Spatial Statistics 21 (2017) 440–459

Contents lists available at ScienceDirect

Spatial Statistics

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

A comparison of estimation methods for multilevel models of spatially structured data

Roger Bivand [a,*], Zhe Sha [b], Liv Osland [c], Ingrid Sandvig Thorsen [d,e]

[a] *Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway*
[b] *University of Oxford, 24-29 St Giles', OX13LB, Oxford, UK*
[c] *Stord/Haugesund University College, Bjørnsonsgt. 45, 5528 Haugesund, Norway*
[d] *Department of Mathematics, University of Bergen, Thormøhlensgate 41A, 5006 Bergen, Norway*
[e] *Analysetjenester AS, Thormøhlensgate 41A, 5006 Bergen, Norway*

ARTICLE INFO

Article history:
Received 24 September 2016
Accepted 9 January 2017
Available online 17 January 2017

Keywords:
Spatially structured random effects
Spatial multilevel models
Estimation of CAR and SAR models

ABSTRACT

Two recent contributions (Dong et al., 2015; Osland et al., 2016) point to the relevance of multilevel models for spatially structured data. In Osland et al. (2016) these models are used to examine the importance of district-level covariates for house prices in Stavanger, Norway, in Dong et al. (2015) similarly for land parcel prices in Beijing; we use these data sets in our comparison. In Osland et al. (2016), a district-level spatial random effect was fitted using an intrinsic CAR model estimated using WinBUGS. Dong et al. (2015) used R code provided in supplementary materials to their article, and subsequently improved in an R package (Dong et al., 2016a); computation there used custom MCMC C++ code to fit a SAR district-level spatial random effect. This article compares approaches to estimating models of this kind, using the R packages R2WinBUGS, HSAR, INLA, R2BayesX, hglm and the new package mclcar for Monte Carlo maximum likelihood estimation (Sha, 2016b). We show that multilevel models of spatially structured data may be estimated readily using a variety of approaches, not only the intrinsic CAR model more typically found in the existing literature. We also point to a range of issues for further research in

* Corresponding author.
E-mail addresses: roger.bivand@nhh.no (R. Bivand), zhesha1006@gmail.com (Z. Sha), liv.osland@hsh.no (L. Osland), ingrid.s.thorsen@analysetjenester.no (I.S. Thorsen).

http://dx.doi.org/10.1016/j.spasta.2017.01.002

situations in which data acquired at different levels of spatial resolution are combined.

1. Introduction

Spatial data are often organized in multilevel structures, and can be characterized by a mixture of two important features: spatial autocorrelation and spatial heterogeneity (Anselin, 1988). For example, when analysing housing markets, individual houses are located in neighbourhoods which are located within census tracts, which are situated in even higher or more aggregated levels such as municipalities, counties and regions. Typical feature of multilevel data are correlations between observations located within one of the spatial zones or defined neighbourhoods. Houses and, hence, housing prices within a neighbourhood may share similar features, inter alia because they are built at the same time, or because there may be a relatively large proportion of wealthy people living there, or because they are located close to environmental amenities or disamenities. The result could be within-zonal correlations at various levels. So when using geographical data, it could be unrealistic to assume that observations within delimited spatial zones are independent. Following, Corrado and Fingleton (2012), analysts should consider using multilevel models in these situations.

Similarly to spatial regression models, but in contrast with the ordinary least squares estimator, multilevel estimators are not based on the assumption that observations are independent. The multilevel models account for unexplained local correlations, by introducing for instance a random intercept, one for each local zone. Frequently, these random intercepts are not the main focus of the research, given that they capture the impact of excluded factors found in the zones. However, by studying the significance, sign and magnitude of the random effects, they may provide important or useful information in the modelling process. This is explained in for instance Rabe-Hesketh and Skrondal (2008) or Osland et al. (2016). We choose not to examine an alternative random slope approach to these models here (Rabe-Hesketh and Skrondal, 2008), but note that a very recent paper by Dong et al. (2016b) takes up the equivalent spatial random slope approach.

In addition to inducing correlations within zones, variants of multilevel models as described in e.g. Rabe-Hesketh and Skrondal (2008), capture unobserved spatial heterogeneity which is a common feature in spatial data. By way of example, coefficients related to housing market structures may vary between the most urbanized area and semi-urban areas. In this way, classical multilevel models are useful in that they via estimation of the variance of random effects also account for the interzonal variations in the data.

One important assumption in the conventional multilevel models is that there is independence between the zonal random effects (Rabe-Hesketh and Skrondal, 2008, p. 61). For many types of data and analyses, this may be a reasonable assumption. For instance, when studying what determines birthweight of children, one may find that birthweights of children of the same mother are correlated. Birthweight of children of different mothers which represents the group level can probably be assumed to be uncorrelated. Another frequently used example is related to academic results in schools: it could be reasonable to assume that there are correlations or more similar results for pupils within the same school. It may also be a reasonable assumption that correlations of educational results between different schools could be ignored. So in these situations the classical multilevel models could be a useful approach, and may provide improved efficiency of estimates at the lower observational level (Dong et al., 2015).

The conventional random effect estimator may not be efficient when analysing spatial data, however. For this type of data it could be highly relevant to account for correlations between random effects located in close geographic proximity (Osland et al., 2016). When studying regional housing prices, for instance, urban residential neighbourhoods may have many similar features. As we move further away from these neighbourhoods, the zones may more or less gradually change character,