Tree-based Models for Poverty Estimation

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at

Manawatu

Penelope A. Bilton

07/11/2016

Abstract

The World Food Programme utilises the technique of poverty mapping for efficient allocation of aid resources, with the objective of achieving the first two United Nations Sustainable Development Goals, elimination of poverty and hunger. A statistical model is used to estimate levels of deprivation across small geographical domains, which are then displayed on a poverty map. Current methodology employs linear mixed modelling of household income, the predictions from which are then converted to various area-level measures of well-being. An alternative technique using tree-based methods is developed in this study. Since poverty mapping is a small area estimation technique, the proposed methodology needs to include auxiliary information to improve estimate precision at low levels, and to take account of complex survey design of the data. Classification and regression tree models have, to date, mostly been applied to data assumed to be collected through simple random sampling, with a focus on providing predictions, rather than estimating uncertainty. The standard type of prediction obtained from tree-based models, a "hard" tree estimate, is the class of interest for classification models, or the average response for regression models. A "soft" estimate equates to the posterior probability of being poor in a classification tree model, and in the regression tree model it is represented by the expectation of a function related to the poverty measure of interest. Poverty mapping requires standard errors of prediction as well as point estimates of poverty, but the complex structure of survey data means that estimation of variability must be carried out by resampling. Inherent instability in tree-based models proved a challenge to developing a suitable variance estimation technique, but bootstrap resampling in conjunction with soft tree estimation proved a viable methodology. Simulations showed that the bootstrap based soft tree technique was a valid method for data with simple random sampling structure. This was also the case for clustered data, where the method was extended to utilise the cluster bootstrap and to incorporate cluster effects into predictions. The methodology was further adapted to account for stratification in the data, and applied to generate predictions for a district in Nepal. Tree-based estimates of standard error of prediction for the small areas investigated were compared with published results using the current methodology for poverty estimation. The technique of bootstrap sampling with soft tree estimation has application beyond poverty mapping, and for other types of complex survey data.

Acknowledgements

To Geoff Jones and Siva Ganesh I wish to express my gratitude for their guidance, advice and encouragement along this PhD journey, with splashes of humour to ease the task. I would also like to thank Stephen Haslett for his contribution to the thesis.

My thanks also to Massey University, including the Institute of Fundamental Sciences, for the provision of scholarships to fund the research.

I would like to extend my appreciation to Timothy Bilton, Jonathan Godfrey and Hannes Calitz for their support with software issues. To Kathryn Stowell, my deepest thanks for your moral and practical support in my time of crisis.

To my Creator, God and Father of my Lord Jesus Christ, without His grace, strength and wisdom, this work would not have been completed.

Contents

List of Figures

List of Tables