characteristics with individual life histories, such as parent's education and mother's age at first birth. Many studies reveal that parent's education and mother's age at first birth are significantly associated with early childbearing of her daughter. Secondly, this study is based on women aged 10–18 years. The BDHS 2014 survey is not considered ever-married women aged 10–14 years and never married women of their age in single years. According to the BDHS 2014 report, 54.8% ($n = 4,485$) of women aged 15–19 years are never married. Hence, it is not possible to know the number of never married women of ages 15–18 years. Thirdly, the survey suffers from recall bias. Women do not report correctly about their age at first birth. It is a common phenomenon for developing countries like Bangladesh. Such incorrect reporting may bias the estimates. Despite these limitations, the BDHS is reliable as it is a nationally representative large data set and mostly used in demographic studies.

References

1. Yavuz, S.: Changes in adolescent childbearing in Morocco, Egypt and Turkey. DHS Working Papers No. 75. ICF Macro, Calverton (2010)
2. Seamark, C.J., Lings, P.: Positive experiences of teenage motherhood: a qualitative study. Br. J. Gen. Pract. **54**(508), 813–818 (2004)
3. World Health Organization.: Adolescent pregnancy, Fact sheet, Sept 2014. WHO (2014). http://www.who.int/mediacentre/factsheets/fs364/en/
4. United Nations Fund for Population.: Motherhood in childhood. Facing the challenge of adolescent pregnancy. State of world population 2013. UNFPA, New York (2013)
5. Zabin, L., Kiragu, K.: The health consequences of adolescent sexual and fertility behavior in Sub-Saharan Africa. Stud. Fam. Plan. **29**(2), 210–232 (1998)
6. Lee, M.C., Suhng, L.A., Lu T.H., Chou, M.C.: Association of parental characteristics with adverse outcomes of adolescent childbearing. Family Pract. **15**(4), 336–342 (1998)
7. Phipps, M.G., Sowers, M.F.: Defining early adolescent childbearing. Am. J. Public Health **92**(1), 125–128 (2002)
8. Branson, N., Ardington, C., Leibbrandt, M.: Trends in teenage childbearing and schooling outcomes for children born to teens in South Africa. University of Cape Town, Cape Town (2013)
9. Doyle, A.M., Mavedzenge, S.N. Plummer, M.L, Ross, D.A.: The sexual behaviour of adolescents in sub-Saharan Africa: patterns and trends from national surveys. Trop. Med. Int. Health **17**, 796–807 (2012)
10. Ozier, O.: The impact of secondary schooling in Kenya: a regression discontinuity analysis. Development Research Group, World Bank, Washington, DC (2011)
11. Zahangir, M.S., Kamal, M.M.: Several attributes linked with child marriage of females' in Bangladesh. Int. J. Stat. Syst. **6**(1), 107–117 (2011)
12. Zahangir, M.S., Karim, M.A., Zaman, M.R., Hussain, M.I., Hossain, M.S.: Determinants of age at first marriage of rural women in Bangladesh: a cohort analysis. Trends. Appl. Sci. Res. **3**(4), 335–343 (2008)

13. Zahangir, M.S.: Patterns in early and very early family formation in Bangladesh. Asian Profile **42**(2), 123–140 (2015)
14. Save the Children.: State of the world's mothers: 2004, Westport (2004)
15. Kamal, S.M.M.: Adolescent motherhood in Bangladesh: evidence from 2007 BDHS data. Can. Stud. Popul. **39**(1–2), 63–82 (2012)
16. Nahar, M.Z., Zahangir, M.S.: Patterns and determinants of age at first birth in Bangladesh. Turk. J. Popul. Stud. **35**, 63–77 (2015)
17. Bosch, A.M., Willekens, F.J., Baqui, A., Ginneken, J.K. van, Hutter, I.: Association between age at menerche and early life nutritional status in rural Bangladesh. J. Biosoc. Sci. **40**(2), 223–227 (2008)
18. Akhter, S.: Complications of adolescent pregnancy and its prevention. The Daily Star Report. 28 July (2013). www.thedailystar.net/news/complications-of-adolescent-pregnancy-and-its-prevention
19. National Institute of Population Research and Training, Mitra and Associates, MEASURE DHS.: Bangladesh Demographic and Health Survey 2011, Dhaka, Bangladesh, and ICF International, Calverton, Maryland: NIPORT, Mitra and Associates, and MEASURE DHS (2013)
20. Suchitta, P.: Motherhood in childhood: facing the challenge of adolescent pregnancy. The daily Star Report. 29 Oct (2013). www.thedailystar.net/news/facing-the-challenge-of-adolescent-pregnancy
21. Desriani, D.: Determinants of teenage motherhood: evidence from the 2007 Indonesia demographic and health survey. Master's dissertation, Flinders University, Adelaide (2009)
22. Nyarko, S.H.: Determinants of adolescent fertility in Ghana. Int. J. Sci. Basic Appl. Res. **5**(1), 21–32 (2012)
23. Mushwana, L., Monareng, L., Richter, S., Muller H.: Factors influencing the adolescent pregnancy rate in the greater Giyani Municipality, Limpopo Province–South Africa. Int. J. Afr. Nurs. Sci. **2**, 10–18 (2015)
24. Nguyen, H., Shiu, C., Farber, N.: Prevalence and factors associated with teen pregnancy in Vietnam: results from two national surveys. Societies **6**(2), 17 (2016)
25. Alam, N.: Teenage motherhood and infant mortality in Bangladesh: maternal age-dependent effect of parity one. J. Biosoc. Sci. **32**(2), 229–236 (2000)
26. Haque, M.N.: Levels, trends and determinants of adolescent childbearing in Bangladesh. Int. J. Curr. Res. **2**(1), 170–175 (2011)
27. Islam, M.M.: Adolescent childbearing in Bangladesh. Asia Pac. Popul. J. **14**(3), 73–87 (1999)
28. Nahar, Q., Min, H.: Trends and determinants of adolescent childbearing in Bangladesh. Calverton: Demographic and Health Research Division, Macro International Inc. DHS Working Paper. Report No: 48. Contract No.: GPO-C-00-03-00002-00. Sponsored by the United States agency for international development through the MEASURE DHS (2008)
29. Islam, M.M., Islam, M.K., Hasan, M.S., Hossain, M.B.: Adolescent motherhood in Bangladesh: trends and determinants. PloS One **12**(11), 1–14 (2017)
30. Grunseit, A.: Impact of HIV and sexual health education on the sexual behaviour of young people: a review update. UNAIDS, Geneva (1997)
31. National Institute of Population Research and Training, Mitra and Associates, The DHS Program.: Bangladesh demographic and health survey 2014, Dhaka, Bangladesh, and ICF International, Rockville, Maryland: NIPORT, Mitra and Associates, and The DHS Program (2016)
32. Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate data analysis: with Readings, 4th edn. Prentice-Hall, Englewood Cliffs (1995)
33. Johnson, R.A., Wichern, D.W.: Applied multivariate statistical data analysis, 6th edn. Prentice Hall, Englewood Cliffs (2002)

Chapter 19
Effect of Women's Education on Skilled Birth Attendants in South and South East Asia: A Cross-Country Assessment on Sustainable Development Goal 3.1

Raaj Kishore Biswas, Nurjahan Ananna, and Jahar Bhowmik

Abstract The Sustainable Development Goal (SDG) 3.1 is to "reduce the global maternal mortality ratio (MMR) to less than 70 per 100,000 live births" by 2030. One of the indicators of MMR is the proportion of births attended by a skilled health personnel. To achieve this goal low- and middle-income countries are required to increase the coverage of skilled birth attendants (SBA) for safe delivery during childbirth. This study used the Demographic and Health Surveys (DHS) data and assessed 1,171,731 women aged 15–49 years from 10 countries selected from South and Southeast Asian (SSEA) region to evaluate the status of SDG 3.1 in this region. This paper also evaluated the contribution of women's education on SBA coverage using surveys conducted during the period 1992–2017. Logistic regression models were fitted adjusting the survey clusters, strata and sampling weights. Meta-analyses were conducted collapsing the effect sizes and confidence intervals of education on SBA coverage. Cambodia, Indonesia and Philippines had over 80% SBA coverage after 2010, whereas Bangladesh had only 44.7% coverage among the selected countries in SEA. Education of women at all levels (primary, secondary and higher) were significantly associated with SBA coverage, suggesting that education is a key to skilled delivery cares in SSEA region.

Keywords Sustainable Development Goal · Maternal mortality ratio · Skilled birth attendants · Education · Meta-analysis · Demographic and Health Surveys

R. K. Biswas (✉)
Transport and Road Safety (TARS) Research Centre, School of Aviation, University of New South Wales, Sydney, NSW, Australia
e-mail: RaajKishore.Biswas@student.unsw.edu.au

N. Ananna
Ibrahim Medical College, Dhaka, Bangladesh

J. Bhowmik
Department of Statistics Data Science and Epidemiology, Swinburne University of Technology, Melbourne, Australia

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_19

19.1 Introduction

Progress in the public health domain in low- and middle-income countries (LMICs) are typically measured using the targets of Sustainable Development Goal (SDG). The aim of SDG 3 is to ensure healthy lives and promote well-being for all at all ages. The first target of this goal (3.1) is to "reduce the global maternal mortality ratio to less than 70 per 100,000 live births" by 2030 [18]. One of the two indicators for this objective is the proportion of births attended by skilled health personnel in that country and the other being maternal mortality ratio (MMR) [19].

A limitation in assessing SDG progress in SSEA countries is data paucity or irregular national level data [20]. According to this UN report, the Asia-Pacific region has made "satisfactory progress on" SDG 3. This includes the proportion of births attended by skilled health personnel, and this progress needs to be maintained to meet the 2030 target [22]. However, performance of the developing countries had varied across goals, countries, and regions in attaining Millennium Development Goals (MDG), which is expected to continue for SDGs as well. Furthermore, consorted collaborative efforts from both government and experts are required to monitor the progress and assess implementation strategies, which demands assessment of national data sets for SGD progress.

SBA is a broad category that encompasses health professionals, particularly doctors, nurses, and midwives, who are certified to attend mothers and newborn babies prior to and during delivery to manage normal deliveries and diagnose, manage or refer obstetric complications [13]. To keep consistency with DHS definition, this study considered a qualified doctor, nurse, midwife, paramedic, family welfare visitor (FWV), community skilled birth attendant (CSBA), or sub-assistant community medical officer (SACMO) as SBA [4]. Orthodox village doctors without academic qualifications, uncertified community workers, and untrained conventional midwives were identified as traditional birth attendants (TBAs).

There has been an increasing number of studies that investigated factors associated with healthseeking behaviors of mothers and children [24]. Common findings across these studies suggested that maternal health care is generally affected by various personal, sociocultural and environmental factors, including individual perceptions of health, self-efficacy, motivation, social values and belief systems [25]. Access to SBA services are found to be associated with various sociodemographic factors, for example education of women/mothers, religion, residency (urban/rural) and household financial capabilities are important predictors for women's access to maternal health services [17].

One important contributor to public health success in LMICs are education of women. Past decades have experienced growth in education, both for men and women in SSEA countries with high socioeconomic return [11]. Multiple governmental and non-governmental programs were conducted using both foreign aid and public funding to increase school enrollments, with particular focus on girls' education [14]. One objective of these programs, and overall literacy rate, is to inform women of their rights and literate them on maternal health care including access to SBA [6].

Literature on past studies demonstrate that improving SBA rate could significantly reduce maternal and child deaths, particularly adopting reinforcement of the programs focusing on training for health personnel and education for mothers [21]. Similarly, it is expected that education of mothers would contribute to the increased access to SBA as educated mothers are more aware of the delivery complications and more likely to access modern health care [12]. However, there is a gap in literature in quantifying the sociodemographic factors that commonly influence SBA accessibility across SSEA region with the objective of attaining SDG 3.1.

This study focuses on the skilled birth attendance of the developing countries in the South and Southeast Asian (SSEA) region, for which representative datasets on the population level are available through the Demographic and Health Surveys (DHS) program. These include multiple surveys from Afghanistan, Bangladesh, Cambodia, India, Indonesia, Maldives, Myanmar, Nepal, Pakistan, Philippines, and Timor-Leste. The primary objective of this study is to assess the progress of SDG 3.1 using proportion of skilled birth attendance and to investigate the sociodemographic factors associated with the gradual increase of skilled birth attendants (SBA) across these countries. More specifically, effects of women's education in the progress of SDG 3.1 will be evaluated through its association with SBA in the selected SEA countries.

Although several studies used DHS data sets from African nations to evaluate maternal health services, evaluation on cross-country assessment in SSEA region is limited [3]. For a consistent nationwide data collection process with similar methodologies in these selected LMICs, this study was limited to DHS surveys between 1990 and 2016. A meta-analysis was undertaken to evaluate the overall association between girls' education on SBA in this region. The outline of the paper is as follows: Sect. 19.2 provides detailed information on the data sets, variables used and statistical methodologies; Sect. 19.3 elucidates the results found from the analyses displayed in tables and figures; and Sect. 19.4 discusses the impact of the findings and compared with the contemporary literally with Sect. 19.5 provides concluding remarks.

19.2 Materials and Methods

19.2.1 Data Overview

DHS is considered a standardized and nationally representative cross-sectional survey, which has been conducted in LMICs since 1984. As the survey methodology are consistent across DHS and collected variables are nearly similar, these surveys allow assessments over multiple populations over time. All DHS follows a standard protocol with consent from the human participants approved by ICF

Macro Institutional Review Board and local research ethics committee. The authors had access to de-identified survey data with permission from Measure DHS and ICF (approval 127313). The secondary data sets analyzed in the current study are freely available upon request from the DHS website at http://dhsprogram.com/data/available-datasets.com.

Every survey under DHS uses two-stage stratified cluster sampling techniques [7]. Sampling frame consists of a list of enumeration areas (EAs) using recent census data. For first stage, EAs (or clusters) are selected using probability proportional to size (PPS) sampling method, where the number of clusters/EAs vary across countries. For example, typically there are 600 clusters in Bangladesh and 28,522 clusters in India. An equal probability systematic sampling method is applied in the second stage to select a pre-specified number of households from each cluster. Generally, the survey focuses on women of reproductive health age group (15–49 years), although some surveys includes men as well. In this current study, only data from female respondents were extracted from the surveys.

19.2.2 Surveys

From selected 10 South and Southeast Asian countries, surveys from 1990 to 2017 were included in this study. Total 37 surveys containing 1,171,731 participants (women) were analyzed. The included surveys of 10 countries are Afghanistan (2015), Bangladesh (1993, 1996, 1999, 2004, 2007, 2011, 2014), Cambodia (2000, 2005, 2010, 2014), India (1992, 1998, 2006, 2015), Indonesia (1997, 2002, 2007, 2012), Myanmar (2016), Nepal (1996, 2001, 2006, 2011, 2016), Pakistan (1990, 2006, 2012, 2017), Philippines (1993, 1998, 2003, 2008, 2013, 2017), Timor Leste (2009, 2016). Surveys that were not typical DHS (e.g., Afghanistan 2010 mortality survey) or had incomplete data (Indonesia 1991 or 1994) lacked the necessary variables and were excluded from the analysis. Data from Maldives were also excluded from the analysis as they had 95% and 100% SBA coverage in 2009 and 2016 survey respectively.

19.2.3 Variable

In this study, access of SBA is the outcome variable. DHS VI standard recode manual were followed while defining SBA. It was recoded as a binary variable with women who had accessed SBA vs those who did not. As explained earlier, qualified doctor, nurse, midwife, paramedic, FWV, MA and SACMO were considered as skilled ANC providers and SBAs. For one respondent who seek multiple services, the one with the highest qualification was considered as birth attendant during delivery.

According to literature and outcome from the pre-analysis results (missing values and consistency of variables in the surveys over the years), seven sociodemographic factors were included in this study as explanatory variables [9]. The selected explanatory variables are age of respondents (continuous measured in in years); residence (urban, rural); education of both respondent and her partner (none, primary, secondary, higher); wealth index (poorest, poorest, middle, richer, richest); age at first birth (years); and age of partner/husband (years). For adjusting models, survey weights, strata and cluster were extracted collected as well.

19.2.4 Statistical Analysis

As the outcome variable was binary, a regression model with binomial family of distributions would be appropriate to find the association between SBA and sociodemographic factors. As DHS data are collected at multiple levels (cluster, strata and individual), generalized linear models (GLMs) with binary outcome were used in this study adjusting for cluster-wise and strata-wise effects. For generalization of the results, survey weights for each individual were adjusted in GLMs commonly used in DHS surveys [1]. The models were fitted using R − *package svyglm (survey).*

Using the adjusted odds ratio of women's education status (primary, secondary and higher compared to no education) a meta-analysis was conducted for all surveys that indicated the association between education and SBA access. R − *package metafor* was used for fitting fixed effect forest plots. All data compilations and analyses were conducted in *R (3.5.0).*

19.3 Results

All 10 countries selected in this study have increased the coverage of SBA over the years (Table 19.1); however, annual improvement for the Afghanistan could not be observed as there were only one survey conducted during the selected period (2015). The highest coverage was noted in Cambodia in 2014 survey (89%), apart from Maldives where 100% coverage was observed but not added in this study. Philippines and Indonesia also had over 80% SBA coverage according to 2012 survey. Only Bangladesh had below 50% SBA coverage (44.7% in 2014) among the countries in their latest survey. The results were cross-checked with individual survey's report.

Most of the sociodemographic factors showed significant associations with access to SBA. As the primary focus of this study was to evaluate the relation between education of participants (women) and access to SBA, the effect sizes of education on SBA were extracted from each survey through GLM outcomes. The association between dichotomous SBA and education of participants categories

Table 19.1 The frequency distribution of skilled birth attendants for the DHSs in 10 selected countries of South and Southeast Asia since 1990

Country	Frequency (%) of women who received SBA services						
				Survey periods			
	1990–1995	1996–1998	1999–2002	2003–2006	2007–2010	2011–2013	2014–2017
Afghanistan							10639 (53.8)
Bangladesh	513 (14.4)	876 (19.0)	1406 (27.0)	968 (18.0)	1171 (23.8)	2433 (33.2)	2009 (44.7)
Cambodia			1973 (32.6)	2602 (42.4)	4690 (72.8)	5246 (88.9)	
India	17555 (47.1)	13123 (45.4)		21683 (58.9)			
Indonesia		7405 (53.9)	9207 (69.2)		11055 (72.2)	12421 (81.7)	
Myanmar							2695 (69.7)
Nepal		346 (10.4)	683 (16.5)	1120 (26.8)		1945 (47.7)	2700 (67.4)
Pakistan	1744 (43.6)			2610 (45.7)		4460 (59.9)	6063 (73.2)
Philippines	5447 (36.2)	2893 (55.3)		3060 (62.3)	2940 (63.7)	3972 (75.1)	6691 (83.7)
Timor Leste					1993 (33.2)		3088 (62.8)

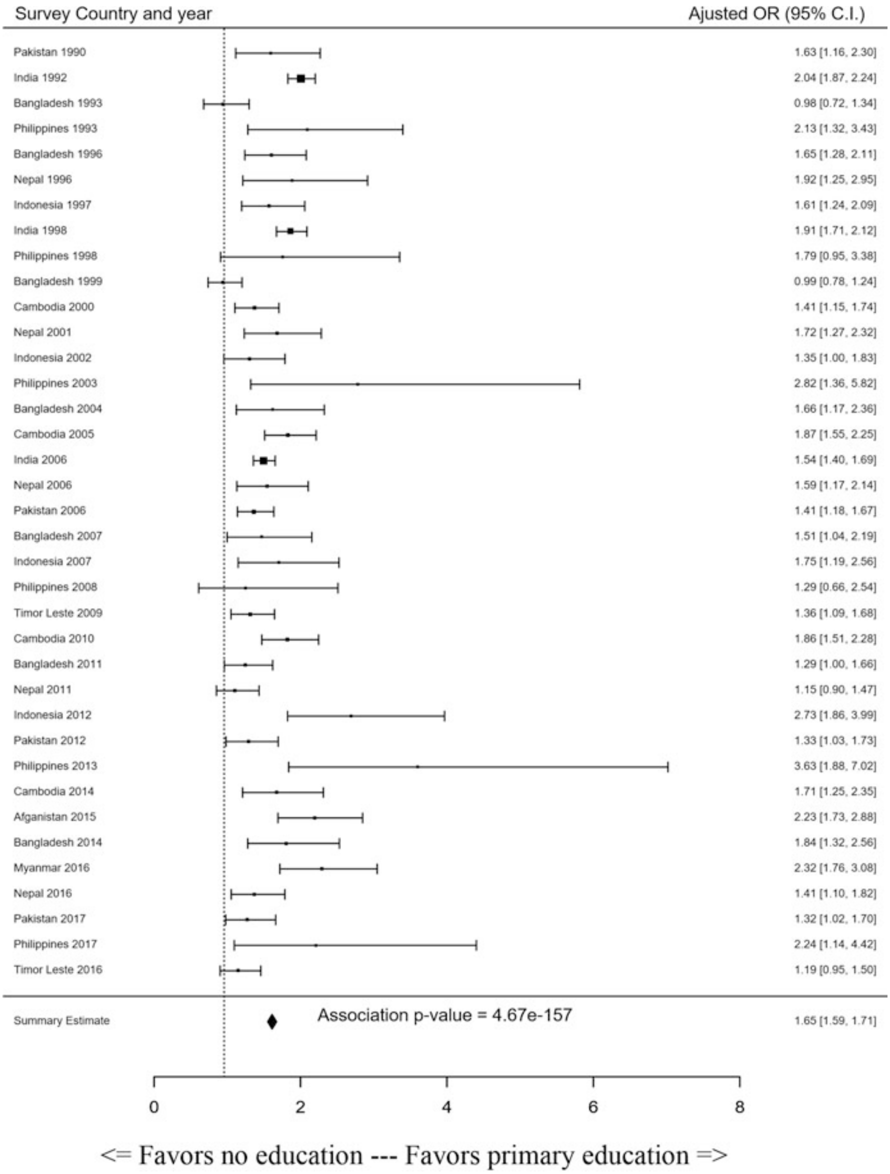


Fig. 19.1 Forest plot of odds of women having primary education on access to skilled birth attendants compared to women with no education (*OR* Odds Ratio, *C.I.* Confidence Interval)

are displayed in Figures 19.1, 19.2 and 19.3 through forest plots. Adjusted odds ratio (AOR) of primary, secondary and higher education compared to no education showed that most of the primary and secondary levels and all higher-level education status were significantly associated with women's access to SBA.

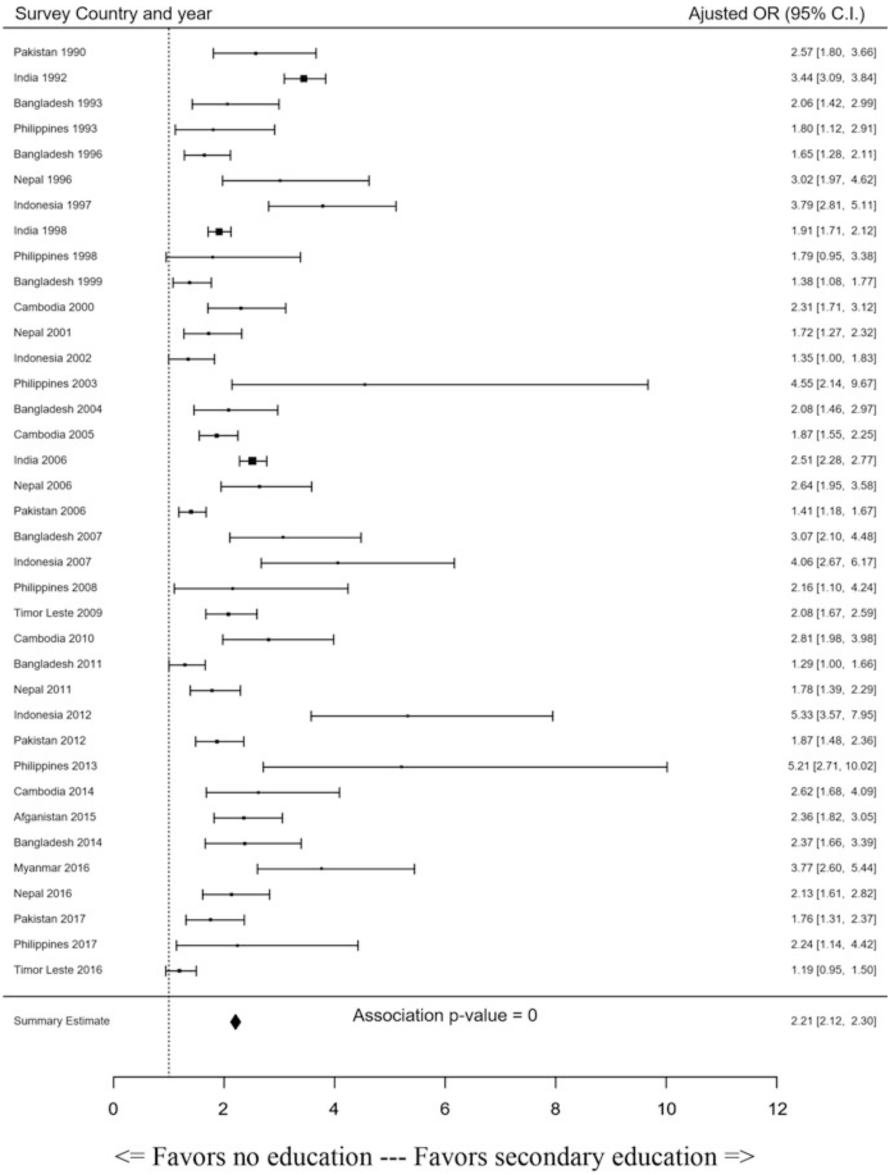


Fig. 19.2 Forest plot odds of women having secondary education on access to skilled birth attendants compared to women with no education (*OR* Odds Ratio, *C.I.* Confidence Interval)

It is evident that the AOR increased with increased level of education (Figs. 19.1, 19.2 and 19.3). Summary estimates indicate that women with primary, secondary and higher level of education were 1.65, 2.21 and 3.14 times significantly more likely to access SBA during childbirth. However, the more variation in the effect

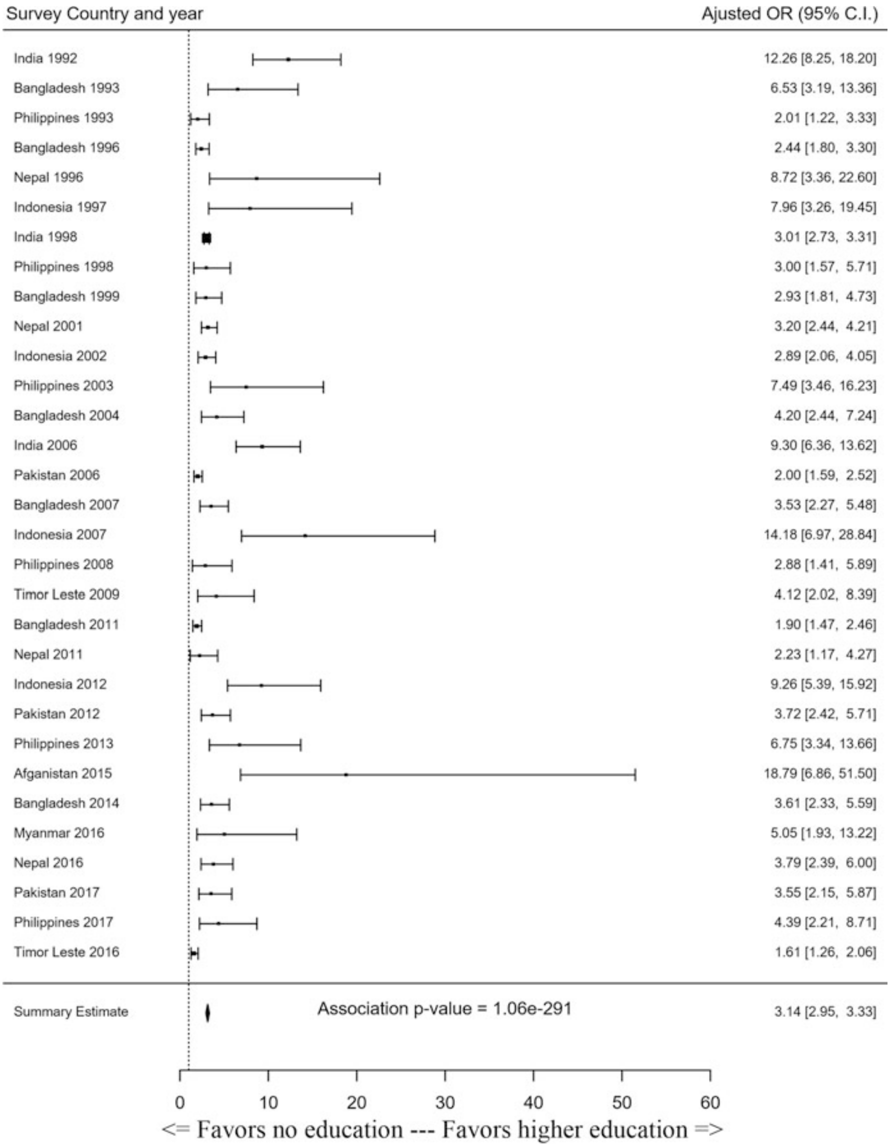


Fig. 19.3 Forest plot of odds of women having higher education on access to skilled birth attendants compared to women with no education (OR = Odds Ratio, C.I. = Confidence Interval). Due to the small sample size of higher education category, inflated AOR and CI were detected in Cambodia (2000, 2005, 2010, 2014), Pakistan (1990), 9 and Nepal (2006), which were not added in this forest plot

sizes for the estimates of higher education were observed, as some surveys estimated increased impact of higher education on access to SBA (e.g., Afghanistan 2015 and Indonesia 2007).

19.4 Discussion

This study aimed to evaluate the SBA coverage in countries of SSEA and evaluated the contribution of women's education in improving the SBA coverage, which ultimately expected to impact SDG 3.1 by decreasing maternal mortality. Results indicate that all selected countries in the SSEA region have improved their maternal health care scenario, particularly regarding access to SBA. However, this improvement is not homogeneous across the region as countries such as Cambodia, Indonesia and Philippines where 80% SBA coverage have already been achieved, according to post 2010 surveys. Only Bangladesh (44%) and Afghanistan (53.8%) had approximately 50% nationwide SBA coverage reported in recent surveys. Further analysis showed that women's education (primary, secondary and higher) were significantly associated with SBA coverage proving the suggestion that education is directly associated with access to SBA coverage.

Autonomy of women plays a key role in deciding her health care. South and Southeast Asian families are typically male donating and male is the house head. In patriarchal society, voice of women is limited to the privileged section of the society, generally higher educated females who contribute to the family economy [2]. This traditional scenario and cultural norms bars women from going beyond typical healthcare (TBA in this case) and avail any modern healthcare on their own. However, it is argued that educated women are likely to firmer in making decisions on their own and keen to reject the ancestral deliveries [15]. Thus, allowing access to education seem to help them partake SBA services and ultimately contribute to lowering maternal deaths.

Health care awareness is commonly associated with level of education. It is supported by the findings in this study as results showed that mothers with greater access to education were more likely to seek SBA services compared to the illiterate counterparts. Women, who completed their secondary and higher education, are expected to be well-informed of the various health issues, particularly the problems of seeking traditional unscientific cheap treatments. An educated individual is more likely to understand the consequences of traditional medicines or services from TBA, usually in rural areas that lead to unattended home deliveries [5, 10]. To rise above the long-established mindset of the community and go against the flow to seek modern medical help, higher education would be necessary to avert maternal deaths [8].

In LMICs, education is typically correlated with family income status. Educated families are likely to be well-off. This limits SBA services to the well-off families with educated women [23]. Insolvent households are less like to afford maternal health care services. Hospital admissions or access any means of health care are considered financial burden for them and compel to use services from TBAs providing cheap service in the locality for a long time [16].

Although this study analyzed a large population, there were few limitations. Firstly, all DHS are cross-sectional, which limits the scope of casual interpretation of the findings. Future studies could consider conducting a counterfactual analysis to determine the life saved due to increase of educated women in SA. Secondly, SBA calculation in surveys varied in terms of number of years preceding the survey. For example, Bangladesh DHS 2014 used 3 years preceding the survey, Afghanistan 2015 used 5 years and India 1992 used 4 years. Thirdly, during data cleaning, some strata in various surveys had complete missing or low sample count, which had to be omitted for model building purpose. Finally, as the data sets were interrupted time series, no prediction was undertaken. However, future studies could avail interrupted time series models to strengthen the meta-analysis.

19.5 Conclusion

With almost 10 years remaining for the 2030 deadline of the SDGs, some LMICs are having difficulties in minimizing the health care gaps, particularly providing access to SBA services. These countries must inject a sense of urgency by changing policies and executing accelerated actions at the national level. The study found that there exists heterogeneous development in SSEA region with some countries already gained above 80% SBA coverage and some are below the halfway mark. Female education, as hypothesized, was found to be significantly associated with access to SBA in all countries over the years, which shows a reasonable link between female literacy and access to health services.

The findings from this study suggest that effective intervention programs focusing on women's education, especially in the regional areas could be useful to improve the SBA coverage. Also, public agencies and non-governmental organization could be involved in regard to SBA training and awareness campaigns to increase the coverage. Finally, figuring out country level sociodemographic factors to identify the most vulnerable groups and then providing them with better facilities including health infrastructure and accessibility would be productive in achieving SDG 3.1.

References

1. Bhowmik, J., Biswas, R., Woldegiorgis, M.: Antenatal care and skilled birth attendance in Bangladesh are influenced by female education and family affordability: Bdhs 2014. Public Health. **170**, 113–121 (2019)
2. Biswas, R.K., Rahman, N., Kabir, E., Raihan, F.: Women's opinion on the justification of physical spousal violence: a quantitative approach to model the most vulnerable households in Bangladesh. PLoS One. **12**(11), e0187884 (2017)
3. Chol, C., Negin, J., Agho, K.E., Cumming, R.G.: Women's autonomy and utilisation of maternal healthcare services in 31 Sub-Saharan African countries: results from the demographic and health surveys, 2010–2016. BMJ Open. **9**(3), e023128 (2019)
4. DHS: Bangladesh Demographic and Health Survey 2014. National Institute of Population Research and Training (NIPORT), Dhaka (2016)

5. Kamal, S.M., Hassan, C.H., Kabir, M.: Inequality of the use of skilled birth assistance among rural women in Bangladesh: facts and factors. Asia Pac. J. Public Health. **27**(2), NP1321–NP1332 (2015)
6. Marphatia, A.A., Ambale, G.S., Reid, A.M.: Women's marriage age matters for public health: a review of the broader health and social implications in South Asia. Front. Public Health. **5**, 269 (2017)
7. Ministry of Health and Family Welfare, India: National family health survey (nfhs-4), International Institute for Population Sciences Deonar, Mumbai (2016)
8. Mullany, B.C., Becker, S., Hindin, M.: The impact of including husbands in antenatal health education services on maternal health practices in urban Nepal: results from a randomized controlled trial. Health Educ. Res. **22**(2), 166–176 (2006)
9. Nababan, H.Y., Hasan, M., Marthias, T., Dhital, R., Rahman, A., Anwar, I.: Trends and inequities in use of maternal health care services in Indonesia, 1986–2012. Int. J. Women's Health. **10**, 11 (2018)
10. Pagel, C., Prost, A., Hossen, M., Azad, K., Kuddus, A., Roy, S.S., Nair, N., Tripathy, P., Saville, N., Sen, A., et al.: Is essential newborn care provided by institutions and after home births? Analysis of prospective data from community trials in rural South Asia. BMC Pregnancy Childbirth. **14**(1), 99 (2014)
11. Psacharopoulos, G., Patrinos, H.A.: Returns to Investment in Education. The World Bank, Washington, DC (2018)
12. Shahi, P., De Kok, B., Tamang, P.: Inequity in the utilization of maternal-health care services in South Asia: Nepal, India and Sri Lanka. Int. J. Health Sci. Res. **7**(1), 271–281 (2017)
13. Shimamoto, K., Gipson, J.D.: The relationship of women's status and empowerment with skilled birth attendant use in Senegal and Tanzania. BMC Pregnancy Childbirth. **15**(1), 154 (2015)
14. Solotaroff, J.L., Pande, R.P.: Violence Against Women and Girls: Lessons from South Asia. The World Bank, Washington, DC (2014)
15. Story, W.T., Burgard, S.A.: Couples' reports of household decision-making and the utilization of maternal health services in Bangladesh. Soc. Sci. Med. **75**(12), 2403–2411 (2012)
16. Talukder, S., Farhana, D., Vitta, B., Greiner, T.: In a rural area of Bangladesh, traditional birth attendant training improved early infant feeding practices: a pragmatic cluster randomized trial. Matern. Child Nutr. **13**(1), e12237 (2017)
17. Tappis, H., Koblinsky, M., Doocy, S., Warren, N., Peters, D.H.: Bypassing primary care facilities for childbirth: findings from a multilevel analysis of skilled birth attendance determinants in Afghanistan. J. Midwifery Womens Health. **61**(2), 185–195 (2016)
18. UN General Assembly: Transforming our World: The 2030 Agenda for Sustainable Development. Technical report resolution A/RES/70/1 (2015)
19. UN General Assembly: Resolution Adopted by the General Assembly on 6 July 2017: Work of the Statistical Commission Pertaining to the 2030 Agenda for Sustainable Development. Technical report resolution A/RES/71/313 (2017)
20. United Nations: Asia and the Pacific SDG Progress Report 2017. Technical report (2017), iSBN: 978-92-1-120776-7
21. United Nations: Asia-Pacific Sustainable Development Goals Outlook. Technical report (2017), iSBN: 978-92-9257-775-9
22. United Nations: The Sustainable Development Goals Report 2018. Technical report (2018), iSBN: 978-92-1-101390-0
23. Wang, W., Hong, R.: Levels and determinants of continuum of care for maternal and newborn health in Cambodia-evidence from a population-based survey. BMC Pregnancy Childbirth. **15**(1), 62 (2015)
24. Woldemicael, G., Tenkorang, E.Y.: Women's autonomy and maternal health-seeking behavior in Ethiopia. Matern. Child Health J. **14**(6), 988–998 (2010)
25. Yaya, S., Bishwajit, G., Ekholuenetale, M.: Factors associated with the utilization of institutional delivery services in Bangladesh. PLoS One. **12**(2), e0171573 (2017)

Part V
Small Area Estimation and Spatial Microsimulation

Chapter 20
Estimation of Child Undernutrition at Disaggregated Administrative Tiers of a North-Eastern District of Bangladesh: An Application of Small Area Estimation Method

Sumonkanti Das, Bappi Kumar, Md. Zakir Hossain, Sabbir Tahmidur Rahman, and Azizur Rahman

Abstract Children of *Sunamganj* district located in the north-eastern part of Bangladesh are highly vulnerable to undernutrition and chronic food insecurity due to its geographic location, long-time waterlog, frequent flash floods, and underdeveloped infrastructure. In this study, child undernutrition indicators stunting and underweight are estimated at district, sub-district (*Upzila*) and union level administrative tiers of *Sunamganj* district employing the World Bank small area estimation (SAE) method to a *Sunamganj* household level survey data collected in 2018 and the census 2011 data of *Sunamganj*. District level prevalence of stunting and underweight are estimated as about 48.5% (95% CI: 45.3–51.7%) and 37.0% (95% CI: 34.6–39.8%) based on the SAE method. At *upzila* level, stunting varied from 41.0% to 54.9% and underweight varied from 24.0% to 53.4%; while the indicators varied over 19.5–59.7% and 20.2–56.8% respectively at union level. A significant number of unions are found as hotspots of higher underweight and stunting over the north, north-eastern and north-western parts of *Sunamganj*. Though the southern part of *Sunamganj* was homogeneous in the *upzila* level maps of stunting and underweight; significant number of heterogeneous unions are found in the union-level maps. The *upzila*s belong to the northern part particularly closer to the Indian border and *haor* areas are mostly vulnerable to stunting and underweight. The study findings on disaggregate level prevalence of stunting and underweight

S. Das (✉)
Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh

Quantitative Economics, Maastricht University, Maastricht, The Netherlands
e-mail: s.das@maastrichtuniversity.nl

B. Kumar · Md. Z. Hossain · S. T. Rahman
Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh

A. Rahman
School of Computing and Mathematics, Charles Stuart University, Wagga Wagga, NSW, Australia

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_20

might help the concerned government and non-government organizations to prepare and implement aid-related programs on public health and nutrition.

Keywords Stunting · *Sunamganj* District · Underweight · *Upzila* · Union · World Bank Method

20.1 Introduction

Sunamganj is a north-eastern district of Bangladesh bounded by Meghalaya state of India on the north. A large part of Sunamganj district is covered by *haors*, a wetland ecosystem characterized by the presence of large bowl-shaped flood plain depression. The low lying areas remain under water for nearly half of the year. Geographic location, long-time waterlog, frequent flash floods, and underdeveloped infrastructure are some major determinants of children undernutrition vulnerability and chronic food insecurity [1, 2]. Sunamganj district belongs to Sylhet division which is known as economically well-off region in Bangladesh, however, the demographic characteristics, child and maternal health care conditions of Sylhet division are very worse than the other divisions of Bangladesh. Currently Sylhet region has been experiencing high fertility rate (2.9% in 2014), low contraceptive use (40.9% in 2014), highest infant and under-5 mortality rates (55 and 67 per 1000 live births respectively), lowest antenatal care visits of pregnant mothers (53.1% ANC from medically trained provider), lower proportion of facility delivery (22.6%), and lowest coverage of postnatal care after delivery (25.1%) [3]. Progress in these indicators in Sylhet division is found very slow in the demographic surveys conducted during the last two decades [3–8]. Such slow-paced improvement scenarios are also persistent in the child health and nutrition sector. For a long time period, the proportions of stunted and underweight children are found markedly higher in Sylhet division. The trends of these two child health indicators shown in Table 20.1 indicate that the prevalence of stunting and underweight are still about 50% and 40% respectively in 2014, while this prevalence was respectively about 61% and 64% in 1996–97. These improvements in Sylhet region during this two-decade are very poor compared to the other regions of the country.

Though division level estimates are available from nationwide household survey data, disaggregated administrative level statistics are not estimable from survey data. Consequently, the demographic statistics of the Sunamganj district are not available from the nationwide survey. For estimating sub-district (*Upzila*) specific estimate of child undernutrition, BBS and UNWFP [9] utilized an indirect statistical technique

Table 20.1 Child undernutrition indicators in Sylhet Division, 1996–2014

Indicators	1996–97	1999–2000	2004	2007	2011	2014
Stunting	61.4	56.8	46.2	44.7	49.3	49.6
Wasting	20.9	11.1	12.2	18.3	18.4	12.1
Underweight	64.0	56.8	49.8	42.1	44.9	39.8

Sources: NIPORT, Mitra and Associates, and ORC Macro [3–8] known as small area estimation (SAE) to the 2000 Child Nutrition Survey data and a 5% sample of 2001 Bangladesh Population and Housing Census (BPHC) data. It is observed that the prevalence of stunting and underweight were more than 60% and 56% respectively for most of the *upzilas* in Sunamganj district. Recently BBS, WFP and IFAD [10] has implemented the same technique to a Child Nutrition data (combination of Child and Mother Nutrition Survey of Bangladesh 2012 and the Health and Morbidity Status Survey 2011) and full census data of 2011 BPHC. Based on this study findings, it is observed that the rates have been reduce to 46.1% and 40.1% respectively at district level (Appendix C, [10]).

Child health and nutrition related issues particularly in the vulnerable areas like *haor*s have been major concerns to both the local and central governments with respect to the achievement of the sustainable development goals by the targeted time-period. In this respect, Bangladesh Government has already started to estimate district level official statistics of different emerging issues like poverty [11] via multiple indicator cluster survey for monitoring and implementing different intervention programmers at district level. For estimating indicators at the finer level below district level, the SAE method is still require to implement by utilizing the survey data and the concurrent census data. In this study, a Household level survey data on child undernutrition in Sunamgnaj district has been combined with the census information of Sunamgnaj district extracted from the 2011 BPHC. As previous studies on *upzila* level child undernutrition [9, 10], the World Bank SAE methodology also known as ELL method after the authors Elbers, Lanjouw, and Lanjouw [12] has been implemented to explore children undernutrition vulnerability and their inequality at finer administrative units of *Sunamgnaj* district. The specific objectives of the research include estimation of district, *upzila*, and union levels prevalence of stunting and underweight with their accuracy measures and then generation of *upzila* and union-level maps of *Sunamganj* district for each of the considered child undernutrition indicators. The outcomes of the research might help to understand the child undernutrition situations at the finer level as well as to understand the spatial distribution of the child vulnerability in *Sunamganj* district.

20.2 Data Sources

A household survey entitled “2018 *Sunamganj* Child Morbidity and Undernutrition Survey” was conducted under a research project “Assessment of Child Morbidity and Undernutrition Status in Different Administrative Tiers of *Sunamganj* District through Small Area Estimation Technique” funded by Ministry of Science & Technology, Bangladesh for collecting data on children health, nutrition, and morbidity in *Sunamganj* district. Children nutrition data have been utilized here to explore the undernutrition vulnerability at finer administrative units of *Sunamganj* district. Full census data of *Sunamganj* District have been extracted from the 2011 BPHC data. Brief descriptions of these datasets are given in the below two subsections.

20.2.1 Sunamganj Population and Housing Census (SPHC) Data, 2011

The full census data of *Sunamganj* district have been collected from BBS for the above mentioned research project. The census data of *Sunamganj* have been referred as 2011 *Sunamganj* Population and Housing Census (SPHC) Data. *Sunamganj* district consists of 11 *upzila*s, 87 unions (rural) and 36 wards (urban), 1599 *Mauza* (rural) and 139 *Mahalla* (urban), and overall 2887 villages. Since full census dataset for *Sunamganj* district is available, the target parameters can be estimated at Mauza level if a good model can be developed. The SPHC data consists of 438,747 general HHs with 2,449,618 population. *Upzila*, *union/ward* and *mauza/mahalla* levels mean number of households were 43,326, 4805, and 3750 respectively. A total of 354,668 under-5 children were found in these general HHs based on which the target parameters at district, *upzila*, and union levels are estimated. The distribution of under-5 (U5) children at *upazila*, *union* and *mauza* levels are given in Table 20.2.

20.2.2 Sunamganj Child Morbidity and Nutrition Survey (SCMNS) Data, 2018

Since child health, nutrition and morbidity information are not available from the census data and also no representative data of *Sunamganj* district is available, a district representative household survey data has been collected covering all the 11 *upzila*s of *Sunamganj* district. A two-stage cluster sampling design has been adopted to select households following the 2011 BDHS sampling plan [3]. *Sunamganj* district was divided into two strata according to rural and urban residence, hence 60 clusters (30 per stratum) are planned to select proportionally from rural and urban strata. The lower administrative unit *mauza* in rural area and *mahalla* in urban area are considered as the primary sampling unit (PSU) and hence the clusters for the study.

At the first stage, 60 clusters were selected in such a way that rural and urban clusters of 11 *upzila*s are covered in the sample. Thus, the clusters were required to select via probability proportional to cluster size and their place of residence

Table 20.2 Administrative units, mean, minimum and maximum number of under-5 children in *Sunamganj* district in the 2011 SPHC and the 2018 SCMNS)

	2011 SPHC			2018 SCMNS		
	Upzila	Union/Ward	Mauza/Mahalla	Upzila	Union/Ward	Mauza/Mahalla
Units	11	123	1738	11	61	62
Mean	34,762	4083	513	122	22	21
Minimum	16,780	177	1	62	14	14
Maximum	53,723	8123	2809	212	43	31

(rural-urban). In the second stage, 20 households having children under 5 years of age were planned to draw randomly from each cluster. Since every household does not have under-5 children, the selected households with under-5 children are random in nature. Thus, the ultimate sample size for the proposed study stands at least 1200 households covering 60 clusters from rural-urban residence of 11 *upzila*s. Considering the issues of undistinguishable urban-rural classification of *Mauza/Mahalla* in some *upzila*s, poor communication circumstances, and adverse weather circumstances, ultimately 1241 households were selected from 62 mauza/mahalla covering 61 *union/ward* and 11 *upazila*s. After scrutinizing, a total of 1339 children were found for which collected data were plausible to analyze. The distribution of surveyed children in SCMNS according to administrative units are given in Table 20.2. Though children undernutrition could be estimated at mauza/mahalla level, only the first three hierarchies district, *upazila* and union/ward are considered in this study.

In the household survey, anthropometric information children age (in month), height (in cm), and weight (in kg) were collected following the recommended methods and tools by WHO [13]. Children age was recorded from his/her Vaccination Card, height was measured using standard height measuring tape, and weight was recorded using digital bathroom scale with one decimal point readabilities. The measuring scales were verified by the Civil Engineering department of Shahjalal University of Science & Technology. Socio-economic information of the households were collected from household head, while mother and children information were collected from mothers.

A questionnaire was prepared both in English and Bangla to record all the collected information. Before interview, an oral consent with mother's signature was taken in the questionnaire. The questionnaire was developed following the BDHS questionnaire pattern but assured the variables are structured similar to the 2011 BPHC questionnaire. The survey was conducted by the university graduates during the period of April–May, 2018. Since it was time consuming to find household with under-5 children, information of such households was collected as snowball sampling technique. As a result, the households have a nested characteristic by nature.

20.3 Statistical Methods

The main response variables height-for-age Z-score (HAZ) and weight-for-age z-score (WAZ) are calculated using the children anthropometric data through WHO Anthro software [3]. Suppose y_{ijk} is the HAZ/WAZ score for the k^{th} child belongs to the j^{th} household (HH) of the i^{th} cluster. It is assumed that the anthropometric index follows either a 2-level or 3-level model as below:

$$y_{ijk} = \mathbf{x}_{ijk}^T \beta + u_{ij} + \varepsilon_{ijk} \quad \text{or} \quad y_{ijk} = \mathbf{x}_{ijk}^T \beta + \eta_i + u_{ij} + \varepsilon_{ijk}$$

where $\mathbf{x}_{ijk}^{T}$ is a vector of explanatory information, β is the vector of regression parameters and η_i, u_{ij} and ε_{ijk} are respectively cluster, HH, and child-specific random errors. It is assumed that the level-specific random errors are identically and independently distributed with mean zero and homoskedastic error variances $\hat{\sigma}^2_{\eta(l)}$, $\hat{\sigma}^2_{u(l)}$, and $\hat{\sigma}^2_{\varepsilon(l)}$ respectively where the suffix (l) for l^{-} level model. Assuming normality of the level-specific random errors, the models can be developed following standard maximum likelihood (ML) or restricted ML (REML) method. The fitted models can be used to estimate mean and distribution functions (say, HAZ < -2.0) of a nutrition index at different hierarchies (as for example, division, district, and *upzila* in this study).

Auxiliary variables at different hierarchical levels (such as cluster- or household-level) are used to explain the respective level variation. If variation at a particular level is explained more by a model, the respective error variance will be smaller. The estimate for a particular small area will typically be the average of the predicted Y_{ijk}s in that area. Since the standard error of a mean gets smaller as the sample size gets bigger, the contribution to the overall standard error of the variation at each level depends on the sample size at that level. The number of households in a small area will typically be much larger than the number of clusters, and the number of children under five larger again, so to get small standard errors for the small area estimates it is of particular importance that, at the highest level, the unexplained cluster-level variance $\hat{\sigma}^2_{\eta(l)}$ should be small. Two important diagnostics of the model-fitting stage, in which the relationship between Y and X is estimated for the survey data, are the R^2 measuring how much of the variability in Y is explained by X, and the ratio $\left(\hat{\sigma}^2_{\eta(3)} + \hat{\sigma}^2_{u(3)} + \hat{\sigma}^2_{\varepsilon(3)}\right)^{-1} \hat{\sigma}^2_{\eta(3)}$ measuring how much of the unexplained variation is at the cluster level.

In this study, the World Bank ELL method instead of other relevant SAE methods like empirical Bayes [14] and M-quantile [15] methods has been applied not only for ease of the calculation but also comparison with the previous studies on child undernutrition in Bangladesh [9, 10]. The ELL procedure is sub-divided into two stages: fitting regression model at first and then prediction of the response variable for all census units based on the fitted model. After fitting the regression model and obtaining the corresponding parameters, the second stage of the ELL method is to conduct either a parametric bootstrap (PB) or a non-parametric bootstrap (NPB) procedure to obtain the area-specific poverty estimates of interest and their corresponding estimated mean square errors (ESMEs). More detail of the ELL procedure are available in ELL [12, 16], Das and Chambers [17]. For both the PB or NPB procedures the basic steps are: **Step 1**: generate regression parameters from a suitable sampling distribution, say the multivariate normal distribution; **Step 2**: generate level-specific random errors using an appropriate parametric distribution or by resampling via simple random sampling with replacement from the estimated level-specific sample residuals; **Step 3**: generate bootstrap response values y^*_{ijk} using the generated regression parameters and the level-specific random errors. The generated response values are used to estimate the area-specific parameter of interest say $P_i^* = N_i^{-1}\sum_{j=1}^{C_i}\sum_{k=1}^{N_{ij}} I\left[y_{ijk}^* < t\right]$ for a specific value of t (say, here $t = -2.00$). These three steps are iterated for a large number of times say $B = 500$ and then the mean and variance of these B estimates are considered as the final estimates and their MSEs respectively as

$$\hat{P}_i^{ELL} = B^{-1}\sum_{b=1}^{B} P_i^* \quad \text{and} \quad se_i^{ELL} = B^{-1}\sum_{b=1}^{B}\left(P_i^* - \hat{P}_i^{ELL}\right)^2.$$

In this study, several selection criteria have been considered for selecting the best model for a response variable. For a linear mixed model, conditional AIC [18], marginal and conditional r-squared values [19], and likelihood ratio test (LRT) of variance components have been considered for selecting a better model. Also, the additive properties of variance components at different hierarchies are also checked when higher level model is selected. As for example, if a 2-level and a 3-level models have been fitted and the sum of level-2 and level-3 variance components of 3-level model is expected to be approximately equal to the level-2 variance component of 2-level model [17]. Also, the level-specific residuals are expected to be approximately normal. If departure from normality is observed, non-parametric bootstrap procedure will be used in the ELL bootstrap procedure so that the estimated level-specific residuals can be utilized.

20.4 Results and Discussion

The explanatory variables common in both survey and census data are used to develop multilevel linear regression models for HAZ and WAZ. A number of *upzila*, union and mauza specific contextual variables created from the Census data are utilized in the model development with the intention of reducing higher-level variations. Several hundred variables including two and three-way interactions are generated from both survey and census data, however some variables have been found significant in the developed models shown in Appendix Tables 20.8 and 20.9. The sub-sections below illustrate (i) development of proper multilevel linear regression models for HAZ and WAZ, (ii) estimation of stunting (HAZ < −2.00 SD) and (iii) underweight (WAZ < −2.00 SD) at district, *upzila* and union levels along with *upzila* and union level spatial maps.

20.4.1 Multilevel Modeling of Child Undernutrition Indicators

A number of multilevel linear regression models including single level linear regression model have been developed to explore the best model for HAZ and WAZ. Summary statistics of the best fitted random intercept linear regression model

Table 20.3 Summary statistics and diagnostics of the fitted linear regression (LM) and random intercept linear regression (LMM) models for Height-for-age (HAZ) and Weight-for-age (WAZ) Z-scores, 2018 SCMNS

Model for Stunting							
Model	DF	Parameters		Marginal R^2	Conditional R^2	AIC	cAIC
		σ_u^2	σ_ϵ^2				
LM	9	–	3.2148	7.48	–	4929.851	–
LMM	9	0.1112	3.1141	7.37	10.57	4938.455	4916.602
LR test of $H_0 : \sigma_\eta^2 = 0$: χ^2=6.2233, P-value = 0.0063 (LM vs. LMM)							
Model for Underweight							
Model	DF	Parameters		Marginal R^2	Conditional R^2	AIC	cAIC
		σ_u^2	σ_ϵ^2				
LM	9	–	2.2590	7.43	–	4810.36	–
LMM	9	0.0242	2.2366	7.33	8.32	4824.329	4811.336
LR test of $H_0 : \sigma_\eta^2 = 0$: χ^2= 0.2706, P-value = 0.3015 (LM vs. LMM)							

denoted hereafter by LMM shown in Table 20.3 indicate that the 2-level *mauza*-specific model perform better than the single level linear regression model (denoted by LM) in terms of conditional AIC (cAIC), R^2, and LRT. For the model of HAZ, the LRT suggests that the *mauza*-specific variance component is significantly greater than zero and the smaller cAIC value (4916.6) of the LMM than its AIC (4938.5) supports the fitted LMM as the better model. Therefore, the 2-level LMM is chosen for implementing the ELL method. For WAZ, the cAIC value of LMM model instead of LRT suggests the LMM performs better than the LM. As a diagnostic of the fitted models, the Q-Q plots, and the histograms with normality and kernel density curves of the estimated *mauza*-specific and children-specific residuals are shown in Fig. 20.1. The Q-Q plots and histograms confirm that both type of residuals are approximately normally distributed. The estimated regression coefficients with their standard errors of the corresponding LMM are utilized as input in the ELL bootstrap procedure. Since the residuals are approximately normally distributed, parametric bootstrap procedure is maintained for estimating stunting and underweight with their mean squared errors.

20.4.2 Prevalence of Stunting among Under-5 Children in Sunamganj District

District and *upzila* level prevalence of stunting are estimated using the ELL method and compared with the design-based direct estimator (DIR) [20]. Union level estimates are not calculated using the DIR estimator for two reasons (i) sample sizes are very small (ranges 14–31) for getting estimates with good accuracy, (ii) only 61 out of 123 union/wards are covered in the household survey. The SAE estimator provides approximately similar estimates of stunting as the DIR estimator at district

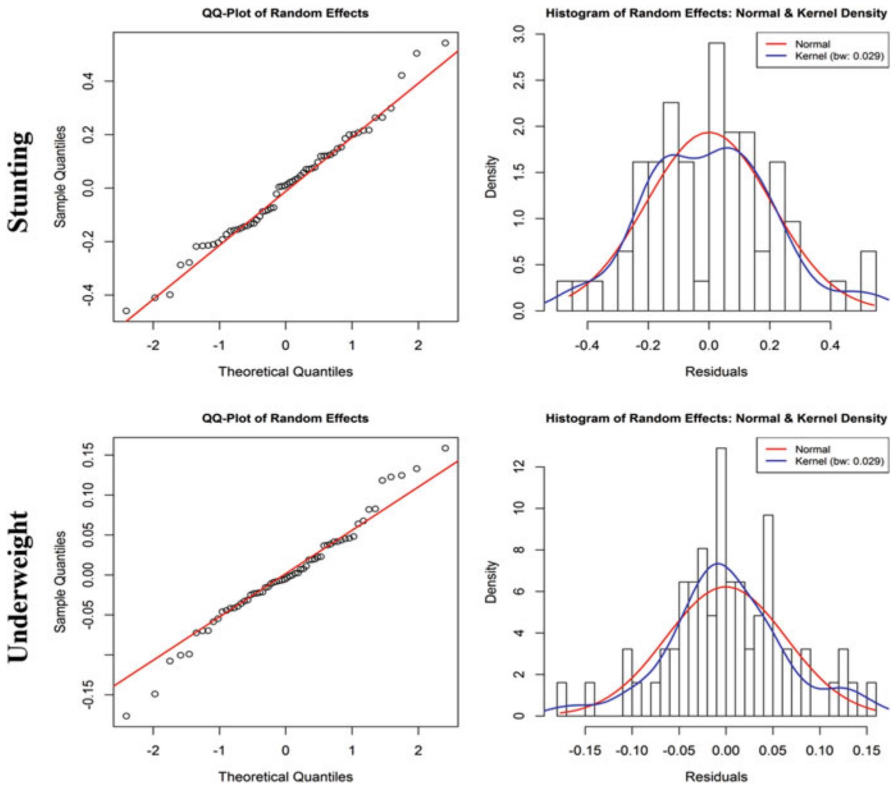


Fig. 20.1 Q-Q plot and histogram of the mauza-level random effects obtained from the random intercept linear regression (LMM) model for Height-for-age (HAZ) and Weight-for-age (WAZ) Z-score, 2018 SCMNS

level (Table 20.4). Both the DIR and SAE estimators indicate that the district level prevalence of stunting was about 49.0% in *Sunamganj* district. At *upzila* level, model-based estimator provides estimates with lower RMSEs and CVs for all *upzila*s. The SAE estimator shows prevalence of stunting varies within the range of 41–55% at *upzila* level (Table 20.4), more specifically the children of *Jagannathpur* were less stunted (95% CI: 36.0–46.1%), while those in *Bishwambarpur* (95% CI: 50.5–59.6%) and *Dharampasha* (95% CI: 49.4–59.5%) were more stunted compared to other *upzila*s. Summary statistics of the estimated union-specific prevalence of stunting with their RMSEs, CVs, and 95% CI are shown in Table 20.5. In the BBS, WFP, and IFAD [10] study, the prevalence of stunting varied within 44.9–48.8%: highest in *Tahirpur* and lowest in *Derai*. The prevalence of stunting varies within 19.5–59.7% at union level: the highest in *Uttar Baradal* (59.7%) of *Tahirpur* upazila and the lowest in Ward-07 (19.5%) of *Derai* Pourashava.

The *upzila* level map for stunting reveals a spatial pattern in the prevalence of stunting. The *upzila*s closer to the border of India and surrounded by *haor*s (east

Table 20.4 Prevalence of stunting among under-5 children at *upzila* level of *Sunamganj* district with standard errors (RMSE), coefficient of variation (CV, %) and 95% confidence interval (CI) using direct (DIR) and SAE (ELL.C) estimators

		Prevalence (%)		RMSE (%)		CV (%)	
Upzila	N	DIR	SAE	DIR	SAE	DIR	SAE
Bishwambarpur	24,380	57.14	54.93	5.40	2.41	9.45	4.38
Chhatak	53,723	40.89	46.31	3.45	2.38	8.44	5.13
Dakshin Sunamganj	27,586	45.90	47.34	4.51	1.89	9.83	4.00
Derai	35,412	58.88	43.65	4.76	2.25	8.08	5.17
Dharampasha	31,894	59.57	54.40	5.06	2.75	8.5	5.06
Dowarabazar	33,971	57.47	52.87	5.30	2.46	9.22	4.66
Jagannathpur	32,113	38.94	41.02	4.59	2.59	11.78	6.31
Jamalganj	25,080	44.30	47.02	5.59	2.15	12.61	4.57
Sulla	16,780	45.00	44.84	6.42	3.12	14.27	6.95
Sunamganj Sadar	37,926	44.38	48.64	3.72	1.86	8.39	3.82
Tahirpur	35,803	57.43	52.54	4.92	2.19	8.57	4.17
Sunamganj	354,668	48.78	48.45	1.43	1.68	2.92	3.47

Table 20.5 Summary statistics of stunting among under-5 children at union level in *Sunamganj* district with standard errors (RMSE), coefficient of variation (CV, %) and 95% confidence interval (CI) using SAE (ELL.C) estimator

				95% CI	
Summary statistics	Prevalence (%)	RMSE (%)	CV (%)	LL	UL
Minimum	19.54	2.2	4.35	9.04	32.74
Q1	41.85	3	6.00	30.98	49.62
Mean	45.59	4.33	10.40	37.39	54.16
Median	47.19	3.61	7.51	39.16	54.1
Q3	51.93	5.57	14.15	45.68	59.51
SD	8.69	1.68	6.10	7.5	10.3
Maximum	59.73	8.85	32.18	53.18	71.39

part) are highly vulnerable to have more stunted children, while the prevalence is lower in the south and southeast region of *Sunamganj* district, which are closer to the Sylhet district. On the other hand, the proportion of stunted children is found higher in *Dharampasha upzila* which actually get no district-level benefits due to its location from the center of the district.

The interactive map of stunting at union level shown in Fig. 20.2 indicates that there are significant inequalities across the unions. A significant number of unions are found as hotspots of higher stunting over the north, north-eastern and north-western parts of *Sunamganj* district. More specifically, about 20 unions had stunting in the range of 54.0–60.0% (dark red-spot in the map). More than 50% children belonging to the unions of *Bishwambarpur* (except only *Fatehpur* union with 47.6%) and *Dharampasha upzilas* (except only *Madhyanagar* union (41.7%) were stunted. A greater inequality has been observed in *Sulla upazila*, where only

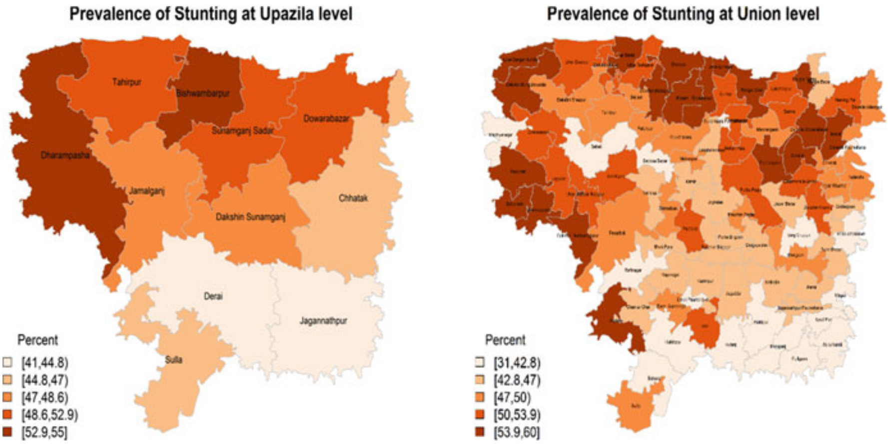


Fig. 20.2 Map of stunting among under-5 children at *Upzila* and Union levels in *Sunamganj* District using SAE estimator

the *Atgaon* union had a prevalence of 55.2% and two other unions had prevalence of around 35.8%. Overall, the prevalence of stunting decreases from north-western part to south-estern parts of *Sunamganj* district.

20.4.3 Prevalence of Underweight among Under-5 Children in Sunamganj District

The district level prevalence of underweight is estimated as about 37.0% (Table 20.6). As expected the SAE estimator provides *upzila* and union level with lower CV and RMSE than the DIR estimator. As per the SAE estimates, the proportion of underweight children was found higher among the children of *Dowarabazar upzila* (53.4% with 95% CI: 44.6–62.3%) and lower among those of *Jamalganj upzila* (24.0% with 95% CI: 17.5–31.2%). Table 20.7 shows that the prevalence of underweight varies within 20.2–56.8% at union level: the highest in *Duhalia* (56.8%) of *Dowarabazar upazila* and the lowest in *Beheli* (20.2%) of *Jamalgonj upazila*. More than half of the unions had prevalence of underweight more than 35%, which is more than the national MDG target of underweight (33.0% in 2015). More specifically, about 18 unions (most are from *Dowarabazar upzila*) had underweight in the range of 39.4–57.0% (dark red-shade in the map).

The *upzila* level map shown in Fig. 20.3 reveals a spatial pattern in the prevalence of underweight similar to stunting. The *upzila*s closer to the border of India (eastern part) and surrounded by *haor*s (except *Tahirpur upzila*) are highly vulnerable to have more underweight children, while the prevalence is lower in the south and southeast region of *Sunamganj* district. On the other hand, the proportions of underweight

Table 20.6 Prevalence of underweight among under-5 children at *upzila* level of *Sunamganj* district with standard errors (RMSE, %), coefficient of variation (CV, %) and 95% confidence interval (CI) using direct (DIR) and SAE estimators

		Prevalence (%)		RMSE (%)		CV (%)	
Upzila	N	DIR	SAE	DIR	SAE	DIR	SAE
Bishwambarpur	24,380	34.88	38.17	5.14	1.97	14.73	5.16
Chhatak	53,723	29.19	36.49	3.14	1.48	10.77	4.05
Dakshin Sunamganj	27,586	33.07	37.62	4.17	1.56	12.62	4.16
Derai	35,412	33.33	34.12	4.30	1.66	12.91	4.87
Dharampasha	31,894	40.37	38.06	4.70	1.72	11.64	4.52
Dowarabazar	33,971	39.78	53.35	5.08	4.70	12.76	8.81
Jagannathpur	32,113	32.03	34.56	4.12	1.30	12.88	3.76
Jamalganj	25,080	25.93	23.95	4.87	3.53	18.78	14.75
Sulla	16,780	35.00	34.57	6.16	1.96	17.59	5.68
Sunamganj Sadar	37,926	30.26	36.75	3.29	1.61	10.87	4.39
Tahirpur	35,803	28.30	35.85	4.38	2.20	15.46	6.15
Sunamganj	354,668	32.42	37.02	1.29	1.30	3.98	3.52

Table 20.7 Summary statistics of underweight among under-5 children at union level in *Sunamganj* district with standard errors (RMSE), coefficient of variation (CV, %) and 95% confidence interval (CI) using SAE estimator

				95% CI	
Summary statistics	Prevalence (%)	RMSE (%)	CV (%)	LL	UL
Minimum	20.15	1.72	4.60	13.18	28.22
Q1	33.15	2.16	5.81	25.70	38.74
Mean	36.00	3.19	9.22	30.02	42.32
Median	35.57	2.73	7.04	30.76	41.20
Q3	38.14	4.14	12.27	33.87	43.13
SD	6.34	1.16	4.10	6.53	6.67
Maximum	56.77	5.62	23.33	47.22	66.42

children are found higher in *Dowarabazar* and *Bishwambarpur upzila*s though they are located very close to the *Sunamganj Sadar* and *Chhatak upzila*s. In BBS, WFP, and IFAD [10] study, the prevalence of underweight varied within the range of 39.3–42.5%: highest in *Dharampasha* and lowest in *Sulla upzila*.

The interactive map of underweight at union level shown in Fig. 20.3 indicates that there are significant inequalities in underweight across the unions. A significant number of unions are found as hotspots of higher underweight over the north-eastern and north-western parts of *Sunamganj* district. Though the southern part was homogeneous (light-shaded region) in the *upazila*-level map, significant number of heterogeneous unions are found at this region in the union-level map.

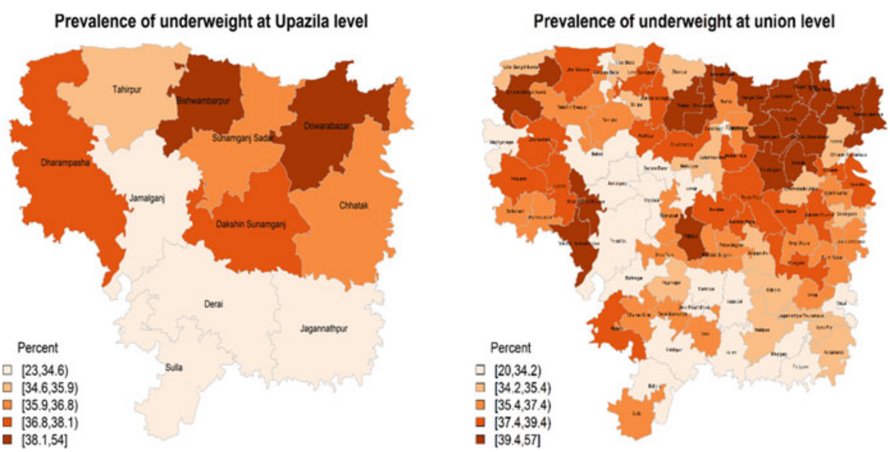


Fig. 20.3 Map of underweight among under-5 children at *Upzila* and Union levels in *Sunamganj* District using SAE estimator

20.5 Conclusion

Children of *Sunamganj* district are highly vulnerable to undernutrition. Near about two-fifths and half of the under-5 children in *Sunamganj* district were underweight and stunted respectively in 2011 [10]. The findings of this study indicate that near two-thirds and half of the under-5 children are underweight and stunted respectively in the *Sunamganj* district in 2018. Only one *upzila Jamalganj* has underweight level below 30%, while all the *upzila*s have stunting level above the threshold of 40%. The children of *Bishwambarpur*, *Dharampasha*, *Dowarabazar*, and *Tahirpur upzila*s were more stunted (more than 50%) than those in other *upzila*s. Most of the unions of these four *upzila*s are highly vulnerable to child stunting. On the other hand, the proportion of underweight children was found higher among the children of *Dowarabazar* (more than 50%) followed by *Bishwambarpur* (near about 40%). Thus, *Bishwambarpur* and *Dowarabazar upzila* are in great concern in terms of the child undernutrition indicators. However, the union level map indicates that the children belonging to the unions of north-eastern part (combination of *Dowarabazar* and *Chhatak*) are more vulnerable to underweight while those belonging to north-western part (mainly *Dharampasha upzila* except *Madhayanagar* union) are vulnerable to stunting. Overall, it can be said that children belonging to the unions of top half of the *Sunamganj* district map are experiencing chronic undernutrition. More than 40% children belonging *Dowarabazar* and *Bishwambarpur upzila*s were found underweight, while near about 50% children of *Tahirpur upzila*s were found stunted followed by the *Bishwambarpur upzila*s in BBS, WFP and IFAD [10] study. Findings of this study seems reasonable in comparison to 2014 study where 2011 household survey data were utilized. Though the findings of these two studies are difficult to compare, the findings of this study do not show any improvement in the undernutrition status among under-5 children of *Sunamganj* district without considering the limitation of this study.

The main limitation of this study is the time gap of recent *Sunamganj* survey data and the 2011 census data. The application of the SAE method to child morbidity and undernutrition data may be in concern due to this time gap since the distributions of a number of explanatory variables have been changed during the time period 2011–2018. Thus, proper care has been taken in model specification so that the district level estimate of child undernutrition indicators calculated by SAE estimator remain very similar to those obtained by the design-unbiased and consistent direct estimator.

Appendix (Tables 20.8 and 20.9)

Table 20.8 Multilevel model for HAZ considering children at level-1 and mauza at level-2, 2018 SCMNS[a]

Variables	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	−5.13	0.90	59.54	−5.68	0.00
Urban	−0.31	0.33	54.61	−0.94	0.35
Household head age	0.01	0.00	1143.78	2.46	0.01
Child age 0 year	0.88	0.13	1218.53	6.73	0.00
Child age 1 year	0.52	0.13	1216.34	4.05	0.00
Katch/jhupry house	−0.40	0.12	928.50	−3.40	0.00
pemp15_un	6.83	3.04	52.45	2.25	0.03
pcpred_un	3.66	1.31	52.60	2.80	0.01
pspucka_un	2.48	1.19	54.34	2.08	0.04

[a]pemp15_un: Proportion of 15+ persons employed in union; pcpred_un: Proportion of household head completed primary school in union; pspucka_un: Proportion of semi-pucka household in union

Table 20.9 Multilevel model for WAZ considering children at level-1 and *mauza* at level-2, *Sunamganj* survey 2018[a]

Variables	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	−1.67	0.21	65.12	−7.86	0.00
Urban	0.04	0.14	58.35	0.30	0.76
Dowarabazar upzila	−0.60	0.19	73.20	−3.09	0.00
Jamalganj *upzila*	0.63	0.18	65.59	3.46	0.00
Child age 0 year	0.62	0.10	1313.71	6.18	0.00
Non-sanitary or no latrine	−0.23	0.09	421.59	−2.45	0.01
Katch/jhupry house	−0.37	0.10	623.16	−3.81	0.00
pscpred_un	11.46	5.46	58.02	2.10	0.04
pnotoilet_mu	1.45	0.62	63.91	2.32	0.02

[a]pcpred_un: Proportion of household head completed primary school in union; pnotoilet_mu: Proportion of hh with unsanitary toilet in mauza

References

1. Kazal, M.H., Villinueva, C.C., Hossain, M.Z., Das, T.K.: Food Security Strategies of the People Living in *haor* Areas: Status and Prospects NFPCSP Final Report PR# 3/8. NFPCSP, Dhaka (2010)
2. FPMU: Chronic Food Insecurity Situation in 10 Districts of Bangladesh, December 2015–2018/20. Food Planning and Monitoring Unit (FPMU). Ministry of Food of Bangladesh in collaboration with country IPC Partners, Dhaka (2016)
3. NIPORT, Mitra and Associates, ICF International: Bangladesh Demographic and Health Survey 2014. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International, Dhaka, Bangladesh and Maryland, U.S.A (2016)
4. NIPORT, Mitra and Associates, Macro International: Bangladesh Demographic and Health Survey 1996–1997. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International. Dhaka, Bangladesh, and Maryland, U.S.A (1997)
5. NIPORT, Mitra and Associates, ORC Macro: Bangladesh Demographic and Health Survey 1999–2000. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ORC Macro, Dhaka, Bangladesh, and Maryland, U.S.A (2001)
6. NIPORT, Mitra and Associates, ORC Macro International: Bangladesh Demographic and Health Survey 2004. National Institute of Population Research and Training; Mitra and Associates; ORC Macro International. Dhaka, Bangladesh, and Maryland, U.S.A (2005)
7. NIPORT, Mitra and Associates, Macro International: Bangladesh Demographic and Health Survey 2007. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International, Dhaka, Bangladesh and Maryland, U.S.A (2009)
8. NIPORT, Mitra and Associates, ICF International: Bangladesh Demographic and Health Survey 2011. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International, Dhaka, Bangladesh and Maryland, U.S.A (2013)
9. BBS, UNWFP: Local Estimation of Poverty and Malnutrition in Bangladesh. Bangladesh Bureau of Statistics and United Nations World Food Programme, Dhaka (2004)
10. BBS, WFP, IFAD: Small-Area Estimation of Child Undernutrition in Bangladesh. Bangladesh Bureau of Statistics, World Food Programme and International Fund for Agricultural Development, Dhaka (2014)
11. BBS: Preliminary Report of the Household Income & Expenditure Survey, vol. 2016. Bangladesh Bureau of Statistics, Dhaka (2017)
12. Elbers, C., Lanjouw, J., Lanjouw, P.: Micro-level estimation of poverty and inequality. Econometrica. **71**(1), 355–364 (2003)
13. WHO Multicentre Growth Reference Study Group: WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development. World Health Organization, Geneva (2006)
14. Molina, I., Rao, J.N.K.: Small area estimation of poverty indicators. Can. J. Stat. **38**(3), 369–385 (2010)
15. Tzavidis, N., Salvati, N., Pratesi, M., Chambers, R.: M-quantile models with application to poverty mapping. Stat. Methods Appl. **17**(3), 393–341 (2008)
16. Elbers, C., Lanjouw, J.O., Lanjouw, P.: Micro-Level Estimation of Welfare World Bank Policy Research Working Papaer, 2911. The World Bank, Washington, DC (2002)
17. Das, S., Chambers, R.: Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw. J. R. Stat. Soc. A. Stat. Soc. **180**(4), 1137–1161 (2017)
18. Greven, S., Kneib, T.: On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika. **97**(4), 773–789 (2010)
19. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. **4**(2), 133–142 (2013)
20. Cochran, W.: Sampling Techniques. Wiley, New York (1977)

Chapter 21
Using a Spatial Farm Microsimulation Model for Australia to Estimate the Impact of an External Shock on Farmer Incomes

Yogi Vidyattama and Robert Tanton

Abstract A greater uncertainty in climate conditions in Australia and external price shocks in commodity prices has posed a real question for communities on the impact of these external factors on farmers. Spatial microsimulation models are ideal for understanding the spatial impacts of various external shocks, including changes in commodity prices; changes in climate conditions; and changes in Government policy. This study demonstrates the building of a spatial microsimulation model to identify farmer financial stress in the Australian State of Victoria, and then shows how this model can be used to estimate the impact of an external shock such as a drop in the price of milk. The model is estimated for the Australian State of Victoria.

Keywords Microsimulation modelling · Farmer wellbeing · Agricultural policy

21.1 Introduction

Agricultural support payments such as the Australian Farm household allowance package help to maintain the wellbeing of farmers and the continuity of the agricultural sector in Australia in periods of extreme drought [7]. The importance of these payments will only increase in the future due to increasing pressure placed on farmers as a result of greater climate uncertainty and changes in prices of commodities. Therefore, there is an increasing need to develop models that can analyse farmer financial stress, how different external factors like climate variability and commodity prices affect financial stress, and how different assistance options may help to reduce this.

Y. Vidyattama (✉) · R. Tanton
National Centre for Social and Economic Modelling (NATSEM), University of Canberra, Canberra, Australia
e-mail: Yogi.Vidyattama@Canberra.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_21

Microsimulation models have been used extensively to test new tax transfer policies in terms of the eligibility rules for transfer policies, the effects on financial stress for different groups in the population, and the cost to the Government of different policy settings. Microsimulation models can help a Government target its policy to the groups of people who most need assistance as it can provide some idea of who will benefit from a change in various external conditions such as climate, markets or government policy. An important aspect of this modelling for farming communities is the spatial aspect of the model because the impact of any change in external factors often varies by location. One reason for this is that there are different conditions in each area so the impact of the external shock is not distributed evenly across Australia. Another reason is that there are various types of farming activity which could be affected differently by the change, for example, a reduction in water availability in an area would affect the output of different crops in different ways, or a policy change may be targeted at particular types of farming or particular locations.

This study aims to develop a microsimulation model at a relatively small area level that will identify where farmers are likely to endure financial stress and also estimate the impact of an external shock that may affect a farming community.

21.2 Spatial Microsimulation and Its Application to the Farm

Since the emergence of microsimulation by Orcutt [29], microsimulation models have been used extensively to measure the impact of government policy on the income distribution and wellbeing. These government policies have included tax and benefit policies [25, 42]. The microsimulation technique has also been used to analyse the agricultural sector. In Australia, Kokic et al. [19] developed a microsimulation model to simulate individual level farm business performance for a large sample of Australian broad acre farms. In particular, they built a farm specific supply model for each farm in the Australian Agricultural and Grazing Industries Survey (AAGIS) with the assumption of profit maximisation of the farm business. The model was able to simulate the changes in farm income due to the change in the determinants that they used. These determinants included changes in commodity prices and the yield produced. Kokic et al. [19] indicated that the model could accommodate other inputs that can affect farm profit such as subsidies and taxes. The model could also be used to measure the direct impact of reducing different farm supports on the income distribution of different types of farm and family as shown by Menon et al. [26] in Italy.

As stated in the introduction, one important aspect of microsimulation is the availability of the spatial distribution of the individual unit records being modelled. In essence, a spatial microsimulation model provides synthetic unit record data for each small area to be analysed. Spatial microsimulation typically starts with a survey that provides individual or household unit records, and a census or other administrative data that provides various known conditions in the small area (based on aggregate small area data) as benchmarks.

The next step is a reweighting process which involves assigning different weights to a unit record survey to make it more representative of the area. Techniques used for this reweighting include iterative proportional fitting [5, 41], generalised regression [45], combinatorial optimisation [50] as well as maximum entropy [21].

Spatial microsimulation methods have increasingly been used to derive small-area estimates of a range of economic and social indicators. In Australia, they have been used to analyse the impact of government policy and the need for government services at a small-area level [13]. Vidyattama et al. [48] estimates not only the spatial distribution of superannuation savings in Australia but also analyses the different level of superannuation ownership between genders in certain age groups. The model has also been further developed to allow demographic projections for small areas. Examples of these types of projections are small area modelling of the need for different types of care [24] the projection of small area obesity among children [31] and small area health-related conditions [6].

Spatial microsimulation models have also used farm unit record data. Ballas et al. [6] tested the capability of a spatial microsimulation model to assess the impact of the common agriculture policy reform in Ireland. This policy breaks the link between direct payments and production for a farm so a single farm payment is paid per hectare. Ballas et al. [6] show how the reforms created winners and losers in Irish agriculture that may affect long term growth plans for Ireland. Shrestha et al. [38] took the simulation further by analysing the impact specifically on beef and sheep farming and introduced some behavioural responses. The results showed that while most beef farmers receive higher incomes, the majority of them could be de-stocking, with the exception of those in the Mideast and Southeast of Ireland.

The development of spatial microsimulation for farms in Ireland has also allowed an assessment of how participation in government programs may affect greenhouse gas emissions in different locations [16, 20]. Hynes et al. [15] has quantified the type of habitats being protected under the Rural Environmental Protection Scheme (REPS) in Ireland, while Murphy et al. [28] analyses how topological conditions may affect a farmer's decision to take part in the program.

The development of farm spatial microsimulation outside Ireland has been much slower. The most notable application is van Leeuwen and Dekkers [46] examining the spatial patterns of off farm income distribution in the Netherlands. Ramilan et al. [32] has also used a spatial microsimulation to show that farmers in New Zealand were likely to increase efficiency to reduce costs and pollution in terms of nitrogen discharge from dairy farming.

The main challenge in the development of farm microsimulation models is the lack of farm level data for the model and the lack of availability of suitable benchmarks at the small area level. This is also the main reason that a spatial microsimulation model using the farm as the unit of observation has not been attempted in Australia.

21.3 Data and Methodology

As mentioned in the introduction, the main part of a spatial microsimulation model distributes unit record data to small areas in a way that accurately reflects the small area totals (benchmarks). In this study, the technique to distribute the unit record data is to assign weights to each unit record (in this case farm) that indicates how much each farm represents all the farms in the small area. This is known as reweighting.

Figure 21.1 shows the structure of the model in our study. It shows that the main part of the model, the reweighting process, needs to be done after the variables within the databases are synchronised. Synchronising the unit record data and the benchmarks is the process of ensuring the definitions and categories used are exactly

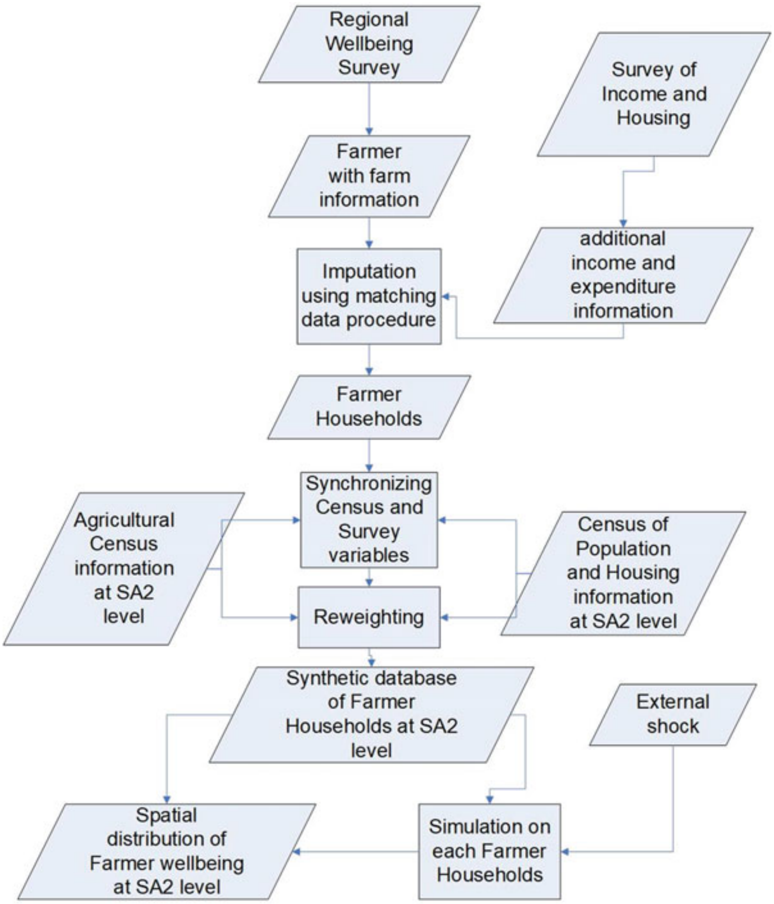


Fig. 21.1 Flowchart for the farm microsimulation model

the same. The synchronisation often involves not only understanding the definition of variables used in the various databases, but also which classifications (eg, age groups) are used.

Figure 21.1 also shows an imputation process in one of the steps. Although it is not a necessary step, imputation can be used to merge indicators from two different datasets to enrich the final database with more information. This is discussed further in Sect. 3.2.

21.3.1 Data

Survey Unit Record File Spatial microsimulation models depend on the availability of unit record data or the possibility to create a synthetic unit record data that can be utilised instead [49]. Ideally for a farm microsimulation model, the unit record survey should contain household information from an agricultural census or a specific agricultural survey. If information on the farm business is available, then a household survey can be linked to the farm business survey, providing information on the socio-demographic characteristics of the households, as well as financial information on the farm business. If a farm business survey is not available, the next best option is to build the household database and impute the farm business information from other sources (eg, publications).

This work uses the Regional Wellbeing Survey (RWS) produced by the University of Canberra [37], for which results were launched on the 18th June 2014. This dataset contains around 8000 observations from 2014, with approximately 2500 responses from farmers. In Victoria, where this model was tested, there were a total of 1605 survey responses, of which 537 were farmers (see [33] to see more discussion on how this database can be used).

The dataset has a wide range of variables including household income, gross value of agricultural output as well as the proportion of off-farm and on-farm income, necessary for the simulation. However, one issue with this survey is that two farmers surveyed may come from the same household and hence the same farm. Currently this cannot be tested as we do not have the address of the farm due to confidentiality restrictions.

The observations used from the RWS are only those where the respondent has identified as a farmer, is managing or co-managing a farm enterprise or is a partner of a farmer and they assist with farm work or management.

Benchmark Dataset For a spatial microsimulation model, the second dataset required is a dataset that can provide socio-demographic estimates reliably for small areas. The ABS Census of Population and Housing is the largest Australian data collection that tries to capture all people living in the country. The census data contain person characteristics such as age and sex, cultural and language diversity, disability and carer status, children, employment and income in ranges, education and qualifications as well as relationship in the household.

The census is used to benchmark the survey dataset in the spatial microsimulation model but can also be used for estimating the characteristics of the farmer or farm labourers who work on a particular farm in the area, using the industry of employment code available on the Census. This data will allow the use of probability tables to allocate workers to different farm businesses. The 4-digit classification includes the industries wheat and other crops, mixed livestock-crops, sheep, beef, combined sheep-beef and dairy farms. The Census used for this work was the 2011 Census.

Another dataset that has been used for the small area benchmarks is the agricultural census. This dataset provides a comprehensive snapshot of the overall condition in the Australian agricultural sector for small areas. This snapshot includes information on agricultural production, water and land management [1]. The main use of this dataset is to assign weights to the different farm businesses in the survey so the weighted aggregate value is the same as the census value. To be able to do this, the benchmark variables on the census need to be defined in the same way as the variables in the survey which is being benchmarked (the synchronisation process described above). Again, the 2011 Ag Census was used.

All the benchmarks were extracted using the Statistical Area Level 2 (SA2) as the spatial unit. These areas are medium-sized general purpose areas that represent a community that interacts together socially and economically. In Australia, SA2 areas can represent suburbs in capital cities, but larger areas outside capital cities. The SA2 is often used for analysis as it is large enough to provide reasonable results but small enough to not suffer from aggregation bias [2].

The benchmarks used for this work are shown in Table 21.1.

For technical reasons, all of the values in each table were adjusted so that the totals matched the total number of Farmers and Farm Managers in the “Age group by Sex” benchmark table for each SA2. This Age and Sex table is used as the benchmark because it is the least perturbed by the Australian Bureau of Statistics confidentialising process, and there is no “not stated” category in this table. The Not Stated category is a category used in most Census variables where the respondent does not answer a question. This adjustment to the Census benchmark tables was fairly minor.

Table 21.1 Benchmarks for the spatial microsimulation

No.	Table
1	Household Income for Farmers and Farm Managers
2	Detailed Industry of Employment for farm classification
3	Estimated Value of Agricultural Operations
4	Age group by Sex for Farmers and Farm Managers
5	Family Household Composition for Farmers and Farm Managers
6	Non School Qualification level for Farmers and Farm Managers

21.3.2 Methodology

The first step in spatial microsimulation is to synchronise the census and survey variables so that the variables from each dataset have the same classifications, for example, matching age groups, income groups, etc. This is essential for correct benchmarking. The benchmarked variables will be available after the modelling and can be estimated from the final benchmarked survey reliably.

The adjustment made to the industry that the farmer is working in is more difficult than the demographic variables. The question from the RWS used is whether certain activities of the farm were the major use of the land. The four-digit detailed Industry of Employment was used to help classify farms. There are 64 groups in the Agriculture, Forestry and Fishing industry that are then summarised into 9 groups. For example, horticulture includes growing mushrooms, various vegetables, fruits and nuts while grains include rice, other grain and other crop growing not elsewhere classified. There are combined classes as well such as Grain-Sheep or Grain-Beef Cattle Farming, while Dairy also has its own industry grouping.

Given this grouping for the census data, the RWS was first checked to see whether the farmer's major land use was grain. If this was the case, and they also had beef or sheep activities as their major land use, then the farmers were classified as grain-beef or grain-sheep. If they didn't have any sheep or beef activity as the major land use, then they were classified as a grain farmer.

The next step was to check whether they had dairy activities or horticulture as their major land use to be able to classify them as dairy or horticulture farmers respectively. After exhausting these options, the specialised sheep and then beef farmers were identified before moving to other cropping and other agriculture activities. This means there is an implicit prioritisation in the identification process from grain-beef or grain-sheep; grain; dairy; horticulture; sheep; beef; other cropping and finally, other agriculture. To ensure the suitability of this classification, the number of farmers allocated to these groupings from the census and RWS were compared. The comparison of the proportion of those identified for Victoria was quite close, and even closer for Australia.

Further explanation also needs to be given on one benchmarking variable that comes not from the Census of Population and Housing but instead from the agricultural census. The variable is the Estimated Value of Agricultural Operations (EVAO). The data contains the number of businesses in a certain EVAO range. The variable from the RWS to match this variable was the Gross value of agricultural production, which was equivalent to the EVAO and was placed into the same ranges as the agricultural census.

The proportion of farmers in each farm type and the value of agricultural production are the two tables where ensuring the proportions from the agricultural census and RWS are similar is crucial. For the farm type table, a similar proportion means that the regrouping of variables that are slightly different between the census and RWS is acceptable. Table 21.2 confirms that the grouping provides similar proportions of farmers in different categories except for dairy. This is likely to be

Table 21.2 Comparison of the proportion of farmers in different categories between the census of population and housing and the RWS

	Census		RWS	
Category	Number	Proportion	Number	Proportion
	Victoria			
Grains	2961	8.6	46	12.2
Horticulture	4241	12.3	36	9.6
Sheep	6084	17.6	72	19.1
Beef	6898	20.0	105	27.9
Grain-beef or sheep	4075	11.8	50	13.3
Dairy	8330	24.1	52	13.8
Other agriculture	1977	5.7	15	4.0
Not Further Defined	853		111	
	Australia			
Grains	14,491	10.5	151	14.4
Horticulture	19,220	13.9	112	10.7
Sheep	21,548	15.6	203	19.4
Beef	34,091	24.7	268	25.5
Grain-beef or sheep	23,152	16.8	216	20.6
Dairy	13,399	9.7	59	5.6
Other agriculture	7472	5.4	37	3.5
Not Further Defined	4579		306	

Note: Farm categories with very low observations in the RWS were removed. This only covered "Other Cropping"

the result of the differences in the definition of the industry of employment category with the activity question in the RWS separating major and minor land use. On some farms, dairy is the major income earner even if it is a small area of the farm. Therefore, even though dairy was prioritised above beef in the classification, there is a possibility that a beef farmer should be classified into dairy according to industry of employment.

A similar proportion between the agricultural census and RWS for the value of agricultural production is also important. Given the value of agricultural production was based on the number of firms rather than farmers, the distribution of farmers in different production ranges should be similar. This is because the number of farmers in the benchmark is calculated using the proportion of firms in each production range to the total number of farmers. This can only be done with the assumption that the distribution of farmers in firms of different production ranges is similar. Table 21.3 shows that the proportion of farms in each range in the agricultural census and the RWS is similar. In the RWS, the proportion of Victorian farms with production value less than $50,000 is slightly higher than the estimate from the agricultural census but this could mostly be attributed to the lower proportion in the next class ($50,000–$100,000).

Table 21.3 Comparison of the proportion of farmers with different production value between agriculture census and RWS

	Census		RWS	
Category	Number	Proportion	Number	Proportion
	Victoria			
Less than 50,000	14,478	37.4	149	43.6
50,000–200,000	11,056	28.6	69	20.2
200,000–500,000	7590	19.6	60	17.5
500,000-1 million	3656	9.4	34	9.9
More than 1 million	1938	5.0	30	8.8
	Australia			
Less than 50,000	55,340	35.6	357	37.9
50,000–200,000	45,074	29.0	217	23.0
200,000–500,000	28,473	18.3	161	17.1
500,000-1 million	15,597	10.0	104	11.0
More than 1 million	10,757	6.9	103	10.9

One crucial process in spatial microsimulation is distributing the observations (in this case farmers) farmers from the survey data to each of the small areas. In this model, this is done by assigning weights to each observation that are calculated to make the weighted total of the observations to be as close possible to the aggregate small area benchmark tables described above. This is done for each of the SA2s.

In the spatial microsimulation method, one test of whether an area provides a reasonable estimate is to estimate the total population for a particular category from the model (eg, age by sex) and test how close it is to the population for the same category from the population Census. If the modelled population is too different from the Census population, the area is rejected.

The test is called the total absolute error (TAE) and it is calculated from all the benchmarks. The accepted error condition for the TAE is when the sum of the absolute differences between the census and the spatial microsimulation results are less than the total population of that area. The TAE has been used in a number of spatial microsimulation models as a criterion for reweighting accuracy [3, 52] and has been supported by other studies such as Smith et al. [40] and Voas and Williamson [50].

In this case, the spatial microsimulation procedure produced reasonable estimates for around 73% of the SA2's. These SA2s cover almost 96% of the farmers in the census (see Table 21.4).

There are another two benchmark tables that could be used – education and household composition. Adding either one of these benchmarks reduced the number of reasonable estimates for SA2's (Table 21.4) and Farmers (Table 21.5) considerably. Adding both these tables meant that the proportion of acceptable SA2s was 41% of all SA2s and around 45% of farmers. Further, these two benchmarks were useful, but not essential, for this project, so they were not used.

Table 21.4 Proportion SA2s passing TAE test

Row Labels	Number of SA2s	Proportion of SA2s (%) using 4 benchmark	Proportion of SA2s (%) using 5 benchmark incl. HH composition	Proportion of SA2s (%) using 5 benchmark incl. Education	Proportion of SA2s (%) using 6 benchmark
Ballarat	17	100.0	76.5	70.6	41.2
Bendigo	16	87.5	75.0	56.3	25.0
Geelong	20	80.0	55.0	45.0	40.0
Hume	22	95.5	90.9	95.5	77.3
Latrobe – Gippsland	27	100.0	92.6	92.6	59.3
Melbourne – Inner	89	66.3	58.4	60.7	50.6
Melbourne – NE	54	68.5	51.9	57.4	46.3
Melbourne – OE	36	61.1	41.7	44.4	36.1
Melbourne – SE	41	58.5	34.1	31.7	24.4
Melbourne – W	42	42.9	35.7	38.1	33.3
Mornington Pen	16	75.0	31.3	31.3	18.8
North West	20	95.0	75.0	65.0	25.0
Shepparton	14	85.7	71.4	71.4	35.7
Warrnambool and SW	14	100.0	100.0	92.9	50.0
Grand total	428	72.9	58.2	57.7	41.8

Another important process in this model was imputation. Schafer [36] defines imputation as filling in missing data with plausible values. While there are many imputation techniques such as categorical average, random selection, classification based and regression based (see [4, 39, 51] for a discussion), this model uses the application imputes data by finding similar observations from different data sources and then using the variables available in those sources [17]. The reweighting described above has provided a database that can estimate the pattern of various RWS variables at the SA2 level, but we often need other variables to conduct our analysis. For example, to calculate poverty rates, we needed income, which was only available in groups on the regional wellbeing survey. An imputation process using the 2009/10 ABS Survey of Income and housing was able to provide this variable to the model.

The imputation process can add a continuous income variable and the exact number of children, both of which were necessary to calculate poverty rates using household equivalised disposable income. Other variables that could be added from the ABS Survey of Income and Housing (SIH) include the amount of government benefit that could be used to calculate how much the current government benefit contributes to farming areas. There are 80 farmers in the SIH in Victoria that can be matched with the 487 observations of Victorian farmers in the RWS.

Table 21.5 Proportion of Farmers living in SA2s that passed the TAE test

Row Labels	Number of farmers	Proportion of farmers (%) using 4 benchmark	Proportion of farmers (%) using 5 benchmark incl. HH composition	Proportion of farmers (%) using 5 benchmark incl. Education	Proportion of farmers (%) using 6 benchmark
Ballarat	1870	100.0	92.5	91.3	70.4
Bendigo	2022	98.2	93.3	81.1	26.8
Geelong	1156	95.9	85.0	72.9	69.5
Hume	4278	99.4	96.4	99.4	92.6
Latrobe – Gippsland	6292	100.0	93.7	94.3	65.7
Melbourne – Inner	416	48.8	22.8	34.9	0.0
Melbourne – NE	638	82.6	44.2	63.5	26.8
Melbourne – OE	858	74.9	31.9	29.1	3.0
Melbourne – SE	1477	84.6	54.6	56.4	46.6
Melbourne – W	508	57.3	2.4	6.1	0.0
Mornington Peninsula	593	70.7	6.9	36.8	0.0
North West	7829	96.9	84.8	72.1	17.5
Shepparton	4491	96.9	86.9	81.1	32.1
Warrnambool and SW	6559	100.0	100.0	98.4	73.6
Grand total	38,987	95.8	85.2	82.1	49.5

Several variables are available in both databases for this matching process, including household income, age, sex, education and family composition. The matching process is done by sequentially merging the data from the SIH to the RWS.

The process starts using all five variables – Household income, age, sex, education and family composition – with the classes or ranges used in the reweighting process. This matched only 23% of the data. The second iteration used the remainder of the SIH that had not been matched but this time, education was taken out as one of the matching variables. The second iteration adds the SIH variables to a further 25% of the RWS data. The merging process is then re-run on the remainder of the RWS but using the initial subset of agricultural managers (farmers) from the SIH. This third iteration only matches the SIH variables to 6% of the RWS data. This means 55% of data has been matched using at least Household income, age, sex and family composition.

These remaining four variables are important in ensuring we achieve a good imputation of the value of government tax and benefits from the SIH. Therefore, no variable was taken out in the fourth iteration. Instead, the farmer age group was adjusted to use only two groups – under 65 and 65 and over. As was done in the previous iterations, the variable was first selected from the remaining SIH observations from the previous iteration before using the entire SIH sample. These steps add the SIH variables to an additional 19% of the RWS. Similar steps are taken for the sixth and seventh iterations, but this time it is the income ranges that are widened from five income groups to three income groups. This adds another 20% of the RWS matched so at this stage 95% of the RWS dataset has been matched. To finish off the process, the income and then the age variables are removed from the matching variables to get a final match for the remaining farmers (only 5% of the farmers).

21.4 Estimating Poverty Rates and Validation

One of the main uses of the model is small area estimation. In Australia, spatial microsimulation has been used to analyse the spatial distribution of socio-economic conditions such as poverty and housing stress [27, 30]. This study will use the base dataset of weighted small area data on farmers with imputed income from the SIH to calculate a poverty rate as an estimate of financial stress.

The poverty rate in Australia is commonly calculated as the proportion of people living in households where the equivalised disposable household income is less than half the national median. Equivalised income is calculated using the modified OECD method which applies an equivalising factor to the disposable income, using a weight of 1 for the first adult in the household; 0.5 for each additional person aged 15 and over; and 0.3 for each person in the household aged under 15. This is a standard poverty rate used in Australia [12, 35, 43]. The median equivalised disposable household income used for this work was a national median of $619.40, and the poverty line was therefore $309.70 per week.

Figure 21.2 shows the poverty rate for dairy farmer households. The figure shows that higher poverty rates for farmers were identified in the Mildura area (the north part of the North West region in the top left corner of the figure), as well as some areas in the Bendigo area. In addition, the area surrounding Melbourne (especially in Melbourne North but also to Geelong in the west and part of Dandenong in the east) also experiences high farmer poverty rates.

The finding of a relatively high poverty rate in areas surrounding Melbourne, especially Melbourne North, is in line with the poverty rate of the overall population. The number of farmer households is relatively small in these areas, while sectors such as health care, retail trade and construction are more dominant. On the other hand, the finding in Mildura indicates that the farmer poverty rate is slightly higher than normal while this area is not known as a poor area.

Another important measure of socio-economic condition is a measure of well-being of the farmer. This can be done by estimating financial stress using the question from the RWS "Overall, how do you feel your farm business is going? -My farm enterprise is under a lot of financial stress at the moment". Responses to this question gave 7 choices ranging from strongly disagree to strongly agree. The farmers in financial stress are those who chose between 5 and 7 while those who chose between 1 and 4 are not in financial stress. The results from the financial stress question estimated using the spatial microsimulation model for small areas across Victoria are shown in Fig. 21.3. The figure shows that the spatial distribution of

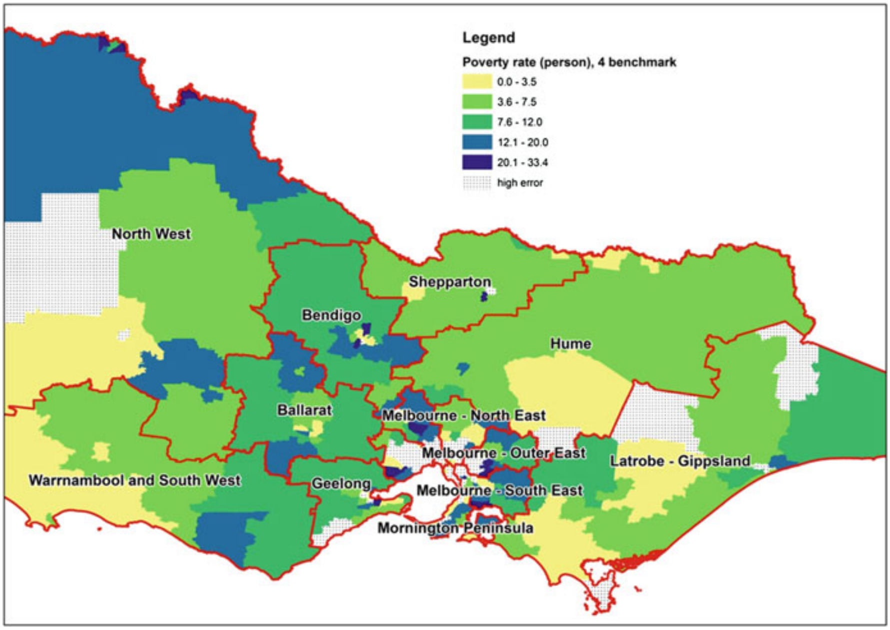


Fig. 21.2 Estimated poverty rate for dairy farmer households

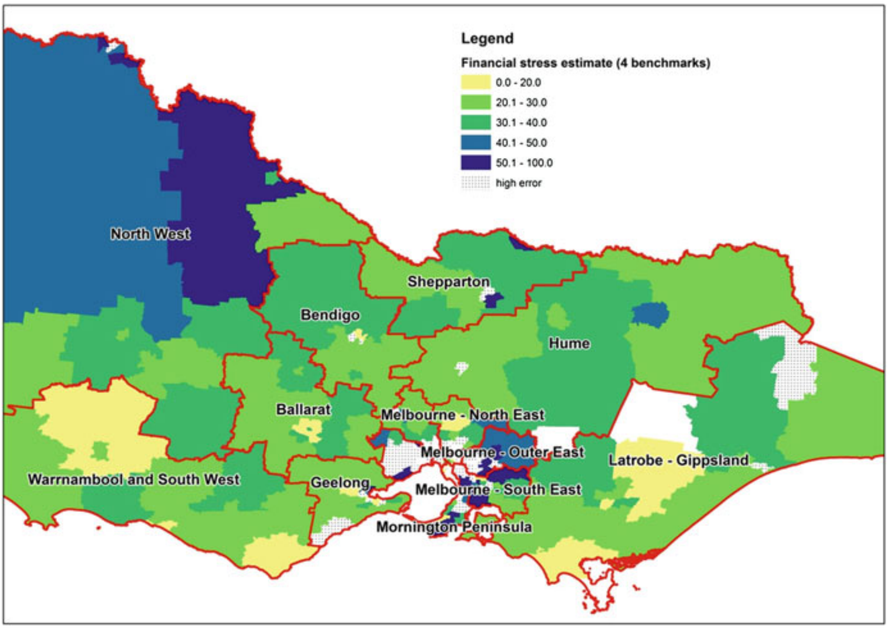


Fig. 21.3 Results for financial stress, Victoria

Table 21.6 Estimates of the proportion of farmers in financial stress at SA4 level

Row Labels	RWS sample (n)	RWS Estimate (%)	4 bench-mark (%)	5 benchmark incl. HH composition (%)	5 benchmark incl. Education (%)	6 bench-mark (%)
Hume	75	32.0	29.5	30.9	27.7	27.0
Latrobe – Gippsland	60	28.3	23.8	20.6	18.1	13.0
North West	62	50.0	40.1	34.3	30.2	5.9[a]

[a]Note: The proportion is estimated from the area that contain less than 50% of farmers (see Table 21.2)
Areas with very low observations in the RWS (n < 50) are not shown in the table

financial stress is relatively similar to the distribution of poverty. However, financial stress in the North West region went beyond Mildura. As mentioned previously, the estimate is based on 2011 data and this was when crop output had been declining for several years. Therefore, it is reasonable that the farming community may be feeling financial pressure even though their income is not below the poverty line yet. Furthermore, given these areas were not considered as poor areas, this financial stress estimate adds to the importance of spatial microsimulation estimates of financial stress for farms as a leading indicator of poverty for farmers.

Validation of these estimates of financial stress can be conducted using the aggregate estimates of financial stress at the SA4 level from the RWS and comparing these to the small area estimates aggregated to the SA4 level. This is a standard way to validate spatial microsimulation results [11]. It is important to note that a high level of non-response to the financial stress question means we have even fewer observations that can be used from the RWS and the estimation at the SA4 level from the RWS for some areas is based on a small number of observations. We have only presented results in this paper for SA4's where we know there are enough observations to provide reliable estimates from the survey, and the number of observations in the SA4 (n) is shown in the table.

Using a spatial microsimulation model, there is a trade-off between the number of benchmarks used in the model and the accuracy of the results. As the number of benchmarks decrease, the accuracy decreases, but the number of areas available to be used (measured through the TAE as outlined above) increases [44]. The results in Table 21.6 show that in three areas where the RWS has enough observations, the difference between our spatial microsimulation estimate using 4 benchmark tables is closer than the 5 and 6 benchmark results.

The validation does show that the model tends to under-estimate financial stress in all areas, by 4–5% points except in North West where it was nearly 10% points lower than the survey estimate. This SA4 was particularly difficult to estimate, and the other models provided even lower estimates than the 4 benchmark model. Therefore, the simulation with 4 benchmark tables is considered the best option for the analysis.

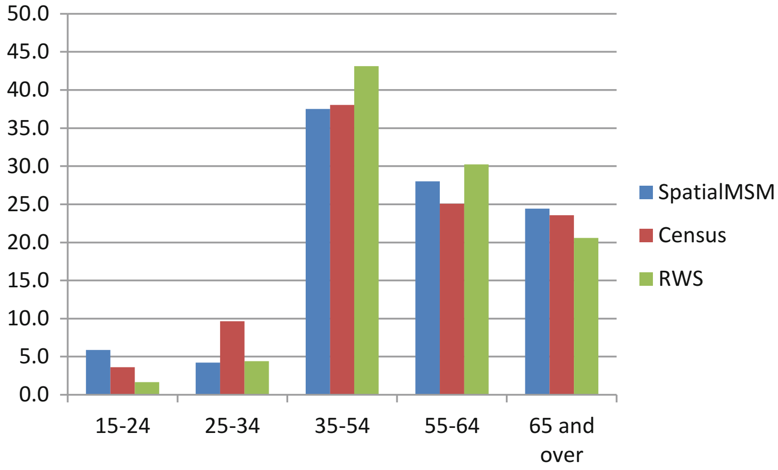


Fig. 21.4 Farmers' age structure in spatial MSM, RWS and census

As an additional step to validate the model, we looked at the RWS, the spatial microsimulation results and data from the population census for the age structure of Farmers in Victoria. The results are shown in Fig. 21.4. It can be seen that the age structure does not differ too much over the three datasets. However, the proportion of people aged 35–64 is higher in the RWS compared to census, possibly reflecting a bias in the RWS sampling (even though weighted estimates were used from the RWS). On the other hand, the proportion of farmers aged 15–24 in the census is higher than the RWS but lower than SpatialMSM. The proportion of people aged 64 and over in the RWS is slightly higher from the Census while the proportion of people aged 35–54 is slightly lower.

The difference between the Census and the modelled population is due to all of the other benchmarks apart from Age/Sex in the microsimulation model. In this case, household income, production value and type of farm are the benchmarks that are contributing to the difference between the census and spatial microsimulation results.

21.5 Applying an External Shock

Another way that the farm spatial microsimulation model can be used is to estimate the impact of various external shocks on the farming community. One way to produce this estimate is to link the unit data that represents the farmers in the microsimulation model as recipients of an external shock from another model or new policy. As mentioned by Hynes et al. [15] and Hopkins et al. [14], this simulation is meaningful if the impacts of these policies or external models will be different for each agent or observations in the micro data. The output from a model like a CGE or input-output based model can be used to supply the shock to the microsimulation model as the impact can be different for those working in different sectors of the economy or occupations or in some case even in different commodities. These impacts can then be directly applied to households in the microsimulation model. Examples where CGE models have been linked to microsimulation models include specific farm microsimulation [10, 18] or more general microsimulation [9, 34].

For this paper, the example of an external shock is a reduction in the milk price. This is a real scenario that has occurred in Australia from 2016, caused by an oversupply of milk and a reduction in demand for milk solids from China [23]. Three scenarios were created for this paper: a price drop of 5%, 10% and 15%. The price drop was based on the drop in milk solids price and was only applied to those with activity in the RWS in dairy farming. The impact of a change in the price of milk will be affected by how much milk the dairy produces for sale, and any off-farm income that can be used to reduce the impact of the price change.

The new income for each farm was calculated using the formula:

$$\text{New income} = \text{old income} - (\text{D\% x production/income})\ (1 - \text{\% off farm income})$$

Where D% is the percentage drop in the milk solids price. From the new and old incomes, we were able to calculate poverty rates before and after this change in milk price.

The construction of the model used several assumptions. The first assumption is that the dairy farms are predominantly dairy as the grain production that is often associated with dairy farms cannot be incorporated into the model. Second, there are no behavioural changes made by farmers (for example, changing to a different type of farming which provides greater profit, or increasing the amount of off farm income). Third, the poverty line (half of median equivalised income) is unaffected by the drop in farm incomes. This is a reasonable assumption as the small number of dairy farmers in Victoria would not affect the national poverty line much. Fourth, price is exogenous to the model. The decreased price is distributed to all dairy farmers given lack of information on the distribution channels and composition of their dairy products.

Figure 21.5 shows the impact of the three different changes in milk prices on poverty rates. This figure shows the original poverty rates (the same map as Fig. 21.2) but superimposes the impact of the milk price drops as a column chart. A lower column means the incomes of farmers in the area are less affected by the change in milk price. It is likely that the area is not a dairy farm area or most of the dairy farmers were already in poverty. There are some areas where the three different scenarios have different impacts while other areas where they have similar impacts. This indicates the variation of margins (income less costs) and size of the farms in those areas as larger farms have greater economies of scale. It can be seen that the impact of the price change on poverty rates for farmers is greatest in

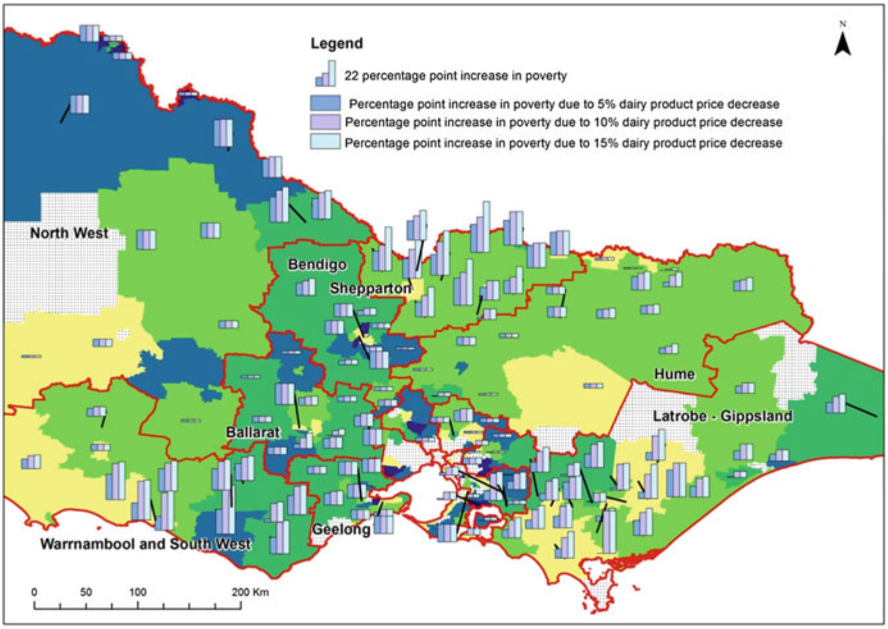


Fig. 21.5 Impact of change in milk price on poverty rates

the regions of Shepparton; Warrnambool and South West; and Latrobe-Gippsland. This is understandable as these regions contribute to approximately 22, 31.5 and 32 percent of dairy farmers in Victoria, respectively. In addition, more than 40 percent of farmers in these regions are classified as dairy farmers. As a comparison, the regions of Melbourne – South East; Hume; Bendigo and North West, where the proportion of dairy farmers among all farmer can be considered high, only have 13, 10.5, 5.8 and 5.5% of their farmers classified as dairy.

Figure 21.5 also shows whether the expected impact affects areas that already have a high proportion of farmers in poverty. Many SA2s in these areas had low poverty rates, but Corangamite-South in South-West near Warrnambool and Swan Hill region in the North-West had high poverty rates and a high increase in poverty. These are the areas where government social policy may need to focus on helping farmers, especially in Swan Hill where not only the margins of dairy products are low but also the financial stress expressed by farmers is estimated to be high. There is a big difference between Corangamite-South and the Swan Hill region. Corangamite-South is a dairy farming area with around 80% of farmers classified as dairy farmers. Therefore, many of the farmers that were considered poor were dairy farmers and the drop in the milk price will affect those who are already in poverty as well as those who previously were not counted as being below the poverty line. On the other hand, only around 5% of farmers in Swan Hill are dairy farmers. The relatively high increase in the poverty rate indicates that these dairy farmers were not considered as being in poverty before the milk price decreases.

The next step in this analysis was to identify farms that were operating at a loss after a drop in the price of milk by 15%. Although farmers in Australia have increased flexibility in facing fluctuating market prices, a large drop in price will be hard to manage [8, 47]. The proportion of farmers operating at a loss (revenue less than expenses) as a result of the milk price dropping by 15% is shown in Fig. 21.6. It can be seen that a high proportion of dairy farmers in West Victoria are operating at a loss after this milk price reduction, while a much lower proportion in East Victoria are operating at a loss.

A closer look at Fig. 21.6 also indicates that farms closer to Melbourne and close to the highway connecting Melbourne and Sydney are less affected by the fall in milk prices. This may partly be driven by the availability of off farm income in areas close to cities like Melbourne or highways, where access to cities and transport is available. One interpretation of this result is that milk production in the West and North-West of Victoria is unsustainable if the price continues to drop, and off farm income is needed to support the farm. Therefore, one important policy in dealing with the persistent drop in milk price is to create more opportunities for off-farm income.

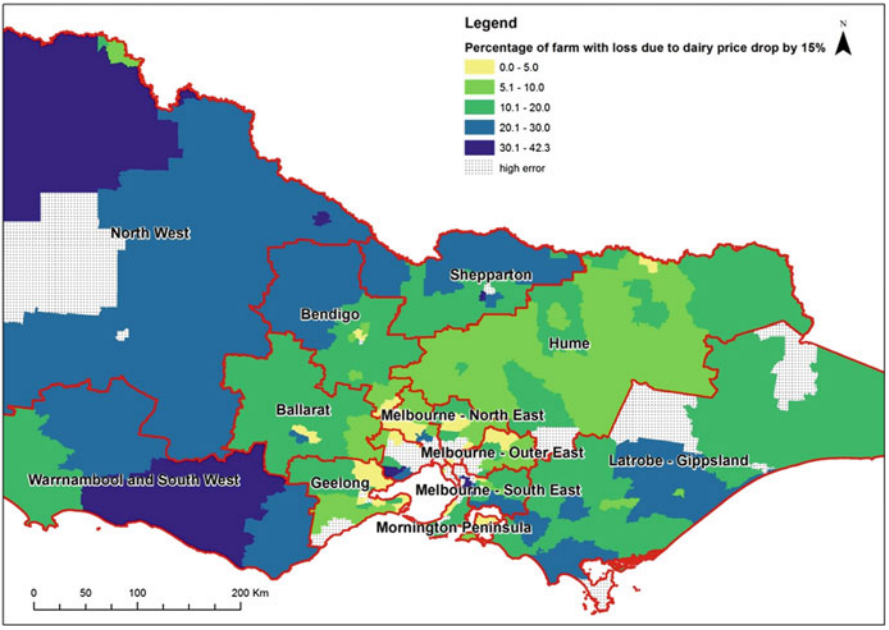


Fig. 21.6 Percentage of farms that are estimated to operate at a loss after a 15% reduction in the price of milk

The results have demonstrated the capabilities of the model in analysing the impact of an external shock on farm income. While the example used in this paper is a drop in the price of milk, there are other external shocks that could be modelled as long as the shock has an impact on the variables in the model. The link could be based on occupation and industry of employment of individuals in the unit record data. Therefore, the drop in production of different commodities due to factors such as climate change, reduced water availability (drought), or flooding can be introduced to the model. However, these would need to be modelled outside the microsimulation model, and then the impacts introduced as an external shock.

The estimation of the impact can also be extended by introducing a dynamic behavioural model. The current static model developed is able to model the short term impact, however there are behavioural changes which a farmer can make; and the decisions made by farmers are usually long term. Dynamic models are ideal for these longer term implications [22], and this is an area of future research for this model.

21.6 Conclusions

This paper has shown how a spatial microsimulation model can be built to analyse farms in Australia. The model has utilised the Regional Wellbeing Survey together with the ABS census of Population and Housing as well as the ABS Agricultural census. The existing variables in the survey also enable us to impute variables from other surveys in Australia. In this case, we have used the 2009/10 ABS survey of Income and Housing to impute income onto our base dataset. This provides us with a synthetic survey data that represents farms at a small area level.

This model is a powerful tool in identifying areas where farmers may need some support from the government or elsewhere. The results show that this is important as farmers' socio-economic disadvantage may be obscured by the condition of the general population. In addition, the model can also estimate the impact of an external shock that may affect a farming community. As an example, we have been able to identify the areas where farms are most are affected by the recent change in the milk price, and then estimate in which areas farms are likely to become financially unviable.

The results show that the simulated impact of the decrease in the price of milk on poverty is greatest in the areas of Shepparton, Warrnambool and Gippsland, which had low farm poverty rates, as well as Corangamite-South in South-West near Warrnambool and Swan Hill region in the North-West, that had high farm poverty rates. Farms closer to Melbourne and the highway between Melbourne and Sydney, which can offer more off farm income, are less affected by the fall in milk prices.

This application has shown that the model can make a significant contribution to farm policy, and with some development, could significantly add to the Commonwealth Government's ability to assist farmers, either through direct financial support, or through assistance in modelling different scenarios.

References

1. ABS.: Agricultural Commodities, Australia, 2010–11, Cat. No. 7121.0. Australian Bureau of Statistics, Canberra (2012)
2. ABS.: Australian Statistical Geography Standard (ASGS): Volume 1 – Main Structure and Greater Capital City Statistical Areas, July 2016, Cat. No. 1270.0.55.001. Australian Bureau of Statistics, Canberra (2016)
3. Anderson, B.: Creating small area income deprivation estimates for Wales: spatial microsimulation modelling. Chimera Working Paper. **11**, (2007)
4. Akande, O., Li, F., Reiter, J.: An empirical comparison of multiple imputation methods for categorical data. Am. Stat. **71**(2), 162–170 (2017)
5. Ballas, D., Clarke, G.P., Wiemers, E.: Building a dynamic spatial microsimulation model for Ireland. Popul. Space Place. **11**(3), 157–172 (2005). https://doi.org/10.1002/psp.359
6. Ballas, D., Clarke, G.P., Wiemers, E.: Spatial microsimulation for rural policy analysis in Ireland: the implications of CAP reforms for the national spatial strategy. J. Rural. Stud. **22**(3), 367–378 (2006)
7. Botterill, L.: Responding to farm poverty in Australia. Aust. J. Polit. Sci. **42**(1), 33–46 (2007). https://doi.org/10.1080/10361140601158534
8. Botterill, L.C., Chapman, B.: A revenue contingent loan instrument for agricultural credit with particular reference to drought relief. Aust. J. Labour Econ. **12**(2), 181 (2009)
9. Buddelmeyer, H., Hérault, N., Kalb, G., de Jong, M.V.Z.: Linking a microsimulation model to a dynamic cge model: climate change mitigation policies and income distribution in Australia. International Journal of Microsimulation. **5**(2), 40–58 (2012)
10. Cockburn, J. (2006). Trade liberalisation and poverty in Nepal: a computable general equilibrium micro-simulation analysis. In Globalisation and Poverty (pp. 189–212). Routledge
11. Edwards, K.L., Tanton, R.: Spatial Microsimulation: a Reference Guide for Users. Springer, London (2013)
12. Harding, A., Lloyd, R., Greenwell, H.: Financial Disadvantage in Australia 1990 to 2000: The Persistence of Poverty in a Decade of Growth. Canberra (2000)
13. Harding, A., Vidyattama, Y., Tanton, R.: Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation. J. Popul. Res. **28**(2–3), (2011). https://doi.org/10.1007/s12546-011-9061-6
14. Hopkins, J. W., Hanson, K., Somwaru, A., & Burfisher, M. E. (2003). Allocation effects of policy reform: a micro-simulation of macro-model results for the United States. (No. 1225-2016-98637)
15. Hynes, S., Farrelly, N., Murphy, E., O'Donoghue, C.: Modelling habitat conservation and participation in agri-environmental schemes: a spatial microsimulation approach. Ecol. Econ. **66**(2–3), 258–269 (2008)
16. Hynes, S., O'Donoghue, C., Morrissey, K., Clarke, G.: A spatial micro-simulation analysis of methane emissions from Irish agriculture. Ecol. Complex. **6**(2), 135–146 (2009). https://doi.org/10.1016/j.ecocom.2008.10.014
17. Iacus, S.M., Porro, G.: Missing data imputation, matching and other applications of random recursive partitioning. Computational statistics & data analysis. **52**(2), 773–789 (2007)
18. Keeney, R. (2009). Transfer Efficiency and Distributional Impacts of US Farm Support: Evidence from a Macro–Micro Simulation
19. Kokic, P., Chambers, R., Beare, S.: Microsimulation of business performance. Int. Stat. Rev. **68**(3), 259–275 (2000)
20. Kruseman, G., Blokland, P.W., Bouma, F., Luesink, H., Mokveld, L., Vrolijk, H.: Micro-simulation as a tool to assess policy concerning non-point source pollution: the case of ammonia in Dutch agriculture. In presentation at the 107th EAAE Seminar "Modelling of Agricultural and Rural Development Policies", vol. 29. Sevilla (2008, January)
21. Leyk, S., Nagle, N.N., Buttenfield, B.P.: Maximum entropy dasymetric modeling for demographic small area estimation. Geogr. Anal. **45**(3), 285–306 (2013)

22. Li, J., O'Donoghue, C., Dekkers, G.: Dynamic models. In: Handbook of Microsimulation Modelling, pp. 305–343. Bingley, Emerald (2014). https://doi.org/10.1108/S0573-855520140000293009
23. Lockhart, J., Donaghy, D., Gow, H.: Milk Price Cuts Reflect the Reality of Sweeping Changes in Global Dairy Market. http://theconversation.com/milk-price-cuts-reflect-the-reality-of-sweeping-changes-in-global-dairy-market-59251 (2016)
24. Lymer, S., Brown, L., Yap, M., Harding, A.: 2001 regional disability estimates for New South Wales, Australia, using spatial microsimulation. Appl. Spat. Anal. Policy. **1**(2), 99–116 (2008)
25. Merz, J.: Microsimulation—a survey of principles, developments and applications. Int. J. Forecast. **7**(1), 77–104 (1991)
26. Menon, M., Perali, F., Salvioni, C.: Microsimulation of the distributional impact of reformed farm support. International Conference Agricultural Policy Reform and the WTO: Where Are We Heading? Capri (Italy), June 23–26 (2003)
27. Miranti, R., McNamara, J., Tanton, R., Harding, A.: Poverty at the local level: national and small area poverty estimates by family type for Australia in 2006. Appl. Spat. Anal. Policy. **4**(3), 145–171 (2011)
28. Murphy, G., Hynes, S., Murphy, E., O'Donoghue, C., Green, S.: Assessing the compatibility of farmland biodiversity and habitats to the specifications of agri-environmental schemes using a multinomial logit approach. Ecol. Econ. **71**, 111–121 (2011)
29. Orcutt, G.H.: Simulation of economic systems. Am. Econ. Rev. **50**(5), 894–907 (1960)
30. Nepal, B., Tanton, R., Harding, A.: Measuring housing stress: how much do definitions matter? Urban Policy Res. **28**(2), 211–224 (2010)
31. Procter, K.L., Clarke, G.P., Ransley, J.K., Cade, J.: Micro-level analysis of childhood obesity, diet, physical activity, residential socioeconomic and social capital variables: where are the obesogenic environments in Leeds? Area. **40**(3), 323–340 (2008)
32. Ramilan, T., Scrimgeour, F., Marsh, D.: Analysis of environmental and economic efficiency using a farm population micro-simulation model. Math. Comput. Simul. **81**(7), 1344–1352 (2011)
33. Peel, D., Berry, H.L., Schirmer, J.: Farm exit intention and wellbeing: a study of Australian farmers. J. Rural. Stud. **47**, 41–51 (2016)
34. Rao, M., Tanton, R., Vidyattama, Y.: Modelling the economic, social and ecological links in the Murray-Darling basin: a conceptual framework. Australas. J. Reg. Stud. **21**(1), 80–102 (2015)
35. Saunders, P., Bradbury, B.: Monitoring trends in poverty and income distribution: data, methodology and measurement. Econ. Rec. **82**(258), 341–364 (2006). https://doi.org/10.1111/j.1475-4932.2006.00344.x
36. Schafer, J.L.: Multiple imputation: a primer. Stat. Methods Med. Res. **8**(1), 3–15 (1999)
37. Schirmer, J., Berry, H.: People and Place in Australia: The 2013 Regional Wellbeing Survey. University of Canberra, Canberra (2014)
38. Shrestha, S., Hennessy, T., Hynes, S.: The effect of decoupling on farming in Ireland : a regional analysis. Ir. J. Agric. Food Res. **46**, 1–13 (2007)
39. Shrive, F.M., Stuart, H., Quan, H., Ghali, W.A.: Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Med. Res. Methodol. **6**(1), 57 (2006)
40. Smith, D.M., Clarke, G.P., Harland, K.: Improving the synthetic data generation process in spatial microsimulation models. Environ Plan A. **41**(5), 1251–1268 (2009)
41. Simpson, L., Tranmer, M.: Combining sample and census data in small area estimates: iterative proportional fitting with standard software. Prof. Geogr. **57**(2), 222–234 (2005). https://doi.org/10.1111/j.0033-0124.2005.00474.x
42. Sutherland, H.: Static microsimulation Models in Europe: a Survey. Microsimulation Unit Discussion Paper No. MU9503, Department of Applied Economics, University of Cambridge, UK (1995)
43. Tanton, R.: Spatial microsimulation as a method for estimating different poverty rates in Australia. Popul. Space Place. **17**(3), 222–235 (2011). https://doi.org/10.1002/psp.601

44. Tanton, R., Vidyattama, Y.: Pushing it to the edge : extending generalised regression as a spatial microsimulation method. Int. J. Microsimul. **3**(2), 23–33 (2010)
45. Tanton, R., Vidyattama, Y., Nepal, B., McNamara, J.: Small area estimation using a reweighting algorithm. J. R. Stat. Soc. A. Stat. Soc. **174**(4), 931–951 (2011). https://doi.org/10.1111/j.1467-985X.2011.00690.x
46. van Leeuwen, E., Dekkers, J.: Determinants of off-farm income and its local patterns: a spatial microsimulation of Dutch farmers. J. Rural. Stud. **31**, 55–66 (2013)
47. Vanclay, F.: Social principles for agricultural extension to assist in the promotion of natural resource management. Aust. J. Exp. Agric. **44**(3), 213–222 (2004)
48. Vidyattama, Y., Cassells, R., Harding, A., Mcnamara, J.: Rich or poor in retirement? A small area analysis of Australian private superannuation savings in 2006 using spatial microsimulation. Reg. Stud. **47**(5), 722–739 (2013)
49. Vidyattama, Y., Tanton, R., Biddle, N.: Estimating small-area indigenous cultural participation from synthetic survey data. Environ. Plan. A. **47**(5), 1211–1228 (2015). https://doi.org/10.1177/0308518X15592314
50. Voas, D., Williamson, P.: An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Int. J. Popul. Geogr. **6**, 349–366 (2000)
51. White, I.R., Royston, P., Wood, A.M.: Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. **30**(4), 377–399 (2011)
52. Williamson, P., Birkin, M., Rees, P.H.: The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environ Plan A. **30**(5), 785–816 (1998)

Chapter 22
A Tax Benefit Model for Policy Evaluation in Luxembourg: LuxTaxBen

Nizamul Islam and Lennart Flood

Abstract This paper presents a new tax benefit microsimulation model for Luxembourg. The main distinguishing feature of the model is that it includes a behavioural module for labour supply. We validate our model by comparing incomes, taxes, and transfers produced with those of EU–SILC and EUROMOD. The labour supply model is validated by studying the fit (comparing observed hours of labour with those predicted by the model). Wage elasticities are reported and compared with similar exercises in the literature. Finally, the model is used for assessing a recent switch of the Luxembourg tax system from joint to individual taxation. By comparing simulated non-behavioural output it is shown that, as a whole, LuxTaxBen produces an output very close to that produced by EUROMOD and EU-SILC. Further, the behavioural simulation suggests that the reform has no impact on the labour supply for males in couple while a significant number of inactive females change their inactivity status and start working due to the reform. It is also shown that the change in the labour income increases by 0.8% while the welfare dependence decreases by −0.6 due to move from joint to individual tax rules.

Keywords Micro-simulation · Distributional and behavioural effect · Welfare

22.1 Introduction

The use of tax-benefit microsimulation models has become (or is about to become) an integrated part of evaluation of (the design of) tax-benefit reforms/policies. In many countries, the core part of this evaluation consists of a tax-transfer program that allows calculating the household net income using representative data for the

N. Islam (✉)
Luxembourg Institute of Socio-Economic Research, LISER, Esch-sur-Alzette, Luxembourg
e-mail: nizamul.islam@liser.lu

L. Flood
University of Gothenburg, Gothenburg, Sweden
e-mail: Lennart.Flood@handels.gu.se

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_22

population. This technique is usually referred micro-simulation and has become popular for its ability to provide a priori assessment of different policy designs [18]. It has been used to simulate the distributional consequences among heterogeneous groups of families and the effect of behavioural changes induced by a reform [5]. Examples of country-specific micro-simulation models include TAXBEN from the Institute for Fiscal Studies (IFS) in London [3] for the UK, SWEtaxben [8] for Sweden, or IZAΨMOD [16] for Germany, the OECD family type Tax-Benefit model, and the micro-simulation model for the European Union countries EUROMOD.[1]

There are few micro-simulation tools for Luxembourg. A micro-simulation model for the social Budget "SOBULUX" has been developed by the Ministry for Social Security (IGSS) aiming to analyze the long-term pension budget in Luxembourg. LIAM2 is another micro-simulation model that has been developed and validated for the first time at the Federal Planning Bureau in Belgium [13] and later on tested by the Luxembourg Team (LISER and IGSS). LIAM2 is a partial, dynamic micro-simulation model which can be used to analyze pensions and social transfers over the longer period for many countries, including Luxembourg. The OECD tax benefit model, a tax benefit model for 38 countries (32 OECD countries and, from 2005, Cyprus1, Latvia, Lithuania, Malta, and, from 2008, Bulgaria and Romania), can be used for policy evaluation in Luxembourg. The problem with this model is that it is strictly limited to various family types and thus does not offer analysis for the whole population. EOROMOD, an excellent instrument for programming the tax-benefit model for Luxembourg, is another example of a micro-simulation model used by several academics and researchers to analyze research questions specific to Luxembourg. For example Berger, Islam and Liegeois [2] analyze the effect of behavioural responses to the significant changes in the tax-benefit system during 2001–2002 in Luxembourg for single women and women in couple. One noteworthy shortcoming of EUROMOD is that the stochastic/behavioural models, such as labour supply model, are not integrated and thus analyzing behavioural aspects hinges exclusively on other software. We can say in another way that though EUROMOD can be convenient for static micro-simulations, it does not accommodate behavioural or stochastic models. Therefore, to calculate the budget set, Berger, Islam and Liegeois [2] and many others used the micro-simulation model EUROMOD, but for analyzing the behavioural model they used STATA, or SAS, or other software.

To the best of our knowledge, no behavioural micro-simulation model has so far been developed exclusively for Luxembourg. Therefore, we develop a Luxembourg Tax-Benefit micro-simulation model – LuxTaxBen. The main aim of this model is to evaluate the effects of tax-transfer policy reforms as well as other changes related

[1]Examples of dynamic micro-simulation, another stream of micro-simulation, include PENSIM2 (a dynamic micro-simulation model which dynamically simulates pension income for the next 50 years in the UK), LIAM [15], SESIM [10, 14], MIDAS [6, 7] and north American dynamic micro-simulation CORSIM, DYNACAN for Canada and POLISIM for the United States.

to poverty, inequality, incentives and the governmental budget in Luxembourg. It contains very precise information on income tax rules, as well as eligibility rules for a number of welfare programs, such as social assistance, housing allowance, etc. It has been built specifically for analyzing the Luxembourgish tax-transfer system whereby one can generate net income for different choices of hours of work and welfare participation. One important feature of LuxTaxBen is that it is constructed in an integrated way so all modules including labour supply can be used together. Another distinguishing characteristic of this model is flexibility, meaning that the model is flexible enough to create a new version of it by updating the database and/or new rules. The model is very simple and handy. It can accommodate either survey or administrative data. One objective of LuxTaxBen is to give the users the opportunity to analyze the effects of planned changes (new rules) in the Luxembourg tax-benefit system. This model has been built at the Luxembourg institute of socio-economic research (LISER) for individual researchers as well as for governmental and non-governmental offices. The current version of LuxTaxBen has been written in SAS-language.

The purpose of this paper is to show the usage of LuxTaxBen where, first, we define the static micro-simulation and simulate various income components, including household net income and taxes. For validation purposes these income components are compared with observed data (EU-SILC) as well as with the output of a similar exercise using EUROMOD. By comparing simulated non-behavioural output it is shown that, as a whole, LuxTaxBen produces an output very close to that produced by EUROMOD and EU-SILC. Second, on 14 December 2016, the Luxembourg parliament approved the law for the 2017 tax reform where one of the adopted measure is the option for individual taxation of couples. The consequences of this policy reform on income distribution has two aspects. (a) A mechanical effect in which individuals do not change their employment behaviour; this effect is usually called pure income effect and can be calculated to perform the distributional analysis. (b) The employment status may be affected by the change of the reform. This effect is typically identified as the behavioural effect or endogenous effect of the reform. Our aim here is to present how LuxTaxBen is used to evaluate the central governmental budget, distinguishing behavioural and non-behavioural effects due to a switch from joint to individual tax systems in Luxembourg. It is also shown that the change in tax revenue is about 6% due to the move from a joint to an individual tax system.

The next section (Sect. 22.2) describes the data used for the empirical exercise. Section 22.3 presents the behavioural and non-behavioural application of LuxTaxBen. Finally, Sect. 22.4 concludes the paper after a short discussion.

22.2 Data

LuxTaxBen can accommodate either survey or administrative data for Luxembourg. In the paper we give an example of application of LuxTaxBen where we use survey

EU-SILC-2010 data (Income year 2009)[2] for Luxembourg. We have presented only the sub-sample of couple household (2276) from EU-SILC. It is worth to mention that for all exercise we have used single male/female household, single male/female parent households, and couple households from two different data sources survey (EU-SILC) and administrative (IGSS). However, other than the couple household sub-sample, these results are not presented in the paper but available on request.

22.3 Empirical Framework

The modelling framework established in LuxTaxBen described in the following subsections has two parts: The first part is the static micro-simulation model that used to evaluate the mechanical or immediate distributional effect of possible policy change on individuals and/or households, while the time dimension and behavioural adjustment are not taken into account. This part is used to define the alternative budget set applied in the second part (the structural labour supply) of the model. The second part is a behavioural or structural labour supply model that is used to estimate the individual preference parameters and to simulate the incentive effects of tax reform on the labour market as well as on individual and household net income.

22.3.1 Static Micro-simulation Model

Static models without behavioural effects are simply tax and benefit simulators that simulate household net income for each household before and after a policy change. The objective of this model is to investigate how changes in the tax-benefit system would affect household incomes. This model can have great value for a reform where the assumption of no behavioural response can be realistic. It allows the simulation of taxes, social security contributions, benefits and net income of several subgroups such as couple household, single male, single female. Using LuxTaxBen, we have simulated all taxes and transfers to yield household net income of each couple household, given their wage, demographic and other characteristics. To validate simulated static output, since the input data of the model originated in EU-SILC, we choose EU-SILC as the reference and compare all its major income components with the output from LuxTaxBen. For cross validation, we compare findings from LuxTaxBen to findings from the same exercise using EUROMOD.[3] For the sake of simplicity, we keep identical names and the definitions of all income

[2]EU-SILC-2010 is the cross-sectional survey "panel socio-economique LiewenzuLetzebuerg (Psell-3)" data collected during the year 2010, which include information on income for 2009.

[3]EUROMOD is a unique source and a recognized standard tool for simulating the effect of tax-transfers reform of all EU countries including Luxembourg. The model enables to simulate all individual and household income components (including net income) using available information from the European Union Statistics on Income and Living Conditions – EU-SILC (for detail about EUROMOD, see for example Sutherland and Figari [19]). Therefore, to control for the validity of

components of LuxTaxBen as given in EUROMOD. Before comparison of all income it is important to know that the two models are different by construction. Table 22.1 provides a short overview of various income and tax-transfer components that have been taken into account for the calculation of household net income for couples.[4] The table shows that the annual average of all income components from EU-SILC, EUROMOD, and LuxTaxBen are quite similar. Original income[5] consists of labour income (employment + self-employment income), investment income, property income, income from child below 16, intra-household transfer received, and maintenance payments. It is clear that these income components (row a–h) are identical although the construction technique of these income components are not same.[6] Row m shows the means-tested benefits that consist of social assistance

Table 22.1 Comparing annual average of all major income components of EUROMOD, EU-SILC and LuxTaxBen

	Income components	EUROMOD	EU-SILC	LuXTaxBen
(a)	Labour income	77476.08	77475.41	77475.41
(b)	Investment income	1255.51	1255.51	1255.51
(c)	Income from child below 16	7.43	7.43	7.43
(d)	Pension from private pension plans	7.23	7.23	7.23
(e)	Property income	1262.30	1262.30	1262.30
(f)	Intra hh transfer received	91.45	91.45	91.45
(g)	Maintenance payments	430.92	430.92	430.92
(h)	**Original income**	**79669.09**	**79668.41**	**79668.41**
(I)	Social assistance (RMG)	654.38	229.03	593.71
(j)	Expensive life allowance	366.95	46.43	292.44
(k)	Housing allowance	326.69	326.69	275.072
(l)	Scholarship for tertiary education	127.98	10.08	10.08
(m)	**Means-tested benefits**	**1475.99**	**612.23**	**1171.30**
(n)	Child benefit	3464.59	3477.80	3104.16
(o)	New school year allowance	76.4763406	183.30	150.22
(p)	Tax bonus for children	1077.27	1036.17	1043.90
(q)	Accident permanent benefit	100.73	100.53	100.53
(r)	Benefit – Care	12.46	12.46	12.46
(s)	Benefit – Dependence	137.75	49.01	49.01
(t)	Primary and post-primary school subsidies	3.78	0.00	0.00

(continued)

the performance of LuxTaxBen, it is appropriate to compare the output of EUROMOD with that of LuxTaxBen.

[4] Various simulated income components for single households are available on demand.

[5] The name (original income) of these income components are used in EUROMOD and the level is ils_origy. A similar rule will be followed for all other names of income components.

[6] For example, in LuxTaxBen the labour income has been constructed by yearly hours of work multiplied by the wage rate while in EUROMOD it is defined according to monthly income.

Table 22.1 (continued)

	Income components	EUROMOD	EU-SILC	LuXTaxBen
(u)	Communal subsidies for schooling	1.61	0.00	0.00
(v)	Education allowances	225.27	225.27	225.27
(w)	Benefit – Disabled Person	37.88	18.23	18.23
(x)	Parental leave allowances	269.35	269.35	269.35
(y)	Antenatal, birth, postnatal ben	54.99	54.99	54.99
(z)	Maternity allowance (lump-sum)	25.24	24.62	24.62
(aa)	Maternity payments	312.85	316.18	316.18
(ab)	Other benefits from the solidarity national fund (FNS)	98.94	37.48	37.48
(ac)	Unemployment benefit	776.85	1807.57	1807.57
(ad)	Sickness benefit	128.89	128.89	128.89
(ae)	Benefit, if fired	60.18	49.04	49.04
(af)	**Non means-tested benefits**	**6865.10**	**7790.92**	**7391.93**
(ag)	Benefit – Early retirement pension	1185.86	1172.00	1172.00
(ah)	Pension – Disability (invalidity)	1467.97	1464.70	1464.70
(ai)	Pension for past education of children	82.66	75.21	75.21
(aj)	Old age pension – Additional from employer (2nd tier)	66.88	66.88	66.88
(ak)	Old age pension – For private sector	3738.87	3461.50	3461.50
(al)	Old age pension – For public sector (state regime)	2115.59	2113.44	2113.44
(am)	Old age pension – End of year allowance	57.56	50.63	50.63
(an)	Survivors pension – Private sector (reversion pension)	416.11	68.15	68.15
(ao)	Survivors pension – Public sector (reversion pension)	6.92	6.92	6.92
(ap)	**Pension**	**9138.42**	**8479.44**	**8479.44**
(aq)	**Total benefit**	**17479.51**	**16882.59**	**17042.67**
(ar)	**Tax**	**10777.92**		**10678.67**
(as)	**Social security**	**10303.50**		**10319.16**
(at)	**Tax and social security**	**21081.42**	**20716.06**	**20997.83**
(av)	**Disposable income**	**76067.18**	**75834.93**	**75713.24**

Labour Income = Employment income + Self-employment income;
h = a + b + c + d + e + f + g; m = I + j + k + l;
af = n + o + p + q + r + s + t + u + v + w + x + y + z + aa + ab + ac + ad +ae;
ap = ag + ah + ai +aj + ak + al + am + an +ao; aq = m + af + ap;
Tax = employment tax + self-employment tax;
Social security = Social security from employment + Social security from self-employment;
av = h + aq -ar -as

(RMG), living expenses allowance, housing allowance, and scholarship for tertiary education. One important characteristic of this income group is that most of them are simulated by tax-benefit rules. For example, RMG, housing allowance, and living expenses allowance are simulated by the tax-benefit rules in both LuxTaxBen and EUROMOD. The difference between actual value (in EU-SILC) and simulated value (from both LuxTaxBen and EUROMOD) can be explained by non-take up ratio. In Luxembourg over 65 percent of all households potentially entitled do not claim RMG because of rational motivation; for example, expected net utility from claiming, as well as stigma, play a major role in explaining the level of non-take up [1].

Non means-tested benefits, another income group, contains child benefit, new school year allowance, tax bonus for children, permanent accident benefit, care benefit, dependency benefit, primary and post primary school subsidies, communal subsidies for schooling, education allowances, benefit for the severely disabled, parental leave allowances, antenatal, birth, and postnatal benefit, maternity allowance (lump-sum), maternity payments, other benefits from the national social security fund (FNS), unemployment benefit, sickness benefit, and benefit if fired. Child benefit, new school year allowance, and tax bonus for children have been simulated in both EUROMOD and LuxTaxBen. These simulations depend on the number and age of children in the household. For example, child benefit for 1 child amounts to €185.6*12/year, for 2 children €440.72*12/year, for 3 children €802.74*12/year, and for more than 3 children €802.74*12 + (361.82*(number of children-3)*12)/year. New school year allowance for 1 child aged 6–11 is €113.15/year, for 2 children it is €194.02*2/year, and for subsequent children €274.82*(number of subsequent children)/year. The child tax credit (*moderation d'impôt pour enfant*) was a benefit that applied only to families who paid income tax. This benefit has been replaced by a bonus for children (*boni pour enfant*) in 2008. This bonus applies to all families with children eligible for family allowances and the bonus is € 922.50/year per child.

Almost all components in the non means-tested income group are rather similar in EUROMOD and in LuxTaxBen as well as in EU-SILC. Nevertheless, some of them are noticeably different such as unemployment benefit. This benefit, on average, when reported in EUROMOD is €776.85/year, in EU-SILC €1807.57/year, and in LuxTaxBen €1807.57/year. This benefit has been simulated in EUROMOD but not in LuxTaxBen. In LuxTaxBen it is the same as in EU-SILC.

Row (ap) presents the pension income which contains benefits from early retirement, disability/invalidity pension, pension for past education of children, additional old age pension from the employer, complementary pension for miners and metal workers, old age pension for private sector, old age pension for public sector (state regime), complementary old age pension for war captivity, old age pension for end of year allowance, survivors pension for private sector (reversion pension), and survivors pension for public sector (reversion pension). LuxTaxBen produces all of these pension components exactly as in EU-SILC, while it is somewhat different with respect to some components produced by EUROMOD. This difference, although not great, can be explained by the structure of household composition. Another source of this variation is the labour income of other members.

As we mentioned earlier and as in other micro-simulation models, the ultimate objective of LuxTaxBen is to generate household net income. Therefore, in the end of the simulation, the model produces household net income that is equal to original income plus total benefit minus the total tax and social security contributions, where total benefit contains pension, means- and non means-tested benefit. Total tax and social security contributions contain employment and self-employment tax and social security contributions respectively. Row (av) presents on average household net income of **€76067.18/**year in EUROMOD, **€75834.93/**year in EU-SILC, and **€75713.24/**year in LuxTaxBen, which are very close to each other.

22.3.2 Behavioural or Structural Labour Supply Model

In order to evaluate the behavioural response of a reform LuxTaxBen adopts a structural model with some innovative features where the simultaneous nature of labour supply and welfare participation decisions are taken into account. It integrates the discrete choice approach proposed by Van Soest [17].[7] The motivation of choosing this approach is that it can analyze the joint decision of spouses and can accommodate the preference heterogeneity inherent in the model.[8]

An important distinguishing characteristic of a micro-simulation labor supply model is that it should be able to replicate the actual values. Thus, to check predictive power, we estimate a series of labour supply models using various specifications and different data sources. In order to check the goodness of fit of the model we compare the actual and predicted hours of work.

The top panel of Table 22.2 presents the percentages of actual and predicted hours' work for a male and female in couple while the bottom panel presents the probability of welfare participation of a couple household. The predicted probability of welfare participation and hours' work are very similar to observed welfare participation and hours' work for both male and female.

Elasticity

After inspecting predictive power of the model, we derive the uncompensated wage elasticities of labour supply for male and female in couple (see Table 22.3).

We find evidence that labor supply elasticity with respect to gross wage are quite small. Of course, the wage elasticity depends on the wage level and otherinstitutional features. For example, a labour supply response of 10% increase

[7]The technical detail of the estimaton of the labor supply model is provided in the Appendix. One important limitation is that it is not an equilibrium model and therefore the model does not take into account demand-side restrictions.

[8]For more information regarding the advantages of this approach, see Flood and Islam [11], Blundell and MaCurdy [4].

Table 22.2 Percentages of actual and predicted hours and welfare participation for male and female in couple

	Male		Female	
	Observed	Predicted	Observed	Predicted
0 h	21.40	21.31	41.17	40.51
0 > h > 1500	5.67	5.05	27.55	28.73
2250 h and more	72.93	73.64	31.28	30.76
	Household			
	Observed		Predicted	
Probability of welfare participation	Not participated	Participated	Not participated	Participated
	93.06	6.94	93.15	6.85

Source: LuxTaxBen

Table 22.3 Wage elasticity (uncompensated) of labour supply for male and female in couple

	10% wage change	
	Female	Male
Female		
Full sample	0.44	−1.40
10th percentile	2.63	−0.88
25th percentile	1.15	−0.29
50th percentile	0.19	−1.46
75th percentile	0.54	−1.07
90th percentile	0.00	0.00
Male		
Full sample	−0.70	0.29
10th percentile	−0.93	3.74
25th percentile	−0.30	1.21
50th percentile	−0.13	0.61
75th percentile	−1.50	0.00
90th percentile	−0.86	0.00

Source: LuxTaxBen

in the 10th percentile wage rate of male (female) varies from the labour supply response of the other percentile wage rate of male (female). In general, female wage elasticity is higher than for males. For example, Van Soest [17] shows that the own wage elasticity for a married woman with average characteristics is 1.03, and for a married man is 0.15. Similarly, using the data from PSID (US Panel Study of Income dynamics) 1976, Hausman and Ruud [12] show that the elasticity for married women is 0.76, while the elasticity for married men is 0.−0.03. Our findings are in line with other previous findings, implying that females have higher wage elasticity than males.

22.3.3 An Evaluation of Individual Tax Reform

To evaluate the individual tax reform LuxTaxBen simulates the net income, choice of hours, taxes and other benefits of each household under joint tax rules. It then simulates the "new" net income, choice of hours, taxes and other benefits under individual tax rules. Tables 22.4 and 22.5 demonstrate the transition matrix of the labour supply due to a switch from a joint to an individual tax system for males and females in couple respectively.

We find almost all males remain on the diagonal, implying that the reform has no impact on the labour supply for males in couple, while the story is different for the females in couple. A notable number (1291) of inactive females change their inactivity status and start working due to reform. Top panel of Table 22.6 shows the non-behavioural effects on household disposable income, labour income, pension income, tax, unemployment benefit, social assistance (RMG), and social security contribution. It is clear that the reform has negative impacts on the household disposable income: 4% decrease after the reform. The reform has positive impact on tax revenue (28.6% increase) because the male, individually, pays more tax after the reform, unrelatedly working hours of his wife. Bottom panel of Table 22.6 shows the effects of the reform with behavioural adjustment. The simulation results show that while the labour income increase by 0.8% the welfare dependence (Social assistance-RMG) decrease by −0.6 due to move from joint to individual tax rules.

22.4 Conclusion

In this paper we present a behavioural micro-simulation model (LuxTaxBen) that can be used to evaluate the impact of changes in the tax-transfer policy on poverty, inequality as well as the central governmental budget system in Luxembourg. It provides the opportunity to simulate the new rules in the Luxembourg tax-transfer system. The model consists of a number of modules such as a module for child benefit, housing allowance, and fees for child care. It is constructed in an integrated way so all the modules can be used together. One important feature of LuxTaxBen is that for behavioural modelling it does not depend on other software while it is constructed in an integrated way so all the modules including labour supply can be used together. Our non-behavioural simulation shows, as a whole, LuxTaxBen produces very similar output to EUROMOD and EU-SILC. In addition, the behavioural simulation provides evidence that the reform has no impact on the labour supply for males in couple while a significant number of inactive females change their inactivity status and start working due to reform. It is also shown that the labour income increase by 0.8% while the welfare dependence (Social assistance-RMG) decrease by −0.6 due to move from joint to individual tax rules. One shortcoming of LuxTaxBen is that at present it includes the labour supply model proposed by Van Soest [17], which in fact does not take into account the demand-

Table 22.4 Transition of economic status switching from joint to individual tax systems for males in couple

Male	Status after								
Status before	Child <17	Old age pensioner	Student	Disability pension	Parental leave	Unemployed	Other (inactive)	Working	Grand total
Child <17	0	0	0	0	0	0	0	0	0
Old age pensioner	0	11,396	0	0	0	0	0	0	11,396
Student	0	0	134	0	0	0	0	0	134
Disability pension	0	0	0	3269	0	0	0	0	3269
Parental leave	0	0	0	0	620	0	0	0	620
Unemployed	0	0	0	0	0	2757	0	0	2757
Other (inactive)	0	0	0	0	0	0	5789	80	5870
Working	0	0	0	0	0	0	81	55,665	55,746
Grand total	0	11,396	134	3269	620	2757	5871	55,745	79,792

Source: LuxTaxBen

Table 22.5 Transition of economic status switching from joint to individual tax system for Females in Couple

Female	Status after								
Status before	Child <17	Old age pensioner	Student	Disability pension	Parental leave	Unemployed	Other (inactive)	Working	Grand total
Child <17	0	0	0	0	0	0	0	0	0
Old age pensioner	0	2727	0	0	0	0	0	0	2727
Student	0	0	1455	0	0	0	0	0	1455
Disability pension	0	0	0	2587	0	0	0	0	2587
Parental leave	0	0	0	0	2331	0	0	0	2331
Unemployed	0	0	0	0	0	2158	0	0	2158
Other (inactive)	0	0	0	0	0	0	26,057	1333	27,389
Working	0	0	0	0	0	0	42	41,103	41,145
Grand total	0	2727	1455	2587	2331	2158	26,098	42,436	79,792

Source: LuxTaxBen

Table 22.6 Behavioural and non-behavioural effects of the Tax Reform

Macro level				
	Joint Tax Rules	Individual Tax Rules	Change	%
Households				
No behavioural adjustment				
Disposable income	5,917,034,919	5,682,868,780	−234,166,139	−4.0
Labour income	5,993,404,021	5,993,404,021	0	0.0
Pension income	676,590,748	676,590,748	0	0.0
Tax	817,975,424	1,052,141,563	234,166,139	28.6
Unemployment benefit	144,158,621	144,158,621	0	0.0
Social assistance (RMG)	56,567,731	56,567,731	0	0.0
Social security	805,070,123	805,070,123	0	0.0
With behavioural adjustment				
Disposable income	5,917,034,919	5,723,230,523	−193,804,396	−3.3
Labour income	5,993,404,021	6,039,572,930	46,168,909	0.8
Pension income	676,590,748	676,590,748	0	0.0
Tax	817,975,424	1,051,653,254	233,677,829	28.6
Unemployment benefit	144,158,621	144,158,621	0	0.0
Social assistance(RMG)	56,567,731	56,228,256	−339,476	−0.6
Social security	805,070,123	810,627,219	5,557,096	0.7

Source: LuxTaxBen

side effect. This is not an equilibrium effect model. Accordingly, an interesting future challenge is to include such a model that allows for a more realistic view of demand-side restrictions.

References

1. Amétépé, F.: The effectiveness of Luxembourg's minimum guaranteed income. Int. Soc. Secur. Rev. **65**(1), 99–116 (2012)
2. Berger, F., Islam, N., Liegeois, P.: Behavioral micro simulation model and labour supply in Luxembourg. Brussels Econ. Rev. **54**(4), 389–420 (2011)
3. Brewer, M., Browne, J., Emmerson, C., Goodman, A., Muriel, A., Tetlow, G.: Pensioner Poverty Over the Next Decade: What Role for Tax and Benefit Reform? IFS Commentary No. 103. London: The Institute for Fiscal Studies (2007)
4. Blundell, R., Macurdy, T.: Labor supply: a review of alternative approaches. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics, vol. 3,. chapter 27, eds edn, pp. 1559–1695. Elsevier (1999)
5. Creedy, J., Duncan, A.: Behavioural micro simulation with labour supply responses. J. Econ. Surv. **16**(1), 1–39 (2002)
6. Dekkers, G., Belloni, M.: Micro Simulation, Pension Adequacy and the Dynamic Model MIDAS: An Introduction. Project AIM – Deliverable 4.10 (2009)
7. Dekkers, G., Buslei, H., Cozzolino, M., Desmet, R., Geyer, J., Hofmann, D., Raitano, M., Steiner, V., Tanda, P., Tedeschi, S.: What are the consequences of the European AWG-projections on the adequacy of pensions? An application of the dynamic micro simulation

model MIDAS for Belgium, Germany and Italy. In: Life-Cycle Microsimulation Modelling. Constructing and Using Dynamic Microsimulation Models, pp. 325–357. Germany LAP LAMBERT Academic Publishing, Saarbrücken (2010)
8. Ericson, P, Flood, L.R., Wahlberg, R.: SWEtaxben: A Swedish Tax/benefit Micro Simulation Model and an Evaluation of a Swedish Tax Reform. IZA DP No. 4106 (2009)
9. EUROMOD: https://www.iser.essex.ac.uk/euromod
10. Flood, L.: Chapter 3: SESIM: a Swedish micro-simulation model. In: Klevmarken, A., Lindgren, B. (eds.) Simulating an Ageing Population: A Microsimulation Approach Applied to Sweden. Contributions to Economic Analysis, vol. 285, pp. 55–83. Emerald Group Publishing Limited, Bradford (2008)
11. Flood, L., Islam, N.: A Monte Carlo evaluation of discrete choice labour supply models. Appl. Econ. Lett. **12**(5), (2005)
12. Hausman, J., Ruud, P.: Family labor supply with taxes. Am. Econ. Rev. **74**, 242–248 (1984)
13. de Menten, G., Dekkers, G., Bryon, G., Liégeois, P., O'Donoghue, C.: LIAM2: a new open source development tool for discrete-time dynamic microsimulation models. J. Artif. Soc. Soc. Simul. **17**(3), 9 (2014)
14. Klevmarken, N.A.: Microsimulation for Public Policy: Experiences from the Swedish Model SESIM, Economics and Social Research. Institute Discussion Paper 242. Cabinet Office, Tokyo, Japan (2010)
15. O'Donoghue, C., Lennon, J., Hynes, S.: The Life-cycle Income Analysis Model (LIAM): a study of a flexible dynamic microsimulation modelling computing framework. Int. J. Microsimul. **2**, 16–31 (2009)
16. Peichl, A., Schneider, H., Siegloch, S.: Documentation Izaψmod: The IZA Policy Simulation Model IZA DP No. 4865. Institute for the Study of Labor, Bonn (2010)
17. Van Soest, A.: Structural models of family labor supply: a discrete choice approach. J. Hum. Resour. **30**, 63–88 (1995)
18. Spadaro, A.: Microsimulation as a tool for the evaluation of public Policies: methods and applications. In: Spadaro A (ed), Bilbao: Fundación BBVA, 357 p. 24 cm; isbn 978-84-96515-17-8 (2007)
19. Sutherland, H., Figari, F.: EUROMOD: the European Union tax-benefit microsimulation model. International Journal of Microsimulation. **1**(6), 4–26 (2013)

Part VI
Business Analytics and Managements Policy Analysis

Chapter 23
Finding Significant Determinants and Impacts of Farm-Level Integrated Pest Management Practices Using Statistical Tools

Md. Sadique Rahman

Abstract The study aims to identify the determinants of adoption of integrated pest management (IPM) and its impacts on productivity and pesticide applications. The study employed two-part poisson hurdle regression, propensity score matching (PSM), and inverse probability weighted regression adjustment (IPWRA) techniques to achieve the objectives. The findings indicated that level of IPM adoption was low. Around 15% and 10% of the respondents adopted poultry refuse and sex pheromone trap, respectively while about 11% adopted both the practices. Decision to adopt significantly influenced by training ($p < 0.01$), neighbour farmers adoption ($p < 0.01$) and distance from highway ($p < 0.05$) while extent of adoption depend on extension contact ($p < 0.05$) and neighbour farmers adoption ($p < 0.05$). Adoption of IPM significantly reduced pesticide applications and increased productivity compared to non-adopters based on kernel and radius matching. Adoption analysis suggested that more research and field demonstrations are required to improve the adoption level. Reduced pesticide applications may have some environmental benefits. Due to higher productivity, there is scope to boost the role of IPM in anti-poverty policies in Bangladesh.

Keywords Impact evaluation · Pesticide application · Poisson hurdle model · Propensity score matching · Vegetable farming

23.1 Introduction

Vegetable farming is viewed as one of the most income generating activities in Bangladesh but overuse and misuse of pesticides in vegetables farming creates many negative consequences including farmers' health, environment and development of

Md. S. Rahman (✉)
Department of Management and Finance, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh
e-mail: sadique@sau.edu.bd

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_23

resistance [1–4]. There are even instances of daily application of pesticides in the vegetable fields in Bangladesh [5]. As a result, IPM emerged as a policy concern in Bangladesh. IPM is an environment friendly and cost-effective approach to pest control [6].

Vegetable IPM was first introduced in Bangladesh during 1996 with support of the Food and Agriculture Organization (FAO) and subsequently a number of donor and non-government agencies (NGOs) started working in Bangladesh [7]. The Integrated Pest Management Innovation Lab (IPM IL), funded by United States Agency for International Development (USAID), is one of the programme started working in Bangladesh from 1998 in collaboration with the Bangladesh Agricultural Research Institute (BARI) to develop and disseminate the vegetable IPM practices. The present study used the data from a subproject of the IPM IL that involved IPM technology transfer for tomato (*Solanum lycopersicum* L.) in some selected regions of Bangladesh.

Tomato, one of the major vegetables of Bangladesh, is widely grown in winter season but susceptible to various pests and diseases [8–9]. To combat pests and diseases, IPM IL recommends a variety of methods, including use of virus resistant varieties, use of pheromone traps, yellow sticky traps, soil amendment with poultry refuse, and use of bio-pesticides [5]. IPM IL used field days, trainings and small group discussion as intervention to disseminate the recommended practices in the south-western parts of Bangladesh. The locations had been chosen by the IPM IL to correspond with USAID "Feed-the-Future" (FtF) programme areas.

Adoption of these IPM practices has created a wide range of impacts that need to be evaluated to understand the contributions of research and development. But farm-level impact assessments of tomato IPM have been relatively limited in Bangladesh. Many studies used discrete choice model to identify the factors affecting the adoption but failed to consider the differences in level of adoption among the adopters [10–12]. Few studies found that adoption of IPM reduced the frequency of pesticide applications and increased yield but the studies treated all IPM practices equal whereas farmers exhibited different adoption attitudes towards different IPM practices [12–14]. Several other studies estimated the farm-level impact of IPM adoption without considering the selection bias problem that may arise due to non-random sampling of households [15–16]. Thus, the causal effect of tomato IPM adoption on pesticide use and productivity is yet to be studied empirically in Bangladesh. The current study was undertaken to identify the determinants of adoption and impacts of tomato IPM practices on productivity, and pesticide applications.

Rest of the paper is organized as follows. The next section describes how data were collected and analyzed. Section 23.3 represent the results and discussion in the wider context of IPM impacts and adoption globally. Section 23.4 concludes with policy recommendations.

23.2 Methodology

23.2.1 Data Sources

The study was conducted in four regions: Jashore, Magura, Barishal and Jhalokati of Bangladesh. The IPM IL randomly selected 104 villages from these four regions and disseminated the IPM practices through a 3-year training programme by selecting eight or nine farmers from each village. Thus, a total of 838 farmers were selected by IPM IL from the villages of which 50% of the farmers received the intervention (adopters). To assess the adoption and impacts of IPM, a survey was conducted in 2015 and the sample was divided into IPM adopter and non-adopter based on the IPM practices: pheromone traps for mass trapping, and soil amendment with poultry refuse to manage soil diseases and improve soil fertility. Out of 838 farmers, a total of 751 farmers were interviewed in 2015 (a few farmers had migrated to other villages and a few were not interested in being re-interviewed), of which the farmers who did not cultivated tomato during 2015 were dropped from the sample, leaving 235 tomato farmers, of which 85 farmers were IPM adopters and 150 were non-adopters.

23.2.2 Analytical Techniques

23.2.2.1 Factors Affecting Adoption

To assess the factors affecting the adoption of new technologies/practices several studies used binary probit/logit model where a farmer who adopted at least 1 practice, he/she was considered as adopter and assigned a score of 1, otherwise assigned 0 [10–12]. But the use of binary probit model may not provide robust result because IPM farmers exhibit different adoption decision for different IPM practices. Thus, the farmers take two decisions: whether to adopt and then deciding how many IPM practices to be adopted. The present study used two-part poisson hurdle regression to answer these two questions [17]. The first part of the model was a binary response model (logit) and the second part was a zero truncated count model [18]. The poisson hurdle model can handle both under and over dispersion problem [19]. The first part of the model deals with adoption decision ($y_i = 1$ if adopted or 0 otherwise) known as hurdle, once the hurdle cross, second part of the model deals with the positive count ($y_i > 0$). The probabilities in the first part of the model were as follows [20]:

$$\Pr(y_i = 0|x_i) = \exp(-u_{1i})$$

$$\Pr(y_i = 1|x_i) = 1 - \exp(-u_{1i})$$

Where, y_i is the count variable and x_i is the explanatory variables. For the second part of the model, zero truncated poisson distribution was given by:

$$\Pr\left(y_i|x_i, y_i > 0\right) = \frac{e^{-\mu_i}{\mu_i}^{y_i}}{\left(1 - e^{-\mu_i}\right) y_i!}, y_i = 1, 2 \ldots$$

23.2.2.2 Impact Assessment

Assessment of impact in the absence of baseline data is always challenging due to potential selection bias and endogeneity. Past studies suggested either matching techniques or instrumental variable (IV) based regression approach to deal with selection bias or endogeneity [21–23]. But finding a valid instrument is difficult. Thus, the present study used matching techniques to estimate the causal effect of tomato IPM adoption on selected outcome variables.

Propensity score matching (PSM), which is commonly used in drawing causal inferences [24–25], helps in creating a counterfactual from control group based on two conditions; a conditional independence assumption (CIA) and overlap in propensity scores across the adopters and non-adopters [14, 26–27]. Estimation of PSM model generally follows three steps procedure: first, use of binary probit/logit model to estimate the propensity scores. If any farmer adopted any one of the IPM practices mentioned earlier, he or she was considered an IPM adopter and given a score of one, otherwise given 0. In the second step, balancing test was executed and region of common support was selected. Finally adopters and non-adopters of IPM were matched based on the estimated propensity scores and the average treatment effect on the treated (ATT) was computed as follows;

$$ATT = E\left(Y^1 – Y^0|X_i, T = 1\right) = E\left(Y^1|X_i, T = 1\right) – E\left(Y^0|X_i, T = 1\right)$$

Where, T indicates treatment status (adopted =1, otherwise =0), X_i is observed characteristics, $E(Y^1| X_i, T = 1)$ is the mean outcome of the adopters conditioned on X_i in the treated situation and $E(Y^0| X, T = 1)$ is the mean outcome of the non-adopters conditioned on X_i in the treated situation.

So far, a farmer was considered as IPM adopters if he/she adopted any one of the IPM practices but there are several IPM practices available and a farmer may adopt more than one. Thus, the impact of adoption on productivity and pesticide applications may differ depending on the types of IPM practices adopted. To address this issue, adoption was defined based on the types of IPM practices a farmer adopts and thus, three separate PSM models were estimated to measure the impacts of adoption. For sex pheromone trap model, a farmer was considered as IPM adopter if he/she adopted the practice and given a score of 1, otherwise 0. Similarly, in the case of poultry refuse model, a farmer was considered as IPM adopter if he/she adopted the practice and given a score of 1, otherwise 0. For the third model, a farmer was considered as adopter if he/she adopted both sex pheromone trap and poultry refuse. Two matching algorithms (radius and kernel matching) were used to estimate the treatment effect.

To check the robustness of the PSM analysis, the present study used inverse probability weighted regression adjustment (IPWRA) which provides consistent results in the presence of mis-specification in the treatment or outcome model, but not both [28]. Following [29], ATT in the IPWRA model was estimated in two steps. In the first step, propensity scores were estimated using multinomial logit regression by generalizing three different and mutually exclusive categories (0,1,2) of IPM adoption and in second step, linear regression was used to estimate the ATT. The treatment and outcome model under IPWRA was computed as follows;

$$\Pr\left(T_i = 1, 2\right) = f\left(J_i \alpha\right) + v_i$$

$$Y_i = g\left(x_{\mathrm{i}} \beta\right) + \mathrm{u_i}$$

Where, J_i is a set of explanatory variables explaining treatment assignment T_i, x_i is a set of explanatory variables that influence the outcome Y_i. Descriptions of the explanatory variables used in the models are given in Table 23.1.

Table 23.1 Definition of the explanatory variables used in the different models

Variable	Notation	Description
Experience (years)	z_1	Tomato cultivation experience of the primary farmer, a proxy for willingness to adopt
Max. education (years)	z_2	Highest years of schooling of household member, representing the knowledge of the household
Farm size (ha)	z_3	Total amount of land owned by the farmer, calculated as: Farm size = own land + rented in+ sharecropped in – rented out – sharecropped out
Credit access (yes/no)	z_4	One if the farmer received credit from any formal source, otherwise 0
Extension contact (no.)	z_5	Total number of visit to local extension agent during last 12 months regarding information on tomato cultivation
Severity of diseases (score)	z_6	Zero if no disease infestation, 1 if severity is low, 2 if medium low, 3 if medium high, 4 if severity is high
IPM training (no.)	z_7	Number of training received on IPM during last 12 months
FAI (no.)	z_8	Total number of neighbour farmers who adopted IPM practices near the primary farmer's field
Distance from highway (km)	z_9	Distance in kilometers of highway from the primary farmer's field

23.3 Results and Discussion

23.3.1 Descriptive Statistics

Table 23.2 present the summary statistics of the independent variables that were used in the estimation process. Differences in maximum education, extension contact, number of IPM training, FAI and distance from highway were significant. IPM adopters received more training on IPM compared to non-adopters. IPM adopters were also more likely to maintain frequent contact with extension officers than non-adopters. These differences indicate that the two groups of tomato growers are not directly comparable which justifies the use of different treatment effect models [14, 27].

23.3.2 Adoption of IPM Practices

Among the IPM practices suggested by IPM IL, highest portion of farmers adopted soil amendment with poultry refuse (14.89%) followed by sex pheromone traps (Table 23.3). Among different IPM practices most farmers in Bangladesh adopted sex pheromone traps and soil amendment methods due to availability, and effective in controlling insects and soil borne diseases compared to pesticides [7]. Level of knowledge is a prerequisite for adoption of IPM practices [30], thus more training and awareness building programmes are warranted to increase the adoption status of IPM.

Table 23.2 Descriptive statistics for variables used in the analysis

	Adopters (n = 85)		Non-adopters (n = 150)		Mean difference	t-value
Variable	Mean	Standard deviation	Mean	Standard deviation		
Experience (years)	19.31	10.20	21.17	11.89	-1.87^{ns}	−1.21
Max. education (years)	11.02	2.86	10.21	3.23	0.81*	1.92
Farm size (ha)	0.98	1.00	0.85	1.04	0.12^{ns}	0.89
Credit access (yes/no)	0.62	0.49	0.55	0.50	0.07^{ns}	1.04
Extension contact (no.)	13.47	12.89	9.51	9.62	3.96***	2.67
Severity of diseases (score)	1.14	1.11	0.97	1.10	0.17^{ns}	1.16
IPM training (no.)	2.80	4.14	0.14	0.48	2.66***	7.78
FAI (no.)	2.07	2.78	0.73	1.61	1.33***	4.67
Distance from highway (km)	0.69	0.84	0.50	0.51	0.18**	2.09

Note: *, ** and *** indicate significant differences at 10%, 5% and 1% level; t-test was used to test the mean differences; ns indicates not significant

Table 23.3 Percentage of farmers adopting different IPM practices

IPM practices	No. of farmers	Percentage
Sex pheromone trap	24	10.21
Soil amendment with poultry refuse	35	14.89
Both	25	10.63

Table 23.4 Determinants of integrated pest management adoption

	Poisson hurdle regression			
	Logit (hurdle)		Poisson	
Variables	Coefficients	SE	Coefficients	SE
Constant	−3.26***	0.83	−1.83*	1.04
Experience (years)	−0.01	0.02	−0.01	0.02
Max. education (years)	0.10*	0.06	0.07	0.07
Farm size (ha)	0.03	0.15	0.06	0.15
Credit access (yes/no)	0.27	0.36	−0.14	0.40
Extension contact (no.)	0.02	0.02	0.02**	0.01
Severity of diseases (score)	0.24	0.15	0.01	0.18
IPM training (no.)	0.97***	0.22	−0.03	0.05
FAI (no.)	0.24***	0.08	0.15**	0.06
Distance from highway (km)	0.53**	0.26	0.21	0.22
Log likelihood	−160.54			
Wald χ^2	35.76***			
Number of obs.	235			

Note: *, ** and *** indicate significant at 10%, 5% and 1% level

23.3.3 Factors Affecting Adoption

The first part of the Table 23.4 presents the result of logit regression (hurdle) of poisson hurdle model. The findings revealed that decision to adopt IPM practices positively influenced by a wide range of variables. Highest education of the household member positively influenced the adoption decision. Education plays a vital role in understanding and decision making. Farmers having educated family members would have perceived favorably the issues of IPM which may augmented the adoption decision. Number of trainings on IPM ($p < 0.01$) played a significant role in the adoption decision, similar to the findings of [7, 31]. The farmers with IPM training adopted more than those without training because training increases the exposure to and knowledge about IPM. Training also enables farmers to come in contact with specialists who have diversified knowledge which may encourage farmers to adopt more. Positive and significant value of FAI ($p < 0.01$) indicates that adoption decision of neighbour farmers positively influenced the IPM adoption of primary farmers. According to Nakano et al. [32] the farmers who were a neighbour of a new technology adopter were more likely to adopt more than those who were not. This may also warrant field level demonstration in the locality to influence farmers regarding IPM. Positive and significant value of distance from highway ($p < 0.05$) indicates that probability of adopting tomato IPM practices goes up as the distance from highway to primary farmers' field increases which is opposite of general expectation. However, one explanation could be that IPM IL introduced the IPM practices in areas which are far away from highway to enhance the productivity and improve the malnutrition situation of the rural people. This may augment the adoption of IPM in rural areas which are located away from highway.

The second part of Table 23.4 shows the poisson regression result regarding extent of IPM adoption. The findings revealed that after deciding to adopt the tomato IPM, most of the factors were no longer decisive in determining the number of IPM practices adoption. Only extension contact and FAI is positive and significantly associated with the number of IPM practices adoption. Positive and significant coefficient of extension contact ($p < 0.05$) implies that the farmers who are frequently visiting extension personnel have higher level of adoption. Adoption of IPM requires additional knowledge regarding different IPM practices and the farmers who are able to visit extension personnel frequently get their required extension services and adopt more practices. Farmers adopt improved practices if they are aware of the benefits and inherent characteristics of the improved practices [33, 34]. Local extension personnel provide advice and training regarding the benefits of using IPM practices which may play a vital role in increasing adoption level. Neighbour farmers adoption ($p < 0.05$) positively influence both adoption decision and level of adoption which implies that more demonstration and motivational field visit are warranted to enhance the extent of IPM practices adoption. Overall findings may imply that farmers are not well aware about the benefits of IPM practices. More efforts including training, demonstration, and field visit are needed to increase the level of adoption to a great extent.

23.3.4 Impact of IPM on Productivity and Pesticide Application

The tomato farmers who adopted either sex pheromone trap or poultry refuse received higher yield than non-adopters based on kernel and radius matching (Table 23.5). It is also evident from the table that the adopters of both the practices received significant higher yield (4463–5702 kg/ha) compared to non-adopters. The result confirms the findings of [12] which reported that IPM adoption increased tomato yield in Bangladesh. Adoption of sex pheromone trap and poultry refuse practices significantly reduced the frequency of pesticide applications based on kernel and radius matching technique which is consistent with findings of other studies [14, 35, 36] indicated that IPM is helpful in reducing frequency of pesticide applications. The farmers who adopted both the practices received higher yield and also be able to reduce pesticide applications to some extent compared to non-adopters (Table 23.5). Tomato IPM adopters used sex pheromone traps and poultry refuse for mass

Table 23.5 Impact of IPM on productivity and number of pesticide application: PSM estimates

IPM practices	Outcome variables	ATT	
		Kernel matching	Radius matching
Pheromone trap	Productivity	5566** (2648)	4761*(2697)
	Pesticide application	−5.40**(2.47)	−3.97*(2.25)
Soil amendment using poultry refuse	Productivity	5073*(2682)	4215*(2684)
	Pesticide application	−3.95*(2.31)	−4.14*(2.25)
Both	Productivity	5702**(2781)	4463*(2701)
	Pesticide application	−2.49*(1.51)	−3.79*(2.25)

Note: * and *** indicate significant differences at 10% and 1% level, respectively; ns indicates not significant; figures in the parentheses indicates standard error

Table 23.6 Impact of IPM practices on productivity and number of pesticide applications

Outcome variable	Adoption status	ATT	SE
Productivity	1 vs 0	2801^{ns}	2727
	2 vs 0	6483*	3490
	2 vs 1	3081*	1831
Number of pesticide applications	1 vs 0	-1.65^{ns}	2.44
	2 vs 0	−4.80**	2.43
	2 vs 1	−4.03*	2.40

Note: *, and ** indicates significant at 10%, and 5% level, respectively; 0 = Not adopted any of the IPM practices; 1 = adopted any one of the IPM practices; 2 = adopted both the IPM practices

trapping of male moth and to prevent soil borne diseases and improve soil fertility which may increase the yield and reduced the number of pesticide applications.

To check the robustness of PSM analysis, adopters were categorized based on the number (0, 1, or 2) of IPM practices adopted and IPWRA technique was used to compare the results with non-adopters (Table 23.6). The results indicate that there was no significant difference in terms of yield and pesticide applications between adopters of 1 practice and non-adopters. The farmers who adopted 2 IPM practices were received significantly higher yield and reduced the pesticide applications compared to non-adopters. The findings also indicate that the farmers who adopted 2 IPM practices were showing better performance in terms of yield and pesticide applications compared to the farmers who adopted 1 IPM practice (Table 23.6). This suggests that more efforts in terms of training and field days are needed to familiarize the tomato IPM package in the local areas. The author observed limited availability of bio-pesticides and the farmers were found very reluctant to adopt the full package of IPM, this may warrant more research to optimize the tomato IPM package.

23.4 Conclusions and Policy Implications

This study measured the determinants of IPM adoption and farm-level impacts on productivity and pesticide applications using data associated with a subproject of the IPM IL. Findings indicate that training on IPM, neighbour farmers' adoption decision and distance of primary farmer's field from highway played a significant role in initial adoption decision but extent of adoption depends on extension contact and neighbour farmers adoption decision. The farmers who maintained frequent communication with extension personnel adopted more practices. The availability of IPM practices like bio-pesticides at a reasonable price need to be ensured at farmers' field level, if high adoption of IPM practices is to occur. There is need to develop and disseminate an appropriate tomato IPM package with emphasis on biological pest control technique. Extension approaches need to be modified to some extent for better result. Field visit and demonstrations on IPM practices should be arranged frequently to increase the awareness and level of adoption. Neighbour farmers can be used as a media to influence other farmers for higher adoption. Reduction in pesticide applications may have some health and environmental benefits. Higher productivity indicates higher return which may have a positive effect on reducing poverty and malnutrition in the country. Thus, there is need for promoting the role of IPM in anti-poverty policies of Bangladesh. Public sector investment is warranted to provide incentives to institutions involved in transferring IPM practices to farmers.

Acknowledgements The author expresses his profound regard to IPM IL for providing the data and to the farmers who supplied the relevant data and enumerators for their heartfelt co-operation.

References

1. Bhattacharjee, S., Chowdhury, M.A.Z., Fakhruddin, A.N.M., Alam, M.K.: Impact of pesticide exposure on paddy farmers' health. Jahangirnagar Univ. Environ. B. **2**, 18–25 (2013)
2. Islam, M.S., Alam, M.S., Uddin, M.N., Zabir, A.A., Islam, M.S., Haque, K.A., Islam, M.A.S., Hossain, S.A.A.M.: Farm level pesticides use in Patuakhali and Comilla region of Bangladesh and associated health risk. J. Health Environ. Res. **2**(4), 20–26 (2016)
3. Jayasooriya, H.J.C., Aheeyar, M.M.M.: Adoption and factors affecting on adoption of integrated pest management among vegetable farmers in Sri Lanka. Procedia Food Sci. **6**, 208–212 (2016)
4. Muriithi, W.B., Affognon, D., Hippolyte, D.M.G., Kingori, W.S., Tanga, M.C., Nderitu, W.P., Mohamed, A.S., Ekesi, S.: Impact assessment of integrated pest management (IPM) strategy for suppression of mango-infesting fruit flies in Kenya. Crop Prot. **81**, 20–29 (2016). https://doi.org/10.1016/j.cropro.2015.11.014
5. Mian, M.Y., Hossain, M.S., Karim, A.N.M.R.: Integrated pest management of vegetables crops in Bangladesh. In: Muniappan, R., Heinrichs, E.A. (eds.) Integrated Pest Management of Tropical Vegetable Crops, pp. 235–249. Springer, Dordrecht (2016). https://doi.org/10.1007/978-94-024-0924-6
6. Ehler, L.E.: Integrated pest management (IPM): definition, historical development and implementation, and the other IPM. Pest Manag. Sci. **62**(9), 787–789 (2006)

7. Kabir, M.K., Rainis, R.: Integrated pest management farming in Bangladesh: present scenario and future prospect. Environ. Dev. Sustain. **17**, 1413–1429 (2015)
8. Dey, M., Das, S., Kamal, M.M., Sarkar, R.: Performance of different management practices on tomato fruit borer (Helicoverpa armigera Hubner) abundance and infestation. J. Bangladesh Agric. Univ. **14**(2), 161–166 (2016)
9. Hasan, M.N., Hasan, M.M., Haque, M.Z., Howlader, M.H.K., Shanta, U.K.: Adaptibility of tomato genotypes suitable for coastal region of Patuakhali in Bangladesh. Progress. Agric. **28**(2), 84–91 (2017)
10. Ghimire, R., Wen-chi, H., Shrestha, B.R.: Factors affecting adoption of improved rice varieties among rural farm households in Central Nepal. Rice Sci. **22**(1), 35–43 (2015)
11. Chuchird, R., Sasaki, N., Abe, I.: Influencing factors of the adoption of agricultural irrigation technologies and the economic returns: a case study in Chaiyaphum province, Thailand. Sustainability. **9**(9), 1524 (2017)
12. Rahman, S., Norton, G.W., Rashid, M.H.: Economic impacts of integrated pest management on vegetables production in Bangladesh. Crop Prot. **113**, 6–14 (2018)
13. Ahsanuzzaman, A.: Three essays on adoption and impact of agricultural technology in Bangladesh. Department of Agricultural and Applied Economics, Virginia polytechnic institute and state university, Blacksburg, Virginia, PhD thesis (2015). https://vtechworks.lib.vt.edu/handle/10919/53510
14. Gautam, S., Schreinemachers, P., Uddin, M.N., Srinivasan, R.: Impact of training vegetable farmers in Bangladesh in integrated pest management (IPM). Crop Prot. **102**, 161–169 (2017). https://doi.org/10.1016/j.cropro.2017.08.022
15. Dasgupta, S., Meisner, C., Huq, M.: Health effect and pesticide perception as determinant of pesticide use: evidence from Bangladesh. World Bank policy research working paper 3776 (2005)
16. Akter, M., Islam, M.N., Afrin, H., Shammi, S.A., Begum, F., Haque, S.: Comparative profitability analysis of IPM and non-IPM technology on vegetable cultivation in selected areas of Kishoreganj district in Bangladesh. Progress. Agric. **27**(3), 311–319 (2016)
17. Mullahy, J.: Specification and testing of some modified count data models. J. Econ. **33**, 341–365 (1986)
18. Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press, Cambridge (1998)
19. Saffari, S.E., Adnan, R., Greene, W.: Handling of over-dispersion of count data via truncation using poisson regression model. J. Comput. Sci. Comput. Math. **1**(1), 1–4 (2011)
20. Hellstrom, J.: Count data modelling and tourism demand. UMEO economic studies No. 584, UMEO university, Sweden (2002)
21. Yen, S.T., Andrews, M., Chen, Z., Eastwood, D.: Food stamp program participation and food insecurity: an instrumental variable approach. Am. J. Agric. Econ. **90**(1), 117–132 (2008). https://doi.org/10.1111/j.1467-8276.2007.01045.x
22. Rejesus, R.M., Palis, F.G., Lapitan, A.V., Chi, T.T.N., Hossain, M.: The impact of integrated pest management dissemination methods on insecticide use and efficiency: evidence from rice producers in South Vietnam. Rev. Agric. Econ. **31**, 814–833 (2009)
23. Yorobe, J.M., Rejesus, R.M., Hammig, M.D.: Insecticide use impacts of integrated pest management (IPM) farmer field schools: evidence from onion farmers in the Philippines. Agric. Syst. **104**(7), 580–587 (2011)
24. Islam, A.H.M.S., Barman, B.K., Murshed-e-jahan, K.: Adoption and impact of integrated rice-fish farming system in Bangladesh. Aquaculture. **447**, 76–85 (2015)
25. Khan, M.A., Alam, M.F., Khan, J.I.: The impact of co-management on household income and expenditure: an empirical analysis of common property fishery resource management in Bangladesh. Ocean Coast. Manage. **65**, 67–78 (2012). https://doi.org/10.1016/j.ocecoaman.2012.04.014
26. Khandker, S.R., Gayatri, B.K., Hussain, A.S.: Handbook on Impact Evaluation: Quantitative Methods and Practices. The World Bank, Washington, DC (2010). http://openknowledge.worldbank.org/handle/10986/2693

27. Schreinemachers, P., Wu, M., Uddin, M.N., Ahmed, S., Hanson, P.: Farmers training in off-season vegetables: effects on income and pesticide use in Bangladesh. Food Policy. **61**, 132–140 (2016). https://doi.org/10.1016/j.foodpol.2016.03.002
28. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data, 2nd edn. The MIT Press, Cambridge, MA (2010)
29. Imbens, G., Woolridge, J.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**(1), 5–86 (2009). https://doi.org/10.1257/jel.47.1.5
30. Allahyari, M.S., Damalas, C.A., Ebadattalab, M.: Farmers' technical knowledge about integrated pest management (IPM) in olive production. Agriculture. **7**, 101 (2017)
31. Haque, M.M., Kabir, M.H., Nishi, N.A.: Determinants of rice farmers' adoption of integrated pest management practices in Bangladesh. J. Exp. Agric. Int. **14**(4), 1–6 (2016)
32. Nakano, Y., Tsusaka, T.W., Aida, T., Pede, V.O.: Is farmer-to-farmer extension effective? The impact of training on technology adoption and rice farming productivity in Tanzania. World Dev. **105**, 336–351
33. Adegbola, P., Gardebroek, C.: The effect of information sources on technology adoption and modification decisions. Agric. Econ. **37**(1), 55–65 (2007)
34. Anik, A.R., Salam, M.A.: Determinants of adoption of improved onion variety in Bangladesh. J. Agric. Environ. Int. Dev. **109**(1), 71–88 (2015)
35. Pretty, J., Bharucha, Z.P.: Integrated pest management for sustainable intensification of agriculture in Asia and Africa. Insects. **6**, 152–182 (2015)
36. Rahman, S., Norton, G.W.: Farm-level impacts of eggplant integrated pest management: a stochastic frontier production function approach. Int. J. Veg. Sci. https://doi.org/10.1080/19315260.2019.1566188

Chapter 24
Consumers Adoption Behavior Prediction through Technology Acceptance Model and Machine Learning Models

Xinying Li and Lihong Zheng

Abstract This paper is to uncover the key factors that influence purchase intention of customers through analysing technology acceptance theories/models, in the current online-to-offline (abbreviated as O2O) mobile commerce, and to improve the prediction accuracy of consumers' adoption behaviour by utilizing machine learning based methods. With a huge amount of smart phone users, O2O mobile commerce derived from electronic commerce (abbreviated as e-commerce) has been growing vastly. There are many research interests has been attracted on online banking, digital wallet, E-tickets, order tracking, supply chain and so on. However, there is little specific study about O2O mobile APP consumers' adoption behaviour. Motivated from the commonly used technology acceptance theories/models, especially, the Unified Theory of Acceptance and Use of Technology (UTAUT) model, this paper is to identify key influencing factors of O2O mobile APP consumers' adoption behaviour. Then, a new model is proposed as an extended version of UTAUT. The new model has been validated through a survey questionnaire conducted in target groups. More significantly, treating consumers adoption behaviour as a binary classification problem, we apply two different types of machine learning based approaches(Linear Discriminant Analysis(LDA) and Logistic Regression(LR)) to predicate the possible action result by taking into consideration of all influencing factors from the collected survey data. Comparing against several other conventional approaches, Logistic regression shows the better predication accuracy. Hence, it will provide better guidance for promotion strategies in a more productive way.

X. Li
Changchun University of Technology, Jilin, China
e-mail: ccgydxlxy@126.com

L. Zheng (✉)
School of Computing and Maths, Charles Sturt University, Wagga Wagga, NSW, Australia
e-mail: lzheng@csu.edu.au

© Springer Nature Singapore Pte Ltd. 2020
A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_24

Keywords The unified theory of acceptance and use of technology · Machine learning · Linear discriminant analysis(LDA) · Logistic regression(LR) · O2O · APP · Adoption behavior

24.1 Introduction

The rapid development of new technologies and mobile online payment system has lifted business of mobile online shopping. Meanwhile, the integration of online and offline services is expedited [1]. O2O is a business strategy that draws potential customers from online channels to make purchases in physical stores [2, 3]. O2O is the embodiment of the integration of online and offline services. Consumers are able to find a shop, choose products then pay for them using the online component, and acquire products or services using the offline component. A mobile application or mobile APP for short, refers to an application software running on mobile devices (such as smartwatch, smartphones, PDA). O2O mobile APP here is defined for the business mode, which combines online and offline, software development companies develop an application running on mobile devices for a business domain, an industry, or a merchant. A survey report by China Internet Network Information Center (abbreviated as CNNIC), an administration and service organization upon the approval of the competent authority and undertakes the responsibilities as the national Internet network information center [4], showed that as of December 2017, the number of mobile internet users in China reached 753 million, increased by 35% when compared the amount of mobile users in 2015. Among them, there are 506 million people, or 67.2% of mobile internet users, conducted online shopping on their phones [1]. The report also showed that the number of mobile online payment users was up to 527 million [1]. It is predicted by research firm Canalys that there is three-fifths of the smartphones shipped in China will be made up of 5G-enabled handsets by 2023.

Using O2O mobile APP, consumers can obtain relevant information about merchants and products for their own needs, and order and pay for them at anytime and anywhere. Mobile Apps have replaced traditional PC programs due to simplify, convenience [5]. It also helps businesses to boost their sales in an easy and fast way. With the rapid growing of O2O mobile commerce, whether consumers adopt a certain O2O mobile APP is very import for merchants and software development companies.

Commonly, SPSS has been used as a main computational tool to explain the data in a sensible way. However, SPSS usually tries to discover the linear function between the inputs and an output. Such a linear way affects the prediction accuracy in the end. Recent advances in machine learning provide much powerful data analysis capability that can discover more explicit relationship among the data non-linearly. Thus, the prediction accuracy can be improved massively.

In this paper, we are going to address the following two main questions:

(1) What factors influence consumers adoption behavior of O2O mobile APP?

(2) Which approach of machine learning is better to predict consumers adoption behavior of O2O mobile APP, logistic regression or linear discriminant analysis?

Through a proposed new Technology Acceptance Theories/Model and machine learning approaches, the key influencing factors contributing to positive customers adoption behavior will be identified through a customized sample survey data. Therefore, more accurate customer's adoption behavior can be predicted accordingly.

The structure of this paper is as follows: In Sect. 24.2, we review the related literature in e-commerce, discuss and compare four technology acceptance theories/models, as well as present some existing machine learning applications for e-business. We examine seven influencing factors individually and propose a new O2O mobile APP consumers adoption behavior model by revising the UTAUT model in Sect. 24.3. Two machine learning methods are elaborated in Sect. 24.4. We then show the experimental results in Sect. 24.5 and finally we conclude in Sect. 24.6

24.2 Literature Review and Theory Foundation

24.2.1 Technology Acceptance Theories/Models

Given that there is not much research has been done on O2O mobile APP consumers adoption behavior analysis, we investigate some existing technology acceptance theories/models and select one of them to analyze O2O mobile APP consumers adoption behavior in order to support our research ideas. Generally speaking, O2O is a type of e-commerce and mobile commerce (abbreviated as m-commerce), while O2O mobile APP is the tool of O2O m-commerce. Therefore, we mainly review the existing literature on e-commerce, m-commerce and m-commerce APP in the recent years. There are many different types of theories/models to be used to analyze the influencing factors of adoption behavior in e-commerce field. The technology adoption theories/models are the most widely used. Within this type, Theory of Reasoned Action (abbreviated as TRA, Fishbein and Ajzen, 1975) is the very first one to gain widespread acceptance in technology acceptance research [6, 7]. To improve on the predictive power of TRA, Ajzen [8, 9] added a perceived behavioral control component to TRA, named Theory of Planned Behavior (abbreviated as TPB, 1985) [10]. Moreover, Technology Acceptance Model (abbreviated as TAM, Davis, 1989) [11] was also developed from TRA. It was the first used to discuss psychological factors affecting technology acceptance [7]. The two important components of the TAM were perceived usefulness and perceived ease of use.

The Unified Theory of Acceptance and Use of Technology (abbreviated as UTAUT, Venkatesh et al., 2003) [12] is one of the latest developments in the field of general technology acceptance models [13]. Venkatesh et al. [12] discussed eight models (Theory of Reasoned Action, Theory of Planned Behavior, Technology Acceptance Model, Model of PC Utilization, Motivational Model, Social Cognitive Theory, Extension of the Technology Acceptance Model, Diffusion of Innovation Model) that each attempted to predict and explain user behavior using a variety of independent variables [13]. UTAUT was created based on the above eight models and it had higher explanatory power (UTAUT was able to account for as much as 70 percent of the variance in usage intention/behavior. In comparison, the maximum of the eight original models was around 40 percent.) [7, 12, 13]. UTAUT provides a solid base to explain why users accept or reject a technology in a specific perspective and it has much potential in enhancing our understanding of technology acceptance [7]. To conclude, we choose UTAUT as the base model of the research model. More details can be found in the following Sect. 24.3.1. Jaradat and Rababaa [14] modified UTAUT to examine which factors affected Jordanian consumers to accept and use m-commerce. Rodröguez and Trujillo [15] analyzed the determinants of purchasing flights from low-cost carrier websites based on UTAUT. The explanatory powers of online purchase intention and use were relatively high (around 60 percent). Lu, Lin and Lin [16] used UTAUT to study the purchase intention of mobile game APP. The result showed that performance expectancy et al. had effects on behavioral intention.

24.2.2 Machine Learning in E-Commerce

Machine learning is “a field of computer science that uses statistical techniques to give computer systems the ability to” learn ‘with data, without being explicitly programmed’ and is used in computer vision, nature language processing, hand writing recognition, object recognition, information retrieval, and social network, etc. [17]. In the e-commerce domain, machine learning mainly applies to product recommendation, product search, customer support, et al. [18]. Zhao et al. [19] used recurrent neural networks and gradient boosting trees to recommend products from e-commerce Web sites to users at social networking sites in “cold-start” situations. Shankar et al. [20] proposed a deep learning approach to build a large scale visual search and recommendation system for e-commerce. Spens and Lindgren [21] built models, using support vector machines and neural networks, to classify voice calls in a customer support center.

But there is rarely research to predict consumers adoption behavior using machine learning in the field of O2O m-commerce which is a newer type of e-commerce. This paper presents that we use UTAUT to construct the research model, then apply machine learning approaches to analyse data, aiming at predicting O2O mobile APP consumers adoption behavior. We choose logistic regression and linear discriminant analysis to respectively analyze data and compared results of the two approaches. The two machine learning approaches always are used to predict behavior. To and Ngai [22] established a model to predict the adoption of online retailing by organizations and used logistic regression to test the data collected. Shi and Marini [23] used linear discriminant analysis to develop a mood recognition tool in order to predict the mood of online customers.

24.3 Model of O2O Mobile APP Consumers Adoption Behavior

In order to get the deep insight of the key factors leading to a good model for O2O mobile APP consumers adoption behavior, we introduce UTAUT model firstly and then discuss each possible influencing factor to construct our proposed model.

This section will discuss the influencing factors of O2O mobile APP consumers adoption behavior based on UTAUT, the existing research results and the actual condition. Then the research model will be constructed.

24.3.1 UTAUT

UTAUT is a kind of technology acceptance theories/models. In recent years, it has been widely used to analyze consumers adoption behavior in e-commerce domain. There are four core components in UTAUT: performance expectancy, effort expectancy, social influence and facilitating conditions [12]. To fits our research purpose, we define these four terms for O2O mobile APP especially in the Table 24.1.

In UTAUT, performance expectancy, effort expectancy and social influence affect behavioral intention [12]. This conclusion has been proved in many other studies [14, 15, 24–27]. Research has found that behavioral intention is the most direct and important influencing factor of actual use/adoption behavior [6, 8, 11, 12]. In addition, our research focuses on the actual use/adoption behavior of O2O mobile APP. Therefore, we consider that each of performance expectancy, effort expectancy and social influence has an influence on the adoption behavior of O2O mobile APP, which is consistent with the findings of Alshehri [13], Zhou [28] et al.

Furthermore, as shown in the original UTAUT [12] and the research results from other scholars [13, 15, 26–29], it has been concluded that facilitating conditions affect the actual adoption behavior. If consumers haven't got any external resources

Table 24.1 Four core components of UTAUT modle

Component	Definition
Performance expectancy	The degree to which consumers believe that using O2O mobile APP will improve their efficiency of work or life
Effort expectancy	The degree to which consumers perceive O2O mobile APP as easy to understand and use
Social influence	Consumers' perceived support level by those who are important to them or those who are influential in their choice of O2O mobile APP
Facilitating conditions	The degree of support of resources required toward O2O mobile APP, including external resources and consumers' personal competency

(e.g. mobile devices, mobile APP, the Internet service quality) and personal competency (e.g. use of mobile devices) of using O2O mobile APP, they will not show good adoption behavior. Hence, in this paper, we propose that facilitating conditions have an influence on the adoption behavior of O2O mobile APP.

24.3.2 Additional Influencing Factors for the New Proposed Model

In addition to the above mentioned four main core factors based on UTAUT model, taking into account the research results of other scholars and the actual condition we take consideration of another three new factors into our new model. Namely, they are information quality, sales promotion and consumer innovativeness. Those three factors have been used in other applications elaborated as follows individually.

24.3.2.1 Information Quality

Information quality can be defined here as a consumers' perception degree to which the information on O2O mobile APP is believed to be accurate, complete, timely, to have content, and to match the expectations for consumers [15]. High-quality information is beneficial to increase consumers satisfaction [30]. This embodies the importance of quality information. Previous researches have investigated that information quality directly or indirectly affect the user acceptance of a new information technology. Filieri and McLeay [31] measured eight dimensions of information quality in order to predict travelers' adoption of information. Rodröguez and Trujillo [15] proved that "the consumers" perception of the quality of website information has a positive effect on trust', and trust affected consumers behavior by influencing behavioral intention. Both Koivumäki et al. [32] and Albashrawi et al. [33] considered that information quality affected user satisfaction, thereby impacting behavior. In conclusion, this paper proposes that information quality has an influence on the adoption behavior of O2O mobile APP.

24.3.2.2 Sales Promotion

In this paper sales promotion can be defined as consumers' perception of the preferential degree on O2O mobile APP. It has two categories. One is price sales promotion. Examples include cents-off deals, sales, coupons, free-gift-with-purchase, rebates/refunds, bonus packs and so on [34, 35]. Another is non-price sales promotion. This category includes free samples, sweepstakes and so on [34, 35]. The most important function of sales promotion is incenting consumers to purchase [36].

Xiao [37] confirmed that sales promotion was positively related to college students' intention of online shopping. Weng and Run [38] found that sales promotion techniques preferences had an influence on consumers' behavioral intention. Neha and Manoj [39] deemed sales promotion was a valuable tool to affect consumers' purchase decision. After depth interviews with consumers who always buy products through O2O mobile APP, we found that they were sensitive to sales promotion. The more promotions and discounts offered by the APP, the more they are willing to use it. Hence, we propose that sales promotion has an influence on the adoption behavior of O2O mobile APP.

24.3.2.3 Consumer Innovativeness

Consumer innovativeness is defined as "the tendency to buy new products more often and more quickly than other people" [40]. Two commonly known aspects are considered in our research. They are innate innovativeness and domain-specific innovativeness.

The definition of innate innovativeness is "the degree to which an individual is receptive to new ideas and makes innovation decisions independently of the communicated experience of others" [41]. The kind of innovativeness is a nature personality trait and doesn't depend on his or her field [42]. Consumers with a high level of innate innovativeness are much easily adopt to new technologies. Bauer et al. [43] considered the higher degree of innate innovativeness the more willing to learn new knowledge. By an empirical study, Ho and Wu [44] proved that innate innovativeness influenced the adoption of new products.

On another hand, adoption behavior of new things is also related to consumers' field, in addition to innate innovativeness [42]. The definition of domain-specific innovativeness is "tendency to learn about and adopt innovations (new products) within a specific domain of interest" [40]. Chao, Reid and Mavodo [45] found that domain-specific innovativeness had a positive and direct influence on really new product adoption. Moreover, Lassar, Manolis and Lassar [46] confirmed that both innate innovation and domain-specific innovativeness significantly and positively affected the adoption behavior.

O2O m-commerce is a relatively new e-commerce type. Thus, it is necessary to consider consumer innovativeness in the analysis of O2O mobile APP consumers adoption behavior. In other words, we believe that consumer innovativeness is another important influencing factor to be considered on the adoption behavior of O2O mobile APP.

Hence, based on the above rationale, the new proposed model which is an extended version of the UTAUT model is shown in Fig. 24.1. We have included another three new influencing factors mentioned above to the original UTAUT model. Finally, the new model (see Fig. 24.1) has seven influencing factors that contribute differently to the consumers adoption behavior.

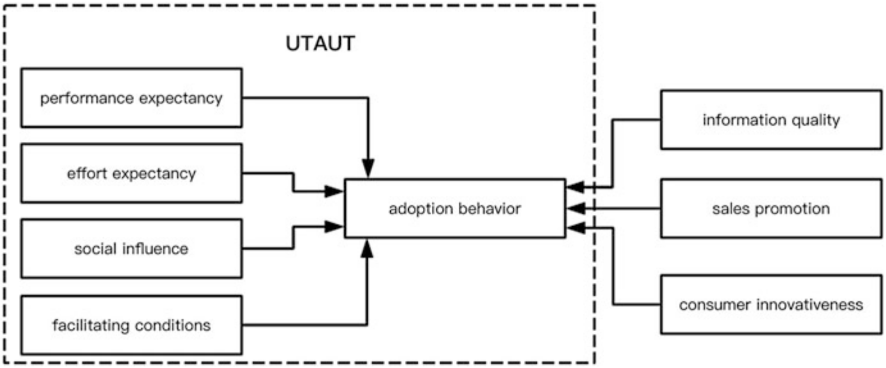


Fig. 24.1 Proposed new model for adoption behavior of O2O mobile APP

24.4 Machine Learning Based Methods

In order to get an objective interpretation of the collected data, in this paper, we apply two machine learning methods respectively predict the consumers adoption behavior. Commonly analysis tool used for such research is SPSS. It is trying to produce a model that represented by a polynomial equation. However, sometimes it is hard to produce the best results due to correlation between input variables. Thus, we are trying to use machine learning based approaches to interpret collected data from a designed model. In this paper, two well-known methods: Linear Discriminant Analysis and Logistic Regression are selected as examples. We explain both methods individually in the following sections.

24.4.1 Linear Discriminant Analysis

Linear discriminant analysis (abbreviated as LDA) [47] is one of linear transformation techniques that are commonly used for dimensionality reduction. LDA is a "supervised" approach. Similar to principal component analysis (abbreviated as PCA), LDA is to combine variables linearly to best present the given data. However, LDA explicitly cares more about the difference between the classes while PCA on the other hand does not consider any difference. LDA can be derived from simple probabilistic models which model the class conditional distribution of the data $p(Xy = k)$ for each class k. For given M training samples for K classes with n features, $M = \sum_{i=1}^{K} N_i$, N_i is the number of training sample of the class K_i, K_i is the n-dimensional mean vectors for the class $K_i(i = 1, , K)$ and is the overall mean of all the training samples. LDA assumes that the data is Gaussian, in other words, each variable is shaped like a bell curve when plotted. More specifically, LDA models $p(X|y)$ as a multivariate Gaussian distribution with density. The following formula can represent LDA in a mathematical way:

$$p(X|y=k) = \frac{1}{(2\pi)^n(|\Sigma_k|)(1/2)} exp(-1/2(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)) \tag{24.1}$$

In order to get the final class output based on the formula above, we need to know the class priors p($y = k$), the class means μ_k and the covariance matrices Σ. The class prior can be calculated by the proportion of instances of class k. The class means μ_k is to average the features of given training samples of class k. Moreover, the covariance matrices include those values of each feature vary around the mean by the same amount via the following formula:

$$\Sigma_i = \sum_{j=1}^{N_i}(X_{ij}-\mu_i)(X_{ij}-\mu_i)^T \tag{24.2}$$

LDA assumes that each class shares the same covariance matrix: $\Sigma_k = \Sigma$ for all k classes, so that, predictions can then be obtained by using Bayes' rule:

$$p(y=k|X) = \frac{p(X|y=k)p(y=k)}{p(X)} = \frac{p(X|y=k)p(y=k)}{\Sigma_i p(X|y=1)p(y=1)} \tag{24.3}$$

In the end, the class k which gets the highest conditional probability is selected as the output class. Thus, LDA makes predictions by estimating the probability that a new set of variables belongs to each class based on Bayes Theorem.

24.4.2 Logistic Regression

Logistic regression (abbreviated as LR), proposed by statistician David Cox in 1958 [48], was such called due that it uses sigmoid function in its representation. Sigmoid function is an S-shaped curve. It can convert any real-valued number into a value between 0 and 1. LR is initially designed for two-class or binary classification problems. It can be extended for multi-class classification but is rarely used for this purpose. LR is an alternative to linear discriminant analysis. If the assumptions of linear discriminant analysis hold, the conditioning can be reversed to produce logistic regression. LR can be seen as a special case of the linear regression.

LR models the probability of an output value (y) via combining linearly input values (x) using weights or coefficient values. The coefficients are to be estimated by maximum likelihood in the training process. Either in a binary classification or in the case of multi-class, the sum of all the posterior probabilities of each class must equal to one.

Given that the conditional distribution $y|x$ is a Bernoulli distribution

$$p(y|x,\theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y} \tag{24.4}$$

Thus, the predicted values are probabilities of (0,1) through the logistic distribution function.

24.5 Experimental Results and Analysis

As discussed previously, there are seven factors to influence consumers adoption behavior of O2O mobile APP. Based on our previous analysis, we have designed a questionnaire and conducted an online survey to the target group whose majority are the University students. The survey questions consist of 27 items to reflect the influences from the above mentioned seven factors. Each question is measured by 5-point Likert scales [49] ranging from strongly disagree to strongly agree. There are about 366 samples collected were valid from the survey. The data collected are used as the main source to feed two different types of machine learning methods. We follow the general procedure of data pre-processing, training and testing. Firstly, the collected scores of the survey are quantitated from 1 to 5 to 0 to 1 as the results of data normalization. In addition, the individual results of potential adoption action as the output have been categorized into two classes. We here use 1 to present stronger adoption behavior, while 0 represents neural or negative adoption behavior. Thus, the pre-processed data can be fed into the machine learning models for training and testing purpose. Generally, we select 70% of the data for training and the rest of 30% is used to validate the prediction accuracy of different models. The accuracy here is calculated by the percentage of the true labels out of the total number of given samples.

The experimental results are given in the following Table 24.2. Compared to the conventional approaches like SPSS, machine learning based approaches take into consideration of all the possible inputs and discover the more complicated relationship between inputs and output in a better way. From the results it clearly shows that LR has higher prediction accuracy.

We have therefore addressed the two guiding questions mentioned at the beginning of this paper clearly.

Table 24.2 Accuracy results of two machine learning based approaches

Method	Accuracy	LogLoss
LDA	75.51%	0.439
LR	78.91%	0.413
SPSS	48.6%	N/A

24.5.1 What Factors Influence Consumers Adoption Behavior of O2O Mobile APP?

This paper takes UTAUT as the theoretical basis, combined with the research results of other scholars and the actual situation of the problem studied, and then we identify seven influencing factors of consumers adoption behavior of O2O mobile APP, and construct a new research model. These seven factors are respectively performance expectancy, effort expectancy, social influence, facilitating conditions, information quality, sales promotion and consumer innovativeness. This new expanded UTAUT model proposed can be used to study similar problems in other research fields such as social networks, e-health and e-government.

24.5.2 Which Approach of Machine Learning Is Better to Predict Consumers Adoption Behavior of O2O Mobile APP?

This paper applies LR and LDA to respectively analyze data and then the results of the two approaches are compared. The experimental results show that LR performs better and the accuracy nearly achieved 80%.

24.6 Conclusions

In this paper, we make two contributions in terms of proposing a new model for consumers adoption behavior of O2O mobile APP, as well as applying machine learning based methods to provide a better consumers adoption behavior analysis. Hence, we have shown better capability of machine learning based approaches in interpreting e-business consumers behavior. Apparently, this paper has some limitations to be considered for future work. Firstly, more data needs to be collected to improve the predictive accuracy. Secondly, we should scale up the scope of survey to include different groups of people, as well as considering genders, ages, job types and so on. Thirdly, this new model does not provide enough information about the importance of each influencing factor. In future, we are going to identify the importance level of each factor. Lastly, we will also investigate the reasons why those consumers don't adopt O2O mobile APP, then construct a model and analyze the importance of each influencing factor. Comparison between the adoption behavior with the non-adoption behavior will provide more meaningful insight for better marketing strategies.

References

1. The 41st China Statistical Report on Internet Development. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249.htm (2018)
2. Online- to- Offline Commerce. https://www.investopedia.com/terms/o/onlinetooffline-commerce.asp, 6 Feb 2018
3. O2O: Why China leads the "online to offline" revolution. https://www.innovationiseverywhere.com/o2o-why-china-leads-the-online-to-offline-revolution/, 15 Feb 2018
4. CNNIC Introduction. http://cnnic.com.cn/AU/Introduction/Introduction/201208/t20120815_33295.htm, 1 Mar 2018
5. Mobile Application (Mobile APP). www.techopedia.com/definition/2953/mobile-application-mobile-app, 2 Feb 2018
6. Fishbein, M., Ajzen, I.: Belief, attitude, intention and behavior: an introduction to theory and research. Addison-Wesley, Reading (1975)
7. Samaradiwakara, G.D.M.N., Gunawardena, C.G.: Comparison of existing technology acceptance theories and models to suggest a well improved theory/model. Int. Tech. Sci. J. **1**(1), 21–36 (2014)
8. Ajzen, I.: From intentions to actions: a theory of planned behavior. In: Kuhl, J., Beckmann, J. (eds.) Action Control. SSSP Springer Series in Social Psychology, vol. 2, pp. 11–39. Springer, Berlin/Heidelberg (1985)
9. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**(2), 179–211 (1991)
10. Theory of planned behavior. https://en.wikipedia.org/wiki/Theory_of_planned_behavior#cite_note-Aizen1991-1 #cite-note-Aizen1991-1 2018/4/6
11. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**(3), 319–340 (1989)
12. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. MIS Q. **27**(3), 425–478 (2003)
13. Alshehri, M., Drew, S., Alhussain, T., Alghamdi, R.: The effects of website quality on adoption of e-government service: an empirical study applying UTAUT model using SEM. In: 23rd Australasian Conference on Information Systems, pp. 1–13. Deakin University, Geelong (2012)
14. Jaradat, M.I.R.M., Rababaa, M.S.A.: Assessing key factor that influence on the acceptance of mobile commerce based on modified UTAUT. Int. J. Bus. Manag. **8**(23), 102–112 (2013)
15. Rodröguez, T.E., Trujillo, E.C.: Online purchasing tickets for low cost carriers: an application of the unified theory of acceptance and use of technology (UTAUT) model. Tour. Manag. **43**, 70–88 (2014)
16. Lin, P.C., Lin Y.C.: A study of the factors affecting the purchase intention on mobile game apps. J. Adv. Inf. Technol. 7(4), 239–244 (2016)
17. Machine learning. https://en.wikipedia.org/wiki/Machine_learning, 8 May 2018
18. Top 5 machine learning applications for e-commerce. https://techblog.commercetools.com/top-5-machine-learning-applications-for-e-commerce-268eb1c89607, 10 May 2018
19. Hao, W.X., Li, S., He, Y., et al.: Connecting social media to e-commerce: cold-start product recommendation using microblogging information. IEEE Trans. Knowl. Data Eng. **28**(5), 1147–1159 (2016)
20. Shankar, D., Narumanchi, S., Ananya, H.A. et al.: Deep learning based large scale visual recommendation and search for e-commerce. ARXIV eprint arXiv:1703.02344 (2017)
21. Spens, H., Lindgren, J.: Using cloud services and machine learning to improve customer support. Uppsala University, Uppsala (2018)
22. To, M.L., Ngai, E.W.T.: Predicting the organizational adoption of B2C e-commerce: an empirical study. Ind. Manag. Data Syst. **106**(8), 1133–1147 (2006)
23. Shi, F., Marini, J.L.: Can e-commerce recommender systems be more popular with online shoppers if they are mood-aware? In: Proceedings of the 12th International Conference on

Web Information Systems and Technologies (WEBIST 2016), vol. 2, pp. 173–180. Science and Technology Publications, Setúbal, Portugal (2016)
24. Abrahão, R.S., Moriguchi, S.N., Andrade, D.F.: Intention of adoption of mobile payment: an analysis in the light of the Unified Theory of Acceptance and Use of Technology (UTAUT). RAI Revista de Administração e Inovação 13(3), 221–230 (2016)
25. Goswami, A., Dutta, S.: E-commerce adoption by women entrepreneurs in India: an application of the UTAUT model. Bus. Econ. Res. **6**(2), 440–454 (2017)
26. Lee, D.C., Lin, S.H., Ma, H.L., Wu, D.B.: Use of a modified UTAUT model to investigate the perspectives of internet access device users. Int. J. Hum. Comput. Interact. **33**(7), 549–564 (2017)
27. Tak, P., Panwar, S.: Using UTAUT 2 model to predict mobile app based shopping: evidence from India. J. Indian Bus. Res. **9**(3), 248–264 (2017)
28. Zhou, T., Lu Y., Wang, B.: Integrating TTF and UTAUT to explain mobile banking user adoption. Comput. Hum. Behav. **26**(4), 760–767 (2010)
29. Yu, C.S.: Factors affecting individuals to adopt mobile banking: empirical evidence from the UTAUT model. J. Electron. Commer. Res. **13**(2), 104–121 (2012)
30. Chae, M., Kim, J., Kim, H., Ryu, H.: Information quality for mobile internet services: a theoretical model with empirical validation. Electron. Mark. **12**(1), 38–46 (2002)
31. Filieri, R., McLeay, F.: E-WOM and accommodation: an analysis of the factors that influence travelers' adoption of information from online reviews. J. Travel Res. **53**(1), 44–57 (2013)
32. Koivumäi, T., Ristola, A., Kesti, M.: The effects of information quality of mobile information services on user satisfaction and service acceptance-empirical evidence from Finland. Behav. Inform. Technol. **27**(5), 375–385 (2008)
33. Albashrawi, M., Motiwalla, L.: When IS success model meets UTAUT in a mobile banking: measuring subjective and objective system usage. In: SAIS 2017 Proceedings, pp. 1–6. AIS, Georgia (2017)
34. Han, R., Tian, Z.: Effects of alternative promotion types on consumers' value perception and purchase intentions. Manag. Sci. China **18**(2), 85–91 (2005)
35. He, L.: Research on price promotion strategy based on consumer perception. Southwest Jiao Tong University, Chengdu (2008)
36. Raghubir, P., Inman, J.J., Grande, H.: The three faces of consumer promotions. Calif. Manag. Rev. **46**(4), 23–42 (2004)
37. Xiao, C.: Research on factors and their effect on college student behavior intention on online shopping. Shanghai Jiao Tong University, Shanghai (2007)
38. Weng, J.T., Run, E.C.: Consumers' personal values and sales promotion preferences effect on behavioural intention and purchase satisfaction for consumer product. Asia Pac. J. Mark. Lofistics **25**(1), 70–101 (2013)
39. Neha, S., Manoj, V.: Impact of sales promotion tools on consumer's purchase decision towards white good(refrigerator) at Durg and Bhilai Region of CG, India. Res. J. Manag. Sci. **2**(7), 10–14 (2013)
40. Roehrich, G.: Consumer innovativeness concepts and measurements. J. Bus. Res. **57**(6), 671–677 (2004)
41. Midgley, D.F., Dowling, G.R.: Innovativeness: the concept and its measurement. J. Consum. Res. **4**(4), 229–242 (1978)
42. Gui, M.J.: Empirical study on the influential factors of the using intention of individual online banks. Zhejiang University, Hangzhou (2007)
43. Bauer H.H., Barnes, S.J., Reichardt, T., Neumann M.M.: Driving consumer acceptance of mobile marketing: a theoretical framework and empirical study. J. Electron. Commer. Res. **6**(3), 181–192 (2005)
44. Ho, C.H., Wu, W.: Role of innovativeness of consumer in relationship between perceived attributes of new products and intention to adopt. Int. J. Electron. Bus. Manag. **9**(3), 258–266 (2011)
45. Chao, C.W., Reid, M., Mavondo, F.T.: Consumer innovativeness influence on really new product adoption. Australas. Mark. J. **20**(3), 211–217 (2012)

46. Lassar, W.M., Manolis, C., Lassar, S.S.: The relationship between consumer innovativeness, personal characteristics, and online banking adoption. Int. J. Bank Mark. **23**(2), 176–199 (2004)
47. Trevor, H., Robert, T., Jerome, H.F.: The elements of statistical learning: data mining, inference, and prediction. Math. Intell. **27**(2), 83–85 (2005)
48. Cox, D.R.: The regression analysis of binary sequences. J. R. Stat. Soc. **20**(2), 215–242 (1958)
49. Jamieson, S.: Likert scales: how to (ab)use them. Med. Educ. **38**(12), 1217–1218 (2004)

Chapter 25
Shaping the Future of Multidimensional Project Management in Retail Industry Using Statistical and Big-Data Theories

Jennifer Hayes, Azizur Rahman, and Md. Rafiqul Islam

Abstract IT projects are by nature, complex and chaotic with a significant proportion failing, assessed as 'not meeting requirements', experiencing overruns in time, budget or scope or not determined acceptable by sponsors and stakeholders. This paper presents a literature review, focused on defining a project initiation and governance framework rooted in complexity theory and bound to the Liminal Cynefin framework with the potential to transform IT project management by understanding projects from the intersection of chaos, complexity and constraints theories. The findings would assist decision makers in the project management industry to assess the potential complexity of a project in the Concept, Validate and Plan stages, matching these results with adaptive governance models and adaptive project management leadership in order to improve project outcomes. With the traditional hard paradigm of the project management industry advocating quantitative measures of project success criteria and projects still failing against these measures, an analysis of historical projects against an amalgamation of current developments in chaos, complexity and constraints theories, combined with alignment to the Cynefin framework is proposed.

Keywords Retail · Complexity theory · Theory of constraints · Chaos theory · Liminal Cynefin

25.1 Introduction

The retail sector in Australia is characterised by small profit margins, high customer expectations and stringent legislative and governance requirements. Coupled with competitive practices and implicit requirement for speed to market in order to

J. Hayes (✉) · Md. R. Islam
Charles Sturt University, Bathurst, NSW, Australia
e-mail: jhayes@csu.edu.au; mislam@csu.edu.au

A. Rahman
School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW, Australia
e-mail: azrahman@csu.edu.au

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_25

capitalise on consumer trends, the IT projects which enable and support this industry need to enable rather than hinder market agility whilst operating in a highly governed and complex IT environment characterised by large numbers of tightly, and often unexpectedly, integrated applications and platforms. The overarching challenge in IT project management in this sector lies in being able to make sense of one's own space in project delivery, relating and understanding how ones projects fit and are constrained within the organisation; along with negotiating and managing across projects within the same business units, applications, and platforms.

The assumption that traditional project management methodologies and governance frameworks are adequate for dealing with a new era of complexity and oftentimes chaotic IT project environments will be analysed in relation to the sense and decision-making theories often deemed applicable to project management – namely Complexity in which Burnham [1] states, "Explicit technical knowledge, factual documentation, and deterministic rational approaches are of least help"; Chaos Theory which "... focuses on how simple systems give rise to complicated and unpredictable behaviour" [1, 2]; Game Theory, dealing with "... establishing and planning your project to be a 'game' that allows you to maximize the gains and minimize the losses ... " [3]; and the Theory of Constraints focusing on "... removing the most important limiting factor." [4].

The introduction of a relatively new framework, that of Cynefin, and more recently the revised Liminal Cynefin, a "... conceptual framework for making sense of the different landscapes faced within and by projects." [5], will be applied as a potential alternative, with the opportunity to apply this framework both to the IT project landscape itself [6], as well as to the individual project manager competencies. The potential for developing a project initiation model that assesses and ranks projects prior to execution, assigning the relevant delivery methodology and governance structure will be investigated, alongside the development of a ranking and matching tool to align projects with project managers based on the soft paradigm [7] as opposed to the traditional hard paradigms associated with the project management industry to date.

The IT project management landscape, while maturing and improving with a wealth of research in relation to success factors and execution best practices, continues to see the same issues and problems arising daily. As initiatives are rolled out in an effort to improve processes, performance and results, more organisations are increasingly interested in understanding why such initiatives are failing to provide the elusive promise of successful IT project delivery and ultimately realisation of business benefits.

IT project process improvements within the study organisation are focused on the quality of project artefacts with little discernible impact on the identification and prevention of risks, nor outcomes of projects. The complexity of the IT landscape within this organisation and the wider industry is a contributing factor however the challenge ultimately lies within the organisation's response to, and ability to adapt to complexity and uncertainty.

The aim of this study is to determine the optimal alignment of project governance, project manager, and project methodology in order to improve initiation, management and governance of complex and chaotic IT projects in Retail. The study objectives are:

- To critically evaluate historical projects and their artefacts, aligning each to the Cynefin framework
- To assess suitability of governance methodology on historical projects
- To determine extent of project manager hard v soft skills and the impact on project success
- To make recommendations to improve project categorisation, adaptive project governance structures, adaptive project leadership, classifying and evaluating project improvements

The rest of the paper is organised as follows. Section 25.2 identifies the significance of the research. Section 25.3 provides a report from the literature review. Section 25.4 discusses the current state of study within this area and is followed by Sect. 25.5, highlighting the research challenges. The final section provides an overview of the proposed research questions and methodological approach that will be undertaken.

25.2 Research Significance

Within the organisation, decision making on IT projects is predominantly considered the remit of the project manager and the Project Steering Group (PSG), with resultant success or failure of the project resting on key decisions made during the project life cycle. How well equipped the project manager is to be able to make these decisions is a variable often left to chance and dependent on their individual capabilities to make decisions in the Chaotic/Unknowable domain [5]. Project Managers are often ill-equipped to understand complexity and chaos in project management [8] and apply, under the auspices of the organisational project governance structure, tools and techniques inherently more suited to Simple projects [8].

Steps to move away from a 'one size fits all' governance approach have been introduced, however differentiation is bound to PMBOK success criteria of schedule, scope and cost [9], with the expectation that projects follow a predictable pattern and methodology. However under complexity, a "Project has within itself the capacity to interact with its environment resulting in a whole that cannot be understood by analysing its constituent parts [8]. The necessity to align towards a new paradigm and "Utilise the complex responsive process of relating for influencing project outcomes instead of implementing rigid control structures that historically have failed" [8, 10] is necessary to improve project outcomes. A number of authors have refuted the notion of predictability in project management, arguing that complexity, chaos, uncertainty and constraints in the IT project management domain necessitate a different type of leadership than that understood under the hard paradigm, a paradigm characterised by quantitative tools and techniques and controlling management methods [7]. Pollack further posits that a shift towards the interpretivist soft paradigm encompassing qualitative measurements and embracing ambiguity and participation [7], is required in order to improve project outcomes in relation to complex projects, in particular those projects categorised as such using the Cynefin framework [5].

The soft skills of leadership go beyond management and allow the project manager to thrive in complex and chaotic environments where "Leading is more important than managing ..." [2], and it is not "... possible to calculate the chaos that is inevitable in a project..." but it is possible to "... develop better strategies for dealing with the changes that are brought about through disorder and unpredictable circumstances [8]. This different skill set requires increased maturity in being able to deal with uncertainty and decision making [7].

25.3 Literature Review

A moderate amount of research on Cynefin as applicable to the IT project management field has been published and is predominantly theoretical. Empirical studies are limited and not undertaken in a large, high volume retail organisation. There appears very little attention to the field of project complexity and the alignment of the Cynefin framework to this field in order to improve project initiation and adaptive governance structures. The recency of the Liminal Cynefin framework has seen no current research available pertaining to Retail IT project management, although it is expected that work is progressing in this regard and will begin to appear in publications and journals as the framework is formalised and published. The following literature review focuses on the individual theories and frameworks as bringing them together has seen very little substantial research undertaken.

25.3.1 Chaos & Complexity

An understanding of the differentiation between chaos and complexity is essential to positioning this research. Chaos theory helps to explain observable phenomenon, an unpredictable set of occurrences on a project or undertaking that "... threaten to throw a project off its planned track." [2]. While it may be desired to separate the two concepts, they are in truth fundamentally and inextricably linked. Singh et al., offer an explanation of how this may be visualised, "... the more general name for the field of chaos is "complexity theory", under which "chaos" is a particular mode of behaviour" [2].

In a somewhat philosophical discussion, unavoidable when attempting to understand random and unpredictable behaviour, Singh et al. raise the practice of attribution of project success to a project managers advanced skills and management while blaming the organisation or external factors for failures. The unpredictability of events that caused the failures, even though the same failures often repeat, is not determined as chaotic or complex. Organisational tendency is to replicate the thinking of the project manager of a successful project, basing governance and project frameworks on such successes, reasoning that the patterns that existed to create the success must therefore have contributed directly to the outcomes. Kurtz and Snowden term this 'the assumption of order' where the implications of causal behaviour and a 'right choice' are made. [11]. Even though chaos and complexity teaches us that random events are unpredictable, structures put in place in organisations would suggest otherwise. [2] Similarly, the tools and techniques used to deal with chaos and complexity assume that of order, rational choice and intentional capability are universally true and this is not the case. [5]. This additionally applies to project management methodologies wherein the application of methodologies designed for use in ordered contexts is not suitable in unordered contexts.

Chaos theory destabilises management approaches to strategic planning and management by providing theoretical reasons for why traditional approaches do not work in modern organisations. Filtered down to the project management level of an organisation, Singh et al. provide the following insight for implications to projects "...(PM) must begin to pay greater attention to the non-linear and subtle influences ... shift away from ... quantitative analysis and project controls." [2] Mikulecky, referenced in Burnham [8], quotes, "Complexity is the property of a real world system that is manifest in the inability of any one formalism being adequate to capture all its properties." Central to the concept of complexity is an anti-reductionist approach. [12]. Traditional project management tools and methodologies are reductionist. This approach has been attributed to Rene Descartes and refers to breaking down into component parts [11], for example a project Work Breakdown Structure (WBS). Complexity cannot be managed with tools and methodologies designed for reductionist approaches. Dahlberg [12] argues that complex systems further cannot be reduced to their component parts and measured or understood in isolation. Kurtz et al. contrasts the approach to managing in the domain of order where optimisation is achieved by optimising the component parts, with management in the domain of unorder where "...the whole is never the sum of the parts" [5]. Without the interactions and interdependencies that are created due to the emergent properties, complex systems cannot be understood [5]. This viewpoint will have a direct bearing on the effectiveness of governance models designed and enforced with the aim to formalise and standardise project management within an organisation, expecting to improve project outcomes by employing a one size fits all strategy. However Dahlberg [12] warns that failures may "...be seen as the outcome of continuous application of linear predictive methods on unpredictable complex systems ... have produced project disasters ...". Burnham concludes that an understanding of chaos and complexity perspectives pertaining to project management is able to offer a "...more profound view of the dynamic nature of projects and can offer empowerment to the PM's leading them." [8]. Such theories appear to offer a panacea to current issues in IT project management, however the lack of application in real world scenarios and the reasons behind this slow uptake require investigation.

25.3.2 The Theory of Constraints

The Theory of Constraints system of business management led, in 1997, to Dr. Eli Goldratt's Critical Chain theory, described in a book of the same name. (Hitchner, Housden & Kania [4] Depicting urgency and focus as the keys to gaining the elusive 'competitive edge', the theory seeks to highlight key tasks or constraints within projects that are pivotal at a point in time to the success or failure of a project. These tasks form the system constraints around which the theory revolves [13] with all other decisions being subordinated to these constraints. The premise behind this approach focuses on the phenomena of everything being urgent subsequently resulting in nothing being urgent due to the lack of real focus on those few key areas that have the potential to derail a project without warning. According to Rand [13], this theory is a 'breakthrough solution' to familiar project problems including late completion, overspending and under-delivering. The lack of research evidence available suggests a lack of practical application of this method in modern organisations, such observation being supported by Hitchner et al. [4], wherein it is stated that critical chain project management "Does not appear to be widely practised."

While the method appears sound, the lack of adoption may result more from the challenges in behavioural management than any shortcomings of the project management approach. The 'human' element warrants robust discussion in the literature, specifically on the difficulties inherent in removing the 'safety buffers' [4] added in automatically by resources targeting achievement of performance management metrics in lieu of minimising project schedule and cost to business. Rand [13] cites the 'psychology of the workforce' in being a deterrent to widespread adoption, with a need to change management approaches to both resource behaviour, and to the technical aspects of project management in order to extract maximum value from Goldratt's approach. Hitchner et al. [4], propose that the benefits of critical chain project management have the potential to significantly improve project outcomes, although the authors note that a significant cultural change would be required in order to remove the reliance on safety buffer estimation practices and the practice of multitasking as a measure of productivity. Application of the Theory of Constraints to improvement of IT Project Management via employing the Critical Chain approach has the potential to significantly influence project outcomes. Real world application in a large-scale retail organisation may provide the opportunity to review suitability of this approach across a broad range of IT projects.

25.3.3 Cynefin

Cynefin is a phenomenological framework based on three distinct ontological states (order, complexity and chaos) [5], which aims to make sense of situations and challenges via alignment across four different domains:

- Simple (Known) domain
- Complicated (knowable) domain
- Complex (Unknowable) domain
- Chaotic (Unknowable) domain [5, 10]

Cynefin aims to bring chaos, complexity and constraints together into one simple to visualise and apply framework. The literature largely agrees on the concept of Cynefin as a sense-making framework with little variation across authors, and all positive concerning the potential of Cynefin to improving decision-making. French [14], does state that the Cynefin framework provides nothing that is new, but does view the framework as providing a valuable basis for discussions to be had. A detailed review of the different domains and their characteristics will not be undertaken within this paper as all sources reviewed are in agreement on the basics of Cynefin and its domains. All authors who reference Cynefin have provided an outline of the domains, all of which are close to identical. The framework itself is not under review or question. The research does however, vary in descriptions and treatment of Disorder, the fifth domain of Cynefin placed at the centre of the framework. It is described by Shalbafan & Leigh, as the "Cause of fleeing to the 'safe' context of the Obvious while failing to realise that uncertainty and chaos are merely learning tools . . . " while also calling it the 'central pivot – an indicator of its potential for destruction' [10]. Further research will aim to understand how to keep projects from starting out or falling into this domain.

Vasilescu [15] situates Cynefin within the historical evolution of decision-making frameworks. Post Dewey, Military and Mintzberg, Cynefin is discussed as responding to a modern business environment that is "... fluid, ambiguous and uncertain." [15]. Fierro, Putino and Tirone [16] additionally qualify modern organisations as complex structures, requiring an understanding and management of complexity. Central to their journal article is the position that a "Better construct to understand and manage complexity is the Cynefin framework" [16]. Key to Vasilescu's discussion on Cynefin is the notion that businesses/projects must make ask the right questions and make decisions according to the prevailing environment. This stands in contrast to traditional, rational sense-making models which are ineffective where instability and unpredictability are evident. Detailed discussion on the Cynefin framework and its domains is not a focus of Vasilescu's research, with the paper focusing on attributing failures to the attempt to apply a standardised leadership model regardless of the "... prevailing operative context" [10], while promoting the ability of Cynefin to assist in identifying and applying contextually appropriate management methods in each situation.

Cynefin is not a categorisation tool yet it hasn't prevented it being used as such, for example ISQUA uses Cynefin to "categorise and evaluate . . . " Snowden himself, in a key paper on the Cynefin framework, also refutes the notion of Cynefin as a categorisation tool, but does however submit that he does "... sometimes use the contextualized Cynefin framework for categorisation within a particular context . . . " [5] Other authors such as French [14] discuss the framework as, "... saying little that is new, provides an intuitive backdrop for discussing many analytical processes . . . " Snowden agrees with the view of non-uniqueness of the framework, relating his team's research back to ancient origins and a bringing together of a number of phenomena, ideas, philosophies, and other aspects to create the Cynefin framework in its present state [5].

Utilising the Cynefin framework requires understanding of the domains and the way in which movement across domains occurs. The Liminal Cynefin framework is a work in progress and aims to represent the movement across domains, clearly defining that projects are expected to move across domains as they are initiated and progressed. Kurtz et al. note that it is this movement that has the potential to create issues in projects due to the requirement for " . . . a different model of understanding and interpretation as well as a different leadership style" [5].

25.3.4 The Project Management Paradigm

In discussing leadership styles in project management, Pollack [7] raises the question of whether the traditional hard paradigm of project management, namely quantitative measurements and reductionist techniques, is still appropriate for modern organisations steeped in complexity and uncertainty. Pollack's discussion on the traditional philosophical basis of project management as aligned to the hard paradigm and resulting in 'problem solving, rather than problem structuring', points to the need for control in these methodologies as being more suited to simple projects and project management where clear and stable goals are presumed [7]. French [14] highlights the use of analytical tools used in various analysis situations, including one would assume, in project management, and determines that such tools are useful only in the Known and Knowable Cynefin domains. While this argument has been made of management and leadership styles, French is one of only a handful who have called out the inadequacy of existing tools for the Complex and Chaotic Cynefin domains.

Adopting a style of project management based on the 'soft' paradigm is, Pollack asserts, more likely to result in success in complex project environments and organisations where collaboration is more important than 'command and control' and where projects 'emerge' rather than commence fully planned. This style of project management requires a major shift in thinking for the leadership team, requiring relearning and acceptance of the change and uncertainty inherent in IT projects [7]. This new type of leadership is defined by Fierro et al. [16] as an active form of leadership in comparison to a passive effort, a leadership which "transcend(s) authoritative expertise to mobilise discovery and generate new capacity . . . "

The research on Cynefin raises the need for different leadership or management styles to be employed dependent on the domain under which a project falls [16]. This ability to adapt managerial response and shift decision-making styles is challenging to expect of project managers trained in traditional 'hard' paradigm methodologies, requiring a mindset of flexibility and willingness to embrace complexity. With a predisposition to innate behaviours that bias decision-making [17], many responses to complexity and uncertainty is to default to an intuitive approach rather than a cognitive approach [18]. The default behaviour for project managers is to gravitate towards the management or leadership style they prefer, which may or may not be suited to the particular Cynefin domain prevailing at the time in their project. This risk is one of many identified throughout the literature and is one of the primary factors to be considered when assessing the reasons behind project successes and failures.

25.4 Current Study

The application and adoption of Cynefin in organisations has not received the detailed level of research expected from what appears to be a significant movement forward on the study of complexity in modern organisations and with the recency and ongoing refinement of Cynefin now taking shape as Liminal Cynefin, further research is required to assess the potential for project improvement and management in a high volume, complex IT project environment.

Review of the literature has identified very little insofar as an empirical, real-world, end to end study of primary sources of information. Literature focus has been towards theoretical repetition of the origins, intricacies and potential applications of the Cynefin framework to business management and decision-making scenarios in general. Case studies and empirical analyses that are academically focused and researched are not available. This review has revealed a lack of understanding and guidelines on how the application of the framework may be practically applied to large, high project volume Australian organisations.

Defining a project initiation and governance framework rooted in complexity theory and bound to the Liminal Cynefin framework has the potential to transform IT project management by understanding projects from the intersection of chaos, complexity and constraints theories, and adjusting management and leadership styles as the projects dictate. The management of projects in times of complexity and for which complexity is inherent in the delivery requires a methodology that has itself been derived from complexity. The application of methodologies designed for projects that follow a defined structure and delivery is expected to be determined as a pivotal reason for the failure of a large number of projects undertaken, where failure is taken to include abandoned projects, as well as those where key decisions have resulted in cost, schedule or scope impacts that have significantly changed since project inception. Such a model would result in improved project initiation and execution, with an understanding of the potential pitfalls and concerns surrounding the initiative prior to execution allowing for adequate oversight, management and focus on required mitigations to ensure and improve the quality of project outcomes.

25.5 Research Challenges

The recency of development of the Liminal Cynefin framework aligns to a correspondingly small scope of literature and research discussing the application of the updated framework to project management generally, and IT project management specifically. While Cynefin, prior to the addition of ‘Liminal’ has generated academic, and to a small extent, corporate, research and examples of applications in some industries, discussion has focused predominantly on an explanation of the framework and potential applications. Practical modelling by way of the undertaking of an empirical study in a large retail organisation in Australia that seeks to plot completed projects on the respective domains in order to establish a baseline and initiation framework on which to measure in-flight and future IT projects against has yet to be undertaken.

Correspondingly, there is a lack of detailed research on the potential for a combination of the available theories and frameworks to be applied in order to generate the project success outcomes desired. While some research has proposed suitable project management methodologies that correspond to each of the Cynefin domains, there is no significant study defining governance frameworks for each of the domains that is flexible and adaptable as a project moves through the various domains during its lifecycle. Theories of chaos, complexity, and constraints will be assessed with a view to consideration of an alignment of approaches combined with right-sized governance models that succeed in breaking through the project status quo currently preventing improved results and benefits.

This research will develop a project initiation and governance model incorporating suitable project management methodologies, adaptable governance structures, and guidance on aligning the assignment of project managers, underpinned by historical assessment of results against relevant theories and application of a combination of processes grounded in complexity theory and Liminal Cynefin.

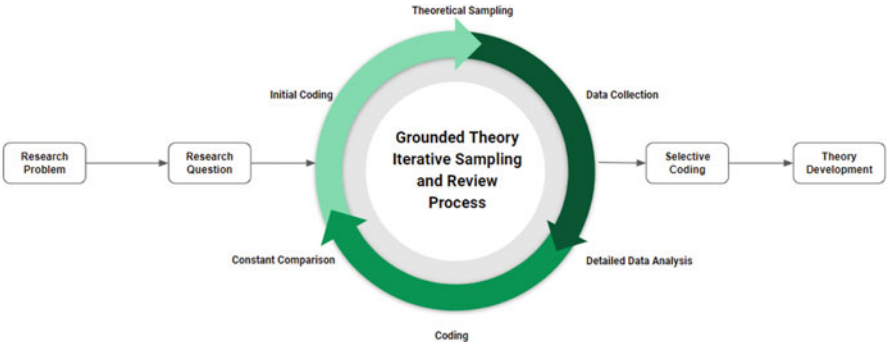
25.6 Research Design and Methodological Approach

This research proposal will explore the following questions:

Will the application of a combination of theoretical approaches based on complexity science that are more aligned to complex organisational realities than the traditional quantitative project management approaches currently in use, result in improved project success outcomes?

This research will focus on the application of qualitative methods based on an underlying interpretive epistemological philosophy [1, 19, 20] in which the researcher is deeply embedded within the social and organisational construct of the IT project management business unit, able to understand the pre-structured words and data existing in the organisational project environment in question [19]. The double hermeneutic principle [19] recognises the researcher as subject, making subjective interpretations in the same manner as those being studied or interviewed, focusing on contextualised meaning-making for situation definition and understanding.

Randomised, multi-stage cluster sampling of projects in various stages of the project life cycle will be undertaken, underpinned by the qualitative research method of grounded theory which will assist in associating outcomes from the data with explanations of the organisational phenomena [19]. The Grounded Theory concept allows a theory to emerge from successive conceptual analyses of empirical data [20], requiring iterative data mining and analysis of IT project artefacts, determining successes and failures based on original baselined scope, schedule and cost against number of variations and project closure scope, schedule and cost.



Decision making within the projects will be assessed via an analysis of the 'Decisions Made' and 'Assumptions' artefacts, relying on interpretive generalisations derived from researcher experience within the organisation for over a decade, allowing acceptable reconstruction of facts based on a shared understanding of meanings and intentions in the organisation [19]. The plausibility of the logical reasoning employed in this research will be validated with the original authors of the documents and artefacts if such persons remain within the organisation.

Complexity Theory, Theory of Constraints, and Chaos Theory will then be applied against this historical data to evaluate the strengths and weaknesses of each in relation to application for sense and decision-making in IT projects in the Retail sector and the potential for improving complex project outcomes by employing a combination of theories and frameworks beginning with initial categorisation of both the project and the project manager using the Cynefin framework. A series of interactive workshops will be held to facilitate the classification of projects across the Cynefin domains, eliciting shared group understanding of project outcomes and further contextualising the research within the organisational construct.

The study will assess projects closed within the previous 3 years, and in progress IT projects, reviewing the past projects to determine whether commonly associated theories or frameworks make sense when applied to the project. Assessment of failures will be undertaken with potential application of the model as a preventative on in progress and future initiatives. The projects and key decisions within each will be mapped onto the Cynefin framework. A parallel assessment of a subset of IT project managers will be undertaken via stakeholder questionnaires and in-flight project satisfaction surveys, with positive correlation between leadership ranking according to the Cynefin framework and projects reported with successful outcomes expected. This research will develop and deliver a framework for new project initiation and governance in the organisation.

25.7 Conclusion

The integration of a variety of approaches to project management, informed by relevant theories of chaos, complexity, and constraints, when combined with projects that are defined and initiated in line with a sense-making framework such as Cynefin and combined with right-sized governance structures, is expected to result in significantly improved IT project outcomes across the organisation.

It is proposed that by developing standardised assessment during project concept and initiation phases against the Cynefin framework and developing governance structures that support projects within and transitioning between each of the domains, improved IT and business expectations for delivery of the projects will result. Coupled with flexible governance structures that adapt alongside the project as it moves through the project lifecycle, the expectation is that IT project outcomes will significantly improve whilst reducing delays and cost overruns when uncertainty and complexity are present. Implications for this study will be to evaluate the effectiveness of the resulting model in a variety of industries and project management organisations.

References

1. Ali, I.: Methodological approaches for researching complex organizational phenomena. Informing Science. Int. J. Emerg. Transdiscip. **17**, 59–73 (2014). Retrieved from http://www.inform.nu/Articles/Vol17/ISJv17p059-073Ali0476.pdf
2. Singh, H., Singh, A.: Principles of complexity and chaos theory in project execution: a new approach to management: a publication of the American association of cost engineers a publication of the American association of cost engineers. Cost Eng. **44**(12), 23–32 (2002). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/220447111?accountid=10344
3. Bockova, K.H., Slavikova, G., Porubcanova, D.: Game theory as a tool of conflict and cooperation solution between intelligent rational decision-makers in project management. Int. J. Econ. Perspect. **10**(4), 147–156 (2016). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/1964460558?accountid=10344
4. Hitchner, K., Housden, G., Kania, E.: The theory of constraints: a unique alternative to traditional project management. Drug Inf. J. **36**(3), 611–621 (2002). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/57586952?accountid=10344
5. Kurtz, C.F., Snowden, D.J.: The new dynamics of strategy: sense-making in a complex and complicated world. IBM Syst. J. **42**(3), 462 (2003). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/222428634?accountid=10344
6. Shalbafan, S., Leigh, E., Pollack, J., Sankaran, S.: Decision-making in project portfolio management: using the Cynefin framework to understand the impact of complexity International Research Network on Organizing by Projects (INROP) 2017, pp. 1–20. UTS ePRESS, Sydney (2017). Retrieved from https://doi.org/10.5130/pmrp.irnop2017.5775
7. Pollack, J.: The changing paradigms of project management. Int. J. Proj. Manag. **25**(3), 266–274 (2007). Retrieved from http://www.sciencedirect.com/science/article/pii/S0263786306001220
8. Burnham, R.: An overview of complexity theory for project management (n.d.). Retrieved from https://www.academia.edu/6908046/An_Overview_of_Complexity_Theory_for_Project_Management?email_work_card=view-paper

9. Pich, M.T., Loch, C.H., De Meyer, A.: On uncertainty, ambiguity, and complexity in project management. Manag. Sci. **48**(8), 1008–1023 (2002). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/213249064?accountid=10344
10. Shalbafan, S., Leigh, E., [UTS]: Design thinking: project Portfolio management & simulation – a creative mix for research. In: Simulation gaming. Applications for sustainable cities and smart infrastructures (2018). Retrieved From: https://www.researchgate.net/publication/325375205_Design_Thinking_Project_Portfolio_Management_and_Simulation_-_A_Creative_Mix_for_Research
11. Ward, J.L.: Untying the Gordian knot of complex projects—a structured approach to complexity. Paper presented at PMI® Global Congress 2005—Asia Pacific, Singapore. Project Management Institute, Newtown Square, PA (2005)
12. Dahlberg, R.: Resilience and complexity conjoining the discourses of two contested concepts. Cult. Unbound J. Curr. Cult. Res. **7**, 541–557 (2015). https://doi-org.ezproxy.csu.edu.au/10.3384/cu.2000.1525.1572541
13. Rand, G.K.: Critical chain: the theory of constraints applied to project management. Int. J. Proj. Manag. **18**, 173–177 (2000). Retrieved from https://www-sciencedirect-com.ezproxy.csu.edu.au/search/advanced?docId=10.1016/S0263-7863(99)00019-8
14. French, S.: Cynefin: uncertainty, small worlds and scenarios. J. Oper. Res. Soc. **66**, 1635–1645 (2015)
15. Vasilescu, C.: Strategic decision making using sense-making models: the Cynefin framework. Defense Resources Management in the 21st Century. (2011). Retrieved from https://search-proquest-com.ezproxy.csu.edu.au/docview/1494348989?accountid=10344
16. Fierro, D., Putino, S., Tirone, L.: The Cynefin framework and the technical leadership: how to handle the complexity CIISE 2017, INCOSE Italia conference on systems engineering, vol. 2010, (2017). Retrieved From: http://ceur-ws.org/Vol-2010/
17. Higgins, G., Freedman, J.: Improving decision making in crisis. J. Bus. Contin. Emer. Plan. **7**(1), 65–76 (2013). Retrieved from http://search.ebscohost.com.ezproxy.csu.edu.au/login.aspx?direct=true&db=tsh&AN=91896039&site=ehost-live
18. Higgins, G., Freedman, J.: Improving decision making in crisis. J. Bus. Contin. Emer. Plan. **7**(1), 65–76 (2013)
19. Myers, M.D.: Qualitative research in business & management, 2nd edn. Sage, London (2013). Myers, Michael, D., (2009)
20. Denzin, N., Lincoln, K., Yvonna, S.: The SAGE handbook of qualitative research, 5th edn. Sage, Los Angeles (2018)

Chapter 26
Technical Efficiency and Value Chain Analysis of Potato in a South-East Asian Country

Mahfuza Afroj, Mohammad Mizanul Haque Kazal, Imtiaz Faruk Chowdhury, and Md. Mahfuzar Rahman

Abstract This study was conducted to assess the technical efficiency, capital financing and value chain analysis of potato in Bangladesh from November 2015 to April 2016. In total 252 farm households were selected as sample from top four potato growing districts Rangpur, Munshigonj, Bogra and Rajshahi. Additionally, 12–13 small and medium rural traders were selected from each district. In the case of retailers, 61 were chosen, 10 from each district while 21 were selected from the capital, Dhaka. Similarly for urban wholesalers 8 were selected from each district and 18 from Dhaka. Stochastic frontier model was used to analyse data and found that among the different variables K fertilizer, irrigation and seed cost is statistically significant at different levels. Additionally, technical inefficiency model find education, farm size and credit is statistically significant at 5% level of significant. These variables have great impact on efficiency of potato cultivation. The govt. should support on N fertilizer, pesticide and other statistically insignificant variables to increase the efficiency of the farmers at higher level. In potato value chain, farmers, small traders and large traders *(Bapari)* share was 28.57%, 13.07% and 11.58% respectively. The share of *Aratdar* was 15.87% and 8.44% for commission agent. The rural and urban retailers share is 9.23% and 11.48%. The share of the processing company and cold storage was 42.31% and 33.33% respectively. The information obtained in this study will help the policy maker to take such policy that helps to increase the efficiency at farmer's level and develop efficient potato value chain in Bangladesh.

M. Afroj (✉)
Department of Agribusiness and Marketing, Faculty of Agribusiness management, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh

M. M. H. Kazal
Department of Development & Poverty Studies, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh

I. F. Chowdhury · Md. M. Rahman
Department of Agronomy, Faculty of Agriculture, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh

© Crown 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_26

Keywords Technical efficiency · Marketing channel · Value chain

26.1 Introduction

Potato (*Solanum tuberosum* L.) is one of the major food crops of the world. It is both a vegetable crop as well as cash crop. Potato is cultivated as a staple food crop in at least 40 countries [1]. It is an important food crop from the very beginning of human civilization and occupying its position just after wheat and rice both in respect of production and consumption [2]. Bangladesh is the 8th potato producing country in the world [3]. The total area under potato crop, national average yield and total production in Bangladesh are 444,534.41 hectares, 19.35 t ha^{-1} and 8,603,000 metric tons, respectively [3]. The total production is increasing day by day as such consumption also rapidly increased in Bangladesh [4]. However, due to lack of proper marketing facilities farmers do not get fair price even sometime they cannot afford to recover production cost [5]. The growers may sell major part of their produces immediately after harvesting at a very low price due to lack of storage facilities and cash need of the farmers [6]. Farmers are compelled to spoil potato in the most potato growing areas of Bangladesh [7]. If farmers fail to sell their produce at an incentive price they are likely to discontinue its production, which may adversely affect the economy [8]. Thus, there is a strong need for an efficient marketing system in order to accelerate and sustain potato production and thereby promote agricultural growth in the country. The main objective of this study is to know about the socio-economic status of potato growers and different value chain actors in Bangladesh, to analyze the technical efficiency of potato production and to analyze value chain of potato in Bangladesh.

26.2 Methodology

This study was conducted in Bangladesh from November 2015 to April 2016. In total 252 farm households were selected as sample from 16 villages of top four potato growing districts; Rangpur, Munshigonj, Bogra and Rajshahi. From each district, two upazilas (sub-district) and from each upazila we chose two villages were chosen. The criteria for choosing the upazilas were that one upazila would be from the more developed region of the district while the other would be from the less developed, more remote part. At the same time, 12–13 small and medium rural traders were selected from each district along with the same number of *Aratdars* from rural and semi-rural areas. In the case of retailers, 61 were chosen, 10 from each district while 21 were selected from the capital, Dhaka. Samples were selected by using random sampling technique and a structured questionnaire was used to collect information.

The stochastic frontier model was used to analyse technical efficiency of potato value chain which consists of two parts technical efficiency and inefficiency. General functional form of the model is as follows.

$$Y_i = f(X_i, \alpha) + \varepsilon_i \tag{26.1}$$

Where, Y_i is the level of output of the ith sample farm, X_i is the value of input of the ith sample farm, α is unknown parameters to be estimated and ε_i is the error term that is composed of two independent elements V_i and U_i, such that $\varepsilon_i = V_i - U_i$. The composite error term V_i is the two-sided error term, and U_i is the one-sided error term. The components of the composed error term are governed by different assumptions about their distribution. The random (symmetric) component V_i is assumed to be identically and independently distributed as $N\ (0,\ \sigma_{v2})$ and is also independent of U_i. U_i is associated with the technical inefficiency of the farmers and are assumed to be independently and identically distributed truncations of the half normal distribution as $N\ (0,\ \sigma u2)$ and also independently distributed of V_i.

The parameters of the stochastic frontier model can be consistently estimated by the maximum-likelihood estimation method. The variance of the parameters of the likelihood function are estimated as:

$$\sigma_S{}^2 = \sigma_v{}^2 + \sigma_u{}^2$$

and

$$\gamma = \sigma_u{}^2/\sigma_s{}^2 = \sigma_u{}^2/\left(\sigma_u{}^2 + \sigma_v{}^2\right)$$

Following their studies, the error specification in Eq. (26.1) is

$$\varepsilon_i = g(X_i, \beta)\,[V_i - U_i] \tag{26.2}$$

Thus, from Eqs. (26.1) and (26.2) we have

$$Y_i = f(X_i, \alpha) + g(X_i, \beta)\,[V_i - U_i] \tag{26.3}$$

Equation (26.3) is the specification of the stochastic frontier production function with flexible risk properties [9]. The mean and variance (risk function) of output of the ith farmer given the values of inputs and technical inefficiency effect can be estimated as

$$E\,(Y_i \backslash X_i, U_i) = \mathrm{f}\,(X_i, \alpha) - \mathrm{g}\,(X_i, \beta)\,U_i \tag{26.4}$$

and

$$Var\,(Y_i \backslash X_i, Ui) = g2\,(X_i, \beta) \tag{26.5}$$

Using this variance (risk function), the marginal production risk can be obtained by partial derivative of variance of production with respect to inputs which can be either positive or negative. That is

$$\frac{\delta Var\,(Y_i X_i/U_i)}{\delta X_{ij}} \quad > 0 \text{ or } < 0 \tag{26.6}$$

Accordingly, the technical efficiency of the ith farmer (TE_i) is defined by the ratio of the mean production for the ith farmer (given the values of the inputs, X_i, and its technical inefficiency effect, U_i) to the corresponding mean maximum possible production (production with no technical inefficiency) can be specified as

$$\mathrm{TE_I} = \frac{E\,(Y_i/X_i, U_i)}{E\,(Y_i/X_i, U_i = 0)} = 1 - \mathrm{TI_i} \tag{26.7}$$

Where TI_i is technical inefficiency defined as potential output loss and represented as

$$\mathrm{TE_i} = \frac{U_i.g\,(X_i, \beta)}{E\,(Y_i/X_i, U_i = 0)} = \frac{U_i.g\,(X_i, \beta)}{E\,(Y_i/X_i, U_i = 0)} \tag{26.8}$$

If the parameters of the stochastic frontier production function are known, the best predictor of U_i would be the conditional expectation of TE_i, given the realized value of the random variable $E_i = V_i – U_i$ [10]. It can be shown that $U_i/(V_i - U_{i)}$ is distributed as $N(\mathrm{m_{i,}}\ \mathrm{s_i}^2)$.

Where $\mu_i{}^*$ and $\sigma_*{}^2$ are defined by

$$\mu_i^* = -\frac{(V_i - U_i)\,\sigma\mathrm{u}2}{(1 + \sigma\mathrm{i}2)} \tag{26.9}$$

$$\sigma^2 = \frac{\sigma u2}{1 + \sigma u2} \tag{26.10}$$

It can also be shown that $E[U_i(V_i - U_i)]$ denoted by,

$$U_i = \mu_i{}^* + \sigma_i{}^* \left[\frac{\varnothing\left(\frac{\mu i *}{\sigma *}\right)}{\phi\left(\frac{\mu i *}{\sigma *}\right)} \right] \tag{26.11}$$

Where, φ(.) and ϕ(.) represent the density and distribution functions of the standard normal random variable. Eq. (26.11) can be estimated using the corresponding predictors for the random variable, Ei, given by

$$E_i = \frac{Yi - f\,(X_i, \alpha)}{g\,(X_i.\beta)} \tag{26.12}$$

After estimating Eq. (26.11), Eq. (26.8) can be estimated as

$$TI_i = \frac{U_i g\,(X_i, \beta)}{f\,(X_i, \alpha)} \tag{26.13}$$

The technical efficiency of the *i*th farmer is predicted by $TE_i = 1 - TI_i$. Technical efficiency of the *i*th farmer can also be calculated as TE_i = exp. $(-U_i)*100$ (TE is converted into percentage through multiplying this equation by 100). It is calculated using the conditional expectation of the above equation, conditioned on the composite error($\varepsilon_i = V_i - U_i$).

The empirical Cobb-Douglas frontier production function with double log form can be expressed as:

$$\begin{aligned} LnY_i = \beta_0 + \beta_1 LnX_{1i} + \beta_2 LnX_{2i} + \beta_3 LnX_{3i} \\ + \beta_4 LnX_{4i} + \beta_5 LnX_{5i} + \beta_6 LnX_{6i} + v_i - u_i \end{aligned}$$

Where, Ln = Natural logarithm, Y_i = Yield of potato of the i-th farm (kg/ha), X_{1i} = Human labor cost by the i-th farm (Tk/ha), X_{2i} = Nitrogen used by the i-th farm (Tk/ha), X_{3i} = Potassium used by the i-th farm (Tk/ha), X_{4i} = Phosphorus used by the i-th farm (Tk/ha), X_{5i} = Pesticide used by the i-th farm (Tk/ha), X_{6i} = Irrigation used by the i-th farm (Tk/ha), $v_i - u_i$ = Error term, v_i's were assumed to be independently and identically distributed random errors, had $N(0,\sigma_v^2)$ distribution.

26.2.1 Technical Inefficiency Effect Model

The u_i' s in equation were non-negative random variables, assumed to be independently distributed such that the technical inefficiency effect for the ith farmer, u_i, were obtained by truncation of normal distribution with mean zero and variance, σ_u^2, such that,

$$U_i = \sigma_0 + \sigma_i Z_{1i} + \sigma_i Z_{2i} + \sigma_i Z_{3i} + \sigma_i Z_{4i} + W_i$$

Where,

Z_{1i} = Farmers age of the i-th farm operator (years), Z_{2i} = Farmers education of the i-th farm operator (year of schooling), Z_{3i} = Family size of the i-th farm operator (years), Z_{4i} = Occupation of i-th farm household head, Z_{5i} = Farm size of i-th farm, Dummy = Credit (0, if not taken lone and 1, if take lone)

W, s were unobservable random variables or classical disturbance term which are assumed to be independently distributed, obtained by truncation of the normal distribution with mean zero and unknown variance, σ^2, such that u_i is non negative.

The β, η and η co-efficient are unknown parameters to be estimated, together with the variance parameters which are expressed in terms of,

$$\sigma^{2=} \sigma_u{}^2 + \sigma_v{}^2$$

$$\gamma = \sigma_u{}^2/\sigma^2$$

γ is the ratio of variance of farm specific technical efficiency to the total variance of output and has a value between zero and one.

The estimates for all parameters of the stochastic frontier and inefficiency model were estimated in a single stage by using the Maximum Likelihood (ML) method. The econometric computer software package STATA was applied to estimate the parameters of stochastic frontier models using the ML method.

26.3 Results

26.3.1 Demographic Profile

The demographic profile of the households indicates a small family size (less than 5) with low levels of child, and especially adult dependency ratios. Table 26.1 shows that average age of family head for the marginal farmer is 44.5, 46.3 for small farmers and 53.7 for large farmers. Females constitute some 48% of the household population. There is some variation across land ownership size groups, particularly for the 'marginal' category (where the female ratio is low, along with child and adult dependency rates).

Additionally, the average education level with less than 5 years is 33.4, with 5–8 years is 32.6, with 9–10 years is average 20.2, household members with 11–12 years education level is 9.2 and 4.6 in case of 12+ years of education. Furthermore, more than 50% of household members reported doing agricultural work within the village. However, very few reported agricultural work outside the village (3%) or nonfarm work (3%). Students or those with 'no occupation' formed more than 40% of the household members. Wage employment was few (around 6%) while another 6% accounted for those who were engaged in non-agricultural self-employment.

Table 26.1 Socio-demographic profile of potato farmers

	Marginal	Small	Medium 1	Medium 2	Large	Total
Age of household						
Total members	4.3	4.6	4.7	4.8	5.6	4.8
Age of head	44.5	46.3	46.5	49.6	53.7	47.7
Female ratio	0.40	0.50	0.48	0.50	0.47	0.48
Child dep. ratio	0.13	0.20	0.16	0.15	0.17	0.16
Adult dep. ratio	0.03	0.06	0.05	0.08	0.08	0.06
Education level						
Less than 5 years	38.5	44.7	31.4	32	24.6	33.4
5–8 years	32.1	31.3	34.6	30.4	32.2	32.6
9–10 years	19.2	15.1	19.1	23.6	24.6	20.2
11–12 years	10.3	5.6	9.4	9.1	13.6	9.2
12 + years	0	3.4	5.5	4.9	5.1	4.6
Total	100	100	100	100	100	100
Household occupation						
Agriculture(in village)	22	57	204	105	37	425
Agriculture(Outside vill.)	1	1	6	19	0	27
Non-farm(In village)	4	2	2	4	1	13
Non- farm(Outside vill.)	0	1	6	1	5	13
No occupation	2	5	23	11	12	53
Student	25	53	133	82	25	318
All	54	119	374	222	80	849

Source: Field Survey 2015
Note: Marginal = <100 decimal; Small = 100–200 decimal; Medium1 = 200–300 decimal; medium2 = 300–400 decimal; Large= > 400 decimal

26.3.2 Land Holding Capacity of Household

26.3.2.1 Land of Household

The households in sample are predominantly agricultural households whose members operate within the village in family-owned farms. About 70.7% of the total cultivated land is owned by the farmer and another 25.6% is rented in (with cash) while just 2.4% is shared-in (Table 26.2).

26.3.2.2 Cropping Pattern

Potatoes are cultivated in 22,090 decimal land and 345 plots and then rice. Small quantities of other crops are also cultivated. This picture is very similar to that obtained from the village data. It should be observed that many of the really small plots combine potato farming with rice crops. In other words, potato cultivation is a very major activity among these households (Table 26.3).

Table 26.2 Land owned, operated, leased or shared-in

Land type	Mean (decimals)	Number of plots	Decimals	%
Owned	34.14	1081	36, 905	70.7
Rented with money	42.76	312	13, 341	25.6
Share-in	29.52	42	1240	2.4
Free	231	1	231	–
Leased out	33.5	8	268	–
Other	39.6	5	198	–
Total		1450	52, 183	100

Source: Field Survey 2015

Table 26.3 Cropping pattern

Crops	Plot	Area (dec.)	Decimal (share)	Plot (share)
Potato	345	22, 090	42	0.24
Boro	24	11, 570	22	0.02
Aman	18	5684	11	0.01
Other	23	4526	09	0.02
Potato mixed	456	5436	10	0.31
(Boro/aman)				
Total	1450	52, 244	1	1

Source: Field Survey 2015
Note: Around 45–50% of the area is used for potato cultivation

26.3.3 Socio Demographic Profile of Traders

26.3.3.1 Profile of Traders

The distribution of sample traders by type and location is indicated below. There are a total of 205 traders in the sample spread across five districts and four categories (small, medium, *Aratdar* and wholesaler).

26.3.3.2 Trade Specialization

While a degree of specialization in trade exists, it is interesting to note that each trader often wears multiple hats. For example, almost all small traders engage in wholesaling but a significant subset (32.1%) also does retailing. Similarly, 94.3% of *Atatdars* are commission agents although 32.1% are also wholesalers. Indeed, the categories of wholesalers and *Aratdars* seem very similar with 64% reporting engaging in commission work (Tables 26.4 and 26.5).

Table 26.4 Trader sample by districts and category

Trader Type	Munshiganj	Rajshahi	Bogra	Rangpur	Dhaka City	Joypurhat	Total
Small	13	13	13	14	0	0	53
Medium	13	13	13	10	0	0	49
Aratdar	13	13	13	14	0	0	53
Wholesaler	8	8	7	8	18	1	50
Total	47	47	46	46	18	1	205

Source: Field Survey 2015

Table 26.5 Trade specialization

Traders type	Number and percent of Traders			Total res.	Total sample
	Business				
	Commission	Wholesale	Retail		
Small	2	51	17	70	53
%	3.8	96.2	32.1		
Medium	11	49	2	62	49
%	22.4	100.0	4.1		
Aratdar	50	17	1	68	53
%	94.3	32.1	1.9		
Wholesaler	32	27	4	63	50
%	64.0	54.0	8.0		
Total responses	95	144	24	263	205
%	46.3	70.2	11.7		

Source: Field Survey 2015
Note: Total responses (percent) can be more than 100% due to multiple responses

26.3.3.3 Age and Education of Traders

The small traders are on average 44.09 years old and medium traders, *Aratdar*, wholesalers have average age of 44.41, 43.04 and 42.88. The maximum education year is 9.1 years which is contained by wholesalers, then *Aratdar,* which is 8.68 years of education. Medium traders had lowest education years which is 7.27 years.

26.3.3.4 Initial Investments

Table 26.6 shows the initial investment of different types of traders. Here, mean of small trader's investment is 68349.06. It is 166224.49 and 192987.74 is for medium traders and *Aratdar*. The wholesaler's initial investment mean is 316449.22 (Table 26.7).

Table 26.6 Age and education of traders

Trader type	Age (year)	Education (year)
Small	44.09	7.83
Medium	44.41	7.27
Aratdar	43.04	8.68
Wholesaler	42.88	9.1
Total	43.6	8.22

Source: Field Survey 2015

Table 26.7 Size of initial investments

Trader Type/Period	Mean	N	Sum
Small	68349.06	53	3, 622, 500
Medium	166224.49	49	8, 145, 000
Aratdar	192987.74	53	10, 228, 350
Wholesaler	316449.22	49	15, 506, 012
Total	744011	204	37, 501, 862

Source: Field Survey 2015

Table 26.8 Investment and assets

Trader type	Stall	Truck or smaller vehicles	Processing investment	Other	Total
Small	48	2	14	3	67
Medium	43	5	17	0	65
Aratdar	51	7	19	2	79
Wholesaler	50	0	3	2	55
Total	192 (93.4%)	14 (6.8%)	53 (25.6%)	7	266

Source: Field Survey 2015

26.3.3.5 Investment and Assets

Over 93.4% of traders (192 out of 205) reported investing in a stall. Another 25.6% invested in avenues related to potato processing. 6.8% (14 traders) also reported investing in transport equipment (Table 26.8). In addition a few cases of investment in cold storage and warehouses were also reported.

26.3.3.6 Source of Investment

Small traders invest BDT 52022 as own money and them borrowed BDT 25666 for investment. Medium traders invest BDT 102076 as own money and BDT 111153 and BDT 273333 was invested as own money by *Aratdar* and wholesalers. The average investment amount was BDT 625514 with over 91% financed by own money (Table 26.9).

Table 26.9 Source of investment and total investment

Trader type		Investment	Own Money	Borrowed	r/i
Small	Mean	52,679	52,022	25,666	12
	N	67	66	3	3
Medium	Mean	106,415	102,076	200,000	0
	N	65	65	1	1
Aratdar	Mean	111,778	111,153	550,625	9.44
	N	79	74	8	8
Wholesaler	Mean	233,825	216,866	273,333	14.25
	N	55	51	12	12
Total	Mean	625,514	482,117	1,381,374	35.69
	N	266	256	24	24

Source: Field Survey 2015

Table 26.10 Characteristics of traditional potato retailers

Characteristics	
Gender (% male)	100
Age(mean) (years)	41
Education (%)	
No schooling	6.56
Primary (1–5 years) only	54.1
More than 5 years but not more than 9	26.23
More than 9 years	13.11
Year started potato retail business	2001
Also sold other food products (%)	100
Share of potato in his/her total retail sales (%)	42.4

Source: Field Survey 2015

26.3.4 Structure of Traditional Potato Retail

Table 26.10 provides information on the characteristics of traditional potato retailers. The retailers surveyed were all males with an average age of 41 years. The majority of them had only primary level education while only about 13.11% of them had schooling over 9 years. All the retailers reported selling other food products and vegetables besides potato and on average potatoes made about 42.4% of their total retail sales. 42.4% of the retailers reported that they grew potato while none were found to be involved with potato processing. Additionally, none of the retailers had stalls at wholesale markets.

26.3.5 Technical Efficiency Measurement

26.3.5.1 Human Labor

It can be seen from Table 26.11 shows that human labor using cost in the potato production process was not statistically significant. This result indicates that human labor could not change the result of yield and co-efficient of labor is positive. It indicates that 1% increase in the cost of using labor in their farming activities will increase yield at 0.009. It has no effect on technical efficiency.

26.3.5.2 N Fertilizer

Table 26.11 shows that N fertilizer using cost in the potato production process was not statistically significant. The result indicates that N fertilizer use cost could not change the result of yield. The co-efficient of N fertilizer is positive. It indicates that higher cost of N fertilizer in their farming activities, they can increase technical efficiency.

Table 26.11 Technical efficiency of potato production

Variable	Parameters	Co-efficient	Standard Error	p-value
Constant	β_0	10.206	0.544	0
Labor	β_1	0.009	0.017	0.583
N fertilizer	β_2	0.021	0.031	0.499
K fertilizer	β_3	−0.082***	0.031	0.007
P fertilizer	β_4	−0.001	0.022	0.982
Pesticide	β_5	0.024	0.055	0.664
Irrigation	β_6	−0.03**	0.014	0.03
Seed	β_7	0.053***	0.016	0.001
Technical inefficiency model				
Constant	σ	−3.754	1.129	0.001
Family size	σ_1	0.05	0.096	0.605
Age	σ_2	0.014	0.013	0.285
Education	σ_3	−0.066*	0.038	0.084
Occupation	σ_4	−0.351	0.617	0.57
Farm size	σ_5	1.205*	0.709	0.089
Credit (dummy): 1, if taken; 0, otherwise)	σ_6	−6.171*	3.775	0.102
Log likelihood value	5.792333			

Source: Field survey, 2015

*** 1% level of significance, ** 5% level of significance, * 10% level of significance

26.3.5.3 K Fertilizer

Table 26.11 represents that K fertilizer using cost in the potato production process was statistically significant at 1% level of significance. The co-efficient of K fertilizer using cost is (−.082) which reveals that 1% increase in the cost of K fertilizer will reduce the yield of potato. The result indicates that N fertilizer cost could not change the result of yield. The co-efficient of K fertilizer is negative. It higher K fertilizer cost in their farming activities can reduce technical efficiency. So the farmer should reduce the cost of K fertilizer.

26.3.5.4 P Fertilizer

It can be seen from Table 26.11 that P fertilizer using cost in the potato production process was not statistically significant. The result indicates that P fertilizer cost could not change the result of yield. The co-efficient of labor is negative and it is (−.001). It indicates that 1% increase in the cost of using P fertilizer in their farming activities will decrease yield at 0.001. It has no effect on technical efficiency.

26.3.5.5 Pesticide

It can be seen from Table 13 that pesticide using cost in the potato production process was not statistically significant. The result indicates that pesticide use cost could not change the result of yield. The co-efficient of pesticide is positive which is 0.024. It indicates that 1% increase in the cost of using pesticide cost in their farming activities will increase yield at 0.024. It can increase technical efficiency.

26.3.5.6 Irrigation

It can be seen from Table 26.11 that irrigation cost in the potato production process was statistically significant at 5% level of significance. The co-efficient of irrigation cost is (−.03) which reveals that 1% increase in the cost of irrigation will reduce the yield of potato. The result indicates that irrigation cost could not change the result of yield. It indicates that increase of irrigation cost in their farming activities can reduce technical efficiency. So the farmer should reduce the cost of irrigation.

26.3.5.7 Seed

Table 26.11 shows that seed cost in the potato production process was statistically significant at 1% level of significance. The co-efficient of seed cost is .053 which reveals that 1% increase in the cost of seed will increase the yield of potato. It indicates that increase of seed cost in their farming activities can increase technical efficiency.

26.3.5.8 Family Size

The co-efficient of family size is positive but not statistically significant. It implies the negative relationship with farming efficiency. Family size has no significant impact on farm efficiency.

26.3.5.9 Age

The co-efficient of age was positive which is 0.014 but not statistically significant. It implies the negative relationship with farming efficiency. Age has no significant impact on farm efficiency.

26.3.5.10 Education

The co-efficient of education was negative and it is (−0.066). It is significant at 10% level of significant. It reveals that technical inefficiency decrease with the increase in education level. Thus, the higher educated farmers were technically more efficient in potato production.

26.3.5.11 Occupation

The co-efficient of age was negative which is (−0.351) and it has no statistical significant. It implies the negative relationship with farming efficiency. Occupation has no significant impact on farm efficiency.

26.3.5.12 Farm Size

The co-efficient of education was positive and it is 1.205. It is significant at 10% level of significant. It reveals that technical efficiency decrease with the increase in farm size. Thus, the higher farm size owned farmers were technically more efficient in potato production.

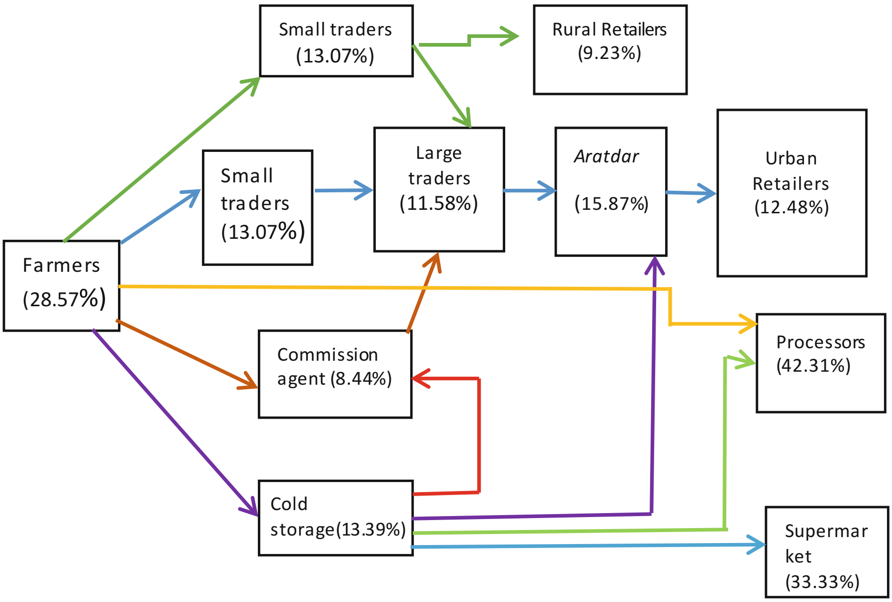
26.3.6 Frequency Distribution of Technical Efficiency

Table 26.12 shows the frequency distribution of technical efficiency estimate from frontier stochastic model. It reveals that the technical efficiency varies from 40% to 99%. The maximum economic efficiency has been observed to be in the range of 80–89%. Here, the maximum farms (135) are efficient as they get proper capital financing in cultivation of potato. However, govt. should support in those variables that are insignificant to make the farmers more efficient.

Table 26.12 Frequency distribution of technical efficiency estimates from stochastic frontier model

Efficiency level (%)	Technical efficiency
40–49%	1
50–59%	0
60–69%	13
70–79%	38
80–89%	135
90–99%	65
Total number of farmer	252
Mean	.857
Minimum efficiency	.489
Maximum efficiency	.995

Source: Field survey, 2015



Flowchart 26.1 Value chain of potato
Source: Field survey, 2015

26.3.7 Potato Value Chain

Flowchart 26.1 shows the potato value chain in Bangladesh and different actors involved in potato value chain has been identified with their share. Here the center point is the producer of potato. Farmers sell potato at 1125 taka per quintal after harvest. From the farmer's commission agent, *Paiker, Baparies, Aratdar*, Cold storage, processing companies purchase potato. Among these purchasers, processors company purchase at high rate and then the cold storage. Here, the farmers share is 28.57%. 13.07 and 11.58% share is for small traders and large traders *(Bapari)*. The share of *Aratdar* is 15.87% and 8.44% for commission agent. The rural and urban retailers share is 9.23 and 11.48%. The share of the processing company and cold storage is 42.31 and 33.33%.

26.4 Discussion

This study aims to analyses technical efficiency and value chain of potato in Bangladesh. Findings from this study contribute to the current literature in technical efficiency and value chain of potato. This study suggest that in potato value chain, farmers share is 28.57%. About 13.07 and 11.58% share is for small traders and large traders *(Bapari)*. The share of *Aratdar* is 15.87% and 8.44% for commission agent. The rural and urban retailers share is 9.23 and 11.48%. The share of the processing company and cold storage is 42.31 and 33.33%. This study confirm by the previous findings by [11] who also find retailers hold lowest share in potato value chain. However, [12] demonstrated that retailers hold highest share in potato value chain which is not similar with the findings of this study. The main reason behind this may include different study area selection and also time gap. Furthermore, this study also explained that among the different variables K fertilizer, irrigation and seed cost is statistically significant at different levels which are confirmed by [13].

26.5 Conclusion

Bangladesh produces a huge amount of potato every year. A relatively high yield and low cost of production of the crop with the introduction of modern technologies have perhaps provided an incentive to the farmers to increase the production of potato which thereby raise the marketable surplus of potato in Bangladesh. However, due to lack of proper marketing facilities farmers do not get fair price even sometime they cannot afford to recover production cost. The growers have to sell major part of their produces immediately after harvesting at a very low price due to lack of storage facilities and cash need of the farmers. The present study intends to find out some of the shortcomings of the existing potato marketing system so that continuous increase in its production can be maintained. This study will therefore, have a scope to deliver the information to the potato growing farmers and value chain actors as well as researcher and policy makers so that they can take proper initiatives to enhance potato production and its marketing and sustainable growth in future.

References

1. Behjati, S., Choukan, R., Hassanabadi, H., Delkhosh, B.: The evaluation of yield and effective characteristics on yield of promising potato clones. Ann. Biol. Res. **4**(7), 81–84 (2013)
2. Karim, M.R., Rahman, H., Ara, T., Khatun, M.R., Hossain, M.M., Islam, A.K.M.R.: Yield potential study of meristem derived plantlets of ten potato varieties (*Solanum tuberosum* L.). Int. Biosci. **1**(2), 48–53 (2011)
3. FAOSTAT (FAO, Statistics Division): Statistical database. Food and Agricultural Organization of the United Nations, Rome (2013)
4. BBS (Bangladesh Bureau of Statistics): Agricultural statistics Yearbook-2013. Statistics Division, Ministry of Planning, Government of the People's Republic of Bangladesh, Dhaka (2013)
5. Hajong, P.: Marketing and storage system of potato in some selected areas of Rangpur district. M.S. thesis, submitted to the Department of Agribusiness and Marketing, Bangladesh Agricultural University, Mymensingh, Bangladesh (2011)
6. Huq, A.S.M.A., Alam, S., Akter, S.: An analysis of marketable surplus of potato in Bangladesh. Asia Pac. J. Rural Dev. **15**(2), 95–102 (2005)
7. Saiyem, M.A.: Marketing system and price behaviour of potato in selected areas of Rangpur district. M.S. thesis. Department of Cooperation and Marketing, Bangladesh Agricultural University, Mymensingh, Bangladesh (2007)
8. Hossain, A.M.J.: Potato marketing in a selected area of Bogra district. M.S. Thesis, Submitted to the Department of Cooperation and Marketing. Bangladesh Agricultural University, Mymensingh, Bangladesh (2004)
9. Battese, G.E., Rambaldi, A.N., Wan, G.H.: Stochastic frontier productions function with flexible risk properties. J. Product. Anal. **8**, 269–280 (1997)
10. Just, R.E., Pope, R.D.: Stochastic specification of production functions and economic implications. J. Econ. **7**, 67–86 (1978)
11. Reddy, K.P.A., Achoth, L.: Channels of potato marketing under irrigated condition in Chikkaballapur and Bangalore markets. J. Indian Potato Assoc. **21**(3–4), 222–225 (1994)
12. Saklayen, M.G.: Marketing of potato in some selected areas of Munshigonj district. M.S. thesis. Department of Cooperation and Marketing, Bangladesh Agricultural University, Mymensingh, Bangladesh (1990)
13. Fuglie, K.O., Khatana, V.S., Ilangantileke, S.G., Singh, D., Kumar, D., Scott, G.I.: Economics of potato storage in northern India International Potato Center Social Science Department Working Paper No. 1997-5. International Potato Center (CIP), Lima (1997)

Chapter 27
Modelling and Analysis of Computer Experiments Using a Simple Pendulum Model

Kazeem Adewale Osuolale

Abstract A computer experiment is an experiment conducted using data obtained from a computer model or simulator in lieu of the physical process. A physics-based experiment known as a simple pendulum experiment was performed to demonstrate a computer experiment. A true model called a computer model of a simple pendulum was used to simulate a real life pendulum experiment. The inputs to the computer code were varied in order to determine the effect of different inputs on the output(s) of a pendulum experiment. The output of such computer model is used as a proxy for the real life observations of the study. The focus of this study is to determine the output, that is, the time it takes the pendulum bob to return to rest. This time is also called the stoppage time in this study. The accuracy of the estimated time of the pendulum based on the calculation of the standard error was checked and 95% bootstrapped Confidence Intervals (CIs) were estimated. MATLAB 2016a computer package (www.mathworks.com/) was used for the development of the program that generates the time it takes the pendulum to return to rest.

Keywords Computer experiment · Simulator · Simple pendulum

27.1 Introduction

Computer experiments are becoming more commonly used in science and engineering. This is mainly because some complex physical experiments may be time consuming, expensive, or even impossible to perform. The rapid growth in computer power has made it possible to perform experiments on simulators. Since the emergence of the first computer experiment conducted by Enrico Fermi and colleagues [1] in Los Alamos in 1953, scientists in diverse fields such as engineering, cosmology, particle physics and aircraft design have turned to computer experiments as a powerful tool to understand their respective processes. For instance, in the

K. A. Osuolale (✉)
Nigerian Institute of Medical Research, Lagos, Nigeria
e-mail: ka.osuolale@nimr.gov.ng

© Springer Nature Singapore Pte Ltd. 2020

A. Rahman (eds.), *Statistics for Data Science and Policy Analysis*,
https://doi.org/10.1007/978-981-15-1735-8_27

design of a vehicle, computer experiments are used to study the effect of a collision of the vehicle with a barrier before manufacturing the prototype of the vehicle, see Bayarri et al. [2]. A simple pendulum is used in this work as a demonstrative example for a computer experiment. A simple pendulum may be described ideally as a point mass suspended by a weightless string from some point about which it is allowed to swing back and forth in a place, Parks [3]. A simple pendulum can be approximated by a small metal sphere which has a small radius and a large mass when compared relatively to the length and mass of the light string from which it is suspended. This work presents the results of a computer experiment that was performed using a simple pendulum model to demonstrate a computer experiment. The pendulum spent the same minimum time to return to rest at different pendulum lengths and displacement angle of 180°. This is a novel application area in the field of computer experiment different from the stochastic simulation experiment. The MATLAB code was written to numerically solve for the time the pendulum stops by iteratively varying the pendulum length and initial displacement angle and calculate the time it takes for the pendulum to stop. A simple pendulum experiment discussed in Osuolale et al. [4] has been modified in this study with some additional statistical estimates to convey the accuracy of the estimated time of the pendulum.

27.2 Materials and Methods

An experimental design is a matrix of input variable values ($\boldsymbol{X}$), where each column of $\boldsymbol{X}$ represents a variable and each row represents the combination of input variable values for a single run of an experiment. Conventional experimental designs according to Montgomery [5] were originated from the theory of Design of Experiments when physical experiments are performed. Computer experiments are different from physical experiments in that they have no random variable and they deal with functions that are thought to have more complex behaviour. Properly designed experiments are essential for effective computer utilization. After the computer experimentation had been carried out based on the simulation model, the next step is to choose an approximate model to fit the process. Motulsky and Christopoulos [6] reported G.E.P. Box to have said that "all models are wrong but some are useful" while many models and methods still exist. A complex mathematical model that produces a set of output values given a set of input values is commonly referred to as a computer model. This model is used to mimic the real life system or physical experiment. More often than not, everybody wants computers to do the extensive computations especially for situations where the intending model does not result to a closed form solution or the one that requires an iterative solution. Computer models are distinct from models of data from physical experiments in that they are often not subjected to random error. In a computer experiment, observations are made on a response function y by running a complex computer model at various settings of input factors X. For instance, in the proposed simple pendulum experiment here, X can be a set of initial displacement angles in a system of differential equation and *y* can *be* the time it takes the pendulum to return to rest. Solving the resulting differential equations numerically for specified X gives a value for y. The goal of this model is to estimate the relationship between X and *y* from a moderate number of runs so that y can be predicted at untried inputs. Observations obtained from this experiment can be used to build up a computationally cheap surrogate model to the selected simulator or computer code. This surrogate model is used to approximate the computer simulator or model efficiently and inexpensively to investigate the behaviour of the function. A computer always produces the same output when it is fed with the same input. Due to the lack of random error, classical statistical modelling approaches are not appropriate**.** For example, one of the features of design of experiments is replication which simply results to redundant information in computer experiments. A major contribution to this area was made by Sacks et al. [7]. In many scientific investigations, complex physical phenomena are represented by a mathematical model:

$$Y = f(X), X \in [0, 1]^m, \tag{27.1}$$

where X consists of m input variables, f is the functional form of the relationship and Y represents the output of a computer experiment. This model is a solution to a set of equations, which can be linear, nonlinear, ordinary or partial differential in nature. Due to the complex nature of $f(X)$ for many real life cases, the solution to Eq. (27.1) is often impossible to obtain analytically. Therefore, the scientists often study the complex relationship between the inputs X and outputs y by varying the inputs to the computer code and observing how the process outputs are affected. The pendulum has the following non-linear differential equation as its model:

$$ML^2\frac{d^2\theta}{dt^2} + b\frac{d\theta}{dt} + MgLsin(\theta) = 0$$

$$\frac{d^2\theta}{dt^2} = \frac{-b\frac{d\theta}{dt} - MgL\sin(\theta)}{ML^2} \tag{27.2}$$

where M is the mass (kg) of the bob, L is the length (metre) of the pendulum, θ is the pendulum displacement angle (degrees), b is the coefficient of friction (Newton) and g is the acceleration (in ms^{-2}) due to gravity. The MATLAB code was written to numerically solve for the time the pendulum stops by iteratively varying the pendulum length and initial displacement angle and calculate the time it takes for the pendulum to return to rest. This model represents a system with two inputs (pendulum length and initial displacement angle) and one output (the stoppage time of the pendulum). The mass, M is specified while the friction (b) is also fixed.

27.3 Implementation

A computer program using MATLAB 2016a package was written to generate values for the input variables and concurrently calculate the output values. The acceleration due to gravity g was fixed at 9.81 ms^{-2} mass was fixed at 0.1 kg and the friction, b was set to be 0.25 N. A complete diagrammatic model for a simple pendulum experiment to represent the non-linear differential equations used in this study is shown in Fig. 27.1 to aid the understanding of the input and output values generated from the pendulum experiment.

The simulink environment visually solves the model using a chosen solver (ode45) to numerically compute the value of θ (pendulum displacement angle). The simulation is run infinitely until the angular displacement becomes zero and the time it occurs determined. The program loads the simulink model and passes the input parameters into the model and returns the stoppage time during each experimental iteration. The Pendulum Length L was varied between 0.2 and 1 m while the initial displacement angle θ was varied between 5° and 180° at regular increment of 5° intervals. The results from this experiment are presented in Table 27.1.

27.4 Results and Discussion

Graphs are presented in this section in order to see at glance, the time the pendulum stops at different degrees of displacement angle when the pendulum length L is 0.2, 0.4, 0.6, 0.8 and 1 m. The line graphs showed separately the time it took the pendulum to stop at each pendulum length used in the experiment. The minimum time it took the pendulum to return to rest was 0.2 s at 180° and the average

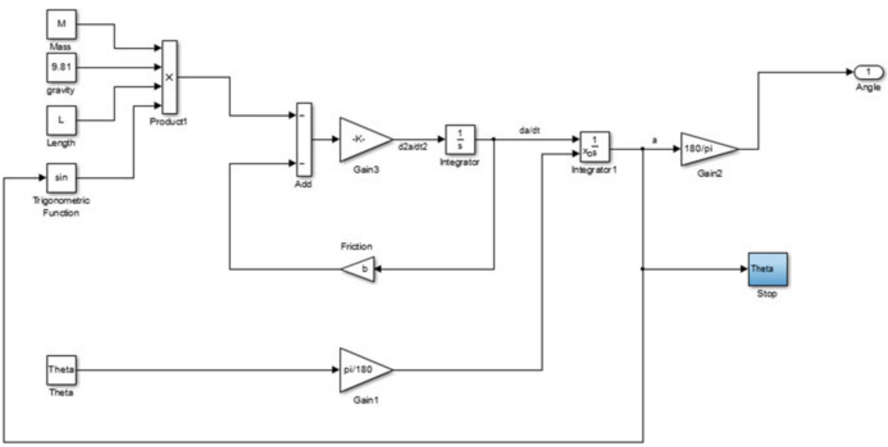


Fig. 27.1 Simple pendulum model build-up

Table 27.1 The estimated stoppage time of the pendulum (in seconds) at various displacement angles for five different pendulum lengths

Initial displacement angle	Stoppage time at pendulum length of 0.2	Stoppage time at pendulum length of 0.4	Stoppage time at pendulum length of 0.6	Stoppage time at pendulum length of 0.8	Stoppage time at pendulum length of 1.0
5	5.19	3.40	4.27	7.49	10.28
10	5.98	5.85	4.25	8.41	11.28
15	4.46	4.23	5.11	8.41	12.32
20	5.79	1.77	5.10	10.25	14.35
25	5.03	3.42	5.93	9.34	14.35
30	5.40	5.05	5.11	9.35	14.35
35	5.97	5.86	5.94	10.34	14.37
40	5.36	6.68	5.93	10.28	15.40
45	5.73	7.53	5.12	10.29	15.40
50	6.16	5.88	5.11	10.29	16.43
55	5.55	6.71	5.11	10.29	15.43
60	6.98	6.72	5.95	10.32	16.45
65	5.50	6.74	6.81	10.31	15.44
70	7.06	5.12	6.82	12.18	15.45
75	5.58	3.48	6.00	10.34	15.48
80	6.83	2.69	6.01	11.36	18.55
85	5.67	2.69	6.85	12.21	16.52
90	6.91	4.35	6.02	11.32	17.56
95	5.75	7.64	6.03	11.33	15.55
100	7.31	5.20	6.90	11.34	17.60
105	5.83	3.58	6.09	11.44	16.61
110	6.66	2.80	6.11	11.47	18.68
115	7.17	2.82	6.11	12.33	18.70
120	6.01	5.31	6.15	11.50	17.72
125	7.25	6.98	6.17	11.46	18.77
130	6.09	4.55	6.19	12.41	18.81
135	6.61	2.96	7.08	12.46	18.85
140	7.45	2.98	6.27	11.56	18.88
145	6.42	6.31	6.32	12.54	18.93
150	6.39	6.38	6.38	12.58	19.00
155	6.50	3.99	6.41	11.73	18.05
160	6.61	3.25	6.52	12.73	18.13
165	8.00	6.62	6.61	12.83	18.23
170	7.00	4.33	6.73	12.11	18.39
175	7.83	4.57	7.00	12.33	18.64
180	0.20	0.20	0.20	0.20	0.20

displacement angle using harmonic mean was 43.12° for all the pendulum lengths used. The maximum time it took the pendulum to return to rest when L = 0.2 m was 8 s at 165° and the average time it took the pendulum to stop using harmonic mean was 3.38 s. It took the pendulum the maximum time of 7.64 s to return to rest when L = 0.4 m at 95° and the average time was 2.71 s. The pendulum spent the maximum time of 7.08 s to stop when L = 0.6 m at 135° and the average time was 3.30s. The pendulum also spent the maximum time of 12.83 s and 19 s to stop when L = 0.8 m at 165° and L = 1.0 m at 150° and the average times were 4.38 s and 5.03 s, respectively. The accuracy of the estimated time of the pendulum based on the length of the pendulum was checked. The stoppage time of pendulum at the length of 0.2 gave the estimated mean of 6.1 with the standard error 0.2 and 95% bootstrapped CI (5.7–6.5) while at the length of 0.4, the mean was 4.7 with the standard error 0.3 and 95% bootstrapped CI (4.1–5.3). The stoppage time of pendulum at the length of 0.6 gave the estimated mean of 5.8 with the standard error 0.2% and 95% bootstrapped CI (5.4–6.1) while at the length of 0.8, the mean was 10.6 with the standard error 0.4 and 95% bootstrapped CI (9.8–11.1). The stoppage time of pendulum at the length of 1.0 gave the estimated mean of 16.0 with the standard error 0.6 and 95% bootstrapped CI (14.8–17.1). All the estimated results were significant based on the bootstrapped confidence interval estimates. The pendulum length at 0.2 and 0.6 gave the smallest estimate of the standard error.

Box plot was also used to compress the information contained in Fig. 27.2. This enables easy comparisons of the median stoppage times at different chosen pendulum lengths. From the boxplots, it can be observed that the minimum and maximum stoppage times of the pendulum at all the displacement angles were attained at the pendulum lengths of 0.4 and 1.0 m, respectively (Fig. 27.3).

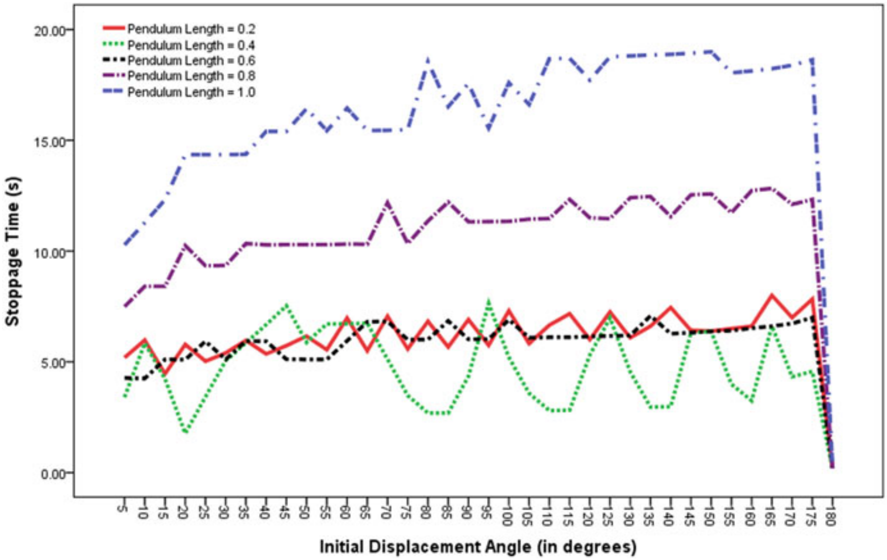


Fig. 27.2 Line graphs of the stoppage time against initial displacement angles (degrees) for five different pendulum lengths

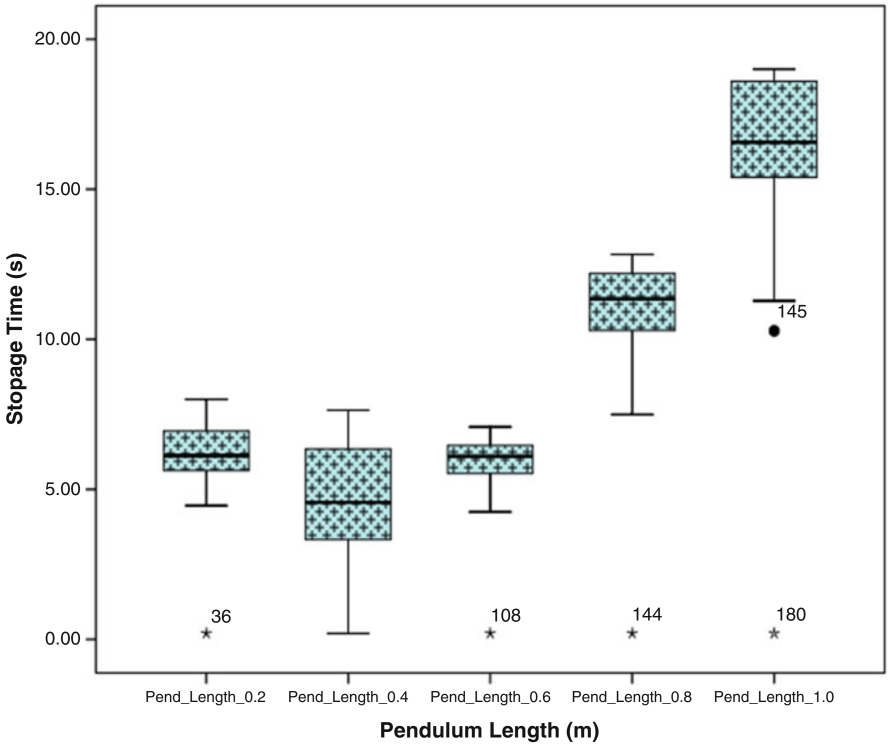


Fig. 27.3 Box plots of the stoppage times at the five chosen pendulum lengths over all the displacement angles

27.5 Conclusion

This study presents the results of a computer experiment that was performed using a simple pendulum model to demonstrate a computer experiment. The pendulum spent the same minimum time to return to rest at different pendulum lengths and displacement angle of 180°. The maximum time of 19 s was spent by the pendulum to stop when L = 1 m at displacement angle of 150°. This study has taken a novel application area in the realm of computer experiment different from the stochastic simulation experiment. The mass can also be varied as done for pendulum length and pendulum displacement angle to simulate the stoppage time. All the estimated results were significant based on the bootstrapped confidence interval estimates. The pendulum length at 0.2 and 0.6 gave the smallest estimate of the standard error. The results obtained in this study show that this methodology indeed drastically reduces the computation time and still provides reasonably accurate results.

References

1. S. Strogatz.: The real scientific hero of 1953. New York Times, March 4, Editorial/Op-Ed. (2003)
2. Bayarri, M., Berger, J.O., Higdon, D., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Tu, J.: A framework for the validation of computer models. In: Pace, D., Stevenson, S. (eds.) Proceedings of the Workshop on Foundations for V&V in the 21st Century. Society for Modelling and Simulation International, United States (2002)
3. Parks, J.E.: The Simple Pendulum. Department of Physics and Astronomy The University of Tennessee Knoxville, Knoxville (2000)
4. Osuolale, K.A., Yahya, W.B., Adeleke, B.L.: Performing a simple pendulum experiment as a demonstrative example for computer experiments. Ann. Comput. Sci. Ser. **12**(2), 33–38 (2014)
5. Montgomery, D.C.: Design and Analysis of Experiments, 5th edn. Willy, New York (2001)
6. Motulsky, H.J., Christopoulos, A.: Fitting Models to Biological Data Using Linear and Nonlinear Regression. A Practical Guide to Curve Fitting. GraphPad, Software Inc., San Diego (2003). www.graphpad.com
7. Sacks, J., Schiller, S.B., Welch, W.J.: Designs for computer experiments. Technometrics. **31**, 41–47 (1989)