Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Classification trees for poverty mapping☆

Penny Bilton [a], Geoff Jones [a,*], Siva Ganesh [b], Steve Haslett [a,c]

[a] Institute of Fundamental Sciences (Statistics), Massey University, Palmerston North, New Zealand
[b] AgResearch, Bioinformatics Maths & Stats, Palmerston North, New Zealand
[c] Statistical Consulting Unit, Australian National University, Canberra, Australia

HIGHLIGHTS

- Adapts the use of classifications trees for small area estimation.
- Produces standard errors for classification tree models fitted to survey data.
- Important applications involving the allocation of millions of dollars of aid to Third World countries.

ARTICLE INFO

Article history:
Received 13 October 2016
Received in revised form 22 May 2017
Accepted 23 May 2017
Available online xxxx

Keywords:
Small area estimation
Sustainable Development Goals
Complex survey data
Clustered data

ABSTRACT

Poverty mapping uses small area estimation techniques to estimate levels of deprivation (poverty, undernutrition) across small geographic domains within a country. These estimates are then displayed on a poverty map, and used by aid organizations such as the United Nations World Food Programme for the efficient allocation of aid. Current methodology employs unit-level regression modelling of a target variable (household income, child weight-for-age). An alternative modelling technique is proposed, using tree-based methods, that has some practical advantages. Alternative ways of amalgamating the unit-level predictions from classification trees to small area level are explored, adapting the trees to account for the survey design, and resampling strategies are proposed for producing standard errors. The methodology is evaluated using both real data and simulations based on a poverty mapping study in Nepal. The simulations suggest that amalgamation of posterior probabilities from the tree gives approximately unbiased estimates, and standard errors can be calculated using a cluster bootstrap approach with cluster effects included in the predictions. Small area estimates of poverty incidence for a region in Nepal, generated using the proposed tree based method, are comparable to the published results obtained by the standard method.

1. Introduction

Elimination of poverty and undernutrition, the first two of the United Nations Sustainable Development Goals (United Nations, 2016b), is addressed through the distribution of billions of dollars in assistance each year to third world countries. Poverty mapping is promoted by the World Bank (World Bank, 2015) for predicting regional variations in the levels of deprivation in a particular country, to facilitate efficient allocation of food aid by agencies such as the United Nations World

☆ Specimen R code for classification tree estimates is provided as supplementary material in the electronic version of the paper (see Appendix B).

* Correspondence to: Institute of Fundamental Sciences (Statistics), Massey University, Private Bag 11222, Palmerston North 4100, New Zealand. Fax: +64 6 3557953.
E-mail address: g.jones@massey.ac.nz (G. Jones).

http://dx.doi.org/10.1016/j.csda.2017.05.009
0167-9473/

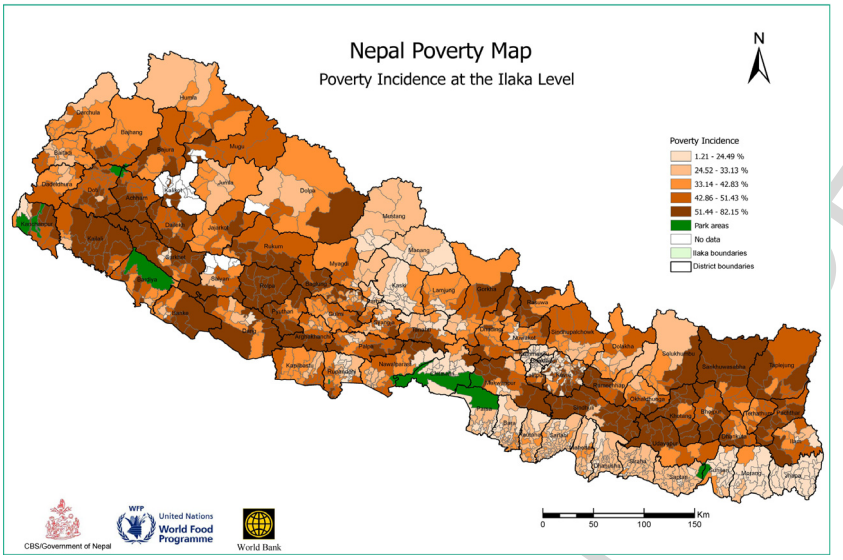

Fig. 1. Poverty map of poverty incidence in Nepal (United Nations, 2016a).

Food Programme (WFP). Statistical techniques are used to generate within-country estimates of deprivation, which can then be combined with Geographic Information System (GIS) data to produce poverty maps, displaying disaggregated measures of poverty and other indicators of well being at low geographical levels. Fig. 1, an example of a poverty map for Nepal, displays small area estimates of the proportion of individuals below a specific expenditure level (the "poverty line") across geographic domains called "ilakas". A poverty line is usually based on the income or expenditure required to enjoy a minimal level of goods and services (Ravallion, 1992). The head-count index is more correctly called *poverty prevalence*; however, the proportion of poor people is referred to in this paper as *poverty incidence*, the term generally used in the poverty mapping literature.

Estimation of poverty status at low geographical level requires the use of small area estimation methodology. Current methods for poverty mapping are usually based on multiple regression models, relating the variable of interest to a set of predictors. An alternative approach for modelling poverty incidence is to use classification trees (Breiman et al., 1984), which offer several advantages. Firstly, a classification tree does not require parametric assumptions (Hastie et al., 2001) and provides a simple, direct and easily understood method of estimating poverty incidence. Multiple regression typically uses a stepwise method, at a preliminary stage, for selection of model predictors, but there can be major problems with this approach (Harrell, 2015); some skill and experience is required on the part of the modeller to avoid over-fitting while including the important predictors in an appropriate way. The classification tree method in contrast has built-in tools for automatically selecting variables and avoiding over-fitting, and is better able to cater for possible non-linear relationships in the data structure (Chambers and Hastie, 1992). Including interactions in multiple regression can be problematic, since decisions must first be made about which interactions to explore, and then the model attempts to estimate effects for all possible combinations, some of which may be unimportant. In contrast, the classification tree readily incorporates variable interactions, selecting only the important combinations.

The next section reviews some basic concepts in poverty mapping and classification trees, after which Section 3 describes the proposed methodology for adapting the classification tree model for small area estimation of poverty incidence. Results from applying the methodology to simulated data having a simple random sampling structure, simulated data containing clusters, and actual Nepal data (Haslett and Jones, 2006) are provided in Section 4. These results are discussed, and recommendations made, in the final section.

2. Review of current methods

Using classification trees to model poverty incidence requires melding small area estimation methodology and complex survey design with the technique of classification trees. The basic features of poverty mapping, resampling methods in complex surveys and classification tree models are reviewed below.

2.1. Poverty mapping

Poverty mapping combines survey data with other information to estimate poverty measures across small domains. The term "small area" describes a subpopulation for which direct estimates from national survey data cannot be provided

Please cite this article in press as: Bilton, P., et al., Classification trees for poverty mapping. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.05.009.