International Journal of Forecasting 30 (2014) 407–414

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Prediction in a spatial nested error components panel data model

Badi H. Baltagi [a,b,*], Alain Pirotte [c,d]

[a] Syracuse University, Department of Economics and Center for Policy Research, USA
[b] Leicester University, Department of Economics, Leicester, UK
[c] CRED-TEPP, University Panthéon-Assas Paris II, France
[d] IFSTTAR-DEST, French Institute of Science and Technology for Transport, Development and Networks, France

ARTICLE INFO

Keywords:
Spatial nested error components
Unbalanced panels
Forecasting
Linear predictor
BLUP

ABSTRACT

This paper derives the Best Linear Unbiased Predictor (BLUP) for a spatial nested error components panel data model. This predictor is useful for panel data applications that exhibit spatial dependence and a nested (hierarchical) structure. The predictor allows for unbalancedness in the number of observations in the nested groups. One application includes forecasting average housing prices located in a county nested in a state. When deriving the BLUP, we take into account the spatial correlation across counties, as well as the unbalancedness due to observing different numbers of counties nested in each state. Ignoring the nested spatial structure leads to inefficiency and inferior forecasts. Using Monte Carlo simulations, we show that our feasible predictor is better in root mean square error performance than the usual fixed and random effects panel predictors which ignore the spatial nested structure of the data.

1. Introduction

Baltagi and Pirotte (2013) derive the Best Linear Unbiased Predictor (BLUP) for a nested error components panel data model that ignores the spatial correlation along the cross-sectional units. They show that forecasting a nested panel data model with a non-nested error structure leads to forecasts with higher root mean square errors (RMSE). This emphasizes the need to account for the *nested structure* of the data when forecasting. However, Baltagi and Pirotte (2013) did not consider possible *spatial autocorrelation* in the data. That is done in this paper. In fact, Baltagi and Pirotte (2010) emphasized that if the spatial dimension is neglected, tests of hypotheses using the usual panel data estimators like random effects (RE) and fixed effects (FE) estimators perform badly and can lead to misleading inference. Accounting for spatial dependence in forecasting using panel data has been considered by Baltagi and Li (2004, 2006), who forecasted sales of cigarette and liquor per capita for U.S. states over time. However, these applications were for balanced panels and had no nested structure for the data. Spatial correlation arises in many examples, see Anselin (1988) and LeSage and Pace (2009) for several examples and a nice introduction to this literature. The structure of the spatial dependence can be related to location and distance, in both a geographical space and a more general economic or social network space (see Anselin, Le Gallo, & Jayet, 2008, Chap. 19). One application includes forecasting average housing prices in a county nested in a state. For this application, one has to account for the spatial correlation across counties, as well as the unbalancedness due to observing different numbers of counties nested in each state.[1] For a survey of panel data forecasting

* Corresponding author at: Syracuse University, Department of Economics and Center for Policy Research, USA. Tel.: +1 315 443 1630; fax: +1 315 443 1081.
E-mail address: bbaltagi@maxwell.syr.edu (B.H. Baltagi).

[1] It is important to note that this paper does not allow for unbalancedness in the time dimension, but assumes that there are no

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.11.006

that does not include spatial dependence, see Baltagi (2008), and for spatial panel data forecasting that does *not* account for the nested structure in the data, see Baltagi, Bresson, and Pirotte (2012). The latter study considered the case where the true Data Generating Process (DGP) is random effects with a spatial autoregressive (SAR) or a spatial moving average (SMA) remainder error. Using Monte Carlo experiments, Baltagi et al. (2012) find that estimators that ignore heterogeneity/spatial autocorrelation perform badly in RMSE forecasts. Their results also show that accounting for heterogeneity improves the RMSE forecast performance by a big margin, while accounting for spatial autocorrelation improves the forecast performance by a smaller margin. Ignoring both leads to the worst RMSE forecasting performance. Another application is that of Longhi and Nijkamp (2007), who obtain short-term forecasts of employment in a panel of 326 West German regional labor markets observed over the period 1987–2002. The authors find that taking spatial autocorrelation into account by means of spatial error models leads to forecasts that are, on average, more reliable than those from models which neglect regional spatial autocorrelation. Girardin and Kholodilin (2011) obtain multi-step forecasts of the annual growth rates of the real gross regional product (GRP) for a panel of 31 Chinese regions over the period 1979–2007. This is done using a dynamic spatial panel model. They argue that using panel data and accounting for spatial effects improve the forecasting performance substantially compared to the benchmark models estimated for each of the provinces separately. They also find that accounting for spatial dependence is even more pronounced at longer forecasting horizons where the root mean squared forecast error (RMSFE) improves from 8% at the 1-year horizon to over 25% at the 13- and 14-year horizons. They recommend incorporating a spatial dependence structure into regional forecasting models, especially when long-run forecasts are made. Also, Kholodilin, Siliverstovs, and Kooths (2008) consider a dynamic spatial panel model for forecasting the GDP of 16 German Länder (states) over the period 1991–2006, at horizons varying from one to five years. Using root mean squared forecast errors, they show that accounting for spatial effects helps to improve the forecast performance, especially at longer horizons. In fact, they find that this gain in RMSFE is about 9% at the 1-year horizon and exceeds 40% at the 5-year horizon.

This paper focuses on prediction and derives the Goldberger (1962) BLUP for a spatial nested error components panel data model.[2] Using Monte Carlo experiments, this paper shows that this predictor performs well in terms of out-of-sample root mean square errors. The predictions are based on the maximum likelihood estimator, which takes into account the special *unbalanced* aspect of the data, the *spatial autocorrelation* and the *nested structure* of the disturbances. The paper is organized as follows: in Section 2, we derive the BLUP for the spatial nested random effects model with the special *unbalanced* aspect of the data. Section 3 describes the Monte Carlo design, while Section 4 describes the Monte Carlo results. Section 5 concludes with suggestions for further work.

2. The spatial nested error components model

Consider the unbalanced panel data regression model:

$$y_{ijt} = \mathbf{x}_{ijt}\boldsymbol{\beta} + \varepsilon_{ijt}, \tag{1}$$

where $i = 1, \ldots, N$, $j = 1, \ldots, M_i$, and $t = 1, \ldots, T$. The dependent variable y_{ijt} could denote the average house price in county j located in state i at time period t. $\mathbf{x}_{ijt}$ is a $(1 \times K)$ vector of explanatory (exogenous) variables, while $\boldsymbol{\beta}$ represents a $(K \times 1)$ vector of parameters to be estimated. N denotes the number of states, and M_i denotes the number of counties in each state i. This model allows for an unequal number of counties in each state i. However, it does not allow for missing observations across time. Moreover, in contrast to the usual panel data framework, we allow ε_{ijt} to be contemporaneously correlated. A simple and widely used approach to modelling spatial error dependence is to assume a SAR process:

$$\varepsilon_{ijt} = \rho \sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{ij,gh}\varepsilon_{ght} + u_{ijt}, \tag{2}$$

where ρ is the autoregressive parameter to be estimated. The weight $w_{ij,gh} = w_{k,l}$ is the (k, l) element of the matrix W_S, with ij denoting county j within state i, and similarly for gh. Thus, $k, l = 1, \ldots, S$, where $S = \sum_{i=1}^{N} M_i$ and W_S is an $(S \times S)$ known spatial weights matrix which has zero diagonal elements and is usually row-normalized so that for row k, $\sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{k,gh} = 1$. Typically, $\mathbf{W}_S$ is defined as first order contiguity; such elements consist of location pairs that have a common border but no higher order contiguity, or could be based on distances between counties. The error component structure of the disturbances u_{ijt} contains an unobserved permanent unit-specific error component α_i, a nested permanent unit-specific error component μ_{ij}, and a remainder error component υ_{ijt}. More formally,

$$u_{ijt} = \alpha_i + \mu_{ij} + \upsilon_{ijt}, \tag{3}$$

where α_i denotes an unobservable state-specific time-invariant effect which is assumed to be i.i.d.$N\left(0, \sigma_\alpha^2\right)$, μ_{ij} denotes the nested effect of county j within the ith state which is assumed to be i.i.d.$N\left(0, \sigma_\mu^2\right)$, and υ_{ijt} is a remainder disturbance term which is also assumed to be i.i.d.$N\left(0, \sigma_\upsilon^2\right)$. The α_is, μ_{ij}s and υ_{ijt}s are independent of each other and among themselves. In contrast to the classical literature on panel data, grouping the data by periods rather than by units is more convenient when we consider the spatial autocorrelation due to Eq. (2). For a cross-section t, the model in Eq. (1) can be written as:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \tag{4}$$

missing observations across the sample period for all counties and states. This is likely to be the case when forecasting the average price in a county, but not when forecasting individual house prices. The latter most probably exhibit unbalancedness in the time dimension, as not all house prices are observed over the sample period.

[2] While Baltagi and Li (2004) extend the BLUP to spatial panel models, Song and Jung (2002) extend the BLUP to the case of spatially *and* serially correlated error component models.