UNCONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION OF DYNAMIC MODELS FOR SPATIAL PANELS[1]

J. Paul ELHORST

SOM-theme C: Coordination and growth in economies

Abstract

This paper hammers out the estimation of a fixed effects dynamic panel data model extended either to include 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 of both models 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 American states over the period 1963 to 1992.

JEL Classification: C21,C23,R15.

KEYWORDS: Dynamic panels, short panels, spatial error, spatial lag, maximum likelihood, cigarette demand

---

1   INTRODUCTION

In recent years, there has been a growing interest in the estimation of econometric relationships based on panel data. In spatial research, panel data refer to observations made on a number of spatial units over time. In this paper, we focus on dynamic models for spatial panels, a family of models for which according to Elhorst (2001) and Hadinger et al. (2002) no straightforward estimation procedure is yet available. This is (as will be explained below) because existing methods developed for dynamic but non-spatial and for spatial but non-dynamic panel data models might produce biased estimates when these methods/models are put together.

The panel data literature has extensively discussed the dynamic (non-spatial) panel data model (Hsiao, 1986, Ch.4; Baltagi, 2001, Ch.8; Sevestre and Trognon, 1996); a linear regression model extended with a serially lagged dependent variable and a variable intercept μ_i,

$$y_{it} = \tau y_{it-1} + \beta' x_{it} + \mu_i + \varepsilon_{it}, \tag{1}$$

where i (= 1, ..., N) refers to an individual unit, t (= 1, ..., T) to a given time period, y_{it} is the variable to explain, x_{it} is a K×1 vector of exogenous explanatory variables, and ε_{it} are i.i.d. error terms for all i and t with zero mean and variance σ^2. The scalar τ and the K×1 vector β are the response parameters of the model. Furthermore, it is assumed that the initial observations y_{i0} and x_{i0} are observable and that the data are first sorted by time and then by individual unit, i.e., we have T sets of N observations. The properties of μ_i are explained below.

There are a number of reasons why serial lags appear in econometric equations. A household may not change its consumption level and labor supply immediately in response to a change in prices or its income. Similarly, a firm may react with some delay to changes in costs and to changes in demand for its product. Moreover, lags can arise from imperfect information. Economic agents require time to gather relevant information, and this delays the decision-making process. Institutional factors can also result in lags. Households may be contractually obliged to supply a certain level of labor hours, though other conditions would indicate a

1

reduction or increase in labor supply. The reason to consider a spatial dynamic panel data model, instead of non-spatial model, is that in the case that i refers to spatial instead of individual units, spatial dependence can be expected when relative location matters (Bell and Bocksteal, 2000). The main reason that one observation associated with a location depends on observations at other locations is that distance affects household and firm behavior. A similar problem when having panel data on individuals or firms over time is usually not considered. When specifying the spatial dependence between observations, the model may incorporate a spatial autoregressive process in the error term, or the model may contain a spatially autoregressive dependent variable. The first model is known as the *spatial error* model and the second as the *spatial lag* model (for the introduction of these terms, see Anselin and Hudak, 1992). To avoid repetition, we apply to the spatial error specification in the main text, while the spatial lag specification is explained in the appendix.

To describe the spatial arrangement of the spatial units we introduce the matrix W:

Definition 1: The N$\times$N spatial weight matrix W is non-negative with zeros on the diagonal. W has real characteristic roots, which implies that W is symmetric (before row-normalizing). It is assumed that the characteristic roots, denoted by ω_i (i=1,...,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 are (see Griffith, 1988: 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.

Starting with W, the dynamic panel data model extended to include spatial error autocorrelation can be specified as (in stacked form)

2

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

where $Y_t = (Y_{1t},...,Y_{Nt})'$, $X_t = (X_{1t}',...,X_{Nt}')'$, $\mu = (\mu_1,...,\mu_N)'$, $\varphi_t = (\varphi_{1t},...,\varphi_{Nt})'$, $\varepsilon_t = (\varepsilon_{1t},...,\varepsilon_{Nt})'$, and δ is called the spatial autocorrelation coefficient.

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 i.i.d. distributed with zero mean and variance σ_μ^2. It has been argued that the random effects model may not be an appropriate specification in spatial research, because there is typically no natural order for arranging sample data. The spatial units of observation should be representative of a larger population, and the number of units should potentially be able to go to infinity in a regular fashion. When the random effects model is implemented for a given set of irregular spatial units, such as all counties of a state or all regions in a country, the population is sampled exhaustively (Nerlove and Balestra, 1996), and the individual spatial units have characteristics that actually set them apart from a larger population (Anselin, 1988, p. 51). 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 left aside.

The standard estimation method for the fixed effects model is to eliminate the intercepts β_1 and μ_i from the regression equation by demeaning the variables (that is, by taking each variable in the regression equation in deviation from its average over time, $z_{it} - \frac{1}{T}\Sigma_t z_{it}$ for z=y,x), then estimate the resulting demeaned equation by OLS, and subsequently recover the intercepts β_1 and μ_i (Baltagi, 2001, pp. 12–15).[2] This estimator is called the LSDV (least squares dummy variables) estimator. It should be stressed that only the slope coefficients can be estimated consistently, in the

---

[2] It should be noted that only $(\beta_1 + \mu_i)$ are estimable, and not β_1 and μ_i separately, unless a restriction such as $\Sigma_i \mu_i = 0$ is imposed.

case of short panels, where T is fixed and $N \rightarrow \infty$. The coefficients of the fixed effects cannot be estimated consistently, because the number of observations available for the estimation of μ_i is limited to T observations. Fortunately, the inconsistency of μ_i is not transmitted to the estimator of the slope coefficients in the demeaned equation, since this estimator is not a function of the estimated μ_i. This implies that large sample properties ($N \rightarrow \infty$) do apply for the demeaned equation.

Spatial econometric literature shows that OLS estimation is inappropriate for models incorporating spatial error autocorrelation ($\delta \neq 0$). This is important since the LSDV estimator of the fixed effects models falls back on the OLS estimator of the response coefficients in the demeaned equation. In the case of spatial error autocorrelation, the OLS estimator of the response parameters remains unbiased, but it loses the efficiency property.[3] Anselin (1988) suggests overcoming this problem by using maximum likelihood.

The log-likelihood function corresponding to the demeaned equation incorporating spatial error autocorrelation when $\tau = 0$ is

$$-\frac{NT}{2}\ln(2\pi\sigma^2) + T\ln|I_N - \delta W| - \frac{1}{2\sigma^2}\sum_{t=1}^{T} e_t'e_t, \quad e_t = (I - \delta W)[Y_t - \overline{Y} - (X_t - \overline{X})\beta], \quad (3)$$

where $\overline{Y} = (\overline{Y}_{1.}, ..., \overline{Y}_{N.})'$ and $\overline{X} = (\overline{X}_{1.}', ..., \overline{X}_{N.}')'$. An iterative two-stage procedure can be used to maximize this log-likelihood function (Anselin, 1988, pp. 181–182). Alternately, estimate δ by numerical optimization of the concentrated log-likelihood function of δ, and β and σ^2, given δ, by OLS after the data have been transformed according to

$$Y_t^* = (I - \delta W)(Y_t - \overline{Y}) \text{ and } X_t^* = (I - \delta W)(X_t - \overline{X}), \quad (4)$$

until convergence.

---

[3] 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.

4

Lee (2001a,b) proves that the asymptotic properties of the maximum likelihood estimator depend on the characteristic features of the spatial weight matrix. Two types of spatial weight matrices are commonly used in practice: a binary contiguity matrix and an inverse distance matrix. In a binary contiguity matrix, $w_{ij}=1$ is used to indicate that two spatial units are contiguous, whereas $w_{ij}=0$ is used to indicate separation between two spatial units. In an inverse distance matrix all the off-diagonal elements are positive and defined by $1/d_{ij}$, where d_{ij} denotes the distance between two spatial units i and j. According to Lee (2001a), the row and column sums 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. When the spatial weight matrix is a binary contiguity matrix, this condition is automatically satisfied. Normally, no spatial unit is assumed to be a neighbor to more than a given number, say q, of other spatial units. When the spatial weight matrix is an inverse distance matrix, this condition is also satisfied, which can be seen as follows. Consider an infinite number of spatial units that are linearly arranged (to simulate one particular row of the spatial weight matrix). The distance of each spatial unit to its first left and right hand neighbor is 1, to its second left and right hand neighbor the distance is 2, and so on. When the off-diagonal elements of W are of the form $1/d_{ij}$, the row sum of W equals $\Sigma_{i=1}^{N} 2/d_{ij}$, representing a series that is not finite. By contrast, the ratio $\frac{1}{N}\Sigma_{i=1}^{N} 2/d_{ij} \rightarrow 0$ as N goes to infinity. Another condition that must be satisfied, according to Lee (2000b), is that the model contains at least one spatially varying regressor, implying that its coefficient is unequal to zero. The adoption of a dynamic panel data model with a serially lagged dependent variable ($\tau \neq 0$) has the side effect that this condition is automatically satisfied.

A 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, in this case consisting of τ and β, using the transformation derived in (4) is inconsistent if T is fixed, regardless of the size of N (see Hsiao, 1986, Ch.4; Baltagi, 2001, Ch.8). 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. 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 exact likelihood function has shown to exist when applying this procedure to a standard linear regression model without exogenous explanatory variables (Hamilton, 1994; Johnston and Dinardo, 1997, pp.229-230), and on a random effects model without exogenous explanatory variables (Ridder and Wansbeek, 1990; Hoogstrate, 1998; Hsiao et al., 2002). Unfortunately, the exact likelihood function does not exist when applying this procedure on the fixed effects model without exogenous explanatory variables. The reason is that the coefficients of the fixed effects cannot be estimated consistently, since 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: 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.

If exogenous explanatory variables are included, then the exact log-likelihood function of the standard linear regression model and of the random effects model also does not exist. This is because the log-likelihood 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 second procedure first differences the model to eliminate the fixed effects and then applies GMM (generalized method-of-moments) using a set of appropriate instruments.[4] The objection to GMM from a spatial econometric point of view is that this approach tends to overestimate the coefficient δ in case the fixed effects model

---

[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

6

is extended to include spatial error autocorrelation (or a spatially lagged dependent variable). This is because δ is bounded from 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, $T \ln |I - \delta W|$ (see eq.(3)), which the ML approach takes into account but the GMM approach does not (Anselin, 1988: 81-88).

Recently, Hoogstrate (1998) and Hsiao et al. (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. Hsiao et al. (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 T. It is also shown that the ML estimator is asymptotically more efficient that the GMM estimator.

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 paper.[5] Since 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 asymptotics, it follows from Hsiao et al. (2002) and Lee (2001a,b) 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, provided that the row and column sums of the spatial weight matrix W 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. We recall that the coefficients of the fixed effects cannot be estimated consistently, unless the time

---

involved, are in a state of flux (Arrelano, 1989; Arrelano and Bond, 1991; Blundell and Smith, 1991; Ahn and Schmidt, 1995, 1997; Blundell and Bond, 1998; Hahn, 1999).

[5] Dynamics in space and time within a standard linear regression framework ($\mu = 0$) have been discussed recently by Elhorst (2001).

dimension T also goes to infinity. This problem does not necessarily matter when τ and β are the coefficients of interest and μ_i are not, which is the case in many empirical applications.

The remainder of this paper consists of one technical, one empirical, one concluding section and one appendix. In the technical section, we consider the dynamic panel data model extended to include spatial error autocorrelation. The unconditional likelihood function of this model is derived first excluding and then including exogenous explanatory variables. This is done because exogenous explanatory variables further complicate the analysis due to the fact that different approaches have been suggested in the econometric literature to deal with the pre-sample values of these variables in a dynamic context. In the empirical section, a dynamic demand model for cigarettes is estimated based on panel data from 46 American states over the period 1963 to 1992. The concluding section recapitulates our major findings. In the appendix, we derive the unconditional likelihood function of the fixed effects dynamic panel data model extended with a spatially lagged dependent variable.

2   SPATIAL ERROR SPECIFICATION

2.1   NO EXOGENOUS EXPLANATORY VARIABLES

In this section exogenous explanatory variables are omitted. 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 (2), the dynamic panel data model excluding exogenous explanatory variables ($\beta = 0$) extended to include spatial error autocorrelation changes into

$$\Delta Y_t = \tau \Delta Y_{t-1} + B^{-1}\Delta\varepsilon_t, \tag{5}$$

where $B = I_N - \delta W$. ΔY_t is well defined for t=2,…,T, but not for ΔY_1 because ΔY_0 is not observed. To be able to specify the maximum likelihood function of the complete sample ΔY_t (t=1,…,T), the probability function of ΔY_1 must be derived first. Therefore, we repeatedly lag equation (5) by one period. For ΔY_{t-m} (m≥1) we get

$$\Delta Y_{t-m} = \tau \Delta Y_{t-(m+1)} + B^{-1}\Delta\varepsilon_{t-m}. \tag{6}$$

Then, by substitution of ΔY_{t-1} into (5), next ΔY_{t-2} into (5) up to $\Delta Y_{t-(m-1)}$ into (5), we get

$$\Delta Y_t = \tau^m \Delta Y_{t-m} + B^{-1}\Delta\varepsilon_t + \tau B^{-1}\Delta\varepsilon_{t-1} + ... + \tau^{m-1}B^{-1}\Delta\varepsilon_{t-(m-1)} = \tag{7}$$
$$= \tau^m \Delta Y_{t-m} + B^{-1}[\varepsilon_t + (\tau-1)\varepsilon_{t-1} + (\tau-1)\tau\varepsilon_{t-2} + ... + (\tau-1)\tau^{m-2}\varepsilon_{t-(m-1)} - \tau^{m-1}\varepsilon_{t-m}].$$

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

$$E(\Delta Y_t) = \tau^m \Delta Y_{t-m} \text{ and } Var(\Delta Y_t) = \sigma^2 v_b B^{-1}B'^{-1}, \tag{8}$$

where the scalar v_b is defined as

$$v_b = \frac{2}{1+\tau}(1+\tau^{2m-1}). \tag{9}$$

9

Two assumptions with respect to ΔY_1 can be made (cf. Hsiao et al., 2001):

[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 a N×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 assumption [I] is more general than assumption [II]; the second assumption reduces to the first 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 assumption [I].

Writing the residuals of the model as $\Delta e_t = \Delta Y_t - \tau \Delta Y_{t-1}$ for t=2,…,T and, using assumption [I], $\Delta e_1 = \Delta Y_1 - \pi_0 I_N$ for t=1, we have $Var(\Delta e_1) = \sigma^2 v_b B^{-1} B'^{-1}$, $Var(\Delta e_t) = 2\sigma^2 B^{-1} B'^{-1}$ (t=2,…,T), $Covar(\Delta e_t, \Delta e_{t-1}) = -\sigma^2 B^{-1} B'^{-1}$ (t=2,…,T), and zero otherwise. This implies that the covariance matrix of Δe can be written as $Var(\Delta e) = \sigma^2 (G_{v_b} \otimes B^{-1} B'^{-1})$, by which v_b is given in (9) and the T×T matrix $G_v |_{v=v_b}$ is defined as

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

with its subelement in the first row and first column set to v. The determinant of the matrix G_v is (Hsiao et al., 2002) : $|G_v| = 1 - T + T \times v$. *The inverse of G_v is*

10

$$G_v^{-1} = \frac{1}{1-T+T \times v} \times \left[(1-T)G_0^{-1} + v(G_1^{-1} - (1-T)G_0^{-1}) \right]$$

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. The determinant of the matrix* $G_v \otimes I_N$ *is* $|G_v \otimes I_N| = (1-T+T \times v)^N$. *Let p denote a NT×1 vector, which can be partitioned in T block-rows of length N. When* p_t *denotes the* t^{th} *block-row (t=1,...,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]

$$\log L = -\frac{NT}{2} \log(2\pi\sigma^2) + T \log|B| - \frac{N}{2} \log|G_{v_b}| - \frac{1}{2\sigma^2} \Delta e^{*'}(G_{v_b} \otimes I_N)^{-1} \Delta e^*, \qquad (10a)$$

$$\text{where } \Delta e^* = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N) \\ B(\Delta Y_2 - \tau \Delta Y_1) \\ . \\ B(\Delta Y_T - \tau \Delta Y_{T-1}) \end{bmatrix}, \quad E(\Delta e * \Delta e^{*'}) = \sigma^2 (G_{v_b} \otimes I_N). \qquad (10b)$$

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-

---

[6] The joint probability function is

$$\prod_{i=1}^{N} (2\pi\sigma^2)^{-T/2} |G_{v_b} \otimes B^{-1}B'^{-1}|^{-1/2} \exp(-\frac{1}{2\sigma^2} \Delta e^{*'}(G_{v_b} \otimes I_N)^{-1} \Delta e^*).$$ 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

$$\log|G_{v_b} \otimes B^{-1}B'^{-1}|^{-1/2} = -\frac{1}{2}[N \log|G_{v_b}| - 2T \log|B|] = -\frac{N}{2} \log|G_{v_b}| + T \log|B|.$$

order maximizing condition, $\hat{\sigma}^2 = 1/NT\,\Delta e^{*\prime}(G_{v_b} \otimes I_N)^{-1}\Delta e^*$. On substituting $\hat{\sigma}^2$ in the log-likelihood function and using the matrix properties of [W] and [G_v] given in definition 1 and 2, the concentrated log-likelihood function of π_0, τ and δ is obtained as

$$
\begin{aligned}
\text{LogL}_C = C - \frac{NT}{2}\log\left[\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}G_{v_b}^{-1}(t_1,t_2)\Delta e_{t_1}^{*\prime}\Delta e_{t_1}^{*}\right] \\
+ T\sum_{i=1}^{N}\log(1-\delta\omega_i) - \frac{N}{2}\log(1-T+T\times\frac{2}{1+\tau}(1+\tau^{2m-1})),
\end{aligned}
\tag{11}
$$

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

2.2 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 (2) extended to include spatial error autocorrelation as

$$
\begin{aligned}
\Delta Y_t = \tau^m \Delta Y_{t-m} + B^{-1}\Delta\varepsilon_t + \tau B^{-1}\Delta\varepsilon_{t-1} + ... + \tau^{m-1}B^{-1}\Delta\varepsilon_{t-(m-1)} + \sum_{j=0}^{m-1}\tau^j \Delta X_{t-j}\beta = \\
\tau^m \Delta Y_{t-m} + \Delta e_t + X^*.
\end{aligned}
\tag{12}
$$

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] or [II]. By contrast, $\text{Var}(\Delta Y_1)$ is undetermined, since 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 Bhargava and Sargan approximation and the Nerlove and Balestra approximation.

12

2.2.1   THE BHARGAVA AND SARGAN 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 et al. (2002) apply this approximation on the fixed effects model formulated in first differences.

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

$$\Delta Y_1 = \pi_0 1_N + \Delta X_1 \pi_1 + ... + \Delta X_T \pi_T + \Delta e_1, \text{ where } \Delta e_1 = \xi + B^{-1} \sum_{j=0}^{m-1} \tau^j \Delta \varepsilon_{1-j}, \quad (13a)$$

$$E(\Delta e_1) = 0, \quad E(\Delta e_1 \Delta e_2^{'}) = -\sigma^2 B^{-1} B'^{-1}, E(\Delta e_1 \Delta e_t^{'}) = 0 \quad (t = 3,..., T), \quad (13b)$$

$$E(\Delta e_1 \Delta e_1^{'}) = \sigma_\xi^2 I_N + \sigma^2 v_b B^{-1} B'^{-1} \equiv \sigma^2 B^{-1} (\theta^2 BB' + v_b I_N) B'^{-1}. \quad (13c)$$

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

13

Definition 3: $H_V \equiv$
$$\begin{bmatrix} V & -I_N & 0 & . & 0 & 0 \\ -I_N & 2\times I_N & -I_N & . & 0 & 0 \\ 0 & -I_N & 2\times I_N & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 2\times I_N & -I_N \\ 0 & 0 & 0 & . & -I_N & 2\times I_N \end{bmatrix},$$

with its submatrix in the first block-row and first block-column set to the N×N matrix V. The determinant of the matrix H_V is: $|H_V| = |I_N - T \times I_N + T \times V|$. *The inverse of H_V 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 $D = I_N - T \times I_N + T \times V$. *The matrix* H_V^{-1} *can be partitioned in T block-rows and T block columns, by which the submatrix* $H_V^{-1}(t_1, t_2)$ *(t$_1$,t$_2$=1,…,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 the matrix properties of [W] and [H_V] given in definition 1 and 3, the log-likelihood function is obtained as

$$\log L = -\frac{NT}{2}\log(2\pi\sigma^2) + T\sum_{i=1}^{N}\log(1-\delta\omega_i)$$
$$-\frac{1}{2}\sum_{i=1}^{N}\log(1-T+T\times\frac{2}{1+\tau}(1+\tau^{2m-1}) + T\theta^2(1-\delta\omega_i)^2)$$
$$-\frac{1}{2\sigma^2}\Delta e^{*'}H_{V_{BS}}^{-1}\Delta e^*, \tag{14a}$$

14

$$\text{where } \Delta e* = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N - \Delta X_1 \pi_1 - ... - \Delta X_T \pi_T) \\ B(\Delta Y_2 - \tau \Delta Y_1 - \Delta X_2 \beta) \\ . \\ B(\Delta Y_T - \tau \Delta Y_{T-1} - \Delta X_T \beta) \end{bmatrix}, \; E(\Delta e * \Delta e*') = \sigma^2 H_{V_{BS}}. \;(14b)$$

This log-likelihood function is well-defined, satisfies the usual regularity conditions and contains KT+K+5 unknown parameters to be estimated: $\pi_1,...,\pi_T, \beta, \pi_0, \theta^2, \tau, \delta$ 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

$$\hat{\sigma}^2 = \frac{\Delta e*' H_{V_{BS}}^{-1} \Delta e *}{NT}, \text{ and } \begin{bmatrix} \hat{\pi}_0 \\ \hat{\pi}_1 \\ . \\ \hat{\pi}_T \\ \hat{\beta} \end{bmatrix} = (\tilde{X}' H_{V_{BS}}^{-1} \tilde{X})^{-1} \tilde{X}' H_{V_{BS}}^{-1} \tilde{Y}, \quad (15a)$$

$$\text{where } \tilde{X} = \begin{bmatrix} B & B\Delta X_1 & . & B\Delta X_T & 0 \\ 0 & 0 & . & 0 & B\Delta X_2 \\ . & . & . & . & . \\ 0 & 0 & . & 0 & B\Delta X_T \end{bmatrix} \text{ and } \tilde{Y} = \begin{bmatrix} B\Delta Y_1 \\ B(\Delta Y_2 - \tau \Delta Y_1) \\ . \\ B(\Delta Y_T - \tau \Delta Y_{T-1}) \end{bmatrix}. \quad (15b)$$

On substituting $\hat{\sigma}^2, \hat{\pi}$ and $\hat{\beta}$ in 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.

2.2.2   THE NERLOVE AND BALESTRA APPROXIMATION

Starting with a regression equation formulated in levels, Nerlove and Balestra (1996) and Nerlove (1999 or 2000) suggest replacing the unknown variance of X*, $Var(X*) = \Sigma_{j=0}^{m-1} X_{t-j} \beta$, by $\beta' \Sigma_X \beta$, where Σ_X denotes the covariance matrix of the

15

explanatory variables X, which may be determined from the sample data in advance. Suppose that each explanatory variable X_{tk} (k=1,...,K) follows a well-specified common stationary time series model

$$X_{tk} = \tau_{X_k} X_{t-1k} + \gamma_t, \quad \text{where} \quad \gamma_t \sim N(0, \sigma_{\gamma X_k}^2 I_N). \tag{16}$$

Then the random variable X* in (12) has a well-defined variance Σ_{X^*}, which is a function of β and $\tau_{X_k}, \sigma_{\gamma X_k}^2$ (k=1,...,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 2000) have pointed out that it is not necessary to go that far. Since we are not really interested in the parameters τ_{X_k} and $\sigma_{\gamma X_k}^2$ (k=1,...,K), we can suppress these parameters and restrict the log-likelihood to the remaining parameters. While omitting estimation of τ_{X_k} and $\sigma_{\gamma X_k}^2$ (k=1,...,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.

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

$$Var(\Delta Y_1) = Var(\Delta e_1) + Var(X^*) = \sigma^2 v_b B^{-1} B'^{-1} + (\frac{1-\tau^m}{1-\tau})^2 \beta' \Sigma_{\Delta X} \beta \times I_N$$
$$\equiv \sigma^2 B^{-1} (v_b I_N + (\frac{1-\tau^m}{1-\tau})^2 \frac{\beta' \Sigma_{\Delta X} \beta}{\sigma^2} \times BB') B'^{-1}. \tag{17}$$

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

16

in definition 3. Using the matrix properties of [W] and [H$_V$] given in definition 1 and 3, the log-likelihood function is obtained as

$$\log L = -\frac{NT}{2}\log(2\pi\sigma^2) + T\sum_{i=1}^{N}\log(1-\delta\omega_i) - \frac{1}{2\sigma^2}\Delta e^{*'}H_{V_{NB}}^{-1}\Delta e *$$
$$-\frac{1}{2}\sum_{i=1}^{N}\log(1-T+T\times\frac{2}{1+\tau}(1+\tau^{2m-1})+T(\frac{1-\tau^m}{1-\tau})^2\frac{\beta'\Sigma_{\Delta X}\beta}{\sigma^2}(1-\delta\omega_i)^2),$$

(18a)

where $\Delta e^* = \begin{bmatrix} B(\Delta Y_1 - \pi_0 1_N) \\ B(\Delta Y_2 - \tau\Delta Y_1 - \Delta X_2\beta) \\ . \\ B(\Delta Y_T - \tau\Delta Y_{T-1} - \Delta X_T\beta) \end{bmatrix}$, $E(\Delta e * \Delta e^{*'}) = \sigma^2 H_{V_{NB}}.$ (18b)

This log-likelihood function is well-defined, satisfies the usual regularity conditions and contains K+4 unknown parameters to be estimated: $\beta, \pi_0, \tau, \delta$ 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.

3   CIGARETTE DEMAND IN AMERICAN STATES

Baltagi and Levin (1986, 1992) and Baltagi et al. (2000) estimate a dynamic demand model for cigarettes based on a panel from 46 American states. In Baltagi et al. (2000), the dataset covers the period 1963-1992. We investigate the following dynamic demand equation

$$\ln C_{it} = \alpha + \beta_1 \ln C_{i,t-1} + \beta_2 \ln P_{it} + \beta_3 \ln Y_{it} + \beta_4 \ln Pn_{it} + \mu_i + \lambda_t + \varepsilon_{it},$$
$$i = 1,..., N\,(46\,); t = 1,..., T\,(29), \tag{19}$$

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 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 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 (19).

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 analyses 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 et al. (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

18

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. 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 U.S. in the 1920s. Based on this information, m is set to 63. As $N \rightarrow \infty$ is believed to be the most relevant asymptotics and m and T are fixed, it is not necessary to assume $|\tau B| < 1$ in the estimations. In spite of this, this restriction always appeared to be satisfied.

The spatial weight matrix has been specified as a binary contiguity matrix; its elements are posited as being 1 if two states share a common border and 0 otherwise. The elements of this spatial weight matrix have then been divided by its largest characteristic root, with the effect that the largest characteristic root of this normalized matrix equals 1 and the smallest characteristic root lies between -1 and 0. Note that this normalization makes no difference from a mathematical viewpoint, but only from an interpretative viewpoint; it has the effect that δ will not be greater than 1.

All the econometric results presented in section 2 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 each variable in the regression equation is taken in deviation from its average over all regions within each time period. This can be explained as follows.

---

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

The standard method for estimating the fixed effects model, in most textbooks spelled out for fixed effects in the cross-sectional domain, is to eliminate the fixed effects by taking each variable in deviation from its average over time ($z_{it} - \frac{1}{T}\Sigma_t z_{it}$) and then to estimate the resulting equation. In the case of time period fixed effects, this implies that each variable should be taken in deviation from its average over all cross-sectional units within each time period ($z_{it} - \frac{1}{N}\Sigma_i z_{it}$). First-differencing a regression equation formulated in levels to eliminate the fixed effects in the cross-sectional domain, does not eliminate the time period fixed effects, but the structure of these first-differenced time period fixed effects is such that common time dummies can replace them. In summary, when the regression equation formulated in levels also contains time period fixed effects, the variables in the first-differenced regression equation should be taken in deviation from their first-differenced averages over all cross-sectional units within each time period $z_{it} - z_{it-1} - \frac{1}{N}\Sigma_i(z_{it} - z_{it-1})$. 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 utilize the first cross-section of observations and does not account for spatial error autocorrelation. The results obtained can also be found in Baltagi et al. (2000, table 1) and can easily be reproduced using standard econometric software on panel data. As pointed out in the introduction to this paper, 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 Bhargava and Sargan (BS) approximation and the Nerlove and Balestra (NB) approximation (eq.(14) and eq.(18) with $\delta = 0$), throw more light onto the magnitude of the bias. The bias in the response parameters of $\ln P_{it}$, $\ln Pn_{it}$ and $\ln Y_{it}$ amounts to 3.4, 11.4 and 3.7 percent compared to the BS approximation and 40.8, 74.3 and 16.8 percent compared to the NB approximation.

Spatial scientists might argue that spatial effects must be included since the data has 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 estimator). 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 response parameters of lnP_{it}, $lnPn_{it}$ and lnY_{it} in this case amounts to 25.3, 0.0 and 41.0 percent compared to the BS approximation and 48.8, 82.6 and 33.3 percent compared to the NB approximation.

The estimation results obtained for lnP_{it}, $lnPn_{it}$ and lnY_{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 really 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 sub-sample 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

21

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.

When comparing the NB and the BS approximation, we see one notable difference. The short-term as well as the long-term elasticities in table 2 and table 3 tend to be closer to each other when using the NB approximation.

Another criterion taken from Baltagi et al. (2000) is the forecast properties of the alternative estimators. Table 3 gives the root mean squared error (RSME) of the predictions obtained by applying the parameter estimates reported in table 2. Because the ability of an estimator to characterize short-term as well as long-term responses is at issue, the RSME is calculated across the 46 states at a forecast horizon of one year, five years and ten years.[8] Three results emerge from table 3. First, a substantial improvement in the forecast performance occurs when the first cross-section of observations is utilized. The average reduction of the RSME amounts to almost 50%. 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 RSME is obtained by also accounting for spatial error autocorrelaton.[9] The average reduction amounts to almost 25%. Although none of the spatial autocorrelation coefficients reported in table 2 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 average reduction amounts to 18%. 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.

---

[8] 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.
[9] Correcting the LSDV estimator for spatial error autocorrelation does not appear to be of (much) help.

4   CONCLUSIONS

The possession of spatial panel data and the wish to be able to estimate a dynamic spatial panel data models is now widely recognized. To overcome the inconsistencies associated with the traditional least squares dummy 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, regardless of the size of T, and provided that the row and column sums of the spatial weight matrix W 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 fixed effects cannot be consistently estimated, since 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 Bhargava and Sargan approximation and the Nerlove and Balestra 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 Nerlove and Balestra approximation outperforms the Bhargava and Sargan approximation. Short-term and long-term elasticities obtained from short panel estimations compared to those obtained from long panel estimations appeared to be closer, and the root mean squared error of predictions at a forecast horizon of one year, five years and ten 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 below equation (12), whereas the BS approximation exploits a new set of parameters π independent of β.

23

In the case study on cigarette demand, it has also been found that the spatial autocorrelation coefficient is not statistically different from zero using estimators that utilize the first-cross section of observations. The fact that the cigarette demand model contains the minimum real price of cigarettes in any neighboring state as one of the explanatory variables is apparently sufficient to accommodate the effect of spatial dependence among the observations. The lesson of this finding is that adding explanatory variables, which reflect the market conditions in neighboring regions, is in some cases more promising than to include spatial error autocorrelation. On the other hand, when the model is also used for forecasting purposes, accounting for spatial error autocorrelation is to be recommended even when the spatial autocorrelation coefficient will not be statistically different from zero. The reason is that the RSME of the predictions under these circumstances may be lower.

Finally, it should be noted that the estimators presented in this paper might also be used to estimate the parameters of a random effects dynamic panel data model, as they are consistent. One objection is that the number of time series observations on each spatial unit is 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.

REFERENCES

Ahn S., Schmidt P. (1995) Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68: 5-27.

Ahn S., Schmidt P. (1997) Efficient estimation of dynamic panel data models: alternative assumptions and simplified estimation. *Journal of Econometrics* 76: 309-321.

Anselin L. (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht.

Anselin L., Hudak S. (1992) Spatial econometrics in practice; a review of software options. *Regional Science and Urban Economics* 22: 509-536.

Arellano M. (1989) A note on the Anderson-Hsiao estimator for panel data. *Economics Letters* 31: 337-341.

Arellano M., Bond S. (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277-297.

Baltagi B.H. (2001) Econometric analysis of panel data, second edition. Wiley, Chichester, England.

Baltagi B.H., Levin D. (1986) Estimating dynamic demand for cigarettes using panel data: The effects of bootlegging, taxation and advertising reconsidered. *The Review of Economics and Statistics* 48: 148-155.

Baltagi B.H., Levin D. (1992) Cigarette taxation: Raising revenues and reducing consumption. *Structural Change and Economic Dynamics* 3: 321-335.

Baltagi B.H., Griffin J.M., Xiong W. (2000) To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand. *The Review of Economics and Statistics* 82: 117-126.

Bell K.P., Bockstael N.E. (2000) Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *The Review of Economics and Statistics* 82: 72-82.

Bhargava A., Sargan J.D. (1983) Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51: 1635-1659

Blundell R., Bond S. (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115-143.

Blundell R., Smith R.J. (1991) Conditions initiales et estimation efficace dans les modèles dynamiques sur données de panel. *Annales d'Économie et de Statistique* 20/21: 109-123

Davidson J. (2000) Econometric theory. Blackwell, Oxford.

Elhorst J.P. (2001) Dynamic models in space and time. *Geographical Analysis* 33: 119-140.

Griffith D.A. (1988) Advanced spatial statistics. Kluwer, Dordrecht.

Hadinger H., Müller W.G., Tondl G. (2002) Regional convergence in the European Unio (1985-1999): A spatial dynamic panel analysis. HWWA discussion paper 210, Hamburg.

Hahn J. (1999) How informative is the initial condition in the dynamic panel data model with fixed effects? *Journal of Econometrics* 93: 309-326.

Hamilton J.D. (1994) Time series analysis. Princeton University Press, New Jersey.

Hoogstrate A.J. (1998) Dynamic panel data models: Theory and macroeconomic applications. PhD University of Maastricht, the Netherlands.

Hsiao C. (1986) Analysis of panel data. Cambridge University Press, Cambridge.

Hsiao C., Pesaran M.H., Tahmiscioglu A.K. (2002) Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109: 107-150.

Johnston J., DiNardo J. (1997) Econometric methods, fourth edition. McGraw-Hill, New York.

Lee L.-F. (1981) Efficient estimation of dynamic error component models with panel data. In: Anderson O.D., Perryman M.R. (eds.) Time series analysis. North Holland, Amsterdam.

Lee, L.-F. (2001a) Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models I: Spatial autoregressive processes. Website Ohio State University, Ohio.

Lee, L.-F. (2001b) Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models II: Mixed regressive, spatial autoregressive processes. Website Ohio State University, Ohio.

Magnus J.R., Neudecker H. (1988) Matrix differential calculus with applications in statistics and econometrics. John Wiley, Chichester.

Nickell S. (1981) Biases in dynamic models with fixed effects. *Econometrica* 49: 1417-1426.

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: Hsiao C., Lahiri K., Lee L.-F., Pesaran M.H. (eds.) Analysis of Panels and Limited Dependent Variable Models. Cambridge University Press, Cambridge.

Nerlove M. (2000) Growth rate convergence, fact of artifact? An essay on panel data econometrics. In: Krishnakumar J., Ronchetti E. (eds.) Panel data econometrics: Future directions, pp.3- 33. Elsevier, Amsterdam.

Nerlove M., Balestra P. (1996) Formulation and estimation of econometric models for panel data. In: Mátyás L., Sevestre P. (eds.) The econometrics of panel data, second revised edition, pp. 3-22. Kluwer, Dordrecht.

Ridder G., Wansbeek T. (1990) Dynamic models for panel data. In: F. van der Ploeg (ed.) Advanced lectures in quantitative economics. Academic Press, London.

Sevestre P., Trognon A. (1996) Dynamic linear models. In: Mátyás L., Sevestre P. (eds.) The econometrics of panel data, second revised edition, pp.120-144. Kluwer, Dordrecht.

APPENDIX   SPATIAL LAG SPECIFICATION

The dynamic panel data model extended with a spatially lagged dependent variable reads as

$$Y_t = \tau Y_{t-1} + \delta W Y_t + X_t \beta + \mu + \varepsilon_t, \quad E\varepsilon_t = 0, \quad E\varepsilon_t \varepsilon_t' = \sigma^2 I_N. \tag{A1}$$

First, the exact log-likelihood function for the model excluding exogenous explanatory variables is determined ($\beta = 0$). Taking first differences of (A1), the model changes into

$$\Delta Y_t = \tau \Delta Y_{t-1} + \delta W \Delta Y_t + \Delta \varepsilon_t. \tag{A2}$$

ΔY_t is well defined for t=2,…,T, but not for ΔY_1 because ΔY_0 is not observed. To be able to specify the maximum likelihood function of the complete sample ΔY_t (t=1,…,T), the probability function of ΔY_1 must be derived first. By continuous substitution, we can rewrite (A2) as

$$\begin{aligned} B\Delta Y_t &= \tau^m B^{-(m-1)} \Delta Y_{t-m} + \Delta \varepsilon_t + \tau B^{-1} \Delta \varepsilon_{t-1} + ... + \tau^{m-1} B^{-(m-1)} \Delta \varepsilon_{t-(m-1)} = \\ &\quad \tau^m B^{-(m-1)} \Delta Y_{t-m} + \varepsilon_t + A\varepsilon_{t-1} + A\tau B^{-1}\varepsilon_{t-2} + ... + \\ &\quad A\tau^{m-2} B^{-(m-2)} \varepsilon_{t-(m-1)} - \tau^{m-1} B^{-(m-1)} \varepsilon_{t-m}, \end{aligned} \tag{A3}$$

where $A = \tau B^{-1} - I_N$ and $B = I_N - \delta W$. Since $E(\varepsilon_t) = 0$ (t=1,…T) and the successive values of ε_t are uncorrelated, we have

$$E(B\Delta Y_t) = \tau^m B^{-(m-1)} \Delta Y_{t-m} \text{ and } Var(B\Delta Y_t) = \sigma^2 V_b, \tag{A4}$$

where the N×N matrix V_b is defined as

$$\begin{aligned} V_b &= I_N + A(I_N - \tau^2 (B'B)^{-1})^{-1} A' - \\ &\quad A\tau^{m-1} B^{-(m-1)} (I_N - \tau^2 (B'B)^{-1})^{-1} \tau^{m-1} B'^{-(m-1)} A' + \\ &\quad \tau^{m-1} B^{-(m-1)} \tau^{m-1} B'^{-(m-1)}. \end{aligned} \tag{A5}$$

When the matrix W is symmetric, V_b reduces to

$$V_b = 2*(I_N + \tau B^{-1})^{-1}(I_N + (\tau B^{-1})^{2m-1}). \tag{A6}$$

Just as in the spatial error model, we assume that the process has started in the past not too far back from the 0th period and that the expected *changes* in the initial endowments are the same across all spatial units. Under this assumption, $E(B\Delta Y_1) = \pi_0 1_N$, where 1_N denotes a N×1 vector of unit elements and π_0 is a fixed but unknown parameter to be estimated.

Writing the residuals of the model as $\Delta e_1 = B\Delta Y_1 - \pi_0 I_N$ for t=1 and $\Delta e_t = \Delta Y_t - \tau \Delta Y_{t-1} - \delta W\Delta Y_t = B\Delta Y_t - \tau \Delta Y_{t-1}$ for t=2,…,T, we have

$Var(\Delta e_1) = \sigma^2 B^{-1} V_b B'^{-1}$, $Var(\Delta e_t) = 2\sigma^2 B^{-1} B'^{-1}$ (t=2,…,T),

$Covar(\Delta e_t, \Delta e_{t-1}) = -\sigma^2 B^{-1} B'^{-1}$ (t=2,…,T), and zero otherwise. This implies that the covariance matrix of Δe can be written as $Var(\Delta e) = \sigma^2 [(I_T \otimes B^{-1})H_{V_b}(I_T \otimes B'^{-1})]$, by which the matrix $H_v|_{V=V_b}$ is given in definition 3. In sum, we have

$$\log L = -\frac{NT}{2}\log(2\pi\sigma^2) + T\log|B| - \frac{1}{2}\log|H_{V_b}| - \frac{1}{2\sigma^2}\Delta e'H_{V_b}^{-1}\Delta e, \tag{A7}$$

where $\Delta e = \begin{bmatrix} B\Delta Y_1 - \pi_0 1_N \\ B\Delta Y_2 - \tau\Delta Y_1 \\ . \\ B\Delta Y_T - \tau\Delta Y_{T-1} \end{bmatrix}$, $\quad E(\Delta e\Delta e') = \sigma^2 H_{V_b}$. $\tag{A8}$

σ^2 can be solved from its first-order maximizing condition, $\hat{\sigma}^2 = 1/NT\,\Delta e'H_{V_b}^{-1}\Delta e$. On substituting $\hat{\sigma}^2$ in the log-likelihood function and using matrix properties of [W] and [H$_V$] given in definition 1 and 3, the concentrated log-likelihood function of π_0, τ and δ is obtained as

29

$$\text{LogL}_C = C - \frac{NT}{2} \log\left[\Delta e' H_{V_b}^{-1} \Delta e\right] + T\sum_{i=1}^{N} \log(1 - \delta\omega_i) -$$

$$\frac{1}{2}\sum_{i=1}^{N} \log\left[1 - T + \frac{2T(1-\delta\omega_i)}{(1-\delta\omega_i + \tau)}(1 + (\frac{\tau}{1-\delta\omega_i})^{2m-1})\right]. \tag{A9}$$

where C is a constant ($C = -NT/2\,(1 + \log 2\pi)$).

By taking first differences and continuous substitution, the dynamic panel data model including exogenous explanatory variables and extended with a spatially lagged dependent variable can be rewritten as

$$B\Delta Y_t = \tau^m B^{-(m-1)}\Delta Y_{t-m} + \Delta\varepsilon_t + \tau B^{-1}\Delta\varepsilon_{t-1} + ... + \tau^{m-1} B^{-(m-1)}\Delta\varepsilon_{t-(m-1)} + \sum_{j=0}^{m-1}\tau^j B^{-j}\Delta X_{t-j}\beta$$

$$= \tau^m B^{-(m-1)}\Delta Y_{t-m} + \Delta e_t + X^*. \tag{A10}$$

As X_t is stationary, we have $E\Delta X_t = 0$ and thus $E(B\Delta Y_1) = \tau^m B^{-(m-1)}\Delta Y_{t-m}$. $\text{Var}(B\Delta Y_1)$ is undetermined, since X^* is not observed. We use the Bhargava and Sargan approximation as well as the Nerlove and Balestra approximation below to approach the probability function of $B\Delta Y_1$.

The optimal predictor of X^* when t=1 according to the Bhargava and Sargan approximation is $\pi_0 1_N + \Delta X_1\pi_1 + ... + \Delta X_T\pi_T + \xi$, where $\xi \sim N(0, \sigma_\xi^2 I_N)$. See section 2.2.1 for potential restrictions on the parameters π. This implies that

$$B\Delta Y_1 = \pi_0 1_N + \Delta X_1\pi_1 + ... + \Delta X_T\pi_T + \Delta e_1, \quad \text{where} \quad \Delta e_1 = \xi + \sum_{j=0}^{m-1}\tau^j B^{-j}\Delta\varepsilon_{1-j}. \tag{A11a}$$

$$E(\Delta e_1) = 0, \quad E(\Delta e_1\Delta e_2') = -\sigma^2 I_N, E(\Delta e_1\Delta e_t') = 0 \quad (t = 3,..., T), \tag{A11b}$$

$$E(\Delta e\Delta e_1') = \sigma_\xi^2 I_N + \sigma^2 V_b \equiv \sigma^2(\theta^2 I_N + V_b), \tag{A11c}$$

$$\theta^2 = \sigma_\xi^2 / \sigma^2. \tag{A11d}$$

30

Let $V_{BS} = \theta^2 I_N + V_b$ with V_b specified as in (A4) or (A5), then the covariance matrix of Δe can be written as $Var(\Delta e) = \sigma^2[(I_T \otimes B^{-1})H_{V_{BS}}(I_T \otimes B'^{-1})]$, by which the matrix $H_V|_{V=V_{BS}}$ is given in definition 3. Using the matrix properties of [W] and [H_V] given in definition 1 and 3, the log-likelihood function is obtained as

$$logL = -\frac{NT}{2}log(2\pi\sigma^2) + T\sum_{i=1}^{N}log(1-\delta\omega_i) - \frac{1}{2}\sum_{i=1}^{N}log(1 - T + \frac{2T(1-\delta\omega_i)}{(1-\delta\omega_i+\tau)}(1+(\frac{\tau}{1-\delta\omega_i})^{2m-1} + T\theta^2)$$

$$-\frac{1}{2\sigma^2}\Delta e' H_{V_{BS}}^{-1}\Delta e, \qquad (A12a)$$

where $\Delta e = \begin{bmatrix} B\Delta Y_1 - \pi_0 1_N - \Delta X_1 \pi_1 - ... - \Delta X_T \pi_T \\ B\Delta Y_2 - \tau\Delta Y_1 - \Delta X_2 \beta \\ . \\ B\Delta Y_T - \tau\Delta Y_{T-1} - \Delta X_T \beta \end{bmatrix}$, $E(\Delta e\Delta e') = \sigma^2 H_{V_{BS}}$. (A12b)

This log-likelihood function contains KT+K+5 unknown parameters to be estimated: $\pi_1,...,\pi_T, \beta, \pi_0, \theta^2, \tau, \delta$ and σ^2. σ^2, π and β can be solved from their first-order maximizing conditions

$$\hat{\sigma}^2 = \frac{\Delta e' H_{V_{BS}}^{-1}\Delta e}{NT} \text{ and } \begin{bmatrix} \hat{\pi}_0 \\ \hat{\pi}_1 \\ . \\ \hat{\pi}_t \\ \hat{\beta} \end{bmatrix} = (\tilde{X}' H_{V_{BS}}^{-1}\tilde{X})^{-1}\tilde{X}' H_{V_{BS}}^{-1}\tilde{Y}, \qquad (A13a)$$

where $\tilde{X} = \begin{bmatrix} 1_N & \Delta X_1 & . & \Delta X_T & 0 \\ 0 & 0 & . & 0 & \Delta X_2 \\ . & . & . & . & . \\ 0 & 0 & . & 0 & \Delta X_T \end{bmatrix}$ and $\tilde{Y} = \begin{bmatrix} B\Delta Y_1 \\ B\Delta Y_2 - \tau\Delta Y_1 \\ . \\ B\Delta Y_T - \tau\Delta Y_{T-1} \end{bmatrix}$. (A13b)

31

On substituting $\hat{\sigma}^2, \hat{\pi}$ and $\hat{\beta}$ in the log-likelihood function, the concentrated log-likelihood function of θ^2, τ and δ is obtained.

According to the Nerlove and Balestra approximation, $\text{Var}(B\Delta Y_1)$ might be approached by

$$\text{Var}(B\Delta Y_1) = \text{Var}(\Delta e_1) + \text{Var}(X^*) = \sigma^2 (V_b + \frac{1}{\sigma^2} \Sigma_{X^*}), \tag{A14a}$$

where $\quad \Sigma_{X^*} = (I_N - \tau B^{-1})^{-1}(I_N - \tau^m B^{-m})\beta' \Sigma_{\Delta X} \beta (I_N - \tau^m B'^{-m})(I_N - \tau B'^{-1})^{-1}.$
$$\tag{A14b}$$

When the matrix W is symmetric, Σ_{X^*} reduces to

$$\Sigma_{X^*} = \beta' \Sigma_{\Delta X} \beta (I_N - \tau B^{-1})^{-2} (I_N - (\tau B^{-1})^m)^2. \tag{A15}$$

Let $V_{NB} = V_b + 1/\sigma^2 \Sigma_{X^*}$ with V_b specified as in (A4) or (A5), then the covariance matrix of Δe can be written as $\text{Var}(\Delta e) = \sigma^2[(I_T \otimes B^{-1})H_{V_{NB}}(I_T \otimes B'^{-1})]$, by which the matrix $H_V|_{V=V_{NB}}$ is given in definition 3. Using matrix properties of [W] and [H_V] given in definition 1 and 3, the log-likelihood function is obtained as

$$\text{LogL} = -\frac{NT}{2} \log[2\pi\sigma^2] + T\sum_{i=1}^{N} \log(1 - \delta\omega_i) - \frac{1}{2\sigma^2} \Delta e' H_{V_{NB}}^{-1} \Delta e$$

$$-\frac{1}{2}\sum_{i=1}^{N} \log \left[1 - T + \frac{2T(1-\delta\omega_i)}{(1-\delta\omega_i+\tau)}(1 + (\frac{\tau}{1-\delta\omega_i})^{2m-1}) + \right. \tag{A16a}$$

$$\left. T\frac{\beta'\Sigma\beta}{\sigma^2}\frac{(1-\delta\omega_i-\tau)^2}{(1-\delta\omega_i)^2}(1 - (\frac{\tau}{1-\delta\omega_i})^m)^2 \right]$$

where $\Delta e = \begin{bmatrix} B\Delta Y_1 - \pi_0 1_N \\ B\Delta Y_2 - \tau\Delta Y_1 - \Delta X_2 \beta \\ . \\ B\Delta Y_T - \tau\Delta Y_{T-1} - \Delta X_T \beta \end{bmatrix}, \quad E(\Delta e \Delta e') = \sigma^2 H_{V_{NB}}. \tag{A16b}$

32

This log-likelihood function contains K+4 unknown parameters to be estimated: $\beta, \pi_0, \tau, \delta$ and σ^2. None of these parameters can be solved analytically from the first-order maximizing conditions.

Table 1 Estimation results cigarette demand using the complete sample (T=29)

Model type	LnC$_{i,t-1}$	lnP$_{it}$	LnPn$_{it}$	lnY$_{it}$	δ
1. LSDV estimator	0.830	-0.292	0.035	0.107	-
Excl. first obs. & 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)

Numbers in parentheses denote t-statistics, 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

Table 2 Estimation results cigarette demand using a sub-sample (T=5)

Model type	$LnC_{i,t-1}$	lnP_{it}	$LnPn_{it}$	lnY_{it}	δ
1. LSDV estimator	0.388	-0.529	0.124	0.175	-
Excl. first obs. & 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)

Numbers in parentheses denote t-statistics, numbers in square brackets denote long-run elasticities,

results obtained for π, σ² and θ are left aside,

BS=Approximation of first observations according to Bhargava and Sargan

NB=Approximation of first observations acording to Nerlove and Balestra

35

Table 3 Comparison of forecast performance measured by root mean squared error ($\times 10^2$)

Estimator	1st year	5th year	10th year
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	5.42	9.84	13.41
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