Manfred M. Fischer and Daniel A. Griffith

# Modelling spatial autocorrelation in spatial interaction data

Paper

# MODELLING SPATIAL AUTOCORRELATION IN SPATIAL INTERACTION DATA*

**Manfred M. Fischer**
*Chair for Economic Geography, Institute for Economic Geography and GIScience, Department of Social Sciences, Vienna University of Economics and Business Administration, Nordbergstr. 15, A-1090 Vienna; e-mail: manfred.fischer@wu-wien.ac.at*

**Daniel A. Griffith**
*Ashbel Smith Professor of Geospatial Information Sciences, School of Economic, Political and Policy, University of Texas at Dallas, PO Box 830688, GR31, Richardson, TX, 75083-0688, USA; e-mail: dagriffith@utdallas.edu*

**ABSTRACT.** Spatial interaction models of the gravity type are widely used to model origin-destination flows. They draw attention to three types of variables to explain variation in spatial interactions across geographic space: variables that characterise an origin region of a flow, variables that characterise a destination region of a flow, and finally variables that measure the separation between origin and destination regions. This paper outlines and compares two approaches, the spatial econometric and the eigenfunction-based spatial filtering approach, to deal with the issue of spatial autocorrelation among flow residuals. An example using patent citation data that capture knowledge flows across 112 European regions serves to illustrate the application and the comparison of the two approaches.

**JEL Classification: C13, C31, R15**

**Keywords:** Spatial autocorrelation, spatial interaction models, eigenfunction-based spatial filtering, spatial econometrics

## 1. INTRODUCTION

Spatial autocorrelation in geocoded data can be a serious problem, rendering conventional statistical analysis unsound and requiring specialised spatial analytical tools. Spatial autocorrelation refers to the pairwise correlation of georeferenced observations for a single variable. Correlation retains its classical meaning of association, whereas ‘auto’ means self and spatial describes the manner in which self-correlation arises. Autocorrelation is attributable to the configurational arrangement of observations. The problem arises in situations where the observations are non-independent over space. That is, nearby spatial units (regions) are associated in some way. Sometimes, this association is due to a poor match between the spatial extent of a phenomenon of interest and the administrative units for which data are available. Sometimes it is due to a spatial spillover effect. And, sometimes it is attributable to common underlying factors. The complications are similar to those found in time series analysis, but are exacerbated by the multidirectional, two-dimensional nature of dependence in space, rather than the unidirectional, one-dimensional nature in time.

Spatial interaction or flow data pertain to measurements each of which is associated with a link or a pair of origin-destination locations that represent points or areas (regions) in space. While a voluminous literature exists for spatial autocorrelation with a focus of interest on the specification and estimation of models for cross-sectional attribute data, there is scant attention paid to its counterpart in spatial interaction data. For example, there is no explicit reference to spatial flows data in some of the commonly cited spatial econometric and statistics texts, such as Anselin (1988) and Cressie (1991). But Griffith (1988, pp. 66-79) implicitly addresses this topic, and Griffith and Jones (1980) treat this very problem. Furthermore, some relevant research has been done about network autocorrelation (see Black, 1992, Black and Thomas, 1998, Tiefelsdorf and Braun, 1999); but this work treats flows in an indirect way.

Modelling spatial interactions has a long and distinguished history in geography and regional science (see, Sen and Smith, 1995, for a review). Spatial interaction models focus on dyads of regions rather than on individual regions. They aim to explain the variation of spatial interaction across geographic space. In doing so, they draw attention to three types of push-pull variables: those relating to properties of origin regions (origin factor); those relating to properties of destination regions (destination factor); and, those relating to the spatial separation between origin and destination regions (separation factor). Spatial interaction models are said to be misspecified if the residuals are spatially autocorrelated, violating the independence assumption. This problem has been largely neglected so far, with very few exceptions (see, for example, Brandsma and Ketellapper, 1979, Griffith and Jones, 1980, Baxter, 1987, Bolduc, Laferiere and Santarossa, 1992, 1995, Fischer, Reismann and Scherngell, 2006a, LeSage and Pace, 2007). This neglect may be because spatial interaction models are more complex than models for the geographic distribution of attribute data, with each region being associated with several values as an origin as well as a destination so that specification of the autocorrelation structure is less obvious.

This paper outlines and compares two approaches that could be used to account for spatial autocorrelation in a spatial interaction modelling context. One approach involves directly modelling spatial autocorrelation among flow residuals by introducing a spatial error structure that reflects origin and destination autoregressive dependence among origin-destination flows. This view leads to spatial autoregressive model specifications that represent not only extensions of the conventional spatial interaction models, but also extensions of the spatial regression models, the workhorses of applied spatial econometrics.

The other approach, eigenfunction spatial filtering, starts from the misspecification interpretation perspective of spatial autocorrelation, which assumes that spatial autocorrelation in the disturbances is induced by missing origin and destination variables, which themselves are spatially autocorrelated. The approach itself is a non-parametric technique that accounts for the inherent spatial autocorrelation in spatial interaction models by introducing appropriate synthetic surrogate variates (i.e., spatial filters) for the origin and destination variables, and hereby exploiting an eigenfunction decomposition associated with the Moran's *I* (*MI*) statistic of spatial autocorrelation.

The structure of the paper is as follows. The section that follows sets forth the context and framework for the discussion, with a particular focus on the log-normal spatial interaction model version, one of the most common specifications employed in applied spatial interaction data analysis, as well as the Poisson regression generalised linear model version, today's preferred specification. Section 3 outlines the spatial econometric approach that generalises the classical spatial interaction models to spatial econometric origin-destination flow models. These models are formally equivalent to conventional regression models with spatially autocorrelated error terms. But they differ in terms of the data analysed and the way in which the spatial weights matrix is defined. Section 4 moves attention to the eigenfunction-based spatial filtering approach that accounts for the inherent spatial autocorrelation in spatial interaction models with a composite map pattern component (i.e., a spatial filter), rather than simply identifying a global spatial autocorrelation parameter for a spatial autoregressive process. The aim of this non-parametric approach is to control spatial autocorrelation by introducing appropriate synthetic variables that serve as surrogates for spatially autocorrelated missing origin and destination variables. This shift in focus leads to spatial filter variants of the classical spatial interaction model. Patent citation data that capture knowledge flows across 112 European regions are used in Section 5 to compare the workings of the two approaches. The final section concludes the paper with a final commentary about the two approaches.

## 2. BACKGROUND

Suppose we have a spatial system consisting of $n$ regions, where $i$ denotes the origin region $(i=1,...,n)$ and $j$ the destination region $(j=1,...,n)$. Let $m(i,j)$ $(i, j=1, \ldots, n)$ denote observations on random variables, say $M(i,j)$, each of which corresponds, for example, to flows of people, commodities, capital or knowledge from region $i$ to region $j$. The $M(i,j)$ are assumed to be independent random variables. They are sampled from a specified probability distribution that is dependent upon some mean, say $\mu(i,j)$. Let us assume that no a priori information is given about the row and column totals of the flow matrix $[m(i,j)]$. Then the mean interaction frequencies between origin $i$ and destination $j$ may be modelled by

$$\mu(i,j) = c\ A(i)^{\alpha}\ B(j)^{\beta}\ F(i,j) \quad (1)$$

where $\mu(i,j) = E[M(i,j)]$ is the expected flow, $c$ denotes a constant term, the quantities $A(i)$ and $B(j)$ are called origin and destination factors or variables, respectively, $\alpha$ and $\beta$ indicate the relative importance, and $F(i,j)$ is a separation factor that constitutes the very core of spatial interaction models. Following Sen and Sööt (1981), we specify the separation factor in form of a multivariate exponential deterrence function

$$F(i,j) = \exp\left[\sum_{k=1}^{K} \theta_k\ {}^{k}d(i,j)\right] \quad (2)$$

where ${}^{k}d(i,j)$ are $K$ measures of separation between $i$ and $j$, and $\theta_k$ are the associated parameters. The advantage of this specification stems from its ability to approximate a wide variety of deterrence functions, including the power and the gamma or Tanner function.

Equation (1) is a very general version of the classical spatial interaction model. The multivariate exponential specification (2) of the separation factor yields the exponential spatial interaction model that can be expressed equivalently as a log-additive model of the form[1]

$$y(i,j) = \kappa + \alpha\, a(i) + \beta\, b(j) + \sum_{k=1}^{K} \theta_k \; {}^{k}d(i,j) \; + \varepsilon(i,j) \tag{3}$$

with $y(i,j) \equiv \ln[\mu(i,j)]$, $\kappa \equiv \ln(c)$, $a(i) \equiv \ln[A(i)]$, and $b(j) \equiv \ln[B(j)]$. Of note is that the back-transformation of this log-linear specification results in an error structure of the exponential spatial interaction model being multiplicative. The parameters $\kappa, \alpha, \beta$ and $\underset{\sim}{\theta} = (\theta_1 ..., \theta_K)$ have to be estimated if future flows are to be predicted.

There are $n^2$ equations of the form (3). Using matrix notation we may write these equations more compactly as

$$\mathbf{y} = \mathbf{X}\underset{\sim}{\gamma} + \underset{\sim}{\varepsilon} \tag{4}$$

where **y** denotes the $N(=n^2)$-by-1 vector of observations on the interaction variable (see Table 1 for the data organisation convention). **X** is the $N$-by-($K$+3) matrix of observations on the explanatory variables including the origin, destination, separation variables, and the intercept, $\underset{\sim}{\gamma}$ is the associated ($K$+3)-by-1 parameter vector, and the $N$-by-1 vector $\underset{\sim}{\varepsilon} = [\varepsilon(1,1), ..., \varepsilon(n,n)]^T$ denotes the vectorised form of $[\varepsilon(i,j)]$. Intraregional unit flows can be eliminated by removing the $n$ cases for which the origin and destination IDs are the same (i.e., $ID_{origin}=ID_{destination}$), and this is done in this paper. Thus, we consider the case of interregional flows with $N = n(n-1)$ observations.

Table 1 about here

If spatial interaction model (4) is correctly specified, then provided that the regressor variables are not perfectly collinear, $\underset{\sim}{\gamma}$ is estimable under the assumption that the error terms are *iid* with zero mean and constant variance[2], and the OLS estimators are best linear unbiased estimators. A violation of this assumption may lead to spatial autocorrelation.

Flowerdew and Aitkin (1982) question the appropriateness of the widely used log-normal specification of the spatial interaction model, and suggest instead that the observed flows follow a Poisson distribution with

(5) $$P\{m(i,j)\} = \frac{\exp(-\mu(i,j))\,\mu(i,j)^{m(i,j)}}{m(i,j)!}$$

where $P\{.\}$ denotes probability, and the expected value, $\mu(i,j)$, is given by Equation (1). Equation (5) models flows between origin $i$ and destination $j$ as inter-point movement counts. Hence, this is the specification of a discrete distribution. Later, Flowerdew and Lovett (1988) extend Equation (5) to singly- and doubly-constrained spatial interaction models (see Wilson, 1970), again assuming independent origin and destination factors. In other words, this Poisson probability model formulation does not incorporate spatial dependencies in the origin and destination terms[3]. Consequences of overlooking such spatial structure effects are conceptualised in Curry (1972), with their presence empirically demonstrated by Griffith and Jones (1980). Accounting for spatial autocorrelation in the disturbances, the focus of this paper, corrects for a source of specification error.

## 3. THE SPATIAL ECONOMETRIC PERSPECTIVE

One way to incorporate spatial autocorrelation into a spatial interaction model of type (4) is to specify a spatial process for the disturbance terms, structured to follow a (first-order) spatial autoregressive process[4]. In this framework, the disturbance term $\varepsilon(i, j)$ corresponding to the dyad $(i, j)$ of regions is modelled as a weighted average of disturbances corresponding to other dyads, plus a purely random element, say $\eta(i, j)$. This weighted average involves a scalar parameter, say $\rho$ , and a set of weights that describe the spatial dependencies. Formally,

(6) $$\mathbf{y} = \mathbf{X}\, \underset{\sim}{\gamma} + \underset{\sim}{\varepsilon}$$

i.e., a spatial interaction model with the $N$-by-1 disturbance vector $\underset{\sim}{\varepsilon}$ generated as

(7) $$\underset{\sim}{\varepsilon} = \rho\ \mathbf{W} \underset{\sim}{\varepsilon} + \underset{\sim}{\eta}$$

where $\rho$ is the spatial autoregressive coefficient for the error lag $\mathbf{W}\underset{\sim}{\varepsilon}$ and $\underset{\sim}{\eta}$ is an $N$-by-1 vector of i.i.d. normally distributed random terms with zero mean and variance $\sigma^2$. $\mathbf{W}$ is an $N$-by-$N$ (row-standardised) non-negative spatial weights matrix with zeros on the diagonal where $N = n\,(n-1)$. It is convenient to assume that $|\,\rho\,| < 1$, resulting in the matrix $\mathbf{I}_N - \rho\,\mathbf{W}$ to be non-singular for all $|\,\rho\,| < 1$. Given these assumptions, it follows from Equation (7) that $\underset{\sim}{\varepsilon} = (\mathbf{I} - \rho\ \mathbf{W})^{-1}\, \underset{\sim}{\eta}$. Thus, $E(\underset{\sim}{\varepsilon}) = 0$ and $E(\underset{\sim}{\varepsilon}\, \underset{\sim}{\varepsilon}^T) = \underset{\sim}{\Omega}\,(\rho)$ where

(8) $$\underset{\sim}{\Omega}(\rho) = \sigma^2\ \mathbf{V}(\rho)\ = \sigma^2 [(\mathbf{I}_N - \rho\ \mathbf{W})^T\, (\mathbf{I}_N - \rho\ \mathbf{W})]^{-1}$$

To ensure that the variance-covariance matrix $\underset{\sim}{\Omega}(\rho)$ is positive definite and, thus, non-singular, the autocorrelation parameter $\rho$ has to be within its feasible range $\rho \in ]\lambda_{\min}^{-1}, \lambda_{\max}^{-1}[$ , where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of **W**, respectively, with $\lambda_{\min} < 0 < \lambda_{\max}$ (Hepple 1995). Since the row sums of **W** are bounded uniformly in absolute value by one, the Perron-Frobenius theorem (Cox and Miller 1965) states that $\lambda_{\max} = 1$ and $-1 \leq \lambda_{\min}$ , so that we have the restriction of $|\rho| < 1$ for the stationarity of spatial origin-destination models of type (6)-(7). If $|\rho| > 1$, the model would be explosive and non-stationary.

*Specification of the spatial weights matrix* **W**

While the conventional notion of spatial autocorrelation in a cross-sectional regression context that involves a sample of *n* regions relies on an *n*-by-*n* spatial weights matrix to represent the connectivity structure between regions, in a spatial interaction context where the **y**-vector reflects flows between origins and destinations, there is a need to extend the notion of spatial autocorrelation to a concept of network spatial autocorrelation or spatial connectivity between origin-destination dyads of regions. This requires shifting attention from a two-dimensional space $\{i, j \mid i \neq j;\ i, j = 1, ..., n\}$ to a four-dimensional space $\{i, j, r, s \mid i \neq j, r \neq s;\ i, j = 1, ..., n; r, s = 1, ..., n\}$ : the geographical space in which flow origins (such as *i* and *r*), on the one hand, and flow destinations (such as *j* and *s*), on the other, are located, in either of which there may be spatial dependence in flow levels originating/terminating in proximate regions. Proximity is defined in terms of first-order origin-related and destination-related contiguity relations specified in an *N*-by-*N* spatial weights matrix $\mathbf{W}^* = {}^o\mathbf{W} + {}^d\mathbf{W}$ that reflect the cumulative impact of origin and destination interaction effects[5]. Formally, we define the elements of the origin-based spatial weights matrix ${}^o\mathbf{W}$ by

(9) $$^{o}w(i, j;\ r, s) = \begin{cases} 1 \ \text{ if } \ j = s \text{ and } c(i,r) = 1, \text{ and} \\ 0 \ \text{ otherwise} \end{cases}$$

where $c(i,r)$ is the element of a conventional $n$-by-$n$ first-order contiguity matrix with

(10) $$c(i,r) = \begin{cases} 1 \ \text{ if } \ i \neq r, \text{ and } i \text{ and } r \text{ have a common border, and} \\ 0 \ \text{ otherwise} \end{cases}$$

This spatial weights matrix $^{o}\mathbf{W}$ specifies an origin-based neighbourhood set for each origin-destination pair ($i$, $j$). An element $^{o}w(i, j;\ r, s)$ defines an origin-destination pair ($r$, $s$) as being a 'neighbour' of ($i$, $j$) if the origin regions $i$ and $r$ are contiguous spatial units and $j = s$.

As a parallel to $^{o}\mathbf{W}$ the destination-based spatial weights matrix $^{d}\mathbf{W}$ is defined to consist of elements

(11) $$^{d}w(i, j;\ r, s) = \begin{cases} 1 \ \text{ if } \ i = r \text{ and } c(j,s) = 1, \text{ and} \\ 0 \ \text{ otherwise} \end{cases}$$

where

(12) $$c(j,s) = \begin{cases} 1 \ \text{ if } \ j \neq s, \text{ and } j \text{ and } s \text{ have a common border, and} \\ 0 \ \text{ otherwise} \end{cases}$$

Then the elements of the row-standardised $N$-by-$N$ spatial weights matrix $\mathbf{W} = \{w(i, j; r, s);\ i \neq j,\ i,\ j = 1,\ldots,n \text{ and } r \neq s;\ r, s = 1,\ldots,n\}$ are given by

(13) $$w(i, j;\ r, s) = \frac{w^{*}(i, j;\ r, s)}{\sum\limits_{\substack{r', s' = 1 \\ (r', s') \neq (i, j)}}^{N} w^{*}(i, j;\ r', s')}$$

for $i \neq r$ and $j \neq s$, and by definition $w(i, j; r, s) = 0$ for $i = r$ and $j = s$.

*Maximum Likelihood Estimation*

The parameters to be estimated in the spatial interaction model (6) with errors (7) are $\underset{\sim}{\gamma}$, $\sigma_\eta^2$ and $\rho$. The maximum likelihood estimates can be obtained from the concentrated log-likelihood function of the form

$$\mathcal{L}(\rho) = C + \ln |\mathbf{I}_N - \rho \mathbf{W}| - \tfrac{N}{2} \ln (S(\rho)) \tag{14}$$

where $S(\rho)$ represents the sum of squared errors expressed as a function of the scalar parameter $\rho$ alone after concentrating out the parameters $\underset{\sim}{\gamma}$ and $\sigma$, and $C$ denotes a constant not depending on $\rho$. The optimisation can be performed with a sophisticated optimisation routine, or with a simple grid search.

The computationally troublesome aspect in numerical maximisation of the concentrated log-likelihood function is the need to compute the log-determinant of the *N*-by-*N* matrix $(\mathbf{I}_N - \rho\ \mathbf{W})$. Standard algorithms for maximum likelihood estimation of the model (6)-(7) become difficult as *N* increases. Computational time for computing this determinant increases with the cube of *N* for dense matrices. While $\mathbf{W}$ is an *N*-by-*N* matrix, it is sparse by construction. A sparse matrix is one that contains a large proportion of zero elements. Thus, sparse Cholesky decomposition algorithms may be used to efficiently tackle the *N*-by-*N* log-determinant problem for larger *N*. Sparse algorithms decrease the storage needed for $\mathbf{W}$ and $(\mathbf{I}_N - \rho\ \mathbf{W})$, and greatly accelerate computations (see Pace and Barry, 1997).

## 4. THE EIGENFUNCTION SPATIAL FILTERING APPROACH

The eigenfunction spatial filtering approach represents an alternative methodology to account for a specific type of spatial autocorrelation in the disturbances, namely spatial autocorrelation arising from missing origin/destination variables that are spatially correlated[6]. The primary motivation for this approach in the current context is to allow spatial analysts to compute OLS estimators for the parameters of the log-normal spatial interaction model, as well as generalised linear model Poisson regression spatial interaction parameter estimates, while ensuring that the required model assumptions are met. The approach outlined in this section derives from the eigenfunction spatial filtering approach devised by Griffith (1996, 2000, 2002, 2003, 2004) for attribute data. This approach is semi-parametric in nature, and aims to control for spatial autocorrelation by introducing appropriate synthetic variables that serve as surrogates for spatially autocorrelated missing origin and destination variables. These synthetic variables are derived as linear combinations of eigenvectors that come from the following modified version of the conventional $n$-by-$n$ binary 0-1 contiguity matrix **C**:

$$(\mathbf{I}-\mathbf{1}\mathbf{1}^T \tfrac{1}{n})\ \mathbf{C}\ (\mathbf{I}-\mathbf{1}\mathbf{1}^T \tfrac{1}{n}) \tag{15}$$

where **I** is the $n$-by-$n$ identity matrix, and **1** is an $n$-by-1 vector of ones. This particular matrix expression appears in the numerator of the Moran's $I$ ($MI$) statistic of spatial autocorrelation defined for attribute data. Tiefelsdorf and Boots (1995) show that all of the eigenvalues of expression (15) relate to distinct $MI$ values.

An eigenfunction linked to some geographic contiguity matrix **C** may be interpreted in the context of latent map pattern as follows (Getis and Griffith 2002): The first eigenvector, say

$\mathbf{E}_1$, is the set of numerical values that has the largest *MI* value achievable for any set of numerical values, for the given geographic contiguity matrix. The second eigenvector, $\mathbf{E}_2$, is the set of numerical values that has the largest achievable *MI* for any set of numerical values that is uncorrelated with $\mathbf{E}_1$. This sequential construction of eigenvectors continues through $\mathbf{E}_n$, which is the set of numerical values that has the largest negative *MI* achievable by any set of numerical values that is uncorrelated with the preceding (*n*–1) eigenvectors. These *n* eigenvectors describe the full range of all possible mutually orthogonal and uncorrelated map patterns, and may be interpreted as synthetic map variables that represent specific natures (that is, positive or negative) and degrees (that is, negligible, weak, moderate, strong) of potential spatial autocorrelation. Eigenvector calculations often can be restricted to only those that are prominent (e.g., have an associated predesignated minimum *MI* value of, say, 0.25), and represent the nature of the detected spatial autocorrelation (e.g., positive).

The eigenvector spatial filtering approach, based upon a stepwise selection criterion, adds a minimally sufficient set of eigenvectors as proxies for missing origin and destination variables, and in doing so accounts for spatial autocorrelation among the observations by inducing mutual dyad error independence. This leads to a spatial filter specification of the spatial interaction model (1)-(2) that may be described as

$$\mu(i,j) = c\ \exp[\sum_{q=1}^{Q} E_{iq}\,\psi_q]\ A(i)^{\alpha}\ \ \exp[\sum_{r=1}^{R} E_{jr}\,\varphi_r]\ B(j)^{\beta}\ \exp[\sum_{k=1}^{K}\theta_k\ {}^{k}d(i,j)] \tag{16}$$

where $\mu(i,j)$, ${}^{k}d(i,j)$, $c$, $\alpha$, $\beta$, $\theta_k$ ($k$=1, …, $K$), $A(i)$ and $B(j)$ are defined as above, $Q$ and $R$ denote selected subsets of the *n* eigenvectors that have been chosen by supervised selection to furnish a good description of flows out of the origins and flows into the destinations, respectively, and $\psi_q$ and $\varphi_r$ are the respective coefficients for the linear combinations of eigenvectors that constitute the origin and destination spatial filters, namely $\sum_{q=1}^{Q} E_{iq}\,\psi_q$ and $\sum_{r=1}^{R} E_{jr}\,\varphi_r$ . For these spatial filters, which are linear combinations of the eigenvectors of expression (15) and represent the spatial autocorrelation components of the missing origin and destination variables, $\psi_q\ (q=1,...,Q)$ and $\varphi_r\ (r=1,...,R)$ are regression coefficients that indicate the relative importance of each distinct map pattern in accounting for spatial autocorrelation in the flows structure.

*The Spatial Filtering Model Specification of the Log-Normal Additive Model*

Spatial filter spatial interaction model (16) can be expressed equivalently in log-additive form, in order to link it to a normal probability distribution for the error term, as

(17) $$y(i,j)=\kappa+\sum_{q=1}^{Q} E_{iq}\,\psi_q+\alpha\ a(i)+\sum_{r=1}^{R} E_{jr}\,\varphi_r+\beta\ b(j)+\sum_{k=1}^{K}\theta_k\ {}^{k}d(i,j)+\varepsilon(i,j)$$

OLS can be employed to estimate the model parameters. All conventional diagnostic statistics developed for linear regression analysis can be computed and interpreted without having to develop spatially adjusted counterparts. The major numerical difficulty of the spatial filter model version is that eigenfunctions have to be calculated, a formidable computational task for larger spatial interaction systems (i.e., large $n$)[7].

*Specification of a Conventional Poisson Spatial Interaction Model*

Equation (3) as a mean response, and hence without the error term, can be estimated with the data organisation given in Table 1 ($ID_{origin} \neq ID_{destination}$) via Poisson regression through the use of a generalised linear model algorithm coupled with a Poisson distribution and its appropriate link function. Parameter estimation can be achieved either with iteratively reweighted least squares or maximum likelihood techniques.

*Specification of a Spatial Filter Spatial Interaction Model*

Spatial filter counterparts to the spatial econometric specification can be obtained in one of two ways: (i) by augmenting the set of covariates with the set of candidate eigenvectors – relating to Equation (17); and, (ii) by estimating parameters for this augmented set with a Poisson regression – relating directly to Equation (16). The origin candidate eigenvectors are obtained from $\mathbf{1} \otimes \mathbf{E}_U$, whereas the destination candidate eigenvectors are obtained from $\mathbf{E}_U \otimes \mathbf{1}$, where $\mathbf{E}_U$ is the set of candidate eigenvectors (e.g., those whose associated *MI* value, when divided by the maximum possible *MI* value, exceeds 0.25), and $\otimes$ denotes the Kronecker product.

## 5. AN ILLUSTRATIVE APPLICATION OF THE APPROACHES

Patent citation data are used to illustrate the way the two approaches could be applied to control for spatial autocorrelation among the residuals in a spatial interaction model. Such data recorded in patent documents are widely recognised as a rich and fruitful source for the study of the spatial dimension of knowledge transmission using patent citations (see, for example, Jaffe and Trajtenberg, 2002, Fischer, Scherngell and Jansenberger, 2006b).

*The Context*

We use interregional patent citation flows as the dependent variable in the models. The data specifically relate to citations between European high-tech patents. By European patents we mean patent applications at the European Patent Office assigned to high-tech firms located in Europe. High-technology is defined to include the International Standard Industrial

Classification (ISIC)-sectors of aerospace (ISIC 3845), electronics-telecommunication (ISIC 3832), computers and office equipment (ISIC 3825), and pharmaceuticals (ISIC 3522). Self-citations (i.e., citations from patents assigned to the same firm) have been excluded, given our interest in pure externalities as evidenced by *interfirm* knowledge spillovers.

It is well known that the observation of citations is subject to a truncation bias, because we observe citations for only a portion of the *life* of an invention. To avoid this bias in the analysis, we have established a five-year window (that is, 1985-1989, 1986-1990, …, 1993-1997) to count citations to a patent[8]. The observation period is 1985-1997 with respect to cited patents and 1990-2002 with respect to citing patents. The sample used in this contribution is restricted to inventors located in $n$ = 112 regions, generally NUTS-2 regions, covering the core of 'Old Europe,' including Germany (38 regions), France (21 regions), Italy (20 regions), the Benelux countries (24 regions), Austria (8 regions), and Switzerland (one region), resulting in $N$ = 12,432 interregional flows.

Subject to caveats relative to the relationship between patent citations and knowledge spillovers, these data allow us to identify and measure spatial separation effects for interregional knowledge spillovers in this interaction system of 112 regions. Our interest is focused on $K$ = 3 separation measures: $\mathbf{d}^{(1)}$ is an $N$-by-1 vector that represents geographic distance measured in terms of the great circle distance (in kilometres) between the regions represented by their economic centres; $\mathbf{d}^{(2)}$ is an $N$-by-1 country dummy variable vector that represents border effects measured in terms of the existence of country borders between the regions; and, $\mathbf{d}^{(3)}$, is an $N$-by-1 vector of technological proximity, which is defined next.

As we consider the distance effect on interregional patent citations, it is important to control for technological proximity between regions, as geographical distance could be just proxying for technological proximity. To do this, we use a technological proximity index $s_{ij}$ that defines the proximity between regions $i$ and $j$ in technology space. We divide the high-technology patents into 55 technological subclasses, following the International Patent Code classification system. Each region is assigned a 55-by-1-technology vector that measures the share of patenting in each of the technological subclasses for a region. The technological proximity index $s_{ij}$ between regions $i$ and $j$ is given by the uncentred correlation of their technological vectors. Two regions that patent exactly in the same proportion in each subclass have an index equal to one, while two regions patenting only in different subclasses have an index equal to zero. This index is appealing because it allows for a continuous measure of technological distance by the transformation $d_{ij} = 1 - s_{ij}$. Appropriate ordering leads to the $N$-by-1 vector $\mathbf{d}^{(3)}$.

The product $A(i)\,B(j)$ in Equation (1) may be interpreted simply as the number of distinct ($i$, $j$)-interactions that are possible. Thus, a reasonable way to measure the origin factor is in terms of the number of patents in knowledge producing region $i$ in the time period 1985-1997, and the destination factor in terms of the number of patents in knowledge absorbing region $j$ in the time period 1990-2002, producing the $N$-by-1 vectors **a** and **b**, respectively.

*Application of the Spatial Autoregression Approach*

Table 2 reports the ML estimates of the spatial autoregressive model specification that reflects origin and destination spatial dependence of flows. We used the *spdep package* running on a Sun Fire V250 with 1.28 GHz and 8 GB RAM to create the $(n^2 - n)$-by-$(n^2 - n)$ spatial weights matrix **W** from polygon contiguities, and the *errorsarlm* procedure based on Ng and Peyton’s (1993) sparse matrix Cholesky algorithm to generate the ML estimates for the model. Using this algorithm, computation of the maximum likelihood estimates of the spatial econometric model required 836 seconds, a remarkably short time given that the estimation approach relies on calculating the determinant of a 12,432-by-12,432 spatial weights matrix at each iteration step in the optimisation process[9].

The table contains the parameter estimates of the spatial autoregressive model specification and its associated log-likelihood function value, together with those of the conventional log-additive spatial interaction model. Moving from the conventional spatial interaction model to the spatial econometric flow model reflecting spatial dependence at the origins and destinations increases the log-likelihood from –21,024.13 to –20,212.013. This is to be expected, given the significance of the spatial autoregressive parameter that points to spatial dependence at origin and destination locations ($\hat{\rho}$ = 0.613). Least squares, which ignores spatial dependence and assumes residual flows to be independent, produces a much lower likelihood function value. Capturing the dependencies greatly reduces the residual variance and strengthens the inferential basis affiliated with the model. It is worth noting that models based on the use of separate weight matrices ${}^{o}\mathbf{W}$ and ${}^{d}\mathbf{W}$ have lower log-likelihoods than the model based on ${}^{o}\mathbf{W} + {}^{d}\mathbf{W}$. This seems to support the notion that both origin and destination dependence information is important[10].

Table 2 about here

The ML estimates display the expected signs, as do the OLS estimates. All the estimated coefficients are clearly significant. Maximum likelihood seems to ascribe a greater negative influence to geographical distance and national borders in creating friction that inhibits knowledge flows. But the two sets of estimates are not significantly different from each other, since the estimates of the spatial autoregressive model specification are within the 95 percent confidence limits of the least squares estimates. This reinforces the idea that spatial autocorrelation, which is commonly being regarded as one particular form of heteroscedasticity, does not induce bias in coefficients.

*Application of the Eigenfunction Spatial Filtering Approach*

Table 2 also shows the parameter estimates from the spatial filter specification of the log-normal additive spatial interaction model given by Equation (17). A separate spatial filter is constructed for the origins and for the destinations. The 27 candidate eigenvectors (those, out of a total of 112, whose *MI* value divided by the maximum *MI* value is at least 0.25) were computed with a FORTRAN program using IMSL routines.

The selected eigenvectors that collectively maximise the log-likelihood function, together with their estimated coefficients and associated levels of spatial autocorrelation, are summarised in Table 4. The origin and destination spatial filters for the log-normal additive model respectively contain 20 and 21 eigenvectors, and capture moderate positive spatial autocorrelation contained in the conventional spatial interaction model residuals. Maps of these two spatial filters appear in Figures 1(a) and 1(b).

The spatial filter model estimates reported in Table 2 are not significantly different from the least squares ones. They lie within the 95 percent confidence interval of the least squares estimates. It is also the case that they are within the 95 percent confidence interval for the spatial autoregressive model estimates and vice-versa. Thus, the spatial filter, the spatial error and the least squares model specifications, produce statistically equivalent point estimates. So, in accordance with theory, mere spatial dependence in the disturbances does not impact the point estimates, but just the precision of the estimates.

*Application of the Poisson Model Specifications*

The fat-tailed nature of the distribution of the vectorised flow matrix and the presence of numerous zero in the matrix reflecting a lack of interaction between regions in the sample raise doubts on the appropriateness of the normality assumption that ignores the true integer nature of the flows and approximates a discrete-valued process by an almost certainly misrepresentative continuous distribution. Hence, the Poisson model specifications appear to be more appropriate in the current context.

Table 3 about here

Table 4 about here

Table 3 reports the ML estimates of the conventional Poisson model and its spatial filter model specification counterpart that reflects origin-destination spatial dependence of flows. A separate spatial filter is constructed for the origins and for the destinations. The 27 candidate eigenvectors (those, out of a total of 112, whose *MI* value divided by the maximum *MI* value is at least 0.25) were computed with a FORTRAN program using IMSL routines, and Poisson regression was executed with the SAS PROC GENMOD procedure.

The selected eigenvectors are summarised in Table 4. The origin and destination spatial filters contain 23 and 16 eigenvectors respectively, and capture moderate positive spatial autocorrelation contained in the basic Poisson model residuals. Maps of these two spatial filters appear in Figures 1(c) and 1(d). Pairwise relationships between these and the origin and destination spatial filters for the log-normal additive model (see Figures 1(a) and 1(b)) are portrayed in Figure 2.

Figure 1 and Figure 2 about here

Table 3 shows that the 95 percent confidence intervals do not overlap for the coefficient estimates for the two Poisson model specifications (except the country border). This suggests that explicitly accounting for spatial error autocorrelation in a Poisson context does impact the point estimates and not only the precision of the estimates. It is also the case that the estimates of the conventional Poisson model and its spatial filter model specification counterpart are significantly different from the least squares estimates.

Some important differences arise in parameter estimates and inferences that we would draw from the conventional and the spatial filter specification of the Poisson spatial interaction model. *First*, accounting for spatial autocorrelation effects in a Poisson setting results in the importance of the origin and destination factors decreasing, while a greater negative influence is ascribed to geographical and technological distances in creating friction that inhibits knowledge flows. *Second*, *MI* statistics indicate that spatial autocorrelation among residuals is captured, but only modestly.

Considerable similarities exist between the map patterns captured by each of the four spatial filters. The bivariate correlation between the two origin spatial filters is 0.803 (see Figure 2), both highlighting a Switzerland-southern France-northern Italy focal region. This focus is less conspicuous with the two destination spatial filters, whose correlation is only 0.593. Both pairs of spatial filters suggest that much of the northern part of continental Europe forms a cluster, too. But the log-normal additive model noticeably differs from the Poisson model in terms of southern Italy, for both origin and destination spatial filters.

## 6. SUMMARY AND CONCLUSIONS

Two effective approaches to account for spatial dependence in the disturbances of geographic flow models are described and demonstrated. Both approaches give researchers tools that aid in the proper specification of spatial interaction models. These approaches are somewhat different in the perspective within which each views the problem of spatial autocorrelation.

The spatial econometric approach is derivative of the literature on spatial autocorrelation in a cross-sectional spatial regression context. As such, it expresses spatial autocorrelation through the specification of a spatial stochastic process. But while the notion of spatial autocorrelation in a conventional spatial regression context involving a sample of $n$ regions relies on an $n$-by-$n$ spatial weights (connectivity) matrix, the notion of spatial autocorrelation in a spatial interaction context relies on an $N$-by-$N$ spatial weights matrix. The spatial weights matrix captures origin-based and destination-based dependence relations among the observations that influence flows from origins to destinations in a system of $n$ regions. The resulting spatial econometric origin-destination flow model is formally equivalent to regression models with spatially autocorrelated errors, but differs in terms of the data analysed and the manner in which the spatial weights matrix is defined.

Eigenvector spatial filtering furnishes an alternative methodology that enables spatial autocorrelation effects to be captured within a spatial interaction model. This approach makes use of the misspecification interpretation of spatial autocorrelation, and shifts attention to spatial autocorrelation arising from missing origin and destination factors that is reflected in flows between pairs of these locations. In doing so, it allows for spatial interaction models where the desire is to avoid especially a log-linear spatial autoregressive specification coupled with a log-normally distributed error term, and to employ a generalised linear model formulation coupled with a Poisson distributed response variable.

In conclusion, explicitly accounting for spatial error autocorrelation in a Poisson setting results in statistically significant changes in distance decay parameter estimates, and increases in parameter estimate standard errors. In a log-normal setting, however, the mere spatial dependence in the disturbances does not impact the point estimates, just the precision of the estimates. Both the spatial econometric and the spatial filter approach yield estimates that are not significantly different from each other and lie within the 95 percent confidence limits of the least squares estimates. This finding is in accordance with econometric theory suggesting that spatial error autocorrelation does not lead to bias.

## ENDNOTES

[1] Note in some cases $y_{ij} = 0$, indicating the absence of flows from $i$ to $j$. This leads to the so-called zero problem, since the logarithm then is undefined. There are several pragmatic solutions to this problem, with adding a small constant to the zero elements of $[y_{ij}]$ being widely used. Here we added 0.08.

[2] This assumption implies that the individual flows, $y(i, j)$, from origin $i$ to destination $j$ are independent from each other and that interaction flows between any pairs of regions are independent from flows between any other pairs of regions.

[3] See LeSage, Fischer and Scherngell (2007) for a Bayesian hierarchical Poisson spatial interaction model that includes latent spatial structure effects representing a spatial autoregressive process.

[4] Hence, the spatial dependence resides in the disturbance process $\underset{\sim}{\varepsilon}$, as in the case of serial correlation in time series regression models. An alternative way involves modelling a functional relationship between the spatial interaction variable $\mathbf{y}$ and its associated spatial lag $\mathbf{Wy}$ rather than directly modelling dependence in the errors. LeSage and Pace (2007) adopt this approach that leads to a spatial origin-destination filter specification applied to the vector of origin-destination flows and captures three types of possible spatial dependence that may occur between origin-destination flows: origin-based, destination-based and origin-to-destination based dependence. Estimation relies on the use of moment matrices and a conventional $n$-by-$n$ spatial weights matrix.

[5] The definitions of ${}^{o}\mathbf{W}$ and ${}^{d}\mathbf{W}$ given in Equations (9)-(10) and (11)-(12) lead to dependence structures that are equivalent to those considered in LeSage and Pace (2007).

[6] See Tiefelsdorf and Griffith (2007) for a concise discussion of how a spatial filter model specification addresses spatial autocorrelation in residuals arising from missing spatially correlated explanatory variables.

[7] Eigenvectors for an *n*-by-*n* connectivity matrix can be computed for matrices up to about 10,000 without too much difficulty using standard software packages. These eigenvectors are approximate when intraregional flows are set aside [i.e., $N = n^2$ becomes $N = n(n–1)$]. They also result in enormous data set sizes that require considerable virtual memory in order for software packages such as SAS to execute, resulting in sizeable CPU time requirements. Cressie et al. (1996) suggests some strategies for handling these difficulties.

[8] For details about data construction, see Fischer, Scherngell and Jansenberger (2006b). To obtain citations by any one patent application in year *t*, one needs to search the references made by all patent applications after year *t*. This is called the inversion problem that arises because the original data about citations come in the form of citations made, whereas we need dyads of cited and citing patents to construct interregional patent citations flows. In the case of cross-regional inventor teams, the procedure of multiple full counting has been applied (see Fischer, Scherngell and Jansenberger, 2006b for details).

[9] Note that Pace and Barry (1997) suggest to compute the sparse Cholesky decomposition *once*, not on every trip through the optimisation loop. This would reduce computational costs by a factor of about 100 (personal communication with James LeSage). But this idea is difficult to implement in the context of *spdep package* software.

[10] The estimate for the origin-based spatial autocorrelation parameter is 0.311 and that for the destination-based counterpart 0.365.

## REFERENCES


Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, Boston and London: Kluwer.

Baxter, Mike J. 1987. "Testing for misspecification in models of spatial flows", *Environment and Planning A*, 19, 1153-1160.

Black, William R. 1992. "Network autocorrelation in transportation network and flow systems", *Geographical Analysis*, 24(3), 207-222.

Black, William R., and I. Thomas. 1998. "Accidents on Belgium's motorways: A network autocorrelation analysis", *Journal of Transport Geography*, 6, 23-31.

Bolduc, Denis, Richard Laferriere, and Gino Santarossa. 1992. "Spatial autoregressive error components in travel flow models", *Regional Science and Urban Economics*, 22(3), 371-385.

Bolduc, Denis, Laferriere Richard, and Gino Santarossa. 1995. "Spatial autoregressive error components in travel flow models: An application to aggregate mode choice", in L. Anselin and R.J. Florax (eds.), *New Directions in Spatial Econometrics*. Berlin, Heidelberg, New York: Springer, pp. 96-108.

Brandsma, A.S., and R.H. Ketellapper. 1979. „A biparametric approach to spatial autocorrelation", *Environment and Planning A*, 11(1), 51-58.

Cliff, Andrew D., R.L. Martin R.L., and Keith J. Ord. 1975. "Maps patterns and friction of distance parameters: Reply to comments by R.J. Johnston and by L. Curry, D.A. Griffith and E.S. Sheppard", *Regional Studies* 9, 285-288.

Cox, D.R., and H.D. Miller. 1965. *The Theory of Stochastic Processes*. London: Methuen.

Cressie, Noel A.C. 1991. *Statistics for Spatial Data*. New York: John Wiley.

Cressie, Noel, Anthony Olsen, and Dianne Cook. 1996. "Massive data sets: Problems and possibilities, with application to environmental monitoring," in Committee on Applied and Theoretical Statistics (eds.), *Massive Data Sets: Proceedings of a Workshop*. Washington, DC: Board on Mathematical Sciences, National Research Council, pp. 115-119.

Curry, Leslie. 1972. "A spatial analysis of gravity flows", *Regional Studies*, 6(2), 131–147.

Fischer, Manfred M., Martin Reismann, and Thomas Scherngell 2006a. „From conventional to spatial econometric models of spatial interactions", Paper Presented at the International Workshop on Spatial Econometrics and Statistics, May 25-27, Rome [Italy].

Fischer, Manfred M., Thomas Scherngell, and Eva Jansenberger. 2006b. „The geography of knowledge spillovers in Europe – Evidence from a model of interregional patent citations in high-tech industries", *Geographical Analysis*, 38(3), 288-309.

Flowerdew, Robin, and Murray Aitkin. 1982. "A method of fitting the gravity model based on the Poisson distribution", *Journal of Regional Science*, 22, 191-202.

Flowerdew, Robin, and Andrew Lovett 1988. "Fitting constrained Poisson regression models to interurban migration flows", *Geographical Analysis*, 20(4), 297-307.

Fotheringham, A. Stewart, and Morton E. O'Kelly. 1989. *Spatial Interaction Models. Formulations and Applications*. Dordrecht, Boston, London: Kluwer.

Getis, Arthur, and Daniel A. Griffith. 2002. "Comparative spatial filtering in regression analysis", *Geographical Analysis*, 34(2), 130-140.

Griffith, Daniel A. 2004. "A spatial filtering specification for the auto-logistic model", *Environment and Planning A*, 36(10), 1791-1811.

Griffith, Daniel A. 2003. *Spatial Autocorrelation and Spatial Filtering*. Berlin, Heidelberg, New York: Springer.

Griffith, Daniel A. 2002. "A spatial filtering specification for the auto-Poisson model", *Statistics & Probability Letters*, 58, 245-251.

Griffith, Daniel A. 2000. "A linear regression solution to the spatial autocorrelation problem", *Journal of Geographical Systems*, 2(2), 141-156.

Griffith, Daniel A. 1996. "Spatial autocorrelation and eigenfunctions of the geographic weight matrix accompanying geo-referenced data", *The Canadian Geographer*, 40, 351-367.

Griffith, Daniel A. 1988. *Advanced Spatial Statistics*. Dordrecht: Martinus Nijhoff.

Griffith, Daniel A., and Ken Jones. 1980. "Explorations into the relationship between spatial structure and spatial interaction", *Environment and Planning A*, 12(2), 187-201.

Hepple, Leslie W. 1995. "Bayesian techniques in spatial and network econometrics: 2. Computational methods and algorithms", *Environment and Planning A*, 27(3), 615-644.

Jaffe, Adam B., and Michael Trajtenberg. 2002. *Patents, Citations & Innovations*. Massachusetts: MIT Press.

LeSage, James, and Kelley R. Pace. 2007. Spatial econometric modeling of origin-destination flows (March 5, 2007). Available at SSRN: http://ssrn.com/abstract=924609.

LeSage, James, Manfred M. Fischer, and Thomas Scherngell. 2007. Knowledge spillovers across Europe. Evidence from a Poisson spatial interaction model with spatial effects. *Papers in Regional Science (OnlineEarly Articles). doi:10.1111/j.1435-5957.2007.00125.x*

Ng, Esmond G., and Barry W. Peyton. 1993. "Block sparse Cholesky algorithms on advanced multiprocessor computers", *SIAM Journal of Scientific Computing*, 14(5), 1034-1056.

Pace, R. Kelley, and Ronald Barry. 1997. "Quick computation of spatial autoregressive estimators", *Geographical Analysis*, 29(3), 232-246.

Sen, Ashish, and Tony Smith. 1995. *Gravity Models of Spatial Interaction Behavior*. Berlin, Heidelberg, New York: Springer.

Sen, Ashish, and Sim Sööt. 1981. "Selected procedures for calibrating the generalized gravity model", *Papers of the Regional Science Association*, 48, 165-176.

Stronge, William B., and R.R. Schultz. 1978. "Heteroscedasticity and the gravity model", *Geographical Analysis*, 10(3), 279-286.

Tiefelsdorf, Michael, and Daniel A. Griffith. 2007. "Semiparametric filtering of spatial autocorrelation: The eigenvector approach", *Environment and Planning A*, 39 (5), 1139-1221.

Tiefelsdorf, Michael. 2003. "Misspecifications in interaction model distance decay relations: A spatial structure effect", *Journal of Geographical Systems*, 5(1), 25-50.

Tiefelsdorf, Michael, and Gerhard Braun. 1999. "Network autocorrelation in Poisson regression residuals: Inter-district migration patterns and trends within Berlin", paper presented at the 11th European Colloquium on Quantitative and Theoretical Geography, September 3-7, Durham City [England].

Wilson, Alan. 1970. *Entropy in Urban and Regional Modelling*. London: Pion.

TABLE 1: Data Organisation Convention

| Dyad Label | $ID_{origin}$ | $ID_{destination}$ | Flow | Origin Variable | Destination Variable | Separation (Origin, Destination) |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | $y(1, 1)$ | $a(1)$ | $b(1)$ | $d(1, 1)$ |
| 2 | 2 | 1 | $y(2, 1)$ | $a(2)$ | $b(1)$ | $d(2, 1)$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | $n$ | 1 | $y(n, 1)$ | $a(n)$ | $b(1)$ | $d(n, 1)$ |
| $n$+1 | 1 | 2 | $y(1, 2)$ | $a(1)$ | $b(2)$ | $d(1, 2)$ |
| $n$+2 | 2 | 2 | $y(2, 2)$ | $a(2)$ | $b(2)$ | $d(2, 2)$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2$n$ | $n$ | 2 | $y(n, 2)$ | $a(n)$ | $b(2)$ | $d(n, 2)$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $(n-1)n$ | 1 | $n$ | $y(1, n)$ | $a(1)$ | $b(n)$ | $d(1, n)$ |
| $(n-1)n+1$ | 2 | $n$ | $y(2, n)$ | $a(2)$ | $b(n)$ | $d(2, n)$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n^2=N$ | $n$ | $n$ | $y(n, n)$ | $a(n)$ | $b(n)$ | $d(n, n)$ |

*Note* that the $n$ cases for which $ID_{origin}=ID_{destination}$ are removed in the interregional case where $i \neq j$.

TABLE 2: Log-Normal Additive Spatial Interaction Models: The Conventional Model, the Spatial Autoregressive Model Specification Using **W**, and the Spatial Filter Model Specification with 20 Origin and 21 Destination Eigenvectors

| | The Conventional Log-Normal Additive Model [OLS] | | | The Spatial Autoregressive Model [ML] | | | The Spatial Filter Model[d] [OLS] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimates (Standard Error) | 95% Confidence Limits | | Estimates (Standard Error) | 95% Confidence Limits[c] | | Estimates (Standard Error) | 95% Confidence Limits[e] | |
| Constant | –4.851 (0.236) | –5.315 | –4.388 | -4.658 (0.320) | –5.414 | -4.532 | –4.045 (0.249) | –4.533 | –3.558 |
| Origin Variable [a] | 0.594 (0.007) | 0.580 | 0.608 | 0.593 (0.009) | 0.576 | 0.610 | 0.587 (0.008) | 0.571 | 0.603 |
| Destination Variable [b] | 0.562 (0.007) | 0.548 | 0.576 | 0.553 (0.009) | 0.536 | 0.570 | 0.551 (0.008) | 0.534 | 0.567 |
| Geographical Distance | –0.181 (0.020) | –0.220 | –0.142 | -0.224 (0.038) | –0.296 | -0.152 | –0.238 (0.023) | –0.283 | –0.193 |
| Country Border | –0.592 (0.034) | –0.658 | –0.526 | -0.651 (0.054) | –0.754 | -0.548 | –0.671 (0.036) | –0.742 | –0.600 |
| Technological Distance | –2.364 (0.203) | –2.763 | –1.966 | -2.183 (0.212) | –2.586 | -1.780 | –2.638 (0.206) | –3.041 | –2.235 |
| Spatial Autoregressive Parameter | – | – | – | 0.613 (0.011) | 0.592 | 0.634 | – | | |
| Origin Spatial Filter | – | – | – | – | – | – | 1 (0.374) | 0.626 | 1.374 |
| Destination Spatial Filter | – | – | – | – | – | – | 1 (0.258) | 0.742 | 1.258 |
| Sigma Square | 1.724 | | | 1.442 | | | 1.614 | | |
| Pseudo-$R^2$ | 0.563 | | | 0.597 | | | 0.719 | | |
| Log-likelihood | –21,024.128 | | | –20,212.013 | | | –20,591.301 | | |
| Moran's *I* (*p*-value) | 0.193 (0.000) | | | -0.006 (0.939) | | | 0.145 (0.000) | | |
| Likelihood Ratio Test (*p*-value) | – | | | 1,624.23 (0.000) | | | – | | |

[a] measured in terms of patents (1985-1997) in the cited region *i*; [b] measured in terms of patents (1990-2002) in the citing region *j*; [c] the Hessian analytical estimate for the spatial autoregressive model is used to produce standard deviations of the estimates; [d] pre-test bias associated with stepwise selection of eigenvectors for constructing a spatial filter should be minimal here, given that spatial filters are constructed to account for residual spatial autocorrelation in a non-parametric context, and eigenvector selection is confirmed with simulation experiments; [e] because the spatial filters are linear combinations of eigenvectors, whose coefficients are estimated within a regression, their standard errors are computed as a linear combination of the squared standard errors of the individual eigenvectors

TABLE 3: Poisson Spatial Interaction Models: The Conventional Model and the Spatial Filter Model Specification with 23 Origin and 16 Destination Eigenvectors in the Spatial Filters

| | The Conventional Poisson Model [ML] | | | The Poisson Spatial Filter Model [ML] | | |
|---|---|---|---|---|---|---|
| | Estimates (Standard Error) | 95% Confidence Limits [c] | | Estimates (Standard Error) | 95% Confidence Limits [c] | |
| Constant | –8.983 (0.112) | –9.196 | –8.770 | –7.428 (0.125) | –7.674 | –7.183 |
| Origin Variable [a] | 0.857 (0.006) | 0.846 | 0.868 | 0.817 (0.007) | 0.803 | 0.831 |
| Destination Variable [b] | 0.835 (0.005) | 0.826 | 0.845 | 0.783 (0.006) | 0.771 | 0.795 |
| Geographical Distance | –0.258 (0.012) | –0.281 | –0.235 | –0.583 (0.019) | –0.619 | –0.546 |
| Country Border | –0.364 (0.017) | –0.396 | –0.332 | –0.330 (0.012) | –0.353 | –0.307 |
| Technological Distance | –0.584 (0.064) | –0.706 | –0.462 | –1.553 (0.077) | –1.703 | –1.402 |
| Scale | 1.508 | | | 1.356 | | |
| Origin Spatial Filter | – | – | – | 1 (0.412) | 0.588 | 1.412 |
| Destination Spatial Filter | – | – | – | 1 (0.202) | 0.798 | 1.202 |
| Sigma Square | 57.736 | | | 34.855 | | |
| Pseudo-$R^2$ | 0.764 | | | 0.858 | | |
| Log-likelihood | 40,919.915 | | | 51,973.184 | | |
| Moran's *I* (*p*-value) computed for Pearson residuals | 0.176 (0.000) | | | 0.112 (0.000) | | |

[a] measured in terms of patents (1985-1997) in the cited region *i*; [b] measured in terms of patents (1990-2002) in the citing region *j*; [c] Wald 95% confidence limits based on the large sample chi-square statistic with one degree of freedom, which are standard SAS output for GLMs

TABLE 4: Eigenvectors Used to Construct the Origin and the Destination Spatial Filters

| Eigenvector | Moran Coefficient | Log-Normal Approximation | | Poisson Approximation | |
|---|---|---|---|---|---|
| | | Origin | Destination | Origin | Destination |
| $\mathbf{E}_1$ | 1.11180 | 1.42741 | 0.93952 | 1.6058 | 1.1994 |
| $\mathbf{E}_2$ | 1.08506 | 0.81432 | 0.46405 | 1.5152 | 0.6819 |
| $\mathbf{E}_3$ | 0.99199 | **0** | **0** | -0.6757 | **0** |
| $\mathbf{E}_4$ | 0.98400 | -0.55321 | -0.46187 | -0.4873 | **0** |
| $\mathbf{E}_5$ | 0.93913 | -0.74278 | -0.32688 | -0.5387 | **0** |
| $\mathbf{E}_6$ | 0.88555 | **0** | **0** | 0.5759 | 0.3286 |
| $\mathbf{E}_7$ | 0.85454 | 0.48859 | 0.29019 | 0.8276 | 0.2586 |
| $\mathbf{E}_8$ | 0.81034 | 1.04167 | 0.64129 | 0.9957 | **0** |
| $\mathbf{E}_9$ | 0.78716 | 0.44417 | **0** | 0.8188 | -0.3755 |
| $\mathbf{E}_{10}$ | 0.74424 | 0.75294 | 0.28025 | 1.9399 | 0.9524 |
| $\mathbf{E}_{11}$ | 0.67839 | -0.41013 | -0.47184 | -0.2653 | -0.6082 |
| $\mathbf{E}_{12}$ | 0.65070 | **0** | 0.21490 | -0.3584 | -0.3663 |
| $\mathbf{E}_{13}$ | 0.62828 | 0.55030 | 0.30766 | **0** | **0** |
| $\mathbf{E}_{14}$ | 0.61328 | -0.37897 | -0.61055 | **0** | -0.5046 |
| $\mathbf{E}_{15}$ | 0.56651 | 0.50432 | 0.50477 | **0** | **0** |
| $\mathbf{E}_{16}$ | 0.53836 | -0.41588 | 0.36152 | -0.3242 | 0.2610 |
| $\mathbf{E}_{17}$ | 0.51970 | -0.37527 | **0** | -0.2258 | **0** |
| $\mathbf{E}_{18}$ | 0.51434 | 0.51669 | 0.28619 | 0.3962 | -0.2542 |
| $\mathbf{E}_{19}$ | 0.48048 | 0.91238 | 0.76615 | 0.7568 | 0.4077 |
| $\mathbf{E}_{20}$ | 0.44250 | -0.34447 | **0** | -0.4618 | **0** |
| $\mathbf{E}_{21}$ | 0.42450 | **0** | 0.20912 | 0.6667 | 0.5700 |
| $\mathbf{E}_{22}$ | 0.39132 | **0** | -0.36015 | -0.7781 | -0.7835 |
| $\mathbf{E}_{23}$ | 0.35404 | 0.43412 | **0** | 1.1175 | 0.6778 |
| $\mathbf{E}_{24}$ | 0.34998 | -0.56182 | -0.29573 | **0** | **0** |
| $\mathbf{E}_{25}$ | 0.32231 | -0.78679 | -0.43436 | -0.8296 | **0** |
| $\mathbf{E}_{26}$ | 0.29816 | **0** | -0.28623 | 0.3795 | 0.2891 |
| $\mathbf{E}_{27}$ | 0.28422 | **0** | -0.24553 | 0.3501 | **0** |

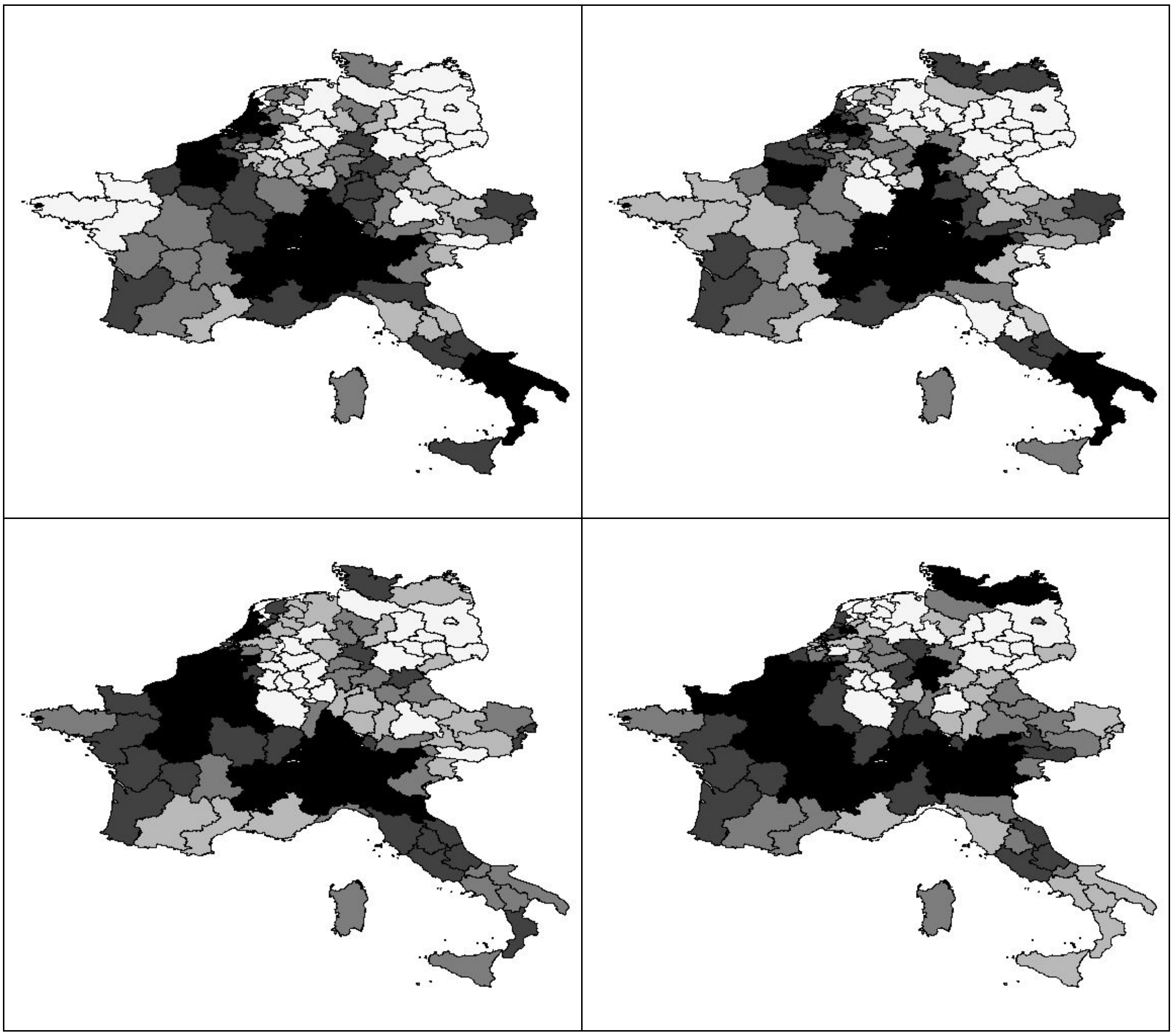

FIGURE 1: Spatial Filters for the Patent Citation Data, whose Relative Values are Proportional to the Darkness of the Gray Scale and are Constrained to Have a Sum of Squares Equal to 1 (Scale: Black – Very High; Dark Gray – High; Medium Gray – Medium; Light Gray – Low; White – Very Low).

(a) Top Left: Log-Linear Additive Model Origin Spatial Filter; Moran Coefficient = 0.778, Geary Ratio = 0.331.

(b) Top Right: Log-Linear Additive Model Destination Spatial Filter; Moran Coefficient = 0.731, Geary Ratio = 0.349.

(c) Bottom Left: Poisson Model Origin Spatial Filter; Moran Coefficient = 0.781, Geary Ratio = 0.328.

(d) Bottom Right: Poisson Model Destination Spatial Filter; Moran Coefficient = 0.751, Geary Ratio = 0.370

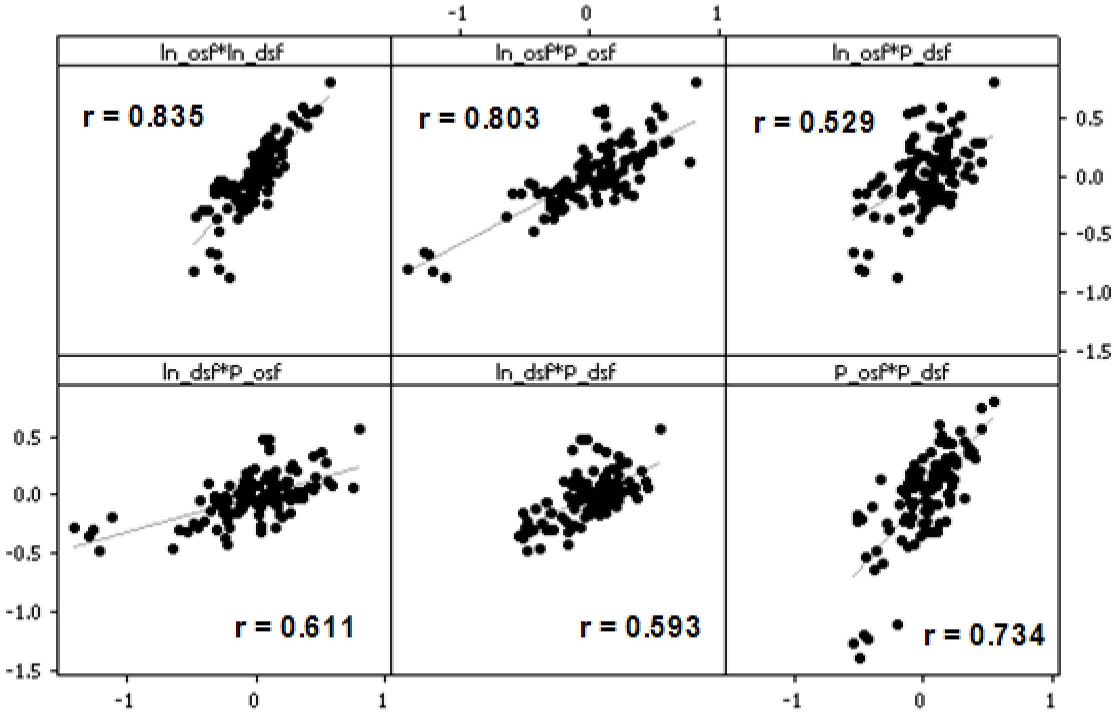


FIGURE 2: Scatterplots for Spatial Filter Cross-Correlations. The Corresponding Bivariate Correlations Range from 0.529 to 0.835.

(a) Top Left: Log-Linear Origin (ln_osf) Versus Log-Linear Destination (ln_dsf) Spatial Filters (SFs).

(b) Top Middle: Log-Linear Origin (ln_osf) Versus Poisson Origin (P_osf) SFs.

(c) Top Right: Log-Linear Origin (ln_osf) Versus Poisson Destination (P_dsf) SFs.

(d) Bottom Left: Log-Linear Destination (ln_dsf) Versus Poisson Origin (P_osf) SFs.

(e) Bottom Middle: Log-Linear Destination (ln_dsf) Versus Poisson Destination (P_dsf) SFs.

(f) Bottom Right: Poisson Origin (P_osf) Versus Poisson Destination (P_dsf) SFs