

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Poverty mapping . . . . .	1
1.2	Development of poverty measures . . . . .	3
1.3	Other measures of deprivation . . . . .	4
1.4	Advantages of poverty mapping . . . . .	5
1.5	Implementation of Poverty Mapping in Nepal . . . . .	5
1.6	Scope of the thesis . . . . .	6
<b>2</b>	<b>Literature Review</b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	Components of poverty mapping . . . . .	9
2.2.1	Incorporating auxiliary information . . . . .	10
2.2.1.1	Borrowing strength . . . . .	10
2.2.1.2	Direct estimators . . . . .	10
2.2.1.3	Traditional indirect estimators . . . . .	11
2.2.1.4	Synthetic estimator . . . . .	11
2.2.1.5	Composite estimator . . . . .	11
2.2.1.6	Model based estimators . . . . .	11
2.2.1.7	Regression-synthetic estimator . . . . .	12
2.2.1.8	Best linear unbiased prediction estimator, BLUP . . . . .	12
2.2.1.9	Empirical best linear unbiased predictor estimator, EBLUP . . . . .	13
2.2.1.10	Empirical Bayes estimator . . . . .	13
2.2.1.11	Hierarchical Bayes estimator . . . . .	13
2.2.1.12	Generalised linear mixed models . . . . .	13
2.2.2	Complex survey design . . . . .	15
2.2.2.1	Simple random sampling . . . . .	16

---

2.2.2.2	Systematic sampling . . . . .	17
2.2.2.3	Survey design weights . . . . .	17
2.2.2.4	Stratified sampling . . . . .	18
2.2.2.5	Cluster sampling . . . . .	19
2.2.2.6	Complex survey design for Nepal . . . . .	20
2.2.3	Variance estimation for complex survey design . . . . .	23
2.2.3.1	Replication method . . . . .	23
2.2.3.2	Balanced repeated replication . . . . .	24
2.2.3.3	Jackknife resampling . . . . .	24
2.2.3.4	Bootstrap resampling . . . . .	26
2.2.3.5	Taylor Series Method . . . . .	27
2.2.3.6	Jackknife and bootstrap for complex data . . . . .	27
2.2.3.7	Inverse sampling . . . . .	29
2.3	ELL methodology for poverty mapping . . . . .	30
2.3.1	Auxiliary information . . . . .	32
2.3.2	Complex survey design . . . . .	32
2.3.3	Variance estimation . . . . .	33
2.3.4	Summary of the ELL methodology . . . . .	34
2.4	Tree based methods . . . . .	35
2.4.1	Introduction . . . . .	35
2.4.2	Building the tree . . . . .	36
2.4.2.1	Distributional structure of tree nodes . . . . .	37
2.4.2.2	Determining the best split in a classification tree . . . . .	37
2.4.2.3	Determining the best split in a regression tree . . . . .	39
2.4.2.4	Pruning the tree . . . . .	41
2.4.3	Assessing model fit . . . . .	41
<b>3</b>	<b>Classification tree models for poverty estimation</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Building the classification tree model . . . . .	44
3.2.1	Obtaining a suitable dataset for modelling . . . . .	44
3.2.2	Construction of an unweighted classification tree . . . . .	44
3.2.3	Incorporating survey weights . . . . .	47
3.2.4	Optimising the tree . . . . .	48

3.2.5	Interpretation of the classification tree model . . . . .	54
3.2.6	Variable importance and surrogates . . . . .	57
3.2.7	Assessing model fit . . . . .	63
3.3	Generating small area estimates of poverty incidence . . . . .	65
3.3.1	Hard and soft predictions . . . . .	65
3.3.2	Small area estimates of poverty incidence for a district in Nepal . . .	66
3.4	Conclusions . . . . .	68
<b>4</b>	<b>Tree instability under resampling</b>	<b>69</b>
4.1	Introduction . . . . .	69
4.2	Variance estimation for poverty incidence in Nepal . . . . .	70
4.2.1	Replicate subsamples . . . . .	70
4.3	Variance under inverse sampling . . . . .	71
4.4	Replicate weights . . . . .	72
4.5	Using the complexity parameter for tree pruning . . . . .	74
4.6	Estimating between replicate variance . . . . .	76
4.7	Jackknife variance estimation of within replicate variability . . . . .	78
4.8	Effect of minimum split and tree depth on tree stability . . . . .	80
4.9	Competing splits . . . . .	82
4.10	Conclusions . . . . .	88
<b>5</b>	<b>A study in stability</b>	<b>90</b>
5.1	Introduction . . . . .	90
5.2	Monte Carlo simulations . . . . .	91
5.3	Source of instability . . . . .	92
5.4	Outline of simulation study . . . . .	92
5.5	Simulating the datasets . . . . .	93
5.6	Simulation process . . . . .	94
5.7	Results of simulations using jackknife and bootstrap resampling . . . . .	95
5.8	Validity of estimated standard errors . . . . .	97
5.9	Experimental design . . . . .	101
5.9.1	Outline of the designed experiment . . . . .	101
5.9.2	ANOVA results for bias . . . . .	103
5.9.3	ANOVA results for relative standard error . . . . .	104

5.9.4	ANOVA results for coverage . . . . .	107
5.10	Conclusion . . . . .	108
<b>6</b>	<b>Adapting classification trees for complex survey data</b>	<b>109</b>
6.1	Introduction . . . . .	109
6.2	Monte Carlo simulation for clustered data . . . . .	110
6.3	Introducing clustering into the model . . . . .	110
6.3.1	Bootstrapping the clusters . . . . .	112
6.3.2	Performance of the bootstrap soft method for clustered data . . . . .	113
6.4	Introducing cluster effects into predictions . . . . .	118
6.5	A non-parametric method for incorporating cluster effects into predictions . . . . .	119
6.5.1	Results of modelling with non-parametric cluster effects in predictions . . . . .	122
6.6	A parametric method for incorporating cluster effects into predictions . . . . .	127
6.6.1	Results of modelling with parametric clusters effects in predictions . . . . .	128
6.7	Classification tree modelling for small area estimation in Nepal . . . . .	131
6.7.1	Setting up the analysis . . . . .	132
6.7.2	Results for classification tree small area estimation in Nepal . . . . .	134
6.8	Conclusion . . . . .	136
<b>7</b>	<b>Regression tree modelling of poverty measures</b>	<b>139</b>
7.1	Introduction . . . . .	139
7.1.1	FGT formula . . . . .	139
7.2	Developing hard and soft regression tree estimates . . . . .	140
7.2.1	Node distribution for a regression tree . . . . .	141
7.2.2	Poverty incidence . . . . .	141
7.2.3	Poverty gap . . . . .	142
7.2.4	Poverty severity . . . . .	144
7.3	Monte Carlo simulation with regression tree modelling . . . . .	146
7.3.1	Results for poverty incidence . . . . .	147
7.3.2	Results for poverty gap and poverty severity . . . . .	148
7.4	Cluster bootstrap soft estimation of poverty measures for Nepal . . . . .	153
7.4.1	Regression tree estimates of poverty incidence . . . . .	153
7.4.2	Regression tree estimates of poverty gap . . . . .	154
7.4.3	Regression tree estimates of poverty severity . . . . .	155



7.5	Conclusion . . . . .	158
<b>8</b>	<b>Discussion</b>	<b>160</b>
8.1	Review of the thesis . . . . .	160
8.2	Weighing tree-based models against ELL . . . . .	161
8.3	Further research . . . . .	164
	<b>References</b>	<b>166</b>
	<b>Appendices</b>	<b>178</b>
<b>A</b>	<b>Auxiliary variables</b>	<b>179</b>
A.1	Household predictors . . . . .	179
A.2	Ward level census means . . . . .	181
A.3	VDC level census means . . . . .	182
A.4	GIS variables . . . . .	182
<b>B</b>	<b>Rpart summary output</b>	<b>183</b>
B.1	Summary for weighted classification tree model on Replicate 1 . . . . .	183
B.2	Summary of weighted classification tree model on jackknife sample # 25 . .	184
<b>C</b>	<b>R code</b>	<b>186</b>
C.1	Code for improve function . . . . .	186
C.2	Code for simulations using a classification tree . . . . .	186
C.3	Code for regression tree estimates for a district in Nepal . . . . .	192
<b>D</b>	<b>Mathematical derivations of soft estimators for poverty gap and poverty severity</b>	<b>200</b>
D.1	Derivation of a soft estimator for poverty gap . . . . .	200
D.2	Derivation of a soft estimator for poverty severity . . . . .	202

# List of Figures

1.1	Poverty map of wasting in children under 6 in Nepal . . . . .	2
1.2	Structure of tree-based modelling to generate poverty estimates . . . . .	7
2.1	Geographical and administrative divisions in Nepal . . . . .	21
3.1	Unweighted classification tree model for poverty incidence in Nepal . . . . .	46
3.2	Cp plot for the weighted classification tree . . . . .	49
3.3	Cp plots for the weighted classification tree . . . . .	50
3.4	Output of cp plot for weighted classification tree model of poverty in Nepal	51
3.5	Weighted classification tree model for poverty incidence in Nepal . . . . .	53
3.6	Weighted classification tree model for poverty incidence in Nepal, omitting <i>tw</i> . . . . .	56
3.7	Competing splits for root node of weighted classification tree for poverty incidence . . . . .	57
3.8	Splitting criterion and surrogate variables for root node in classification tree	58
3.9	Plot of variable importance for classification tree with $cp = 0.005$ . . . . .	60
3.10	Classification tree model for poverty incidence with $cp = 0$ and tree depth 4	61
3.11	Plot of variable importance for classification tree with $cp = 0$ and depth 4 .	62
3.12	Layout of a confusion matrix . . . . .	64
3.13	Aggregated measures of classification accuracy from models based upon replicates . . . . .	64
3.14	ELL predictions compared with hard and soft tree predictions for two $cp$ values . . . . .	67
4.1	Construction of replicate subsamples . . . . .	71
4.2	Table of $cp$ values and associated cross-validation error for different tree sizes	74
4.3	Plot of $cp$ values against cross-validation error for model with cluster weights	75

4.4	Tree diagram for weighted classification model using only data from Replicate 1 . . . . .	77
4.5	Table of estimates of poverty incidence in Ilaka1 using 163 jackknife subsamples of Replicate 1 . . . . .	79
4.6	Contour plot of jackknife standard deviation values for varying minimum split and tree depth . . . . .	80
4.7	Contour plot of jackknife mode estimate values for varying minimum split and tree depth . . . . .	81
4.8	Tree diagram for model using all data from Replicate 1 . . . . .	84
4.9	Tree diagram for model using data from jackknife subsample #25 of Replicate 1 . . . . .	84
4.10	Summary of Node 1 for model on full replicate sample, cp=0, split=3, depth=4 . . . . .	85
4.11	Summary of Node 1 for model on JK #25 subsample, cp=0, split=3, depth=4	85
5.1	Flowchart describing the simulation process . . . . .	96
5.2	Actual coverage of a 100 intervals for a nominal level of 95% for survey size 300 . . . . .	98
5.3	Actual coverage of a 100 intervals for a nominal level of 95% for survey size 3000 . . . . .	99
5.4	Flowchart of algorithms used in the designed experiment . . . . .	101
5.5	ANOVA table for analysis of bias of variance estimation methods . . . . .	103
5.6	Table of coefficients for analysis of bias of variance estimation methods . . .	104
5.7	ANOVA table for analysis of relative s.e. for variance estimation methods .	105
5.8	Table of coefficients for analysis of relative s.e. of variance estimation methods	105
5.9	Relative standard error for BS method for different sample sizes and minimum split values . . . . .	106
5.10	Table of coefficients for analysis of coverage of variance estimation methods	107
6.1	Coverage of cluster and naive bootstrap using fixed small area . . . . .	114
6.2	Coverage of cluster and naive bootstrap using simulated small area . . . . .	116
6.3	Coverage for bootstrap soft intervals including non-parametric cluster effects	123
6.4	Coverage for full tree intervals including non-parametric cluster effects . . .	126
6.5	Empirical coverage for parametric prediction cluster effects, for three types of intervals: . . . . .	129
6.6	Plot of classification tree versus ELL estimates . . . . .	136

- 7.1 Empirical coverage for regression tree estimates of poverty incidence . . . . 149
- 7.2 Empirical coverage for regression tree estimates of poverty gap . . . . . 151
- 7.3 Empirical coverage for regression tree estimates of poverty severity . . . . . 152
- 7.4 Plot of classification and regression tree versus ELL point estimates . . . . 157

# List of Tables

3.1	Scores for the seventeen most important predictors in the weighted classification tree: <i>hh</i> means household . . . . .	59
4.1	Estimates of poverty incidence for Ilaka 1 using replicate subsamples . . . . .	78
4.2	Predictor values for households omitted from jackknife sample #25 . . . . .	86
4.3	Predictor values for households omitted from jackknife sample #25 . . . . .	86
4.4	Class counts and <i>improve</i> functions using <i>skids6w</i> as first split . . . . .	86
4.5	Class counts and <i>improve</i> functions using <i>edulv4w</i> as first split . . . . .	87
5.1	Average prediction bias and s.e. from 100 simulations, for two different survey sizes . . . . .	96
5.2	True standard error for different survey sizes . . . . .	102
5.3	Percentage variability explained by by Method, Type, Survey size and their interactions . . . . .	104
6.1	Cluster effect values and corresponding intracluster correlations . . . . .	111
6.2	P-values for McNemar’s test of coverage for ordinary bootstrap and cluster bootstrap, 95% nominal level . . . . .	117
6.3	Average standard error of predictions for cluster and ordinary bootstrap with small area dataset simulated for each Monte Carlo iteration . . . . .	118
6.4	Comparing the composition of strata and groups in the Nepal modelling . . . . .	132
6.5	Size and total sampling weights for each stratum in the Nepal analysis . . . . .	134
6.6	Comparison of ELL and bootstrap soft tree estimates for an ilaka in one district of Nepal . . . . .	135
7.1	Average bias and s.e. of hard and soft regression tree estimates for poverty incidence . . . . .	147
7.2	Average bias and s.e. of hard and soft regression tree estimates for poverty gap . . . . .	150

7.3	Average bias and s.e. of hard and soft regression tree estimates for poverty severity . . . . .	150
7.4	ELL and cluster bootstrap soft tree estimates of poverty incidence for a district in Nepal . . . . .	154
7.5	ELL and cluster bootstrap soft regression tree estimates of poverty gap for a district in Nepal . . . . .	155
7.6	ELL and cluster bootstrap soft regression tree estimates of poverty severity for a district in Nepal . . . . .	156