Geographical Analysis ISSN 0016-7363

Unconditional Maximum Likelihood Estimation of Linear and Log-Linear Dynamic Models for Spatial Panels

J. Paul Elhorst

Faculty of Economics, University of Groningen, Groningen, The Netherlands

This article hammers out the estimation of a fixed effects dynamic panel data model extended to include either spatial error autocorrelation or a spatially lagged dependent variable. To overcome the inconsistencies associated with the traditional least-squares dummy estimator, the models are first-differenced to eliminate the fixed effects and then the unconditional likelihood function is derived taking into account the density function of the first-differenced observations on each spatial unit. When exogenous variables are omitted, the exact likelihood function is found to exist. When exogenous variables are included, the pre-sample values of these variables and thus the likelihood function must be approximated. Two leading cases are considered: the Bhargava and Sargan approximation and the Nerlove and Balestra approximation. As an application, a dynamic demand model for cigarettes is estimated based on panel data from 46 U.S. states over the period from 1963 to 1992.

Introduction

In recent years, there has been a growing interest in the estimation of econometric relationships based on panel data. In this article, we focus on dynamic models for spatial panels, a family of models for which, according to Elhorst (2001) and Hadinger, Müller, and Tondl (2002), no straightforward estimation procedure is yet available. This is (as will be explained later) because existing methods developed for spatial but non-dynamic and for dynamic but non-spatial panel data models produce biased estimates when these methods/models are put together.

A dynamic spatial panel data model takes the form of a linear regression equation extended with a variable intercept, a serially lagged dependent variable and either a spatially lagged dependent variable (known as *spatial lag*) or a spatially autoregressive process incorporated in the error term (known as *spatial*

Correspondence: J. Paul Elhorst, Faculty of Economics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
e-mail: j.p.elhorst@eco.rug.nl

Submitted: July 11, 2003. Revised version accepted: May 18, 2004.

error). To avoid repetition, we apply to the spatial error specification in this article. The spatial lag specification is explained in a working paper (Elhorst 2003a).[1] The model is considered in vector form for a cross-section of observations at time *t*:

$$Y_t = \tau Y_{t-1} + X_t\beta + \mu + \varphi_t, \quad \varphi_t = \delta W\varphi_t + \varepsilon_t, \quad E(\varepsilon_t) = 0, \quad E(\varepsilon_t\varepsilon_t') = \sigma^2 I_N \quad (1)$$

where Y_t denotes an $N \times 1$ vector consisting of one observation for every spatial unit ($i = 1, \ldots, N$) of the dependent variable in the *t*th time period ($t = 1, \ldots, T$) and X_t denotes an $N \times K$ matrix of exogenous explanatory variables. It is assumed that the vector Y_0 and matrix X_0 of initial observations are observable. The scalar τ and the $K \times 1$ vector β are the response parameters of the model. The disturbance term consists of $\mu = (\mu_1, \ldots, \mu_N)'$, $\varphi_t = (\varphi_{1t}, \ldots, \varphi_{Nt})'$, and $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$, where ε_{it} are independently and identically distributed error terms for all i and t with zero mean and variance σ^2. I_N is an identity matrix of size N, W represents an $N \times N$ non-negative spatial weight matrix with zeros on the diagonal, and δ represents the spatial autocorrelation coefficient. The properties of μ are explained below.

The reasons for considering serial and spatial dynamic effects, either directly as part of the specification or indirectly as part of the disturbance term, have been published earlier (Elhorst 2001, 2004). A standard space–time model, even if it is dynamic, still assumes that the spatial units are completely homogeneous, differing only in their explanatory variables. Standard space–time models include the STARMA/STARIMA (Space Time AutoRegressive [Integrated] Moving Average) model (Hepple 1978; Pfeifer and Deutsch 1980), spatial autoregression space–time forecasting model (Griffith 1996), and the serial and spatial autoregressive distributed lag model (Elhorst 2001). A panel data approach would presume that spatial heterogeneity is a feature of the data and attempt to model that heterogeneity. The need to account for spatial heterogeneity is that spatial units are likely to differ in their background variables, which are usually space-specific time-invariant variables that affect the dependent variable, but are difficult to measure or hard to obtain. Omission of these variables leads to bias in the resulting estimates. One remedy is to introduce a variable intercept μ_i representing the effect of the omitted variables that are peculiar to each spatial unit considered (Baltagi 2001, chap. 1). Conditional upon the specification of the variable intercept μ_i, the regression equation can be estimated as a fixed or a random effects model. In the fixed effects model, a dummy variable is introduced for each spatial unit as a measure of the variable intercept. In the random effects model, the variable intercept is treated as a random variable that is independently and identically distributed with zero mean and variance σ_μ^2.

Whether the random effects model is an appropriate specification in spatial research remains controversial. When the random effects model is implemented, the units of observation should be representative of a larger population, and the number of units should potentially be able to go to infinity. There are two types of asymptotics that are commonly used in the context of spatial observations: (i) the "infill" asymptotic structure, where the sampling region remains bounded as $N \to \infty$. In this case, more units of information come from observations taken from between those already observed and (ii) the "increasing domain" asymptotic structure, where the sampling region grows as $N \to \infty$ and the sample design is such that there is a minimum distance separating any two spatial units for all N. According to Lahiri (2003), there are also two types of sampling designs: (i) the stochastic design where the spatial units are randomly drawn and (ii) the fixed design where the spatial units lie on a non-random field, possibly irregularly spaced. The spatial econometric literature mainly focuses on increasing domain asymptotics under the fixed sample design (Cressie 1991, p. 100; Griffith and Lagona 1998; Lahiri 2003). Although the number of spatial units under the fixed sample design can potentially go to infinity, it is questionable whether they are representative of a larger population. For a given set of regions, such as all counties of a state or all regions in a country, the population may be said "to be sampled exhaustively" (Nerlove and Balestra 1996, p. 4), and "the individual spatial units have characteristics that actually set them apart from a larger population" (Anselin 1988, p. 51). According to Beck (2001, p. 272), "the critical issue is that the spatial units be fixed and not sampled, and that inference be conditional on the observed units." In addition, the traditional assumption of zero correlation between μ_i in the random effects model and the explanatory variables is particularly restrictive. For these reasons, the random effects model is often not used. We will return to the random effects model briefly in the concluding section.

The dynamic spatial panel data model was first considered by Hepple (1978). His conclusion was that empirical studies with a serially lagged dependent variable have a real problem in that ordinary least squares (OLS) is no longer consistent when relaxing the assumption that the disturbance term is homoskedastic and independently distributed (e.g., because of a variable intercept). He also pointed out that estimation would have to be by maximum likelihood (ML) and that this was worth pursuing, but did not explain how to do so. Buettner (1999) has estimated a wage curve for Germany using the non-dynamic fixed effects spatial lag model and using the non-spatial dynamic fixed effects models, but not the dynamic spatial panel data model.

Spatial but non-dynamic panel data model

The standard estimation method for the fixed effects model without a serially lagged dependent variable and without spatial error autocorrelation ($\tau = 0$; $\delta = 0$) is to eliminate the intercept β_1 and the dummy variables μ_i from the regression equation. This is possible by taking each variable in the regression equation in deviation from its average over time $\left(z_{it} - (1/T)\sum_t z_{it} \text{ for } z = y, x\right)$, called demeaning. The slope coefficients β (the $K \times 1$ vector β without the intercept) in the resulting equation can then be estimated by OLS, known as the least-squares dummy variables (LSDV) estimator. Subsequently, the intercept β_1 and the dummy variables μ_i may be recovered (Baltagi 2001, pp. 12–15). It should be stressed that the coefficients of these dummy variables cannot be estimated consistently, because the number of observations available for the estimation of μ_i is limited to T observations. However, in many empirical applications this problem does not matter, because τ and β are the coefficients of interest and μ_i are not. Fortunately, the inconsistency of μ_i is not transmitted to the estimator of the slope coefficients in the demeaned equation, because this estimator is not a function of the estimated μ_i. This implies that increasing domain asymptotics under the fixed sample design ($N \rightarrow \infty$) do apply for the demeaned equation.

Anselin (1988) shows that OLS estimation is inefficient for cross-sectional models incorporating spatial error autocorrelation (τ still 0, but $\delta \neq 0$)[2] and suggests overcoming this problem by using ML. This is important because the LSDV estimator of the fixed effects models falls back on the OLS estimator of the response coefficients in the demeaned equation. Elhorst (2003b) shows that ML estimation of the spatial fixed effects model can be carried out with standard techniques developed by Anselin (1988, pp. 181–82) and Anselin and Hudak (1992) after the variables in the regression equation have been demeaned. The asymptotic properties of the ML estimator depend on the spatial weight matrix. The critical condition is that the row and column sums, before normalizing the spatial weight matrix, should not diverge to infinity at a rate equal to or faster than the rate of the sample size N in the cross-section domain (Lee 2002).[3]

Dynamic but non-spatial panel data model

The panel data literature has extensively discussed the dynamic but non-spatial panel data model ($\tau \neq 0$, but $\delta = 0$; see Hsiao 1986, chap. 4; Sevestre and Trognon 1996; Baltagi 2001, chap. 8). The most serious estimation problem caused by the introduction of a serially lagged dependent variable is that the OLS estimator of the response coefficients in the demeaned equation, as discussed above and in this case consisting of τ and β, is inconsistent *if T is fixed, regardless of the size of N.* Two procedures to remove this inconsistency are being intensely discussed in the panel data literature.

The first procedure considers the unconditional likelihood function of the model formulated in levels (cf. Equation (1)). Regression equations that include variables lagged one period in time are often estimated conditional upon the first observations. When estimating these models by ML, it is also possible to obtain unconditional results by taking into account the density function of the first observation of each time-series of observations. This so-called unconditional likelihood function has been shown to exist when applying this procedure to a standard linear regression model without exogenous explanatory variables (Hamilton 1994; Johnston and Dinardo 1997, pp. 229–30), and on a random effects model without exogenous explanatory variables (Ridder and Wansbeek 1990; Hoogstrate 1998; Hsiao, Pesaran, and Tahmiscioglu 2002). Unfortunately, the unconditional likelihood function does not exist when applying this procedure on the fixed effects model, even without exogenous explanatory variables. The reason is that the coefficients of the fixed effects cannot be estimated consistently, because the number of these coefficients increases as N increases. The standard solution to eliminate these fixed effects from the regression equation by demeaning the Y and X variables also does not work, because this technique creates a correlation of order $(1/T)$ between the serial lagged dependent variable and the demeaned error terms (Nickell 1981; Hsiao 1986, pp. 73–76), as a result of which the common parameter τ cannot be estimated consistently. Only when T tends to infinity does this inconsistency disappear.

The second procedure first-differences the model to eliminate the fixed effects, $Y_t - Y_{t-1} = \tau(Y_{t-1} - Y_{t-2}) + (X_t - X_{t-1})\beta + \varphi_t - \varphi_{t-1}$, and then applies generalized method-of-moments (GMM) using a set of appropriate instruments.[4] Recently, Hoogstrate (1998) and Hsiao, Pesaran, and Tahmiscioglu (2002) have suggested a third procedure that combines the preceding two. This procedure first-differences the model to eliminate the fixed effects and then considers the unconditional likelihood function of the first-differenced model taking into account the density function of the first-differenced observations on each spatial unit. Hsiao, Pesaran, and Tahmiscioglu (2002) prove that this procedure yields a consistent estimator of the scalar τ and the response parameters β when the cross-sectional dimension N tends to infinity, regardless of the size of T. It is also shown that the ML estimator is asymptotically more efficient that the GMM estimator.

Estimation of a dynamic spatial panel data model

The advantage of the last procedure is that it also opens the possibility to estimate a fixed effects dynamic panel data model extended to include spatial error autocorrelation (or a spatially lagged dependent variable), which is the objective of this article. We utilize the ML estimator; the objection to GMM from a spatial point of view is that this estimator is less accurate than ML (see Das, Kelejian, and Prucha 2003). This is because δ is bounded from below and above using ML, whereas it is unbounded using GMM; the transformation of the estimation model from the error term to the dependent variable contains a Jacobian term,[5] which the ML approach takes into account but the GMM approach does not.

In addition to spatial heterogeneity, it is the specification of the generating process of the initial observations that sets this estimation procedure apart from those used previously to standard space–time models (STARMA, STARIMA, and the spatial autoregression space–time forecasting model). As the first cross-section of observations conveys a great deal of information, conditioning on these observations is an undesirable feature, especially when the time-series dimension of the spatial panel is short. It is not difficult to obtain the unconditional likelihood function once the marginal distribution of the initial values is specified. The problem arises in obtaining a valid specification of this distribution when the model contains exogenous variables. This is because the likelihood function under this circumstance depends on pre-sample values of the exogenous explanatory variables and additional assumptions have to be made to approach these values. The panel data literature has suggested different distributions leading to different optimal estimation procedures. We consider the Bhargava and Sargan (1983) (BS) approximation, which is also applied in Hsaio, Pesaran, and Tahmiscioglu (2002), and an approximation recently introduced by Nerlove and Balestra (1996) (NB) and Nerlove (1999) or Nerlove (2000).

As a spatial panel has two dimensions, it is possible to consider asymptotic behavior as $N \rightarrow \infty$, $T \rightarrow \infty$, or both. Generally speaking, it is easier to increase the cross-section dimension of a spatial panel. If as a result $N \rightarrow \infty$ is believed to be the most relevant inferential basis, it follows from the above discussion that the parameter estimates of τ and β derived from the unconditional likelihood function of the fixed effects dynamic panel data transformed into first differences and extended to include spatial autocorrelation (or a spatially lagged dependent variable) are consistent.

The remainder of this article consists of one technical, one empirical, and one concluding section. In the technical section, we derive the unconditional likelihood function of the dynamic panel data model extended to include spatial error autocorrelation first excluding and then including exogenous explanatory variables. In the empirical section, a dynamic demand model for cigarettes is estimated based on panel data from 46 U.S. states over the period from 1963 to 1992. The concluding section recapitulates our major findings.

Spatial error specification

No exogenous explanatory variables

In this section, exogenous explanatory variables are omitted from Equation (1). Although this model will probably seldom be used in applied work, it is still interesting because the exact log-likelihood function exists. Taking first differences of (1), the dynamic panel data model excluding exogenous explanatory variables ($\beta = 0$) extended to include spatial error autocorrelation changes into

$$\Delta Y_t = \tau \Delta Y_{t-1} + B^{-1} \Delta \varepsilon_t \qquad (2)$$

where $B = I_N - \delta W$. It is assumed that the characteristic roots of the spatial weight matrix W, denoted by ω_i ($i = 1, \ldots, N$), are known. This assumption is needed to ensure that the log-likelihood function of the models below can be computed. Additional properties of W, which we call *Griffith's matrix properties* throughout this article, are (Griffith 1988, p. 44, Table 3.1): (i) if W is multiplied by some scalar constant, then its characteristic roots are also multiplied by this constant; (ii) if δI is added to W, where δ is a real scalar, then δ is added to each of the characteristic roots of W; (iii) the characteristic roots of W and its transpose are the same; (iv) the characteristic roots of W and its inverse are inverses of each other; and (v) if W is powered by some real number, each of its characteristic roots is powered by this same real number.

ΔY_t is well defined for $t = 2, \ldots, T$, but not for ΔY_1 because ΔY_0 is not observed. To be able to specify the ML function of the complete sample ΔY_t ($t = 1, \ldots, T$), the probability function of ΔY_1 must be derived first. Therefore, we repeatedly lag Equation (2) by one period. For ΔY_{t-m} ($m \geq 1$), we obtain

$$\Delta Y_{t-m} = \tau \Delta Y_{t-(m+1)} + B^{-1} \Delta \varepsilon_{t-m} \tag{3}$$

Then, by substitution of $\varDelta Y_{t-1}$ into (2), next $\varDelta Y_{t-2}$ into (2) up to $\varDelta Y_{t-(m-1)}$ into (2), we obtain

$$\begin{aligned}\Delta Y_t =& \tau^m \Delta Y_{t-m} + B^{-1} \Delta \varepsilon_t + \tau B^{-1} \Delta \varepsilon_{t-1} + \cdots + \tau^{m-1} B^{-1} \Delta \varepsilon_{t-(m-1)} \\ =& \tau^m \Delta Y_{t-m} + B^{-1} [\varepsilon_t + (\tau - 1) \varepsilon_{t-1} + (\tau - 1) \tau \varepsilon_{t-2} \\ & + \cdots + (\tau - 1) \tau^{m-2} \varepsilon_{t-(m-1)} - \tau^{m-1} \varepsilon_{t-m}]\end{aligned} \tag{4}$$

As $E(\varepsilon_t) = 0$ ($t = 1, \ldots, T$) and the successive values of ε_t are uncorrelated,

$$E(\Delta Y_t) = \tau^m \Delta Y_{t-m} \quad \text{and} \quad \mathrm{Var}(\Delta Y_t) = \sigma^2 v_b B^{-1} B'^{-1} \tag{5}$$

where the scalar v_b is defined as

$$v_b = \frac{2}{1+\tau}(1 + \tau^{2m-1}) \tag{6}$$

Two assumptions with respect to ΔY_1 can be made (cf. Hsiao, Pesaran, and Tahmiscioglu 2002):

(I) The process started in the past, but not too far back from the 0th period, and the expected *changes* in the initial endowments are the same across all spatial units. Note that this assumption, although restrictive, does not impose the even stronger restriction that all spatial units should start from the same initial endowments. Under this assumption, $E(\Delta Y_1) = \pi_0 1_N$, where 1_N denotes an $N \times 1$ vector of unit elements and π_0 is a fixed but unknown parameter to be estimated.

(II) The process has started long ago (m approaches infinity) and $|\tau| < 1$. Under this assumption, $E(\Delta Y_1) = 0$, while v_b reduces to $v_b = 2/(1 + \tau)$.

It can be seen that the first assumption reduces to the second one, when $\pi_0 = 0$, $|\tau| < 1$, and m is sufficiently large so that the term τ^m becomes negligible. Therefore, we consider the unconditional log-likelihood function of the complete sample under the more general assumption (I).

Writing the residuals of the model as $\Delta e_t = \Delta Y_t - \tau \Delta Y_{t-1}$ for $t = 2, \ldots, T$ and, using assumption (I), $\Delta e_1 = \Delta Y_1 - \pi_0 l_N$ for $t = 1$, we have $\mathrm{Var}(\Delta e_1) = \sigma^2 v_b B^{-1} B'^{-1}$, $\mathrm{Var}(\Delta e_t) = 2\sigma^2 B^{-1} B'^{-1}$ ($t = 2, \ldots, T$), $\mathrm{Covar}(\Delta e_t, \Delta e_{t-1}) = -\sigma^2 B^{-1} B'^{-1}$ ($t = 2, \ldots, T$), and zero otherwise. This implies that the covariance matrix of Δe

can be written as $\mathrm{Var}(\Delta e) = \sigma^2(G_{v_b} \otimes B^{-1}B'^{-1})$, by which v_b is given in (6), $\otimes$ denotes the Kronecker product, and the $T \times T$ matrix $G_v|_{v=v_b}$ is defined as

Definition 1.

$$G_v \equiv \begin{bmatrix} v & -1 & 0 & . & 0 & 0 \\ -1 & 2 & -1 & . & 0 & 0 \\ 0 & -1 & 2 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 2 & -1 \\ 0 & 0 & 0 & . & -1 & 2 \end{bmatrix}$$

with its subelement in the first row and first column set to v.

Properties of the matrix G_v used below are: (i) The determinant is $|G_v| = 1 - T + T \times v$; (ii) the inverse is $G_v^{-1} = 1/(1 - T + T \times v) \times [(1 - T)G_0^{-1} + v(G_1^{-1} - (1 - T)G_0^{-1})]$, where the inverse matrices $G_0^{-1} = G_v^{-1}|_{v=0}$ and $G_1^{-1} = G_v^{-1}|_{v=1}$ can easily be calculated and are characterized by a specific structure; and (iii) let p denote an $NT \times 1$ vector, which can be partitioned into T block rows of length N. When p_t denotes the tth block row $(t = 1, \ldots, T)$ of p, then $p'(G_{v_b} \otimes I_N)^{-1}p = \sum_{t_1=1}^{T}\sum_{t_2=1}^{T} G_{v_b}^{-1}(t_1, t_2)p'_{t_1}p_{t_2}$, where $G_{v_b}^{-1}(t_1, t_2)$ represents the element of $G_{v_b}^{-1}$ in row t_1 and column t_2.

In sum, we have[6]

$$\log L = -\frac{NT}{2}\log(2\pi\sigma^2) + T\log|B| - \frac{N}{2}\log|G_{v_b}| - \frac{1}{2\sigma^2}\Delta e^{*'}(G_{v_b} \otimes I_N)^{-1}\Delta e^{*} \qquad (7a)$$

where

$$\Delta e^{*} = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N) \\ B(\Delta Y_2 - \tau\Delta Y_1) \\ . \\ B(\Delta Y_T - \tau\Delta Y_{T-1}) \end{bmatrix}, \quad E(\Delta e^{*}\Delta e^{*'}) = \sigma^2(G_{v_b} \otimes I_N) \qquad (7b)$$

This log-likelihood function is well-defined, satisfies the usual regularity conditions, and contains four unknown parameters to be estimated: π_0, τ, δ, and σ^2. An appropriate value of m should be chosen in advance. σ^2 can be solved from its first-order maximizing condition, $\hat{\sigma}^2 = 1/NT\,\Delta e^{*'}(G_{v_b} \otimes I_N)^{-1}\Delta e^{*}$. On substituting σ^2 into the log-likelihood function and using Griffith's matrix properties and the properties of G_v given below Definition 1, the concentrated log-likelihood function of π_0, τ, and δ is obtained as

$$\begin{aligned} \mathrm{Log}\, L_C = {} & C - \frac{NT}{2}\log\left[\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} G_{v_b}^{-1}(t_1, t_2)\Delta e_{t_1}^{*'}\Delta e_{t_1}^{*}\right] \\ & + T\sum_{i=1}^{N}\log(1 - \delta\omega_i) - \frac{N}{2}\log\left(1 - T + T \times \frac{2}{1+\tau}(1 + \tau^{2m-1})\right) \end{aligned} \qquad (8)$$

where C is a constant ($C = -NT/2(1 + \log 2\pi)$). As the first-order maximizing conditions of this function are non-linear, a numerical iterative procedure must be used to find the maximum for π_0, τ, and δ.

Exogenous explanatory variables

In this section, explanatory variables are added to the model. They are assumed to be strictly exogenous and to be generated by a stationary process in time. By taking first differences and continuous substitution, we can rewrite the dynamic panel data model (1) extended to include spatial error autocorrelation as

$$\begin{aligned}\Delta Y_t =& \tau^m \Delta Y_{t-m} + B^{-1}\Delta\varepsilon_t + \tau B^{-1}\Delta\varepsilon_{t-1} + \cdots + \tau^{m-1}B^{-1}\Delta\varepsilon_{t-(m-1)} \\ &+ \sum_{j=0}^{m-1} \tau^j \Delta X_{t-j}\beta = \tau^m \Delta Y_{t-m} + \Delta e_t + X^*\end{aligned} \tag{9}$$

As X_t is stationary, we have $E\Delta X_t = 0$ and thus $E(\Delta Y_1) = \tau^m \Delta Y_{t-m}$. This expectation is determined under assumption (I). By contrast, $\mathrm{Var}(\Delta Y_1)$ is undetermined, as X^* is not observed. This implies that the probability function of ΔY_1 is also undetermined. The panel data literature has suggested different assumptions about X^* leading to different optimal estimation procedures. We consider two leading cases: the BS approximation and the NB approximation.

The BS approximation

Bhargava and Sargan (1983) suggest predicting X^* when $t = 1$ by all the exogenous explanatory variables in the model subdivided by time over the observation period. In other words, when the model contains K_1 time varying and K_2 time invariance explanatory variables over T time periods, X^* is approached by $K_1 \times T + K_2$ regressors. Lee (1981), Ridder and Wansbeek (1990), and Blundell and Smith (1991) use a similar approach. Hsiao, Pesaran, and Tahmiscioglu (2002) apply this approximation on the fixed effects model formulated in first differences. One of their main conclusions is that there is much to recommend the ML estimator based on this approximation. The results of a Monte Carlo simulation study strongly favor the ML estimator over other estimators (instrumental variables [IV], GMM) and the ML estimator appears to have excellent finite sample properties even when both N and T are quite small.

The predictor of X^* under assumption (I) is $\pi_0 1_N + \Delta X_1 \pi_1 + \cdots + \Delta X_T \pi_T + \xi$, where $\xi \sim N(0, \sigma_\xi^2 I_N)$, π_0 is a scalar, and π_t ($t = 1, \ldots, T$) are $K \times 1$ vectors of parameters. When the kth variable of X is time invariant, the restriction $\pi_{1k} = \cdots = \pi_{Tk}$ should be imposed. In addition to this, the condition $N > 1 + K \times T$ should hold; otherwise, the number of parameters used to predict X^* must be reduced. We thus have

$$\begin{aligned}\Delta Y_1 &= \pi_0 1_N + \Delta X_1 \pi_1 + \cdots + \Delta X_T \pi_T + \Delta e_1 \\ \text{where } \Delta e_1 &= \xi + B^{-1} \sum_{j=0}^{m-1} \tau^j \Delta\varepsilon_{1-j}\end{aligned} \tag{10a}$$

$$E(\Delta e_1) = 0, \quad E(\Delta e_1 \Delta e_2') = -\sigma^2 B^{-1} B'^{-1}, \quad E(\Delta e_1 \Delta e_t') = 0 \ (t = 3, \ldots, T) \tag{10b}$$

$$E(\Delta e_1 \Delta e_1') = \sigma_\xi^2 I_N + \sigma^2 v_b B^{-1} B'^{-1} \equiv \sigma^2 B^{-1}(\theta^2 BB' + v_b I_N) B'^{-1} \tag{10c}$$

Instead of estimating σ_ξ^2 and σ^2, it is easier to estimate θ^2 ($\theta^2 = \sigma_\xi^2/\sigma^2$) and σ^2, which is allowed as there exists a one-to-one correspondence between σ_ξ^2 and θ^2.

Let $V_{\mathrm{BS}} = \theta^2 BB' + v_b I_N = \theta^2 BB' + \frac{2}{1+\tau}(1 + \tau^{2m-1}) I_N$; then, the covariance matrix of Δe can be written as $\mathrm{Var}(\Delta e) = \sigma^2[(I_T \otimes B^{-1}) H_{V_{\mathrm{BS}}} (I_T \otimes B'^{-1})]$, by which the $NT \times NT$ matrix $H_V|_{V=V_{\mathrm{BS}}}$ is defined as

Definition 2.

$$H_V \equiv \begin{bmatrix} V & -I_N & 0 & . & 0 & 0 \\ -I_N & 2 \times I_N & -I_N & . & 0 & 0 \\ 0 & -I_N & 2 \times I_N & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 2 \times I_N & -I_N \\ 0 & 0 & 0 & . & -I_N & 2 \times I_N \end{bmatrix}$$

with its submatrix in the first block row and first block column set to the $N \times N$ matrix V.

Properties of the matrix H_v used below are: (i) The determinant is $|H_V| = |I_N - T \times I_N + T \times V|$; (ii) The inverse is $H_V^{-1} = (1 - T)(G_0^{-1} \otimes D^{-1}) + ((G_1^{-1} - (1 - T)G_0^{-1}) \otimes (D^{-1}V)$, where $D = I_N - T \times I_N + T \times V$; and (iii) H_V^{-1} can be partitioned into T block rows and T block columns, by which the submatrix $H_V^{-1}(t_1, t_2)$ $(t_1, t_2 = 1, \ldots, T)$ equals $H_V^{-1}(t_1, t_2) = (1 - T)G_0^{-1}(t_1, t_2) \times D^{-1} + (G_1^{-1}(t_1, t_2) - (1 - T)G_0^{-1}(t_1, t_2) \times (D^{-1}V)$. The last equation is used to obtain the matrix H_V^{-1} computationally.

Using Griffith's matrix properties and the properties of H_V given below Definition 2, the log-likelihood function is obtained as

$$\begin{aligned} \log L = & -\frac{NT}{2}\log(2\pi\sigma^2) + T\sum_{i=1}^{N}\log(1 - \delta\omega_i) \\ & -\frac{1}{2}\sum_{i=1}^{N}\log\left(1 - T + T \times \frac{2}{1+\tau}(1 + \tau^{2m-1}) + T\theta^2(1 - \delta\omega_i)^2\right) \\ & -\frac{1}{2\sigma^2}\Delta e^{*\prime} H_{V_{\mathrm{BS}}}^{-1} \Delta e^* \end{aligned} \tag{11a}$$

$$\text{where } \Delta e^* = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N - \Delta X_1 \pi_1 - \cdots - \Delta X_T \pi_T) \\ B(\Delta Y_2 - \tau \Delta Y_1 - \Delta X_2 \beta) \\ . \\ B(\Delta Y_T - \tau \Delta Y_{T-1} - \Delta X_T \beta) \end{bmatrix}, \tag{11b}$$

$$E(\Delta e^* \Delta e^{*\prime}) = \sigma^2 H_{V_{\mathrm{BS}}}$$

This log-likelihood function is well-defined, satisfies the usual regularity conditions, and contains $KT+K+5$ unknown parameters to be estimated: $\pi_1, \ldots, \pi_T$, $\beta, \pi_0, \theta^2, \tau, \delta$, and σ^2. An appropriate value of m should be chosen in advance. σ^2, π, and β can be solved from their first-order maximizing conditions

$$\hat{\sigma}^2 = \frac{\Delta e^{*\prime} H_{V_{\mathrm{BS}}}^{-1} \Delta e^*}{NT} \quad \text{and} \quad \begin{bmatrix} \hat{\pi}_0 \\ \hat{\pi}_1 \\ . \\ \hat{\pi}_T \\ \hat{\beta} \end{bmatrix} = (\tilde{X}' H_{V_{\mathrm{BS}}}^{-1} \tilde{X})^{-1} \tilde{X}' H_{V_{\mathrm{BS}}}^{-1} \tilde{Y} \tag{12a}$$

$$\text{where } \tilde{X} = \begin{bmatrix} B & B\Delta X_1 & . & B\Delta X_T & 0 \\ 0 & 0 & . & 0 & B\Delta X_2 \\ . & . & . & . & . \\ 0 & 0 & . & 0 & B\Delta X_T \end{bmatrix} \quad \text{and}$$

$$\tilde{Y} = \begin{bmatrix} B\Delta Y_1 \\ B(\Delta Y_2 - \tau \Delta Y_1) \\ . \\ B(\Delta Y_T - \tau \Delta Y_{T-1}) \end{bmatrix} \tag{12b}$$

On substituting $\hat{\sigma}^2, \hat{\pi}$, and $\hat{\beta}$ into the log-likelihood function, the concentrated log-likelihood function of θ^2, τ, and δ is obtained. A numerical iterative procedure must be used to find the maximum for these parameters.

The NB approximation

Starting with a regression equation formulated in levels (instead of first differences), Nerlove and Balestra (1996) and Nerlove (1999) or Nerlove (2000) suggest replacing the variance of X_{t-j} ($j = 0, \ldots, m-1$) by Σ_X, where Σ_X denotes the covariance matrix of the explanatory variables X. This covariance matrix may be determined from the sample data in advance and then used to calculate the unknown variance of X^*. Suppose that each explanatory variable X_{tk} ($k = 1, \ldots, K$) follows a well-specified common stationary time-series model

$$X_{tk} = \tau_{X_k} X_{t-1\,k} + \gamma_t \quad \text{where} \quad \gamma_t \sim N(0, \sigma^2_{\gamma X_k} I_N) \tag{13}$$

Then the random variable X^* in (9) has a well-defined variance Σ_{X^*}, which is a function of β and $\tau_{X_k}, \sigma^2_{\gamma X_k}$ ($k = 1, \ldots, K$). Although it would be possible to determine the resulting log-likelihood function based on Σ_{X^*}, this covariance matrix depends on so many parameters that its practical value in empirical applications is almost nil (unless K is very small). Nerlove and Balestra (1996) and Nerlove (1999) or Nerlove (2000) have pointed out that it is not necessary to go that far. As we are not really interested in the parameters τ_{X_k} and $\sigma^2_{\gamma X_k}$ ($k = 1, \ldots, K$), we can suppress these parameters and restrict the log-likelihood to the remaining parameters. While omitting estimation of τ_{X_k} and $\sigma^2_{\gamma X_k}$ $(k=1,\ldots,K)$ leads to a loss of efficiency, the ML estimates obtained in this way remain consistent as long as the random variables have well-defined variances and covariances, which they will if the explanatory variables are generated by a stationary process. One of Nerlove's (1999) or Nerlove's (2000) main conclusions is that the conditional LSDV estimator yields low estimates of the parameter τ compared with the unconditional ML estimator, particularly when T is small. Unfortunately, the loss of efficiency as a result of suppressing parameters in the unconditional log-likelihood function is not investigated. For a comparison of the conditional LSDV estimator and the unconditional ML estimator according to this approximation, refer to the next section.

Following Nerlove and Balestra, but then for a regression equation formulated in first differences, $\mathrm{Var}(\Delta Y_1)$ might be approached by

$$\begin{aligned}\mathrm{Var}(\Delta Y_1) &= \mathrm{Var}(\Delta e_1) + \mathrm{Var}(X^*) = \sigma^2 v_b B^{-1}B'^{-1} + \left(\frac{1-\tau^m}{1-\tau}\right)^2 \beta'\Sigma_{\Delta X}\beta \times I_N \\ &\equiv \sigma^2 B^{-1}\left(v_b I_N + \left(\frac{1-\tau^m}{1-\tau}\right)^2 \frac{\beta'\Sigma_{\Delta X}\beta}{\sigma^2} \times BB'\right)B'^{-1}\end{aligned} \tag{14}$$

Let $V_{\mathrm{NB}} = v_b I_N + \left(\frac{1-\tau^m}{1-\tau}\right)^2 \frac{\beta'\Sigma_{\Delta X}\beta}{\sigma^2} \times BB' = \frac{2}{1+\tau}(1+\tau^{2m-1})I_N + \left(\frac{1-\tau^m}{1-\tau}\right)^2 \frac{\beta'\Sigma_{\Delta X}\beta}{\sigma^2} \times BB'$, then the covariance matrix of Δe can be written as $\mathrm{Var}(\Delta e) = \sigma^2[(I_T \otimes B^{-1}) H_{V_{\mathrm{NB}}}(I_T \otimes B'^{-1})]$, by which the matrix $H_V|_{V=V_{\mathrm{NB}}}$ is given in Definition 2. Using Griffith's matrix properties and the properties of H_V given below Definition 2, the log-likelihood function is obtained as

$$\begin{aligned}\log L = &-\frac{NT}{2}\log(2\pi\sigma^2) + T\sum_{i=1}^{N}\log(1-\delta\omega_i) - \frac{1}{2\sigma^2}\Delta e^{*'} H^{-1}_{V_{\mathrm{NB}}}\Delta e^* \\ &-\frac{1}{2}\sum_{i=1}^{N}\log(1 - T + T \times \frac{2}{1+\tau}(1+\tau^{2m-1}) + T\left(\frac{1-\tau^m}{1-\tau}\right)^2 \frac{\beta'\Sigma_{\Delta X}\beta}{\sigma^2}(1-\delta\omega_i)^2)\end{aligned} \tag{15a}$$

where

$$\Delta e^* = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N) \\ B(\Delta Y_2 - \tau\Delta Y_1 - \Delta X_2\beta) \\ \cdot \\ B(\Delta Y_T - \tau\Delta Y_{T-1} - \Delta X_T\beta) \end{bmatrix}, \quad E(\Delta e^* \Delta e^{*'}) = \sigma^2 H_{V_{\mathrm{NB}}} \tag{15b}$$

This log-likelihood function is well-defined, satisfies the usual regularity conditions, and contains $K+4$ unknown parameters to be estimated: β, π_0, τ, δ, and σ^2. An appropriate value of m should be chosen in advance. In contrast to the preceding models, none of the parameters can be solved analytically from the first-order maximizing conditions. This implies that a numerical iterative procedure must be used to find the maximum for all the parameters simultaneously.

Cigarette demand in US states

Baltagi and Levin (1986, 1992) and Baltagi, Griffin, and Xiong (2000) estimate a dynamic demand model for cigarettes based on a panel from 46 U.S. states. In the previous study, the dataset covers the period 1963–1992. We investigate the following dynamic demand equation:

$$\begin{aligned}\log(C_{it}) &= \alpha + \beta_1 \log(C_{i,t-1}) + \beta_2 \log(P_{it}) + \beta_3 \log(Y_{it}) \\ &\quad + \beta_4 \log(\mathrm{Pn}_{it}) + \mu_i + \lambda_t + \varepsilon_{it}, \qquad (16)\\ i &= 1, \ldots, N(46); \quad t = 1, \ldots, T(29)\end{aligned}$$

where C_{it} is real per capita sales of cigarettes by persons of smoking age (14 years and older). This is measured in packs of cigarettes per capita. P_{it} is the average retail price of a pack of cigarettes measured in real terms. Y_{it} is the real per capita disposable income. Pn_{it} denotes the minimum real price of cigarettes in any neighboring state. This last variable is a proxy for the casual smuggling effect across state borders. It acts as a substitute price attracting consumers from high-tax states to cross over to low-tax states. There are reasons given in Baltagi and Levin (1986, 1992) to assume that the state-specific effects (μ_i) and time-specific effects (λ_t) are fixed, in which case one includes state dummy variables and time dummies for each year in Equation (16).

We have decided to investigate this particular model for four reasons. First, the dataset can be downloaded freely from www.wiley.co.uk/baltagi/. Second, the analysis of cigarette consumption is interesting because of the policy importance of the price elasticity of demand in affecting tax revenues and discouraging consumption. Third, an interesting methodological question is to what degree can elasticity differences be attributable to the manner in which applied econometricians analyze a given body of data. Specifically, this study analyzes to what extent the inclusion of the first observation of each time-series of observations and spatial dependence among the observations matter. Baltagi and Levin (1986, 1992) and Baltagi, Griffin, and Xiong (2000) have investigated the effect of the price level in any neighboring state. Although this variable accommodates the effect of spatial dependence among the observations to a certain degree, we want to investigate whether or not this effect has been completely captured by extending the equation with spatial error autocorrelation.[7] Fourth, the time dimension of the spatial panel gives the opportunity to compare the results of short- and long-panel estimations.

We have seen that in each model an appropriate value of *m* should be chosen in advance. The value of *m* is case-specific. In this case, *m* is set to 63. Although 1963 is the first year in which cigarette demand was observed, it is clear that the process of selling packs of cigarettes started prior to 1963. According to the Encyclopædia Britannica, the cigarette industry developed after 1880 when J. A. Bonsack was granted a U.S. patent for the first cigarette machine. Improvements in cultivation and processing, which lowered the acid content of cigarette tobacco and made it easier to inhale, helped bring a major expansion in cigarette smoking during the first half of the 20th century. During World War I, the prejudice against smoking by women was overcome, and the practice became widespread among women in Europe and the United States in the 1920s. Based on this information, a value of 63 is reasonable. We have found that other values of *m* (43 or 83) do not really alter the results.

The regression equation as formulated in (16) has been extended with a spatially autoregressive process incorporated in the error term (cf. Equation (1)). The spatial weight matrix has been specified as a normalized binary contiguity matrix.

All the econometric results presented in ''spatial error specification'' have been derived under the assumption that the regression equation contains regional fixed effects but not time period fixed effects. If the regression equation, just as the cigarette demand equation, also contains time period fixed effects, the econometric results are still applicable, provided that the variables are taken in deviation from their first-differenced averages over all cross-sectional units within each time period $z_{it} - z_{it-1} - 1/N\Sigma_i(z_{it} - z_{it-1})$. There is one difference. This procedure not only eliminates the time period fixed effects but also the intercept π_0. This implies that π_0 cannot be estimated using the transformed equation, but that it must be recovered afterwards.

Table 1 reports the estimation results based on the complete sample of 1334 observations ($T = 29$). The first row shows the results of the LSDV estimator applied on the regression equation formulated in levels. Recall that this estimator does not

Table 1 Estimation Results of Cigarette Demand Using the Complete Sample ($T = 29$)

Model type	Log($C_{i,t-1}$)	Log(P_{it})	Log(Pn_{it})	Log(Y_{it})	δ
1. LSDV estimator	0.830	−0.292	0.035	0.107	—
Excl. first obs. and spatial error	(65.77)	(12.64) [−1.72]	(1.34) [0.21]	(4.58) [0.63]	
2. Incl. first obs.—BS	0.848	−0.282	0.039	0.103	—
Excl. spatial error	(65.02)	(12.24) [−1.85]	(1.60) [0.26]	(4.53) [0.68]	
3. Incl. first obs.—NB	0.897	−0.173	0.009	0.089	—
Excl. spatial error	(53.43)	(5.03) [−1.69]	(0.31) [0.09]	(2.87) [0.87]	
4. Excl. first obs.	0.797	−0.328	0.046	0.144	0.099
Incl. spatial error	(63.34)	(14.53) [−1.61]	(1.76) [0.23]	(6.10) [0.71]	(2.05)
5. Incl. first obs.—BS	0.868	−0.245	0.046	0.085	0.054
Incl. spatial error	(59.35)	(9.73) [−1.86]	(1.77) [0.35]	(3.41) [0.64]	(1.05)
6. Incl. first obs.—NB	0.835	−0.168	0.008	0.096	0.041
Incl. spatial error	(49.80)	(4.96) [−1.02]	(0.30) [0.05]	(3.30) [0.58]	(0.71)

NOTES: Numbers in parentheses denote *t*-statistics, and numbers in square brackets denote long-run elasticities; results obtained for π, σ^2, and θ are left aside. BS, approximation of first observations according to Bhargava and Sargan; NB, approximation of first observations according to Nerlove and Balestra; LSDV, least-squares dummy variables; Excl., excluding; Incl., including; obs., observations.

utilize the first cross-section of observations and does not account for spatial error autocorrelation. The results obtained can also be found in Baltagi, Griffin, and Xiong (2000, Table 1) and can easily be reproduced using standard econometric software on panel data. As pointed out in the introduction to this article, the estimates of the response parameters in a dynamic panel data model using the LSDV estimator are inconsistent. The next two estimators, which utilize the first cross-section of observations successively according to the BS approximation and the NB approximation (Equations (11) and (15) with $\delta = 0$), shed more light onto the magnitude of the bias. The bias in the autoregressive parameter τ of $log(C_{it-1})$ amounts to 2.1% compared with the BS approximation and 7.5% compared with the NB approximation. The bias is relatively small, because T is relatively large.

Spatial scientists might argue that spatial effects must be included as the data have a locational component. The fourth, fifth, and sixth estimators show what happens when the first three estimators are corrected for spatial error autocorrelation. Remarkably, whereas the spatial autocorrelation coefficient appears to be statistically different from zero when the first cross-section of observations is ignored (fourth estimator), it turns insignificant when the first cross-section of observations is utilized (fifth and sixth estimators). Just as the estimates of the response parameters in a dynamic panel data model using the LSDV estimator are inconsistent, so are the response parameters when the LSDV estimator is corrected for spatial error autocorrelation. The bias in the autoregressive parameter τ of $log(C_{it-1})$ in this case amounts to 8.2% compared with the BS approximation and 4.6% compared with the NB approximation.

The estimation results obtained for $\log(P_{it})$, $\log(\mathrm{Pn}_{it})$, and $\log(Y_{it})$ shown in Table 1 reflect short-term elasticities. Long-term estimated elasticities can be obtained from the short-term estimated elasticities by multiplying the latter by $1/(1-\hat{\tau})$, where $\hat{\tau}$ is the coefficient estimate of lagged consumption (see the numbers in square brackets in Table 1). The long-term own price elasticities of the first five estimators appear to range from -1.61 to -1.80. Only the sixth estimator actually produces a different long-term own price elasticity of -1.02. The long-term neighboring price elasticities range from 0.21 to 0.35 using the LSDV estimator or the second or fourth estimator based on the BS approximation, and from 0.05 to 0.09 using the third or fifth estimator based on the NB approximation. Finally, the long-term income elasticities range from 0.58 to 0.87.

In Table 2, the above analysis is repeated but then for $T = 5$ instead of $T = 29$ to simulate the situation that the researcher has the availability over only a short panel. We have found that the precise subsample period in this respect does not really alter the results. The most striking result is that a short panel causes the coefficient on lagged consumption to decline from 0.83 to 0.39 when using the simple LSDV estimator and from 0.80 to 0.34 when using the LSDV estimator corrected for spatial error autocorrelation. These coefficients are no doubt biased, because they are correlated to the demeaned error terms. When the first cross-section of observations is utilized, we find a lagged-consumption estimate that ranges from 0.54 to 0.78.

Table 2 Estimation Results of Cigarette Demand Using a Sub-Sample ($T = 5$)

Model type	$\text{Log}(C_{i,t-1})$	$\text{Log}(P_{it})$	$\text{Log}(\text{Pn}_{it})$	$\text{Log}(Y_{it})$	δ
1. LSDV estimator	0.388	− 0.529	0.124	0.175	—
Excl. first obs. and spatial error	(7.22)	(9.56) [− 0.86]	(1.81) [0.20]	(2.66) [0.29]	
2. Incl. first obs.—BS	0.611	− 0.331	0.156	0.106	—
Excl. spatial error	(4.94)	(3.50) [− 0.85]	(2.10) [0.40]	(1.37) [0.27]	
3. Incl. first obs.—NB	0.775	− 0.235	0.031	0.140	—
Excl. spatial error	(13.26)	(3.88) [− 1.04]	(0.41) [0.14]	(1.57) [0.62]	
4. Excl. first obs.	0.339	− 0.511	0.160	0.218	− 0.017
Incl. spatial error	(6.95)	(10.48) [− 0.77]	(2.65) [0.24]	(3.73) [0.33]	(0.14)
5. Incl. first obs.—BS	0.585	− 0.393	0.133	0.149	− 0.078
Incl. spatial error	(6.34)	(5.37) [− 0.95]	(1.81) [0.32]	(2.07) [0.36]	(0.55)
6. Incl. first obs.—NB	0.543	− 0.295	0.040	0.159	− 0.071
Incl. spatial error	(9.60)	(5.43) [− 0.65]	(0.54) [0.09]	(1.94) [0.35]	(0.47)

NOTES: Numbers in parentheses denote *t*-statistics, and numbers in square brackets denote long-run elasticities; results obtained for π, σ^2, and θ are left aside. BS, approximation of first observations according to Bhargava and Sargan; NB, approximation of first observations according to Nerlove and Balestra; LSDV, least-squares dummy variables; Excl., excluding; Incl., including; obs., observations.

The bias in the autoregressive parameter τ of $log(C_{it-1})$ in this case of $T = 5$ is much larger and falls within the range of 36.5–42.1%, dependent on the type of approximation and whether spatial error autocorrelation is included. These results corroborate Nerlove's conclusion that the conditional LSDV estimator yields low estimates of the parameter τ, particularly when T is small.

In Table 3, the analysis described above is repeated for $T = 5$ but then for non-consecutive years to reflect that while the researcher has availability only over a short panel it is over a longer time span.[8] It appears that the decline of the coefficient on lagged consumption diminishes considerably. As part of this decline can be attributed to the fact that serial dependence between the observations over non-consecutive years tends to be weaker, it may be concluded that the bias in the coefficient on lagged consumption in the non-consecutive panel is significantly smaller than in the consecutive panel. When the LSDV estimator or the LSDV estimator corrected for spatial error autocorrelation is used, it also leads to a significant improvement of the estimated elasticities of the variables P, Pn, and Y; the short- and the long-term elasticities obtained from the non-consecutive panel (Table 3) are closer to those obtained from a long panel (Table 1) than the short- and the long-term elasticities obtained from the consecutive panel (Table 2).[9] Mixed results occur for estimators utilizing the first cross-section of observations. When spatial error autocorrelation is included, the short-term elasticities obtained from the non-consecutive panel outperform their counterparts obtained from the consecutive panel, whereas the performance of the long-term elasticities of both

Table 3 Estimation Results of Cigarette Demand Using a Sub-Sample of Nonconsecutive Points in Time ($T = 5$)

Model type	$\text{Log}(C_{i,t-1})$	$\text{Log}(P_{it})$	$\text{Log}(\text{Pn}_{it})$	$\text{Log}(Y_{it})$	δ
1. LSDV estimator	0.778	−0.257	0.093	0.233	—
Excl. first obs. and spatial error	(32.73)	(5.75) [−1.16]	(1.74) [0.42]	(4.41) [1.05]	
2. Incl. first obs.—BS	0.793	−0.190	0.070	0.190	—
Excl. spatial error	(29.02)	(3.61) [−0.92]	(1.13) [0.34]	(3.10) [0.92]	
3. Incl. first obs.—NB	0.764	−0.410	−0.117	0.416	—
Excl. spatial error	(13.24)	(4.54) [−0.26]	(0.71) [−0.50]	(3.22) [1.76]	
4. Excl. first obs.	0.789	−0.243	0.090	0.262	−0.081
Incl. spatial error	(33.34)	(5.50) [−1.15]	(1.72) [0.43]	(3.67) [1.24]	(0.64)
5. Incl. first obs.—BS	0.776	−0.236	0.060	0.222	−0.086
Incl. spatial error	(28.30)	(4.51) [−1.05]	(0.97) [0.27]	(3.67) [0.99]	(0.60)
6. Incl. first obs.—NB	0.785	−0.189	−0.004	0.228	−0.029
Incl. spatial error	(21.85)	(3.20) [−0.88]	(0.04) [−0.02]	(2.64) [1.06]	(0.18)

NOTES: Numbers in parentheses denote t-statistics, and numbers in square brackets denote long-run elasticities, results obtained for π, σ^2, and θ are left aside. BS, approximation of first observations according to Bhargava and Sargan; NB, approximation of first observations according to Nerlove and Balestra; LSDV, least-squares dummy variables; Excl., excluding; Incl., including; obs., observations.

types of panels is almost the same. When spatial error autocorrelation is not included, although from a spatial point of view this case should perhaps be left aside, the performance of the elasticities obtained from the non-consecutive panel tends to be worse. Although these results indicate that from a spatial viewpoint, the elasticities obtained from the non-consecutive panel are better than those obtained from the consecutive panel, it does not mean that unconditioning on the first cross-section of observations is no longer useful. The results obtained from a consecutive panel when utilizing the first cross-section of observations are clearly better; the NB approximation outperforms the BS approximation, in terms of the deviation of both the short- and the long-term elasticities in Table 2 with respect to those in Table 1, which in turn outperforms the LSDV estimator and the LSDV estimator corrected for spatial error autocorrelation. Mixed results occur again for the non-consecutive panel. However, when spatial error autocorrelation is included, which from a spatial viewpoint is desirable, the NB approximation also outperforms the other estimators.

Another criterion taken from Baltagi, Griffin, and Xiong (2000) is the forecast properties of the alternative estimators. Table 4 gives the root mean squared error (RMSE) of the predictions obtained by applying the parameter estimates reported in Tables 2 and 3, respectively. Because the ability of an estimator to characterize short- as well as long-term responses is at issue, the RMSE is calculated across the 46 states at a forecast horizon of 1, 5, and 10 years.[10] Three results emerge from

Table 4 Comparison of Forecast Performance Measured by Root Mean Squared Error ($\times 10^2$)

Estimator	1st year	5th year	10th year
(A) Estimators based on sub-sample (T = 5) of consecutive points in time			
1. LSDV estimator	4.30	10.40	12.02
2. Incl. first obs.—BS	3.59	6.13	7.45
3. Incl. first obs.—NB	3.36	4.92	5.35
4. Incl. spatial error	4.36	10.37	11.91
5. Incl. first obs.—BS/spatial error	3.28	4.60	4.70
6. Incl. first obs.—NB/spatial error	3.08	3.94	3.85
(B) Estimators based on sub-sample (T = 5) of nonconsecutive points in time			
1. LSDV estimator	4.39	9.27	11.43
2. Incl. first obs.—BS	3.49	5.79	6.92
3. Incl. first obs.—NB	4.19	9.03	10.79
4. Incl. spatial error	4.30	8.93	10.90
5. Incl. first obs.—BS/spatial error	3.50	5.82	6.67
6. Incl. first obs.—NB/spatial error	3.47	5.73	6.56

NOTES: BS, approximation of first observations according to Bhargava and Sargan; NB, approximation of first observations according to Nerlove and Balestra; LSDV, least-squares dummy variables; Excl., excluding; Incl., including; obs., observations.

Table 4 when comparing the different estimators in panel A (consecutive data) and panel B (non-consecutive data). First, a substantial improvement in the forecast performance occurs when the first cross-section of observations is utilized. The average reduction of the RMSE amounts to 49% in panel A and 27% in panel B. We may therefore draw the conclusion that unconditional estimators are preferred to estimators conditional on the first cross-section of observations, especially when panels are short. Second, additional reduction in the forecast RMSE is obtained by also accounting for spatial error autocorrelation. The average reduction amounts to 28% in panel A and 25% in panel B. Although none of the spatial autocorrelation coefficients reported in Tables 2 and 3 appears to be statistically different from zero, the accounting for spatial error autocorrelation apparently still helps to improve the forecast performance of these models. Third, the forecast performance of estimators utilizing the first cross-section of observations according to the NB approximation is better than that according to the BS approximation. The reduction amounts to 18% in panel A and 2% in panel B. In summary, the best forecast performance for all time horizons is obtained by the estimator accounting for spatial error autocorrelation and utilizing the first cross-section of observations according to the NB approximation. The reduction of the RMSE achieved in a short consecutive panel is greater than in a short non-consecutive panel. It is to be expected that the reduction because of accounting for spatial error autocorrelation will increase and may even exceed the reduction because of utilizing the first cross-section of observations in problems where spatial autocorrelation is more severe.

Conclusion

The possession of spatial panel data and the wish to be able to estimate a dynamic spatial panel data model are now widely recognized. To overcome the inconsistencies associated with the traditional least-squares dummy variables estimator, the models have been transformed into first differences to eliminate the fixed effects and then the unconditional likelihood function has been derived taking into account the density function of the first-differenced observations on each spatial unit. This procedure yields a consistent estimator of the response parameters (τ and β) and the spatial autocorrelation coefficient (δ) when the cross-sectional dimension N tends to infinity (increasing domain asymptotics under a fixed sample design), regardless of the size of T, and provided that the row and column sums of the spatial weight matrix W (before normalizing) do not diverge to infinity at a rate equal to or faster than the rate of the sample size N in the cross-section domain. Only the coefficients of the spatial fixed effects cannot be consistently estimated, because the number of these coefficients increases as N increases. To model the pre-sample values of the exogenous variables for the first-differenced observations on each spatial unit, we have worked out and investigated both the BS approximation and the NB approximation.

From the case study on cigarette demand, it appeared that the need to utilize the first cross-section of observations is to be recommended especially when the time-series dimension of the panel is short. We also found that the NB approximation outperforms the BS approximation. Short- and long-term elasticities obtained from short-panel estimations compared with those obtained from long-panel estimations appeared to be closer, and the RMSE of predictions at a forecast horizon of 1, 5, and 10 years appeared to be smaller. The explanation for these empirical findings is that the NB approximation approaches the (variance of the) unobserved pre-sample values of the exogenous variables by the response parameters β consistent with the derivation given in "The NB Approximation," whereas the BS approximation exploits a new set of parameters π independent of β.

Finally, it should be stressed that the estimators presented in this article might also be used to estimate the parameters of a random effects dynamic panel data model, as they are consistent. Recall that the unconditional likelihood function of the random effects model without exogenous explanatory variables formulated in levels, in contrast to its fixed effects counterpart, does exist. Therefore, one objection to the application of this consistent estimator to the random effects model is that the number of time-series observations on each spatial unit is needlessly reduced by one through first-differencing. Consequently, the estimators presented in this paper when μ would really be random, while consistent, are not as efficient as the ML estimators of the random effects model formulated in levels (instead of first differences) and taking into account the joint density function of the first cross-section of observations also in levels. The derivation of these ML estimators is a subject for further research. These estimators are relevant to space–time problems under a random sample design.

Acknowledgements

The author thanks Bertus Talsma for checking the mathematical formulas in this paper and three anonymous referees for constructive comments.

Notes

[1] Software written in Matlab to estimate the models developed in this article can be downloaded from http://www.regroningen.nl (English, Staff, Elhorst, Software).

[2] In the case where the specification contains a spatially lagged dependent variable, the OLS estimator of the response parameters not only loses the property of being unbiased but it is also inconsistent. The latter is a minimal requirement for a useful estimator.

[3] To limit the correlation of sample observations across different spatial units to a manageable degree, Kelejian and Prucha (1999) assumed that the row and column sums of W are uniformly bounded. Griffith and Lagona (1998) established that the correlation between two spatial units should converge to zero as the distance separating them increases to infinity.

[4] Although these instruments can be obtained from the moment conditions in principle, the number and kind of moment conditions, and therefore the number and kind of instruments involved, are in a state of flux (for an overview, see Baltagi 2001, chap. 8).

[5] In the spatial autoregression literature this term is also known as the ''normalizing constant'' (see Griffith 1996). The Jacobian term takes the form of $T\log|I-\delta W|$, which leads to the condition $1/\omega_{\min}<\delta<1/\omega_{\max}$, where $\omega_{\min}$ and $\omega_{\max}$ denote the minimum and maximum eigenvalue of the spatial weight matrix W.

[6] The joint probability function is $\prod_{i=1}^{N}(2\pi\sigma^2)^{-T/2}|G_{v_b}\otimes B^{-1}B'^{-1}|^{-1/2}\exp(-\frac{1}{2\sigma^2}\Delta e^{*\prime}(G_{v_b}\otimes I_N)^{-1}\Delta e^{*})$. We also have $|G_{v_b}\otimes B^{-1}B'^{-1}|=|G_{v_b}|^N|B^{-1}B'^{-1}|^T$ (Magnus and Neudecker 1988, p. 29), so that $\log|G_{v_b}\otimes B^{-1}B'^{-1}|^{-1/2}=-\frac{1}{2}[N\log|G_{v_b}|-2T\log|B|]=-\frac{N}{2}\log|G_{v_b}|+T\log|B|$.

[7] In Baltagi and Levin (1992), the maximum neighboring price and both the minimum and maximum neighboring prices have also been investigated.

[8] At the request of one of the anonymous referees, as this is a common feature of government census data. We took a time interval of 4 years between the cross-sections.

[9] We calculated the difference between the three elasticities of P, Pn, and Y in both tables and then summed these three figures in absolute values.

[10] Predictions were intercept-adjusted for each state. Additionally, it is assumed that all estimators have zero forecast errors in the last year of the sub-sample.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer.

Anselin, L., and S. Hudak. (1992). ''Spatial Econometrics in Practice; A Review of Software Options.'' *Regional Science and Urban Economics* 22, 509–36.

Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*, 2nd edition. Chichester, UK: Wiley.

Baltagi, B. H., J. M. Griffin, and W. Xiong. (2000). ''To Pool or Not to Pool: Homogeneous Versus Heterogeneous Estimators Applied to Cigarette Demand.'' *The Review of Economics and Statistics* 82, 117–26.

Baltagi, B. H., and D. Levin. (1986). ''Estimating Dynamic Demand for Cigarettes Using Panel Data: The Effects of Bootlegging, Taxation and Advertising Reconsidered.'' *The Review of Economics and Statistics* 48, 148–55.

Baltagi, B. H., and D. Levin. (1992). ''Cigarette Taxation: Raising Revenues and Reducing Consumption.'' *Structural Change and Economic Dynamics* 3, 321–35.

Beck, N. (2001). ''Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years?'' *Annual Review of Political Science* 4, 271–93.

Bhargava, A., and J. D. Sargan. (1983). ''Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods.'' *Econometrica* 51, 1635–59.

Blundell, R., and R. J. Smith. (1991). ''Conditions initiales et estimation efficace dans les modèles dynamiques sur données de panel.'' *Annales d'Économie et de Statistique* 20/21, 109–23.

Buettner, T. (1999). ''The Effect of Unemployment, Aggregate Wages, and Spatial Contiguity on Local Wages: An Investigation with German District Level Data.'' *Papers in Regional Science* 78, 47–67.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. New York: Wiley.

Das, D., H. H. Kelejian, and I. R. Prucha. (2003). ''Finite Sample Properties of Estimators of Spatial Autoregressive Models with Autoregressive Disturbances.'' *Papers in Regional Science* 82, 1–26.

Elhorst, J. P. (2001). ''Dynamic Models in Space and Time.'' *Geographical Analysis* 33, 119–40.

Elhorst, J. P. (2003a). *Unconditional Maximum Likelihood Estimation of Dynamic Models for Spatial Panels*, Research School SOM, Groningen (http://som.rug.nl, 03C27).

Elhorst, J. P. (2003b). ''Specification and Estimation of Spatial Panel Data Models.'' *International Regional Science Review* 26, 244–68.

Elhorst, J. P. (2004). ''Serial and Spatial Dependence in Space–Time Models.'' In *Spatial Econometrics and Spatial Statistics*, edited by A. Getis, J. Mur, and H. G. Zoller. New York: Palgrave.

Griffith, D. A. (1988). *Advanced Spatial Statistics*. Dordrecht, The Netherlands: Kluwer.

Griffith, D. A. (1996). ''Computational Simplifications for Space–Time Forecasting within GIS: The Neighbourhood Spatial Forecasting Model.'' In *Spatial Analysis: Modelling in a GIS Environment*, edited by P. Longley and M. Batty. Cambridge, UK: GeoInformation International.

Griffith, D. A., and F. Lagona. (1998). ''On the Quality of Likelihood-Based Estimators in Spatial Autoregressive Models when the Data Dependence Structure is Misspecified.'' *Journal of Statistical Planning and Inference* 69, 153–74.

Hadinger, H., W. G. Müller, and G. Tondl. (2002). ''Regional Convergence in the European Union (1985–1999): A Spatial Dynamic Panel Analysis.'' *HWWA Discussion Paper 210*, Hamburg, Germany.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hepple, L. W. (1978). ''The Econometric Specification and Estimation of Spatio-Temporal Models.'' In *Time and Regional Dynamics*, edited by T. Carlstein, D. Parkes, and N. Thrift. London: Edward Arnold.

Hoogstrate, A. J. (1998). *Dynamic Panel Data Models: Theory and Macroeconomic Applications*. PhD, University of Maastricht, Maastricht, The Netherlands.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu. (2002). ''Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods.'' *Journal of Econometrics* 109, 107–50.

Johnston, J., and J. DiNardo. (1997). *Econometric Methods*, 4th edition. New York: McGraw-Hill.

Kelejian, H. H., and I. R. Prucha. (1999). ''A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model.'' *International Economic Review* 40, 509–33.

Lahiri, S. N. (2003). ''Central Limit Theorems for Weighted Sums of a Spatial Process under a Class of Stochastic and Fixed Designs.'' *Sankhya* 65, 356–88.

Lee, L.-F. (1981). ''Efficient Estimation of Dynamic Error Component Models with Panel Data.'' In *Time Series Analysis*, edited by O. D. Anderson and M. R. Perryman. Amsterdam: North Holland.

Lee, L.-F. (2002). ''Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models.'' *Econometric Theory* 18, 252–77.

Magnus, J. R., and H. Neudecker. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester, UK: Wiley.

Nerlove, M. (1999). ''Properties of Alternative Estimators of Dynamic Panel Models: An Empirical Analysis of Cross-Country Data for the Study of Economic Growth.'' In *Analysis of Panels and Limited Dependent Variable Models*, edited by C. Hsiao, K. Lahiri, L.-F. Lee, and M. H. Pesaran. Cambridge: Cambridge University Press.

Nerlove, M. (2000). ''Growth Rate Convergence, Fact of Artifact? An Essay on Panel Data Econometrics.'' In *Panel Data Econometrics: Future Directions*, edited by J. Krishnakumar and E. Ronchetti. Amsterdam: Elsevier.

Nerlove, M., and P. Balestra. (1996). ''Formulation and Estimation of Econometric Models for Panel Data.'' In *The Econometrics of Panel Data*, 2nd revised edition, edited by L. Mátyás and P. Sevestre. Dordrecht, The Netherlands: Kluwer.

Nickell, S. (1981). ''Biases in Dynamic Models with Fixed Effects.'' *Econometrica* 49, 1417–26.

Pfeifer, P. E., and S. J. Deutsch. (1980). ''A Three-Stage Iterative Procedure for Space–Time Modeling.'' *Technometrics* 22, 35–47.

Ridder, G., and T. Wansbeek. (1990). ''Dynamic Models for Panel Data.'' In *Advanced Lectures in Quantitative Economics*, edited by F. van der Ploeg. London: Academic Press.

Sevestre, P., and A. Trognon. (1996). ''Dynamic Linear Models.'' In *The Econometrics of Panel Data*, 2nd revised edition, edited by L. Mátyás and P. Sevestre. Dordrecht, The Netherlands: Kluwer.