Spatial Methods in Econometrics: An Application to R&D Spillovers

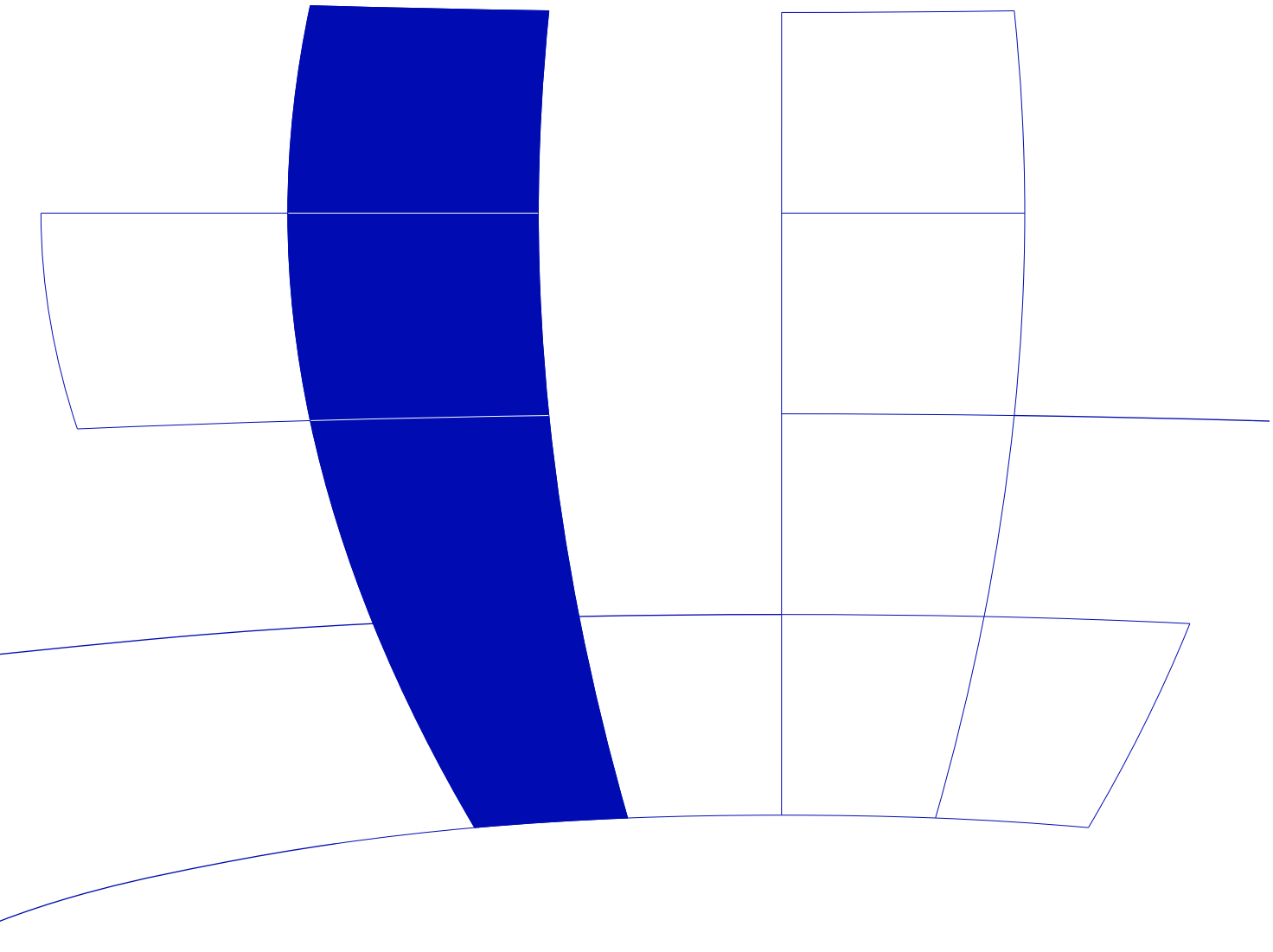
Daniela Gumprecht

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 26
December 2005

http://statmath.wu-wien.ac.at/



Spatial Methods in Econometrics: An Application to R&D Spillovers

Daniela Gumprecht

December 2005

Abstract

In this paper I will give a brief and general overview of the characteristics of spatial data, why it is useful to use such data and how to use the information included in spatial data. The first question to be answered is: how to detect spatial dependency and spatial autocorrelation in data? Such effects can for instance be found by calculating Moran's $\Im$, which is a measure for spatial autocorrelation. The Moran's $\Im$ is also the basis for a test for spatial autocorrelation (Moran's test). Once we found some spatial structure we can use special models and estimation techniques. There are two famous spatial processes, the SAR- (spatial autoregressive) and the SMA- (spatial moving average process) process, which are used to model spatial effects. For estimation of spatial regression models there are mainly two different possibilities, the first one is called spatial filtering, where the spatial effect is filtered out and standard techniques are used, the second one is spatial two stage least square estimation. Finally there are some results of a spatial analysis of R&D spillovers data (for a panel dataset with 22 countries and 20 years) shown.

Keywords: spatial dependency and autocorrelation, Moran's $\Im$, SAR- and SMA process, spatial filtering, S2SLS

1 Introduction: Spatial Analysis - What for?

One reason for using a spatial analysis is the exploitation of regional dependencies (so called information spillover) to improve statistical conclusions. The techniques used in the framework of a spatial analysis originally stem from geological and environmental sciences. These approaches have gained attraction in other fields through the dispersion of Geographical Information Systems (= GIS) and the increasing number of data with geographic coordinates. Especially in the social and economic sciences a growing number of applications can be witnessed. Currently there is a division between two views of what composes a spatial statistical analysis: (1) spatial prediction from continuous random fields (i.e. kriging), (2) processes developing over discrete neighbouring units (analogous to the ARIMA time series literature). However, data in an empirical analysis does not reflect this division, simply being a collection of measurements with attached geographical coordinates, in economics frequently called a spatial panel. One problem when analysing spatial data with standard statistical methods is the following: if the observations are spatially connected or spatially autocorrelated, the standard assumptions of uncorrelated error terms and uncorrelated observations and errors are violated which can lead to inconsistent and biased estimators. Therefore it is crucial to detect a spatial effect - if it is existent - and use adequate estimation techniques for the data.

What spatial dependency and spatial autocorrelation means, can be found e.g. in Fotheringham et al. (2002), they say about spatial dependency: "It (spatial dependency) is the extent to which the value of an attribute in one location depends on the values of the attribute in nearby locations." Griffith (2003) says about spatial autocorrelation: " It (spatial autocorrelation (...) is the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space (...)." Both of them relate their explanations to geographic space or

locations, nevertheless spatial dependency is not necessarily restricted to geographic space - e.g. one can look at spatial dependency in an economic context and use some specific measurement for the distances between locations. However spatial dependency is measured, positive spatial autocorrelation means that nearby values of a variable tend to be similar: high values are near high values, medium values near medium values, and low values near low values; negative spatial autocorrelation means that nearby values of a variable tend to be dissimilar: high values tend to be near low values, medium values near medium values, and low values near high values.

2 Spatial Data

Spatial data have the following characteristics: They contain attribute and locational information (so called georeferenced data). Spatial relationships are modelled with spatial weight matrices. Spatial weight matrices measure the similarities (e.g. neighbourhood matrices) or dissimilarities (distance matrices) between spatial objects.

2.1 Spatial Weight Matrix

Spatial relationships are represented with spatial weight matrices (also called spatial link matrices). A spatial link matrix $W = [w_{ij}]$ is a n by n matrix (n is the number of observations) with the following properties: $w_{ij} = 0$ if i and j are not spatially connected or if $i = j$ by definition, and $w_{ij} \neq 0$ if i and j are spatially connected. There are quite a lot of different forms of such spatial link matrices, which can measure similarity between objects, so called contiguity matrices, or dissimilarity between objects, so called distance matrices. Similarity and dissimilarity matrices are inversely related - the higher the connectivity, the smaller the distance and vice versa.

There are many different possibilities to measure the contiguity between objects, hence there are many different spatial link matrices in use. Neighbourhood matrices are symmetric, binary, n by n spatial link matrices with $w_{ij} = 1$ if two observations are neighbours and $w_{ij} = 0$ if not and if $i = j$. These matrices depend on the definitions of the neighbourship. The most commonly used definitions of neighbourhood are the Rook's criterion, where adjacent areas are neighbours if they share nonzero-length boundaries, the Bishop's criterion, where adjacent areas are neighbours if they share zero-length boundaries, and the Queen's criterion, where adjacent areas are neighbours if they share zero-length or nonzero-length boundaries. Spatial connectivity matrices are similar to neighbourhood matrices, but they are non-binary. They are symmetric n by n matrices, where the elements w_{ij} measure the intensity of the contiguousness. Similar to these connectivity matrices are distance matrices which are again non-binary symmetric n by n matrices, here the elements w_{ij} measure the distance between locations.

The original symmetric spatial link matrices are often converted by using coding schemes to cope with the heterogeneity which is induced by the different linkage degrees of the spatial objects. Tiefelsdorf (2000) defines the linkage degree of a spatial object i by the total sum of its interconnections with all other spatial objects, that is $d_i = \sum_{j=1}^{n} w_{ij}$. There are mainly three different coding schemes used: the globally standardized C-coding scheme, the row-sum standardized W-coding scheme, and the variance stabilizing S-coding scheme (Tiefelsdorf, 2000, p.29-30). E.g. in a row standardized version of a spatial link matrix the sum of each row is equal to one, the elements are simply calculated by $\frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$.

2.2 Spatial Stochastic Processes

Spatial stochastic processes are a functional relationship between a random variable at a given location and this same random variable at other locations. The covariance structure follows from the nature of the process, c.f. Anselin (1999). There are two famous spatial stochastic processes, the first one is called spatial autoregressive (SAR) process, the second one is called spatial moving average (SMA) process. Both use a spatial lag operator, which is a weighted average of random

variables at neighbouring locations (also called a spatial smoother): Wy, where W is a n by n spatial link matrix and y is a n by 1 vector of random variables. We consider centered variables y ($y = y^* - \mu\mathtt{1}$, where μ is the common mean of the random variables y_i^*, and $\mathtt{1}$ is a n by 1 vector of ones). Then these processes can be defined as
simultaneous SAR process:

$$y = \rho W y + \varepsilon = (I - \rho W)^{-1}\varepsilon \tag{1}$$

or SMA process:

$$y = \lambda W \varepsilon + \varepsilon = (I + \lambda W)\varepsilon \tag{2}$$

where I is an n by n identity matrix, ε are i.i.d. zero mean error terms with common variance σ^2, ρ is the autoregressive parameter (in most cases $|\rho| < 1$) and λ is the moving average parameter. The variance-covariance matrices for y are functions of the noise variance σ^2 and the spatial coefficient, ρ or λ. For the processes given in equation (1) and (2) the variance-covariance matrices are:

$$\Omega(\rho) = Cov(y, y) = E[yy'] = \sigma^2[(I - \rho W)'(I - \rho W)]^{-1} \tag{3}$$

for a simultaneous SAR process, and

$$\Omega(\lambda) = Cov(y, y) = E[yy'] = \sigma^2(I + \lambda W)(I + \lambda W)' \tag{4}$$

for a SMA process. For further processes and more detailed explanation see e.g. Anselin (1999) or Tiefelsdorf (2000).

2.3 Spatial Regression Models

In a standard linear regression model spatial dependency can be included as an additional regressor or in the error structure. According to these different possibilities to include spatial dependency in the model, there is a distinction between the spatial lag model and the spatial error model.

In the spatial lag model the spatially lagged dependent variable Wy is included as an additional regressor. This kind of model is used when the aim is to assess the existence and strength of spatial interaction.

$$y = \rho W y + X\beta + \varepsilon \tag{5}$$

where ε are i.i.d. disturbances. In this case the spatially lagged regressor is correlated with the error term and OLS estimation will give biased and inconsistent results due to the simultaneity bias.

The spatial error model is appropriate when spatial data are used and the potential influence of the spatial autocorrelation should be corrected. The spatial error model depends on the specification of the spatial structure. The most commonly used specification is the SAR process. The SAR error model has the following form:

$$y = X\beta + u,\ u = \rho W u + \varepsilon \tag{6}$$

where ε are again i.i.d. disturbances and W is a spatial link matrix. In this case OLS is unbiased but inefficient and the classical estimators for standard errors are biased. The SAR error model can be expressed as a spatial lag model with an additional set of spatially lagged exogenous variables and nonlinear constraints on the coefficients:

$$y = X\beta + \rho W y - \rho W X\beta + \varepsilon \tag{7}$$

The error variance covariance matrix is no longer $\sigma^2 I$ but the one given in equation (3): $E[uu'] = \Omega(\rho) = \sigma^2[(I - \rho W)'(I - \rho W)]^{-1}$. If the restriction is relaxed, the equation can be written as

$$y = X\beta + \rho W y + WX\gamma + \varepsilon \tag{8}$$

where $\gamma \neq -\rho \cdot \beta$ is allowed. The constraint $\gamma = 0$, which can be imposed, reduces the simultaneous AR-process to a spatial lag model, see equation (5). The constraint $\rho = 0$ reduces model (8) to a model with only a spatially lagged exogenous variable:

$$y = X\beta + WX\gamma + \varepsilon \tag{9}$$

2.4 Moran's $\Im$

One of the first questions that raises when analysts have to deal with georeferenced data is, whether there is a spatial effect existent or not. If not, i.e. the observations are spatially independent, there is no need for using special models or methods in the analysis. There are many different possibilities to test spatial autocorrelation, the most commonly used test is based on a statistic developed by Moran (1948, 1950a, 1950b). Spatial autocorrelation can be quantified and tested with Moran's $\Im$ statistic, which is defined as scale invariant ratio of quadratic forms in the normal distributed regression residuals:

$$\Im = \frac{\hat{\varepsilon}' \frac{1}{2}(W + W')\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}} \tag{10}$$

where $\hat{\varepsilon}$ are the normally distributed OLS residuals and W is a spatial link matrix. Expected value and variance of Moran's $\Im$, under the assumption of spatial independence are given by

$$E[\Im] = \frac{tr(MW)}{n-k}$$

$$Var[\Im] = \frac{tr(MWMW') + tr(MW)^2 + \{tr(MW)\}^2}{(n-k)(n-k+2)} - \{E[I]\}^2$$

where tr(.) denotes the trace operator, $M = I - X(X'X)^{-1}X'$ is the projection matrix, n is the number of observations and k is the number of regressors. Inference for Moran's $\Im$ is usually based on a normal approximation, using the standardized z-value:

$$z(\Im) = \frac{I - E[\Im]}{\sqrt{Var[\Im]}} \tag{11}$$

The z-transformed Moran's $\Im$ is for normal distributed residuals and well-behaved spatial link matrices under the assumption of spatial independence asymptotically standard normal distributed, see e.g. Tiefelsdorf (2000). With this z-value parametric hypothesis about the spatial autocorrelation level ρ can be tested. The z-values are simply compared with the well known critical values of the normal distribution.

2.5 Handling of Spatial Data

If there is some spatial dependency existent in the data, there are mainly two possibilities to deal with it. The first alternative is to filter out the spatial effect and use standard statistic methods for the analysis, e.g. use OLS for a regression model. The second one is to use some special spatial estimation techniques, e.g. the spatial two stage least technique or the maximum likelihood estimation technique.

2.5.1 Spatial Filtering

The basic idea of spatial filtering is to separate the regional interdependencies by partitioning the original variable into two parts: a filtered non-spatial (so called "spaceless") variable, and a residual spatial variable, and use conventional statistic techniques that are based on the assumption of spatially uncorrelated errors for the filtered ("spaceless") variables. There are different spatial filtering techniques available, one of these methods is based on the local spatial autocorrelation

statistic $G_i(\delta)$ from Getis and Ord (1992). Other techniques are based on an eigenfunction decomposition related to the global spatial autocorrelation statistic Moran's $\Im$ (Getis and Griffith, 2002). The first method is equally effective but computationally simpler and therefore described in more detail.

The $G_i(\delta)$ statistic was originally developed as a diagnostic to reveal local spatial dependencies that are not properly captured by global measures as the Moran's $\Im$ statistic. It is a distance-weighted and normalized average of observations $(x_1, ..., x_n)$ from a relevant variable x:

$$G_i(\delta) = \frac{\sum_j w_{ij}(\delta) x_j}{\sum_j x_j},\ i \neq j$$

where $w_{ij}(\delta)$ are the elements of a row-standardized spatial link matrix, δ is a locality parameter of the regional weighting scheme (typically δ is a distance parameter and observations which are further apart are down-weighted). Like the Moran's $\Im$, the $G_i(\delta)$ statistic can be standardized to z_{Gi} which is approximately Normal (0,1) distributed and can therefore be directly compared with the well-known critical values. The expected value of $G_i(\delta)$ represents the realization at location i when no spatial autocorrelation occurs.

$$E[G_i(\delta)] = \frac{\sum_i w_{ij}(\delta)}{(n-1)}$$

The ratio of this expected value and the original variable indicates the local magnitude of spatial dependence. The filtered observations are therefore given by:

$$\tilde{x_i} = \frac{x_i E[G_i]}{G_i(\delta)} = \frac{x_i \sum_i w_{ij}(\delta)/(n-1)}{G_i(\delta)}$$

The purely spatial component of the variable is then given by $(x_i - \tilde{x_i})$. If δ is chosen properly, the standardized value of $G_i(\delta)$ corresponding to $\tilde{x_i}$ is insignificant (demonstrated by Getis and Griffith, 2002). This means: filtering all variables (dependent and independent ones) in a regression model removes the spatial dependency and allows one to use a conventional regression model in which the parameters are estimated by ordinary least squares. A practical problem, when using this filtering technique is the choice of the structure of the spatial link matric W and the choice of the locality parameter δ of the regional weighting scheme. One possibility to model the distance decay is to use a negative exponential function, i.e.

$$w_{ij} = exp(-\delta d_{ij}),\ 0 \leq \delta \leq \infty$$

where d_{ij} denotes the (e.g. geographic) distance between the locations i and j. The choice of the structure does not have decisive impact on the outcomes, whereas the choice of δ is more delicate. Several methods to determine δ are discussed in Getis (1995), one of these methods to choose δ properly is: $\tilde{\delta} = Arg\ max_\delta \sum_i |z_{Gi}(\delta)|$.

2.5.2 Spatial Estimation

Another possibility to deal with spatially dependent data is to use spatial estimation techniques. In this case the spatial effect is not excluded from the data, like in the spatial filtering approach, but adequately included in the estimation. There are different estimation methods for spatial data, one can e.g. use the Maximum Likelihood technique (first outlined by Ord, 1975), or a Spatial Two Stage Least Squares method based on Instrumental Variable estimation (see e.g. Kelejian and Robinson, 1993; or Kelejian and Prucha, 1998), or based on a Method of Moments (Kelejian and Prucha, 1999).

Kelejian and Prucha (1999) suggest to use the following procedure to estimate a SAR model: For a spatial autoregressive model, given in equation (6): $y = X\beta + u$ and $u = \rho W u + \varepsilon$, the

covariance matrix is already given by equation (3): $\Omega(\rho, \sigma^2) = \sigma^2[(I - \rho W)'(I - \rho W)]^{-1}$. The auxiliary parameters ρ and σ^2 are estimated via the generalized method of moments (GMM) technique. The generalized moments estimator of ρ and σ^2 is a non-linear least squares estimator:

$$(\tilde{\rho}, \tilde{\sigma}^2) = Arg\ min_{\rho,\sigma^2}\ \{[\Gamma(\rho, \rho^2, \sigma^2)' - \gamma]'[\Gamma(\rho, \rho^2, \sigma^2)' - \gamma]\}$$

where $\rho \in [-a, a]$ with $a \geq 1$ and $\sigma^2 \in [0, b]$, they are elements of the vector (ρ, ρ^2, σ^2). Matrix Γ and vector γ are functions of the OLS residuals derived from the moment conditions, and $(\Gamma(\rho, \rho^2, \sigma^2)' - \gamma)$ can be viewed as a vector of residuals, for detailed specification see Kelejian and Prucha (1999, p.8). The parameter β of the regression model is then a feasible generalized least squares (FGLS) estimator:

$$\tilde{\beta} = [X'\tilde{\Omega}^{-1}X]^{-1}X'\tilde{\Omega}^{-1}y \tag{12}$$

where $\tilde{\Omega} = \Omega(\tilde{\rho}, \tilde{\sigma}^2)$.

3 An Application in Economics: R&D Spillovers

Coe and Helpman (1995) defend the theories of economic growth that treat commercially oriented innovation efforts as a major engine of technological progress and productivity growth (Romer, 1990; Grossman and Helpman, 1991). This means that on one hand innovation profit from knowledge that results from R&D spending and on the other hand innovation contributes to this stock of knowledge. Coe and Helpman (1995) claim that the productivity of a global economy depends on its own stock of knowledge as well as the knowledge of its trade partners and used a panel data set to study the extent to which a country's productivity level depends on the domestic and foreign stock of knowledge. They use the cumulative spending for R&D of a country to measure the domestic stock of knowledge, and the foreign stock of knowledge is calculated as import-weighted sum of cumulated R&D expenditures of the trade partners of the country. The importance of the R&D capital stock is measured by the elasticity of total factor productivity with respect to the R&D capital stock. Coe and Helpmans panel dataset contains 22 countries (21 OECD countries plus Israel) and 20 years (during the period from 1971 to 1990). The variables total factor productivity (TFP), domestic R&D capital stock (DRD) and foreign R&D capital stock (FRD) are constructed as indices with basis 1985 (1985=1). All data are available on the homepage of Elhanan Helpman (Helpman, 2003), which is accessible via the internet address:

`http://post.economics.harvard.edu/faculty/helpman/data.html`

3.1 Non-spatial Approach for the Analysis of R&D Spillover

In their paper Coe and Helpman (1995) used a variety of specifications to model the effects of DRD and FRD on TFP. To simplify the exposition only one of those is regarded here. The following conclusions, however, are not limited to this particular case but rather apply to all of the suggested models (for a more complete analysis see Gumprecht, 2003). The illustrative model contains three variables: total factor productivity (TFP) as the regressand, domestic R&D capital stock (DRD) and foreign R&D capital stock (FRD) as the regressors. The impact of domestic and foreign R&D expenditures is supposed to be the same for all countries. The equation - with regional index i and temporal index t - has the following form:

$$logF_{it} = \alpha_{it}^0 + \alpha_{it}^d logS_{it}^d + \alpha_{it}^f logS_{it}^f + \varepsilon_{it} \tag{13}$$

where F_{it} denotes total factor productivity (TFP), S_{it}^d domestic R&D capital stock (DRD) and S_{it}^f foreign R&D capital stock (FRD), which is defined as a bilateral import-share weighted average of the domestic R&D capital stocks of trade partners:

$$S_{it}^f = \sum_{i \neq j} b_{ijt} S_{jt}^d \tag{14}$$

where b_{ijt} denotes the bilateral import-shares of country i from country j in period t. Note that $b_{ijt} \neq b_{jit}$ and $\sum_j b_{ijt} = 1$. α_{it}^0 stands for the intercepts, which are allowed to vary across countries (for two reasons: first, there may exist country specific effects on productivity that are not included in the variables of this model, and second, all variables are transformed into index numbers, TFP is measured in the country specific currency whereas DRD and FRD are measured in U.S. dollars), α_{it}^d denotes the regression coefficient, which corresponds to the elasticity of TFP with respect to DRD, and finally α_{it}^f determines the elasticity of TFP with respect to FRD. According to standard practice in time series literature Coe and Helpman (1995) used a panel data model with fixed effects for their estimations.

Results of a non-spatial analysis

Coe and Helpman wanted to estimate the long-run relationship between TFP and the domestic and foreign R&D capital stocks. Given this and the fact that the series exhibit non-stationarity (as confirmed by respective tests), they estimate cointegrated equations. The OLS estimate of such a cointegrated equation is said to be "super consistent", that is, the estimate converge to the true parameter value much faster than in the case where the variables are stationary (Stock, 1987). Coe and Helpman (1995) give the following result for the model specified in equation (13), fixed effects model estimated via OLS:

$$\widehat{logF_{it}} = \alpha_{it}^0 + 0.097 logS_{it}^d + 0.0924 logS_{it}^f \tag{15}$$

This is the basic specification where the estimated coefficients on the domestic and foreign R&D capital stocks are constrained to be the same for all countries and the intercepts are allowed to vary between the countries (= fixed effects panel regression). Coe and Helpman (1995) took these estimation results, with both a positive regression coefficient as a confirmation of their hypothesis that TFP of a country depends on both domestic and foreign R&D capital stock. They did not calculate t- or p-values for the parameter estimators because using the standard method leads to biased results, and the asymptotic distribution of the t-values in the case of co-integrated panel data was not known at that time. Therefore this model was estimated once again, now using the Least Squares Dummy Variable (LSDV) method for the estimation and including the tests for the parameter estimators. The coefficients are the same as the ones from Coe and Helpman, the t-value for $\hat{\alpha}^d$ is 10.6834, the one for $\hat{\alpha}^f$ is 5.8673. Both coefficients are positive and significant; *pseudo* $R^2 = 0.5584$. The *pseudo* R^2 is calculated as the squared correlation between $\hat{y}_{it}$ and y_{it}. These results are given in column "Model 2" in the left part of Table 2.

Suggestions for improvement of Coe and Helpman's estimations came - among others - from Kao, Chiang and Chen (1999). They criticized (among other points) that in spite of the super consistency of the time-series estimator, the bias of the estimation can be quite substantial for small samples and there is no reason to assume that this bias becomes negligible by the inclusion of a cross section dimension in panel data. Kao, Chiang and Chen (1999) used different estimation methods for Coe and Helpman's international R&D spillovers regression and compared the empirical consequences from the different estimation methods. They claim that the dynamic OLS (DOLS) estimation is the best solution for this problem because in the given setting the DOLS estimator exhibits no bias and is asymptotically normal. The DOLS estimator is based on a regression including q_1 time lags and q_2 time leads of the regressors, therefore the number of time periods reduces from t to $(t - q_1 - q_2 - 1)$. For the R&D spillover model 2 lags and 1 lead were used for the calculation. The DOLS estimation of Coe and Helpman's Fixed Effects model, given in (13), can be found in column "Model 2" in the left part of Table 3.

As a second major issue, there are many debates in the panel data estimation literature, whether one should regard the region specific or other effects as random. This poses a valuable alternative to the fixed coefficient model. In the present context Müller and Nettekoven (1999) suggest a so called random coefficient model (the parameters are assumed to vary randomly around

a common mean) to analyse the R&D spillovers model of Coe and Helpman (1995) and conclude that, although this alternative specification is well compatible with the data, one astonishingly has to draw contradictory conclusions. Estimates for the random coefficient model differ decisively from the fixed effect model and especially the estimator of the foreign R&D expenditures even changes sign, although this is not statistically significant. Contrary to Coe and Helpman's conclusions, this model indicates that the foreign effect is not significant. See column "Model 2" in the left part of Table 4.

After a detailed examination of the model of Coe and Helpman (1995) and the various critics of it, the following changes and modifications were suggested by Gumprecht et al. (2004): use of a random coefficient model and use of DOLS technique for its estimation. The DOLS random coefficient estimation yields

$$\widehat{logF_{it}} = \alpha_{it}^0 + 0.3529 logS_{it}^d - 0.085 logS_{it}^f \tag{16}$$

The t-value for $\hat{\alpha}^d$ is 7.7946 and is significant, the one for $\hat{\alpha}^f$ is -1.1866 which is not significant; *pseudo* $R^2 = 0.9736$. The results of the panel cointegration model with random coefficient and dynamic regressors do not support Coe and Helpman's hypothesis, that the TFP of a country depends on domestic and foreign R&D knowledge (measured by the R&D expenditures). The effect of the knowledge of the trade partners of a country is not significant. It seems from (16) as foreign R&D do rather not affect the TFP of a country.

3.2 A Spatial Approach for the Analysis of R&D Spillover

The R&D spillover data can also be examined from a spatial point of view, because the countries can be regarded as regions and with an appropriate spatial link matrix a spatial analysis can be done. The first question that raises is: How to measure the distance or contiguity between the observations at different locations in an adequate way? In a global economy not the geographic distance but rather the trade intensity between two countries is relevant for R&D spillovers. To be consistent with Coe and Helpman, the bilateral import shares (of the year 1990) were used as a row-standardized spatial link matrix. This asymmetry in the spatial link matrix is no problem for the calculation of estimators but it is a problem if we want to define some kind of economic distances. Therefore a symmetric kind of trade intensity was used to measure the distance between two economies. In this context the symmetric trade intensity between two countries is defined as the average of the bilateral import-shares of these countries. The elements of the symmetric spatial connectivity matrix are simply calculated by:

$$w_{ij} = \frac{b_{ij} + b_{ji}}{2}$$

if $i \neq j$, and b_{ij} are the bilateral import-shares of country i from country j in period 1990, and by definition $w_{ij} = 0$ for $i = j$. It was assumed that the trade intensity is the same for all periods, this means the same spatial link matrix is used for all years. The distances between two countries are simply the inverse connectivity:

$$d_{ij} = \frac{1}{w_{ij}}$$

and by definition $d_{ii} = 0$. These distances can be used to produce a "trade-intensity" landscape by projecting the distances from the 21-dimensional space to the 2-dimensional space. For this projection a Multidimensional Scaling method is used: the squared sums of the distances between the original and the projected points (the points represent the countries) are minimized. This gives an approximation of all 231 distances between the 22 countries in the 2-dimensional space, and provides a quite good survey of the relationships in the data set, see Figure 1. Here the countries are quite evenly scattered, nevertheless some clusters can be identified, e.g. Australia, New Zealand and Israel are quite far apart from the rest of the countries, this means they have a small trade intensity with other countries and a relative high trade intensity within their group.

The U.S. are settled in the center, it can be interpreted in the way that the U.S. are an important trade partner for all countries. One thing to remember when looking at this landscape is, it is only an approximation and it can never show the true and exact distances.

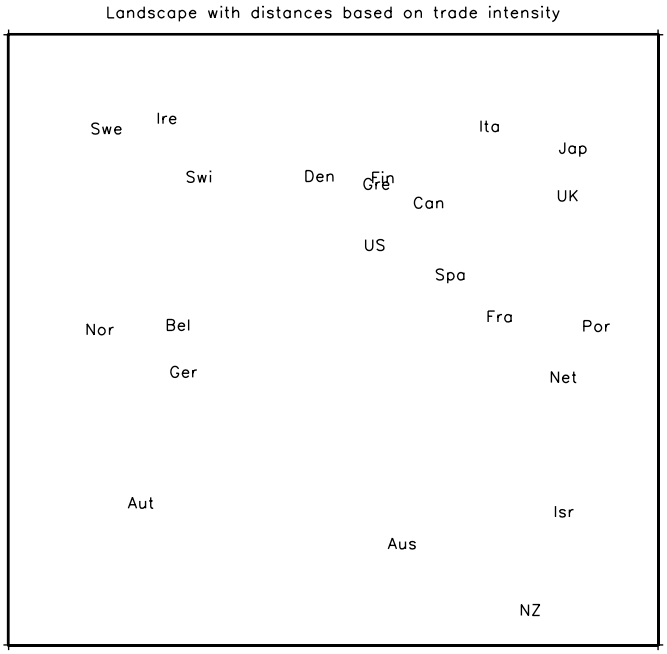


Figure 1: Landscape based on trade-intensities between the countries

Results of a spatial analysis

The spatial link matrix for the spatial regression model is the row-standardized bilateral import-shares matrix V from Coe and Helpmans dataset. The first steps in the spatial analysis are a fixed effect model without any foreign R&D spending and without any spatial structure:

$$logF_{it} = \alpha_{it}^0 + \alpha_{it}^d logS_{it}^d + \varepsilon_{it}$$

which gives an $\hat{\alpha}^d = 0.1362$ and to calculate and test Moran's $\Im$ for the residuals of this model for each period separately (see section 2.4). As spatial link matrix the bilateral import shares (matrix V) are used. Nearly all values are not significant (see Table 1), this means there is no global spatial effect in the error term. Even if there is no global spatial effect, local spatial effects can be included, and e.g. if there are positive and negative local spatial effects in the data these effects can compensate each other and the global Moran's $\Im$ test shows no significant global spatial effect. By the way, as there are only 20 countries in the dataset one should not put too much weight on the Moran's test because $z(\Im)$, given in equation (11) is only approximately normally distributed. Tiefelsdorf (2000, p. 97) recommends to use this test for datasets with at least 100 observations for exploratory statistical analysis and at least 200 observations for confirmatory statistical analysis. The assumption of some spatial effect is legitimate because the effect of FRD, which measures some kind of spatial dependency, is significant in the original model (13). Under the assumption of a spatial effect in the error term, one should use an adequate estimation technique for the SAR regression model, given in (6), e.g. the FGLS estimation from Kelejian and Prucha (1999), see section 2.5.2. This leads to similar results as the non-spatial analysis (the results of the non-spatial

model are given in column "Model 1" in the left part of Table 2, the results of the spatial model are given in column "Model 1" in the right part of Table 2). A fixed effect SAR model including the foreign R&D spending (original definition from Coe and Helpman) is estimated to compare the results with the ones from Coe and Helpman (1995), given in (15). For the results see Table 2, column "Model 2" in the right part.

Another alternative to analyse the R&D spillover dataset is the following: the foreign R&D spending can be seen as spatially lagged domestic R&D spending. To avoid the logarithms of the independent variables (as used by Coe and Helpman) and as all of the values of S_{it}^d are around one, a Taylor series approximation is used for the logarithm. This means:

$$logS = log(1) + \frac{1}{1}(S-1) + \frac{1}{2} - \frac{1}{S^2}(S-1)^2 + ...$$

Therefore $logS_{it}^d$ is substituted by the approximation: $logS_{it}^d \simeq S_{it}^d - 1$ and $log(\sum_{i \neq j} b_{ijt} S_{jt}^d)$ is substituted by $\sum_{i \neq j} b_{ijt} S_{jt}^d - 1$. This leads to the following model with a spatially lagged exogenous variable:

$$logF_{it} = \tilde{\alpha_{it}}^0 + \alpha_{it}^d S_{it}^d + \alpha_{it}^f \sum_{i \neq j} b_{ijt} S_{jt}^d + \varepsilon_{it} \tag{17}$$

where the fixed effects change from α_{it}^0 to $\tilde{\alpha_{it}}^0 = \alpha_{it}^0 - \alpha_{it}^d - \alpha_{it}^f$. In a first approach the fixed effect panel regression, given in equation (17) is estimated by LSDV, which gives positive and significant parameter estimators for the effect of DRD as well as FRD, see column "Model 3" in the left part of Table 2.

Under the assumption of a SAR error model, where a spatial effect is included in the error term, see equation (6), a FGLS estimation based on GM estimators of the autoregressive parameter $\hat{\rho}$ and the noise variance $\hat{\sigma}^2$ leads to:

$$\widehat{logF_{it}} = \tilde{\alpha_{it}}^0 + 0.1410 S_{it}^d - 0.0498 \sum_{i \neq j} b_{ijt} S_{jt}^d$$

The result diverges from the one of the non-spatial analysis, the effect of DRD on TFP is again positive and significant but the effect of FRD on TFP is negative and significant. On the other hand, the fit of this model yields worse *pseudo* $R^2 = 0.2685$, and gives a negative $z(\Im) = -0.5056$ (see column "Model 3" in the right part of Table 2). These values indicate an overcompensation of the spatial effect, due to the fact that the spatial dependency is included twice, once as the spatially lagged variable DRD and once in the error term.

However, as all of the critics of the original, non-spatial R&D spillovers analysis are also legitimate in the spatial context, all different more sophisticated models (namely the dynamic fixed effects, the static random coefficients and finally the dynamic random coefficients one) are estimated via OLS and FGLS. All results can be seen in Tables 2, 3, 4 and 5.

Now, the method of choice should again be the dynamic estimation of the random coefficient model. For the original variables DRD and FRD, the SAR error model should be used to correct for a spatial effect. The FGLS estimation yields

$$\widehat{logF_{it}} = \tilde{\alpha_{it}}^0 + 0.2522 logS_{it}^d - 0.0160 logS_{jt}^f$$

with *pseudo* $R^2 = 0.9564$, and estimates of the auxiliary parameter $\hat{\rho} = 0.3754$ and $\hat{\sigma}^2 = 0.0071$. Concerning the parameters, we have the same result as in the non-spatial case: a positive effect of DRD and no spillover effect of FRD. See column "Model 2" in the right partition of Table 5

Now, using the Taylor Series approximated variables instead of the original variables and running the FGLS estimation yields $\hat{\rho} = 0.7199$ and $\hat{\sigma}^2 = 0.0030$ and

$$\widehat{logF_{it}} = \tilde{\alpha_{it}}^0 + 0.0809 S_{it}^d + 0.0161 \sum_{i \neq j} b_{ijt} S_{jt}^d$$

with *pseudo* $R^2 = 0.9603$ and $z(\Im) = -0.1842$. Neither the effect of DRD nor the effect of FRD is significant. The unusual high value of $\hat{\rho}$ indicates overcompensation. This is caused by the fact, that the spatial effect is already included as spatially lagged independent variable and an additional spatial effect in the error term leads to an overcompensation - like in the case of the fixed effect model. For the results see column "Model 3" in the right partition of Table 5.

Thus, the preferred method is the DOLS estimation of the random coefficient model with Taylor Series approximated variables, which yields

$$\widehat{logF_{it}} = \tilde{\alpha_{it}}^0 + 0.1252 S_{it}^d + 0.1663 \sum_{i \neq j} b_{ijt} S_{jt}^d$$

with *pseudo* $R^2 = 0,9760$ and a standardized Moran's $\Im$ of $z(\Im) = -0.1908$. All results are shown in column "Model 3" in the left part of Table 5. This model has the best fit of all examined models and the result is in consensus with the original conclusions from Coe and Helpman (1995).

4 Conclusions

In general, one of the advantages of using spatial models and methods is, that a spatial dependency that might be inherent in empirical data, can be taken into account and treated correctly. And even if there is already a spatial dependency assumed, one can correct further spatial relationships that might not be captured by the variables in the model, by using a spatial error model. Especially when there is a spatial link matrix available, that describes the relationship between the observations, it is no problem to use adequate models and estimation techniques. The price one pays for running a spatial analysis is much less than the benefit one can earn by getting unbiased and consistent estimates.

In the R&D dataset an adequate spatial contiguity matrix is already given by the bilateral import shares, even if it is not used in this way in the original analysis. Anyway, it is quite simple to use these relationships for correcting an additional spatial dependency that is not properly captured by the given regressors. The aim of the analysis of the R&D spillover data set was to answer the question, whether domestic and foreign R&D spending have an effect on the total factor productivity of a country. Concerning domestic R&D spending the answer is quite obvious, all different reasonable estimation techniques (fixed effects- and random coefficients model) and both non-spatial and spatial approach lead to the conclusion that domestic R&D spending have a positive effect on the total factor productivity of a country. Concerning the foreign R&D spending the answer is not that clear, because different estimation techniques lead to different conclusions. Results for all different models can be found in Table 2, 3, 4 and 5. Some results support Coe and Helpman's (1995) conclusion that the foreign R&D spending have a positive effect on the total factor productivity, some do not. Nevertheless if one takes the DOLS estimation of the random coefficient model with a spatially lagged exogenous variable as the superior specification, the effect of foreign R&D spillovers seems to be existent.

Acknowledgments

I would like to thank Werner Müller from the University of Economics and Business Administration for his help within all steps during the preparation of the manuscript. I would also like to thank the Jubiläumsfonds of the OeNB (Austrian central bank) for funding the project "Common Structures in Spatial Econometrics" (project number 11052) in the course of which this paper arose.

Table 1: Moran's I for residuals of the Fixed Effects model with independent variable $logS_{it}$

Period	Moran's $\Im$	$z(\Im)$	Period	Moran's $\Im$	$z(\Im)$
1990	0.0532	1.3818	1980	-0.0323	0.2101
1989	-0.1777	-1.7822	1979	-0.0860	-0.5263
1988	-0.1184	-0.9693	1978	-0.0426	-0.0694
1987	-0.0607	-0.1789	1977	0.0623	1.5064
1986	-0.0330	-0.1789	1976	-0.1337	-1.1788
1985	-0.0258	0.2993	1975	-0.0584	-0.1480
1984	-0.0702	-0.3088	1974	-0.0473	0.0042
1983	-0.0101	0.5144	1973	0.0198	0.9241
1982	0.0506	1.3451	1972	-0.0053	0.5795
1981	0.0910	1.8989	1971	-0.0328	0.2026

$E(\Im) = -0.1476$ and $Var(\Im) = 0.0053$ for all periods

Table 2: Results for R&D Spillovers: Static Fixed Effects Model.

			Static Fixed Effects, LSDV			Static Fixed Effects, FGLS		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Original Variables	ln(drd)	$\hat{\alpha}$	0.1362	0.0970		0.1383	0.0961	
		t-ratio	21.3317	10.6834		22.2154	10.5393	
		p-value	0.0000	0.0000		0.0000	0.0000	
	ln(frd)	$\hat{\alpha}$		0.0924			0.0956	
		t-ratio		5.8673			6.1200	
		p-value		0.0000			0.0000	
Taylor Series Approximation	1+drd	$\hat{\alpha}$			0.0673			0.1410
		t-ratio			4.1483			6.1766
		p-value			0.0000			0.0000
	1+drd*V'	$\hat{\alpha}$			0.1787			-0.0498
		t-ratio			8.2235			-1.8678
		p-value			0.0000			0.0312
	Moran's $\Im$	z($\Im$)	0.2022	0.3613	0.2551	-0.0430	0.1409	-0.5056
	spatial p.	$\hat{\rho}$				0.1369	0.1636	0.2279
	variance	$\hat{\sigma}^2$				0.0025	0.0023	0.0019
	Model Fit	pseudo R^2	0.5218	0.5584	0.6240	0.5420	0.5799	0.2685
		pseudo adj.R^2	0.4966	0.5339	0.6032	0.5179	0.5566	0.2280

Table 3: Results for R&D Spillovers: Dynamic Fixed Effects Model.

			Dyn. Fixed Effects, LSDV			Dyn. Fixed Effects, FGLS		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Original Variables	ln(drd)	$\hat{\alpha}$	0.1461	0.1078		0.2124	0.0667	
		t-ratio	17.1916	13.6515		8.6564	2.3232	
		p-value	0.0000	0.0000		0.0000	0.0104	
	ln(frd)	$\hat{\alpha}$		0.0464			0.3831	
		t-ratio		3.7133			8.6208	
		p-value		0.0000			0.0000	
Taylor Series Approximation	1+drd	$\hat{\alpha}$			0.1887			0.0227
		t-ratio			27.2654			1.9333
		p-value			0.0000			0.0270
	1+drd*V'	$\hat{\alpha}$			0.0187			0.0800
		t-ratio			1.9464			5.9329
		p-value			0.0262			0.0000
	Moran's $\Im$	z($\Im$)	-0.5460	-0.3964	-0.6359	0.2580	0.4142	0.8376
	spatial p.	$\hat{\rho}$				-0.2180	-0.5336	-0.1424
	variance	$\hat{\sigma^2}$				0.0008	0.0005	0.0002
	Model Fit	pseudo R^2	0.8050	0.8758	0.9468	0.6533	0.8151	0.9001
		pseudo adj.R^2	0.7919	0.8671	0.9431	0.6301	0.8021	0.8931

Table 4: Results for R&D Spillovers: Static Random Coefficients Model

			Static Random Coeff., LSDV			Static Random Coeff., FGLS		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Original Variables	ln(drd)	$\hat{\alpha}$	0.2443	0.2874		0.1826	0.2061	
		t-ratio	9.0446	7.3441		9.3238	6.9246	
		p-value	0.0000	0.0000		0.0000	0.0000	
	ln(frd)	$\hat{\alpha}$		-0.0603			-0.0046	
		t-ratio		-0.9155			-0.0949	
		p-value		0.1802			0.4622	
Taylor Series Approximation	1+drd	$\hat{\alpha}$			-0.0205			0.1871
		t-ratio			-0.4279			3.3408
		p-value			0.3345			0.0005
	1+drd*V'	$\hat{\alpha}$			0.3787			-0.2104
		t-ratio			5.6590			-3.4582
		p-value			0.0000			0.0003
	Moran's $\Im$	z($\Im$)	0.4301	0.3630	0.4095	-0.2893	-0.2752	-0.3779
	spatial p.	$\hat{\rho}$				0.4977	0.5042	0.3669
	variance	$\hat{\sigma^2}$				0.0061	0.0075	0.0034
	Model Fit	pseudo R^2	0.9061	0.9135	0.9164	0.8792	0.8923	0.7054
		pseudo adj.R^2	0.9012	0.9087	0.9118	0.8728	0.8863	0.6891

Table 5: Results for R&D Spillovers: Dynamic Random Coefficients Model

			Dyn. Random Coeff., LSDV			Dyn. Random Coeff., FGLS		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Original Variables	ln(drd)	$\hat{\alpha}$	0.2431	0.3529		0.1631	0.2522	
		t-ratio	9.1011	7.7946		7.4995	6.5971	
		p-value	0.0000	0.0000		0.0000	0.0000	
	ln(frd)	$\hat{\alpha}$		-0.0850			-0.0160	
		t-ratio		-1.1866			-0.2727	
		p-value		0.1181			0.3926	
Taylor Series Approximation	1+drd	$\hat{\alpha}$			0.1252			0.0809
		t-ratio			2.2895			1.4394
		p-value			0.0113			0.0755
	1+drd*V'	$\hat{\alpha}$			0.1663			0.0161
		t-ratio			2.1853			0.2508
		p-value			0.0148			0.4011
	Moran's $\Im$	z($\Im$)	-0.1107	-0.1043	-0.1908	0.0566	-0.0600	0.1842
	spatial p.	$\hat{\rho}$				0.3208	0.3754	0.7199
	variance	$\hat{\sigma^2}$				0.0041	0.0071	0.0030
	Model Fit	pseudo R^2	0.9378	0.9736	0.9760	0.8963	0.9564	0.9603
		pseudo adj.R^2	0.9337	0.9717	0.9743	0.8894	0.9534	0.9575

References

Anselin, L. (1999). Spatial Econometrics. *Bruton Center, School of Social Siences, University of Texas at Dallas, Richardson, TX 75083-0688.*

Coe, D., & Helpman, E. (1995). International R&D Spillovers. *European Economic Review, 39*, 859-887.

Fotheringham, A., Brunsdon, C., & Charlton, M. (2003). *Geographically Weighted Regression.* New York: Wiley.

Getis, A. (1995). Spatial filtering in a regression framework: Experiments on regional inequality, government expenditures, and urban crime. In L. Anselin & R. Florax (Eds.), *New directions in spatial econometrics* (p. 172-188). Berlin: Springer.

Getis, A., & Griffith, D. (2002). Comparative spatial filtering in regression analysis. *Geographical Analysis, 34(2)*, 130-140.

Getis, A., & Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189-206.

Griffith, D. (2003). *Spatial Autocorrelation and Spatial Filtering* (1st ed.). Berlin: Springer.

Grossman, G., & Helpman, E. (1991). *Innovation and Growth in the Global Economy.* Cambridge, MA: MIT Press.

Gumprecht, D. (2003). Ein Panel Kointegrationsmodell für internationale Forschungs- und Entwicklungs Spillovers. *unpublished Master Thesis, University of Vienna.*

Gumprecht, D., Gumprecht, N., & Müller, W. (2004). Some current issues in the statistical analysis of spillovers. In G. Meier & S. Sedlacek (Eds.), *Spillovers and innovations: City, environment, and the economy, interdisciplinary studies in economics and management, vol. 4* (p. 51-70). Wien: Springer.

Helpman, E. (2003). *Professor Elhanan Helpman's data on the web.* Retrieved January 24, 2003, from Harvard University, Economics Department Web site: http://post.economics.harvard.edu/faculty/helpman/data.html.

Kao, C., Chiang, M., & Chen, B. (1999). International R&D spillovers: An application of estimation and inference in panel cointegration. *Oxford Bulletin of Economics and Statistics, 61*, 693-711.

Kelejian, H., & Prucha, I. (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances: a serious problem. *International Regional Science Review, 20*, 103-111.

Kelejian, H., & Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review, 40*, 509-533.

Kelejian, H., & Robinson, D. (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a country expenditure model. *Papers in Regional Science, 72*, 297-312.

Moran, P. (1948). The interpretation of statistical maps. *Biometrika, 35*, 255-260.

Moran, P. (1950a). Notes on continuous stochastic phenomena. *Biometrika, 37*, 17-23.

Moran, P. (1950b). A test for the serial dependence of residuals. *Biometrika, 37*, 178-181.

Müller, W., & Nettekoven, M. (1999). A panel data analysis: research and development spillover. *Economics Letters, 64*, 37-41.

Ord, J. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*, 120-126.

Romer, P. (1990). Endogenous technical change. *Jounal of Political Economy*, *98*, 71-102.

Stock, J. (1987). Asymptotic properties of least squares estimations of co-integrating vectors. *Econometrica*, *55*, 1035-1056.

Tiefelsdorf, M. (2000). *Spatial Autocorrelation and Spatial Filtering.* Berlin: Springer.