Spatial Statistics 18 (2016) 455–473

Contents lists available at ScienceDirect

Spatial Statistics

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

Model-based inference for small area estimation with sampling weights

Y. Vandendijck [a,*], C. Faes [a], R.S. Kirby [b], A. Lawson [c], N. Hens [a,d]

[a] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium

[b] Department of Community and Family Health, College of Public Health, University of South Florida, Tampa, FL, United States

[c] Department of Public Health, University of South Carolina, Charleston, SC, United States

[d] Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

ARTICLE INFO

Article history:
Received 7 May 2015
Accepted 13 September 2016
Available online 14 October 2016

Keywords:
Integrated nested Laplace approximations
Model-based inference
Small area estimation
Spatial smoothing
Survey weighting

ABSTRACT

Obtaining reliable estimates about health outcomes for areas or domains where only few to no samples are available is the goal of small area estimation (SAE). Often, we rely on health surveys to obtain information about health outcomes. Such surveys are often characterised by a complex design, stratification, and unequal sampling weights as common features. Hierarchical Bayesian models are well recognised in SAE as a spatial smoothing method, but often ignore the sampling weights that reflect the complex sampling design. In this paper, we focus on data obtained from a health survey where the sampling weights of the sampled individuals are the only information available about the design. We develop a predictive model-based approach to estimate the prevalence of a binary outcome for both the sampled and non-sampled individuals, using hierarchical Bayesian models that take into account the sampling weights. A simulation study is carried out to compare the performance of our proposed method with other established methods. The results indicate that our proposed method achieves great reductions in mean squared error when compared with standard approaches. It performs equally well or better when compared with more elaborate methods when there

* Corresponding author.
E-mail address: yannick.vandendijck@uhasselt.be (Y. Vandendijck).

http://dx.doi.org/10.1016/j.spasta.2016.09.004

is a relationship between the responses and the sampling weights. The proposed method is applied to estimate asthma prevalence across districts.

1. Introduction

In public health we are often interested in the question whether there are disparities in illness, behavioural risk factors or health conditions across areas. An increasing amount of information on individuals is collected in this respect. Bayesian methods in disease mapping based on census or population registry data are well developed and used in a fairly standard manner (see e.g., Elliott et al., 2001; Waller and Gotway, 2004; Lawson, 2013 for a review of the methods). Such population registry or census data obtains information pertaining to each member of the population of an area. Historically, focus was on the construction of cancer atlases and on mapping rare diseases based on registry data (see e.g., Kemp et al., 1985; Mason, 1995).

Since it is nearly always impossible to measure the health outcome of interest in every individual in the population, a survey is used to record information from a random sample of individuals from the population (Cochran, 1977). Such surveys are often characterized by a complex design, with stratification, clustering and unequal sampling weights as common features. Policy makers are often interested in a specific characteristic, such as the total number of diseased cases or the prevalence, per area. In small area estimation (SAE) one investigates how to obtain these area specific characteristics from survey data covering more than only the area of interest by using spatial smoothing methods.

In SAE, one needs to choose whether to base inference on *design-based*, *model-based* or *design-based model-assisted* approaches. In design-based inference the values of the health outcomes are assumed fixed, and inference is based on the randomization distribution of the sample inclusion indicators. Often a model is used in the construction of a design-based estimator (known as design-based model-assisted approaches). A popular design-based estimator is the Horvitz–Thompson (HT) estimator (1952) and its extensions that weigh sampled individuals with the associated sampling weight. These estimators play a dominant role in sample surveys, however, they often fail in SAE because the sample size per area could be very small or even zero inflating the mean squared error tremendously. This makes design-based estimators unreliable or not feasible to use (Rao, 2011). Additionally, because of the spatial nature of the problem, understanding the geographical distribution of the health outcome is important. Model-based approaches that perform spatial smoothing, both those based on empirical and hierarchical Bayesian methodology, have shown to be more relevant in the handling of spatially correlated health survey data. In model-based approaches one conditions on the selected sample and the inference is based on the underlying model of the health outcome. Examples include Fay and Herriot (1979) which proposed a linear empirical Bayes model to estimate the income for small areas, while Datta and Ghosh (1991) considered a hierarchical Bayesian formulation instead. A number of extensions have been made, see Rao (2003) and Jiang and Lahiri (2006) for an overview. For binary data, MacGibbon and Tomberlin (1989) developed an empirical Bayes model using a logistic regression model with fixed and random effects. Stroud (1994), Ghosh et al. (1998) and Farrell (2000) described hierarchical Bayesian approaches to estimate small area proportions.

While model-based SAE is conceptually appealing, complex survey designs with the accompanying survey weights cause a difficulty in their practical implementation. Only relatively few approaches acknowledge the survey sampling mechanism and account for it in the model. Kott (1989) and Prasad and Rao (1999) described a design-consistent model-based estimator. Kott (1989) proposed an estimator which is a weighted combination of the HT estimator and the sample means of the different areas. Prasad and Rao (1999) proposed a pseudo-empirical best linear unbiased prediction estimator for the small area mean based on area level data. You and Rao (2002, 2003) used unit level data instead. Malec et al. (1997) described a hierarchical Bayesian model for binary survey data. They examined