ARTICLE IN PRESS

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

Small area estimation under a spatially non-linear model

Hukum Chandra [a,*], Nicola Salvati [b], Ray Chambers [c]

[a] Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi 110012, India
[b] Dipartimento di Economia e Management, University of Pisa, Italy
[c] Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia

ARTICLE INFO

Article history:
Received 11 August 2017
Received in revised form 8 April 2018
Accepted 9 April 2018
Available online xxxx

Keywords:
Small area estimation
Nonparametric models
Spatial relationship
Count data
Poverty indicator

ABSTRACT

We describe a methodology for small area estimation of counts that assumes an area-level version of a nonparametric generalized linear mixed model with a mean structure defined using spatial splines. The proposed method represents an alternative to other small area estimation methods based on area level spatial models that are designed for both spatially stationary and spatially non-stationary populations. We develop an estimator for the mean squared error of the proposed small area predictor as well as an approach for testing for the presence of spatial structure in the data and evaluate both the proposed small area predictor and its mean squared error estimator via simulations studies. Our empirical results show that when data are spatially non-stationary the proposed small area predictor outperforms other area level estimators in common use and that the proposed mean squared error estimator tracks the actual mean squared error reasonably well, with confidence intervals based on it achieving close to nominal coverage. An application to poverty estimation using household consumer expenditure survey data from 2011–12 collected by the national sample survey office of India is presented.

1. Introduction

The demand for small area statistics has increased rapidly over the past few years (e.g. measurement of social exclusion and social wellbeing at disaggregate level, see Tzavidis et al., 2008). As a consequence, many small area estimation (SAE) methods based on linear mixed models have been proposed in the literature. In many cases, however, the response variable is not continuously distributed but is binary valued or a count. Such response variables cannot be modelled using standard linear models. When the variable of interest is binary or a count and small area estimates are required for these data, use of standard estimation methods based on linear mixed models becomes problematic. For example, poverty indicators and many other indicators related to socio-economic status and food insecurity usually behave in a non-Gaussian manner at small area levels, and so estimation in these cases is typically based on a generalized linear mixed model (GLMM); see Manteiga et al. (2007) and Ruppert et al. (2003, chapter 10). The most commonly used GLMMs are the logistic-normal mixed model (i.e. GLMMs with logistic link function, also referred as the logistic linear mixed model) and the general Poisson-normal mixed model (i.e. GLMMs with log link function, also referred as the log linear mixed model). Unit level predictions generated by a GLMM are generally used to define the empirical predictor for small areas for such data. In many applications this is not possible, for example poverty mapping where data confidentiality restricts access to unit level survey data with small area identifiers, or where the agency carrying out the small area analysis does not have the resources to analyse unit level data, as in many developing countries. In such situations, an area level version of the GLMM can be used for SAE. In particular,

* Corresponding author.
E-mail addresses: hchandra12@gmail.com (H. Chandra), nicola.salvati@unipi.it (N. Salvati), ray@uow.edu.au (R. Chambers).

https://doi.org/10.1016/j.csda.2018.04.002
0167-9473/

when only area level data are available, an area level version of the GLMM is fitted to obtain the plug-in empirical predictor for the small areas, see for example, Chandra et al. (2011) and Johnson et al. (2010). Other recent work on this topic include Boubeta et al. (2016, 2017) who use area-level Poisson mixed models for estimating small area counting indicators. Other authors have developed SAE under a GLMM using a Bayesian approach. See Torabi and Shokoohi (2015), Rao and Molina (2015), Mourra et al. (2006), Datta et al. (1999), and references therein. Mercer et al. (2014), Liu et al. (2014) and Franco and Bell (2013) consider the use of survey weights in a Bayesian hierarchical model framework when estimating small area proportions. In this context, we note that the hierarchical Bayes (HB) approach to SAE offers considerable promise because of it can accommodate complex small area models and provide "exact" inferences. Unfortunately, however, the choice of noninformative priors that can provide frequentist validity for the Bayesian approach may not be easy in practice, especially when complex sampling designs are involved. Also, caution needs to be exercised in the routine use of popular HB model-checking methods, see Rao (2011). In this paper we focus on the situation where only aggregated level data are available and SAE is carried out under area level small area models. Our development is based on a frequentist approach to SAE. This approach is often easier to explain to practitioners, can be less time consuming and inferential expressions can typically be written out explicitly. For this reason, national statistical offices as well as other Government agencies involved in production of statistics often prefer a frequentist approach to estimation and prediction over a Bayesian approach.

In economic, environmental and epidemiological applications, estimates for areas that are spatially close may be more alike than estimates for areas that are further apart. It is therefore reasonable to assume that the effects of neighbouring areas, defined via a contiguity criterion, are correlated. Chandra and Salvati (2018) and Saei and Chambers (2003) describe an extension of the area level version of GLMM that allows for spatially correlated random effects using a SAR model (SGLMM) and define a plug-in empirical predictor (SEP) for the small area proportion under this model. This model allows for spatial correlation in the error structure, while keeping the fixed effects parameters spatially invariant. Chandra et al. (2017) introduce a spatially nonstationary extension of the area level version of GLMM, using an adaptation of the geographical weighted regression (GWR) concept to extend the GLMM to incorporate spatial nonstationarity (NSGLMM), which they then apply to the SAE problem to define a plug-in empirical predictor (NSEP) for small areas. Non-stationary spatial effects can be also modelled using a spatially non-linear extension of the GLMM. In the GLMM, the relationship between the link function and the covariates is often assumed to be linear. However, when the functional form of the relationship between the link function and the covariates is unknown or has a complicated functional form, an approach based on the use of a non-linear regression model can offer significant advantages compared with one based on a linear model. Torabi and Shokoohi (2015) describe a data cloning approach to fitting a GLMM that uses this idea, but which is based on a unit level GLMM. When geographically referenced area-level responses play a central role in the analysis and need to be converted to maps, we can use bivariate smoothing to fit a spatially heterogeneous GLMM. In particular, we use *P*-splines that rely on a set of bivariate basis functions to handle the spatial structures in the data, while at the same time including small area random effects in the model. We denote this nonparametric P-spline-based extension of the usual GLMM by SNLGLMM. See Ugarte et al. (2009), Opsomer et al. (2008) and Ruppert et al. (2003). We then describe a non-linear version of the plug-in empirical predictor for small areas (SNLEP) under an area level version of SNLGLMM. We also develop mean squared error estimation for SNLEP using the approach discussed in Chandra et al. (2011), Johnson et al. (2010), Opsomer et al. (2008) and Saei and Chambers (2003).

Note that an alternative to computing the plug-in empirical predictor (EP) is to compute the empirical best predictor (EBP, Jiang, 2003). Unfortunately, computing the EBP is generally not straightforward since it does not have a closed form and usually has to be computed via numerical approximation. As a consequence, national statistical agencies tend to favour computation of an analytic approximation such as the EP. It is our understanding that an approximation closely related to the EP is also used in Lopez-Vizcaino et al. (2013) and Lopez-Vizcaino et al. (2015) and Molina et al. (2007).

The rest of this article is organized as follows. Section 2 introduces the area level version of GLMM to define the plug-in empirical predictor for small areas and reviews the SGLMM and NSGLMM and its corresponding plug-in empirical predictors (SEP and NSEP). In Section 3 we describe the spatially non-linear extension of an area level version of GLMM (i.e. the SNLGLMM) and subsequently use this model to carry out SAE. We focus on the GLMM with logistic link function (i.e. logistic-normal mixed model) for binary data and the GLMM with log link function (i.e. general Poisson-normal mixed model) for count data. The development reported in this paper can be easily generalized to other variants of GLMMs. Section 4 then discusses mean squared error estimation for the proposed small area predictor, and develops a corresponding analytic estimator. Empirical results are provided in Section 5 and application of the proposed small area method to poverty mapping is described in Section 6. Finally, Section 7 summarizes our main conclusions and identifies areas where further research is necessary.

2. SAE under a generalized linear mixed model

Consider a finite population U of size N, and assume that a sample s of size n is drawn from this population according to a given sampling design, with the subscripts s and r used to denote quantities related to the sampled and non-sampled parts of the population. We assume that population is made up of m small domains or small areas (or simply domains or areas) $U_i (i = 1, \ldots, m)$, where we use a subscript of i to index those quantities associated with area. In particular, n_i and N_i are used to represent the sample and population sizes in area i, respectively. We also assume that the underlying unit level variable of interest y is discrete, and in particular is either a binary value or is a non-negative integer, and the aim is to estimate the corresponding small area population proportions or population totals (i.e. counts). Let the total of y in area i be denoted y_i,

Please cite this article in press as: Chandra H., et al., Small area estimation under a spatially non-linear model. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.04.002.