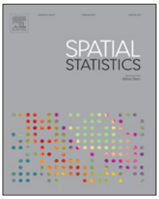
Contents lists available at ScienceDirect

Spatial Statistics

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



A comparison of spatial smoothing methods for small area estimation with sampling weights

Laina Mercer[a], Jon Wakefield[a,*], Cici Chen[b], Thomas Lumley[c]

[a] Department of Statistics, University of Washington, United States
[b] Department of Biostatistics, Brown University, United States
[c] Department of Statistics, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 23 February 2013
Accepted 9 December 2013
Available online 18 December 2013

Keywords:
Complex surveys
Design-based inference
Intrinsic CAR models
Random effects models
Weighting

ABSTRACT

Small area estimation (SAE) is an important endeavor in many fields and is used for resource allocation by both public health and government organizations. Often, complex surveys are carried out within areas, in which case it is common for the data to consist only of the response of interest and an associated sampling weight, reflecting the design. While it is appealing to use spatial smoothing models, and many approaches have been suggested for this endeavor, it is rare for spatial models to incorporate the weighting scheme, leaving the analysis potentially subject to bias. To examine the properties of various approaches to estimation we carry out a simulation study, looking at bias due to both non-response and non-random sampling. We also carry out SAE of smoking prevalence in Washington State, at the zip code level, using data from the 2006 Behavioral Risk Factor Surveillance System. The computation times for the methods we compare are short, and all approaches are implemented in R using currently available packages.

1. Introduction

Small area estimation (SAE) is used in many fields including education and epidemiology, and global, environmental and public health. Often the surveys carried out to inform SAE are complex in nature, with non-random sampling being carried out for reasons of necessity (i.e., logistical reasons) or to ensure that certain populations of interest are well represented. In addition, post-stratification

[*] Corresponding author. Tel.: +1 12066169388.
E-mail address: jonno@uw.edu (J. Wakefield).

2211-6753/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spasta.2013.12.001

Table 1
Summary statistics for population data, and the 2006 Washington State BRFSS data on adult current smokers, across zip codes.

	Mean	S.D.	Median	Min	Max
Population	12 570.0	12 931.0	7208.0	11.0	55 700.0
Sample sizes	46.9	54.8	30.0	1.0	384.0
Number of current adult smokers	7.5	9.5	4.0	0.0	67.0

may be used to reweight the observations in order to recover known population totals. This approach can account for non-response within the strata used in the post-stratification.

There are two approaches to modeling complex survey data that we shall consider in this paper. In the first *design-based* approach weighted estimators are considered, with inference carried out based on the (randomization) distribution of the samples that could have been collected, i.e., the distribution of the individuals that could appear in the sample. In contrast, a *model-based* approach assumes a hypothetical infinite population from which the responses are drawn. While appealing from a conceptual point of view (since standard statistical modeling machinery can be leaned upon), the modeling approach is difficult to implement since one must model the sampling mechanism, if informative, at least to some extent. For example, if non-random sampling is based on particular inclusion variables (e.g., race or geographical area) then these variables must be included in the model if they are associated with the outcome of interest. Similarly, variables that affect the probabilities of non-response must also be included in the model, again if they are related to the outcome. The alternative is to assume that variables upon which sampling is based and non-response depends are unrelated to the outcome of interest, which is a dangerous endeavor. Another impediment to the model-based approach is that the key variables that are required for inclusion may be unavailable in public-use databases. Even if available, the sampling scheme may be highly complex, requiring a model which has a large number of parameters and being therefore difficult to fit. Gelman (2007b) describes the issues, and the accompanying discussion (Bell and Cohen, 2007; Breidt and Opsomer, 2007; Little, 2007; Lohr, 2007; Pfefferman, 2007; Gelman, 2007a) gives a range of perspectives on the use of weighted estimators, regression modeling, or a combination of the two.

In this paper we will consider SAE in the situation in which either the variables upon which sampling was based are unavailable or the scheme is so complex that a simpler approach is desired. SAE has seen a great deal of research interest, with Rao (2003) being a classic text. In the related field of disease mapping, the use of spatial modeling is commonplace (Wakefield et al., 2000), but in this context the data usually consist of a complete enumeration of disease cases in an area, so that no weighting scheme needs to be considered. It is the existence of the weights that causes a major difficulty when one wishes to use spatial smoothing in SAE, and consequently there are relatively few instances of approaches that use spatial smoothing within a model that acknowledges the sampling scheme. In Chen et al. (submitted for publication) a new method of incorporating the weights within a spatial hierarchical model was introduced, and various random effects models were compared via simulation. In this paper we compare the method with a number of other suggested methods for weighting.

As a motivating example, we examine data from the Behavioral Risk Factor Surveillance System (BRFSS). This survey is carried out at the state level in the United States and is the largest telephone-based survey in the world. In the BRFSS survey, interviewees (who are 18 years or older) are asked a series of questions on their health behaviors and provide general demographic information, such as age, race, gender and the zip code in which they live. In this paper we focus on the survey conducted in Washington State in 2006, and on the Centers for Disease Control (CDC) calculated variable *Adults who are current smokers*. With respect to this question, 19,502 respond with "No", 3733 with "Yes" and 132 were classified as "don't know/refuse/missing". In the analysis, we remove these latter values. The response variable is therefore a binary indicator and our objective is to estimate the number of individuals who are 18 or older and who are current smokers, in each of 498 zip codes in Washington State. We also utilize population estimates from 2006. Table 1 summarizes the population and survey data. So far as the survey is concerned, the number of samples per zip code shows large variability with a median of 30 and minimum and maximum values of 1 and 384. The spread is apparent in Fig. 1. Fig. 2