A general framework based on the theory of random field for the specification of spatial linear regression models[1]

Giuseppe Arbia [*]

Department of Business, Statistics, Technological and Environmental Sciences, G. D'annunzio University, Viale Pindaro, 42; I65124 Pescara (Italy)

This version: December, 2005

Abstract

Spatial econometrics has been so far concerned with the specification of regression models that seek to explain the spatial dependence inherent with spatial data in a similar fashion as the time series models try to capture serial dependence. In most of the cases the literature that treated both the purely spatial (cross-sectional) and the spatio-temporal (panel data) cases examined almost exclusively two models: the so-called spatial lag and spatial error models (Anselin and Bera, 1999; LeSage, 1999) whose probabilistic basis are quite weak. When alternatives to the basic linear regression model are considered the framework does not change in its substance. In this paper we wish to present the spatial linear regression model within a framework that is explicitly linked with the theory of random fields introduced by Yaglom (1962. This will allow us to present the problem in a more general way and to introduce a wider variety of possible specifications that can be adapted to the various empirical situations.

JEL Classification: C33 ; C43 ; C53

Keywords: Spatial correlation; Spatial models; Random fields; Regression

1 Introduction

Historically, spatial econometric methods directly stem from the developments that were introduced in the last century in the statistical literature to give consideration to the problem of the violation of the classical sampling model (the urn paradigm) with a big emphasis given to similarities due to spatial proximity. These developments were necessary to provide the right environment for the explanation of spatial diffusion phenomena like those frequently encountered in many applied fields like epidemiology, geography, agricultural studies, geology, image analysis, regional sciences, astronomy, archaeology and many others (for a review see Haining, 2003).

The spatial statistical techniques that are at the basis of spatial econometrics date back to about half a century ago and can be conventionally dated to a seminal paper by Peter Whittle (1954) followed by other important contributions of the same author (Whittle, 1962; 1963), by Bartlett (1964; 1975) and by Besag (1974) amongst the

[1] Paper prepared for presentation at the IGIER, Innocenzo Gasparini Institute for economic research, Università Bocconi, Via Salasco, 5; Milan, 5th December 2005. A previous version of the paper was presented as an Econometrics Seminar of the Faculty of Economics of Cambrdige University, Meade Room, Austin Robinson Building, February 2005.

[*] E-mail address: arbia@unich.it

others. The main results obtained led to a first codification in the seventies with some important publications like the celebrated books by Cliff and Ord (1973) and Bennett (1979). Other well-established textbooks followed in the eighties (Ripley, 1981; 1988; Upton and Fingleton, 1985; 1989; Griffith, 1988; Arbia, 1989 amongst the others), in the nineties (Haining, 1990; Cressie, 1991) and at the beginning of the new century (Haining, 2003).

The term "*spatial econometrics*" was coined by Jean Paelinck in the late seventies (see Paelink and Klaassen, 1979), and it is meant to represent "a blend of economic theory, mathematical formalisation, and mathematical statistics" (page vii). Apart from some remarkable example of the possibility of using spatial statistical methods in economics (like e. g. that of Granger in the sixties and seventies; see Granger, 1969, 1974), an important step forward in the historical development of the discipline is certainly constituted by the book published by Luc Anselin in the late eighties (Anselin, 1988).

However the integration between spatial methods and econometrics is still at an early stage. No mention is made, for instance, in some of the most recent introductory textbooks like Baltagi (1999), Berndt (1991), Davidson (2000), Dougherty (2002), Goldberg (1998), Griffith et. al (1993), Greene (2003), Gourieroux and Montfort (1995), Hayashi (2000), Hendry and Morgan (1997), Kennedy (1998), Peracchi (2001), Ruud (2000), Spanos (1999), Thomas (1997) and Verbeek (2000), Woolridge (2002b).

Kmenta (1997; p. 512) acknowledges the problem of non independence of statistical observations in space. However no mention is made to possible solutions to this problem. Maddala reports only a brief mention to the problem of spatial dependence amongst contiguous residuals of a linear regression (see Maddala, 2001, p. 228). A short mention can also be found in Johnston (1991; page 305), Kennedy (2003) and (Gujarati, 2003; p. 441).

The second edition of Baltagi's well known textbook on panel data includes a short discussion of the problems generated by treating spatial panels. (Baltagi, 2001a; pp. 195-197). Finally Woolridge (2002a) devotes a mention in the very first pages of his book to the issue of spatial dependence (Woolridge, 2002; p. 6) and devotes a short section that develops the idea a bit more thoroughly when dealing with the various forms of dependence amongst (Woolridge, 2002a; p. 134).

In fact it is not however until the years between the two millennia that we experience a growing interest of mainstream econometrics to spatial statistical methods, an interest that is witnessed by the increasing number of spatial econometric papers appeared in the econometric and applied economic journals. Amongst these important contribution are presented by Pinkse at al. (2002); Baltagi and Li (2001), Lee (2002), Conlwy and Topa (2002), Gelfand (1998), Bloomstein and Koper (1998), Pinkse and Slade (1998), Conley (1999), keleijan and Prucha (2001), Chen and Conley (2001), Baltagi at al. (2003) Keleijan and Prucha (2003), Giacomini and Granger (2003), Driscol and Kraay (1998) Bella and Bockstael (2000) Beron at al., (2003). In a recent thorough review Florax and De Vlist (2003) survey 11 articles in econometric journals and 30 in applied economic journals only in the period after the year 2000!

In all these applications the specification of spatial econometric models has been restricted to only two models: the so-called *spatial lag* model and *spatial error* model (Anselin and Bera, 1999; LeSage, 1999) whose probabilistic basis are quite weak. When alternatives to the basic linear regression model are considered (like e.g. in Beron and Vijverberg, 2004) the framework does not change. In this paper we wish to present the spatial linear regression model within a framework that is explicitly linked with the theory of random fields introduced by Yaglom (1957; 1961; 1962) and studied by Matern (1960) and Whittle (1954; 1963) this will allow us to present the problem in a

more general way and introduce a wider variety of possible specifications that can be adapted at the various empirical situations.

In Section 2 we will review the two basic spatial regression models used in spatial econometrics. In Section 3 we discuss the specification of a spatial linear regression model based on a specific random field: the so-called bivariate Conditional AutoRegressive model (CAR) introduced by Besag (1974). In Section 4 we derive the likelihood of this model and we construct the test statistics for the hypothesis of spatial dependence. In Section 5 we extend this to the multivariate case. Finally in Section 6 we draw some tentative conclusions and indicate the path for further generalization and developments.

2. Traditional model specifications in spatial econometrics

2.1 The spatial error model

One of the most commonly used alternatives to the classical a-spatial regression that can be found in the spatial econometric literature is the so-called spatial error model which represents a way to express formally the violation of the unrealistic condition of error independence across spatial units. The spatial error model is based on the hypothesis that, rather than modelling the whole set of variables involved in the specification as a vector random field, the problem of dependence can be eliminated by modelling the error component as a univariate random field. In this way the problem shifts from the direct modelling of the random field, say $\mathbf{Z}_i = (Y_i, \mathbf{X}_i{}^T)^T$, to the simpler problem of postulating a plausible form of spatial dependence for the errors. Of course, at least in principle, any random field model could be used for this aim (on random fields models see e.g. Besag, 1974; Cressie, 1993; for econometric applications see Arbia, 2005). However one of the most popular alternative to the white noise hypothesis adopted in the literature consists in postulating a SAR model for the non-systematic component. This formulation is referred in the literature as the "Spatial Error Model" or SEM (see Anselin, 1988Anselin and Bera, 1998; Anselin et al., 2004).

If we decide to model the non-systematic component of the model as a SAR random field, we need to redefine the linear regression model by supplementing the fundamental equation:

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + e_i \tag{43}$$

with the simultaneous autoregressive expression for the error term:

$$e_i = \rho \sum_{\substack{j=1 \\ i \neq j}}^{n} w_{ij} e_j + u_i \tag{44}$$

where u_i is a Gaussian spatial white noise, $w_{ij} \in \mathbf{W}$ and $\mathbf{W}$ a properly defined weight matrix based on the definition of neighbours

(on the definition of weight matrices see e.g.

In compact matrix notation Equations (43) and (44) become respectively

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e} \tag{45}$$

with **x** an n-by-k matrix of observations and

$$\mathbf{e} = \rho\mathbf{We} + \mathbf{u} \tag{46}$$

Since **u** is Gaussian white noise field, then **e** is also Gaussian. If we make these assumptions the problem is transformed to a situation where it is necessary to make inference on the vector of unknown parameters $\theta \equiv (\beta, \sigma^2, \rho)$ and, with the additional feature of normality introduced through the hypothesis on the random field **u,** the violation of the random sampling hypothesis is reduced to the study of the spatial autocorrelation which is present in the non-systematic component.

We know (see e.g. Cressie, 1993) that a SAR process is characterised by a variance-covariance matrix:

$$\mathbf{V} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma}(\mathbf{I} - \mathbf{B})^{-\mathbf{T}} \tag{47}$$

with $\mathbf{B} \equiv \{\rho_{ij}\}$, e $\rho_{ij} = \rho\, w_{ij}$ and $\mathbf{\Sigma}$ a diagonal matrix of generic element $\sigma_i^2 = \mathrm{Var}(u_i)$, or, in the case of constant variances:

$$\mathbf{V} = \sigma^2(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B})^{-\mathbf{T}} \tag{48}$$

From Equation (45) we then derive **e = y - Xβ** and, since **e** is assumed to be distributed as a Gaussian SAR random field, we easily obtain the likelihood function given by:

$$L(\rho, \sigma^2, \boldsymbol{\beta}; \mathbf{e}) = c(\mathbf{e})\left|\mathbf{V}(\rho, \sigma^2)\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{e}^T\mathbf{V}(\rho, \sigma^2)^{-1}\mathbf{e}\right\} \tag{49}$$

By substituting the expression **e** = **y** - **Xβ** and the explicit expression for the matrix **V** in this last equation and by taking the log we finally obtain the log-likelihood function:

$$l(\rho, \sigma^2, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) =$$
$$= c(\mathbf{y}, \mathbf{X}) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln\left|(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B})^{-\mathbf{T}}\right| - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X\beta})^T\left[(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B})^{-\mathbf{T}}\right]^{-1}(\mathbf{y} - \mathbf{X\beta}) \tag{50}$$

It is known (Anselin, 1988; Le Sage, 1999) that Equation (50) cannot be maximized analytically due to the high degree of nonlinearity in the parameters and the computational procedures employed in the available softwares (e. g. Lesage, 1998, Geoda,) are based on a partial likelihood function version of it. GLS estimators (Anselin, 1988) and approximate iterative techniques (Hordijk, 1974, Bartels, 1980 and Anselin, 1980) were also proposed in the literature.

Starting from the likelihood function reported in Equation (50) it is possible to build up tests of spatial independence. Indeed, if we consider that the non-systematic

component of the model obeys a SAR random field as postulated in this section, the test can be constructed considering the null hypothesis ρ=0 against the alternative hypothesis ρ≠0. Thus the likelihood-ratio test for the hypothesis of spatial independence can be easily obtained as:

$$LRT = -\ln\left|(\mathbf{I}-\mathbf{B})^{-1}(\mathbf{I}-\mathbf{B})^{-\mathrm{T}}\right| - \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\left[(\mathbf{I}-\mathbf{B})^{-1}(\mathbf{I}-\mathbf{B})^{-\mathrm{T}}\right]^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \quad (51)$$

The above test is distributed as a χ^2 with one degree of freedom. Two further tests of spatial independence based on the general expressions of the Wald test and of the Lagrange Multiplier were also proposed in the literature (Anselin, 1988; Arbia, 2005).

2.2 The Spatial Lag model

A second alternative, that is particularly popular in the spatial econometric literature, is not based on any specific random field model. It rather consists in a technical expedient that seeks to account for the spatial dependence amongst data by adding the spatially lagged value of *y* as an extra independent variable in a similar fashion to the inclusion of a serially autoregressive term in a time series context. This model is often referred as the *spatial lag* model (e.g. in Anselin and Bera, 1998), or as the *mixed regressive spatial autoregressive* model (Anselin, 1988), or finally, as *spatial autoregressive* or *SAR* model (LeSage, 1999). This last definition is, however, particularly misleading because with the acronym SAR in spatial statistics we indicate the Simultaneous AutoRegressive field.

The model can be written as the set of the following hypotheses:

$$y_i = \boldsymbol{\beta}^T\mathbf{x}_i + \rho\sum_{j=1}^{n} w_{ij}y_j + u_i \quad (52)$$

where $\boldsymbol{\beta}$ is the usual k-by-1 vector of regressive parameters, $\mathbf{x}_i$ the k-by-1 vector of explicative variables at site *i*, ρ an autoregressive parameter, $w_{ij} \in \mathbf{W}$ the elements of a (possibly row-standardized) weight matrix and **u** a Gaussian spatial random field such that:

$$f_u(\mathrm{u}|\mathbf{X}) \sim \mathrm{N}(0, \sigma^2\mathbf{I}_\mathrm{n}) \quad (53)$$

with $\mathbf{I}_\mathrm{n}$ an n-by-n identity matrix.

The presence of the spatially lagged term amongst the esplicative variables induces (unlike the time series analogous specification) a correlation between the error term and the lagged variable itself (see Anselin and Bera, 1999). Thus Ordinary Least Squares do not provide consistent estimators in this specification. It is important to remark that this specific result does not depend on assumption A2 and it is irrespective of the properties of the non-systematic component.

Let us write assumption (52) in a more compact matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \mathbf{u} \quad (54)$$

with **X** now indicating a n-by-k matrix of observations.

In what follows we will provide a probabilistic justification to Equation (54) using concepts of the random fields theory and we will introduce the derivation of the likelihood function for this alternative spatial regression model. Equation (54) can be interpreted as a non-stochastic linear regression where the matrix of observations **X** is assumed to be a fixed set of numbers. As a consequence of the lack of probabilistic assumptions on the dependent variables **X**, Equation (54) can be interpreted as a differential equation leading to a simultaneous Autoregressive (SAR) random field in which it appears the additional constant term **Xβ**. The introduction of this term only affects the expected value of the random field: neither its variance nor its structure of dependence. We can therefore exploit the results related to the SAR field (and specifically use the variance-covariance matrix defined in Equation (30)) after introducing the necessary amendments.

Formally, let us re-formulate our model as:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \qquad (55)$$

with a non-systematic component $\boldsymbol{\varepsilon}$ defined as $\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ and with **u** a spatial white noise field such that $\mathbf{u} \approx N(0;\sigma^2\mathbf{I}_n)$. As a consequence of the Gaussian assumption on the white noise component we have that $\boldsymbol{\varepsilon}$ is also Gaussian, but with non-zero expected values such that $\boldsymbol{\varepsilon} \approx N(\mathbf{X}\boldsymbol{\beta};\sigma^2\mathbf{I})$.

Let us now isolate the variable **y** in Equation (55) and reformulate the model as:

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$$

It is now possible to derive the properties of the random field **y** thus generated, and consequently the likelihood associated to a set of empirical observations, as follows. To start with the expected value of the random field **y** is given by:

$$E(\mathbf{y}) = E\left[(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}\right] = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} \qquad (56)$$

Secondly (from Equation (30)) the variance covariance matrix of the field is given by:

$$E(\mathbf{y}^T\mathbf{y}) = \mathbf{V}(\sigma^2,\rho) = \sigma^2(\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I} - \rho\mathbf{W})^{-T} \qquad (57)$$

From (56) and (57) we then derive the Gaussian log-likelihood of a sample of observations and obtain:

$$l(\sigma^2,\rho,\boldsymbol{\beta};\mathbf{y}) =$$

$$= c(\mathbf{y}) - \frac{1}{2}\ln\left|\mathbf{V}(\sigma^2,\rho)\right| - \frac{1}{2}\left[\mathbf{y} - (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\right]^T\mathbf{V}^{-1}\left[\mathbf{y} - (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\right] \qquad (58)$$

The determinant of the matrix $\mathbf{V}(\sigma^2,\rho)$, remembering Equation (57), is equal to $\left|\mathbf{V}(\sigma^2,\rho)\right| = \sigma^{2n}\left|(\mathbf{I}-\rho\mathbf{W})\right|^{-2}$. If we use this result and substitute (57) into (58) after some algebra we obtain:

$$l(\sigma^2,\rho,\boldsymbol{\beta};\mathbf{y}) =$$
$$= c(\mathbf{y}) - \frac{n}{2}\ln\sigma^2 - \ln\left|\mathbf{I}-\rho\mathbf{W}\right| - \frac{1}{2\sigma^2}\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]^T\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right] \qquad (59)$$

that represents the formal expression of a *spatial lag* linear regression model. (For details see Arbia, 2005).

Once the log-likelihood of the *spatial lag* linear regression model has been derived, we can maximize it in order to obtain maximum likelihood estimates of the parameters of interest. Unfortunately likewise the spatial error model Equation (4.86) cannot be maximized anlitically. Anselin (1988) proposed an approximate solution based on the idea of the profile likelihood and LeSage (1998) uses it to derive the informatic procedures for its computation.

From the log-likelihood thus derived it is immediate to define a test statistic for the hypothesis of spatial independence. In fact under the alternative hypothesis of a linear regression with an additional spatial lag the log-likelihood assumes the expression

$$l(\sigma^2,\rho,\boldsymbol{\beta};\mathbf{y}) =$$
$$= c(\mathbf{y}) - \frac{n}{2}\ln\sigma^2 - \ln\left|\mathbf{I}-\rho\mathbf{W}\right| - \frac{1}{2\sigma^2}\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]^T\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right] \qquad (60)$$

whereas, under the null hypothesis we have that $H_0:\ \rho = 0$ and, hence, the log-likelihood can be expressed as:

$$l_0(\sigma^2,\boldsymbol{\beta};\mathbf{y}) = c(\mathbf{y}) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]^T\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right] \qquad (61)$$

As a consequence the likelihood ratio test statistics after some straightforward algebra can be written as:

$$LRT = \left\{2\ln\left|\mathbf{I}-\rho\mathbf{W}\right| + \frac{1}{\sigma^2}\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]^T\left[(\mathbf{I}-\rho\mathbf{W})\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right] + \frac{1}{\sigma^2}\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]^T\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right]\right\} \qquad (62)$$

Equation (62), as it is known, is distributed asymptotically as a χ^2 random variable with one degree of freedom and can be used to test the hypothesis of spatial dependence within the framework of the linear regression model treated in this section.

3. The Bivariate CAR re-specification of the linear regression model

In this section we wish to present the spatial linear regression model within a framework that is explicitly linked with the theory of random fields. We claim this specification to be more probabilistically grounded than the two specifications

presented in Section 2 and that it can help in introducing a wider variety of possible models that can be adapted to the various empirical situations.

Let Y_i be the dependent variable of the model at location $\mathbf{s}_i$, $\mathbf{X}_i$ a vector of esplicative variables of dimension k (including a constant term), and $\mathbf{Z}(\mathbf{s}_i) = \mathbf{Z}_i = [Y_i, \mathbf{X}_i^T]^T$ a collection of random variables belonging to the vector random field $\{\mathbf{Z}(\mathbf{s}_i) \ \mathbf{s} \in \delta\}$ defined on the probability space (Ω, B, P(.)) which generates a set of data observed in n locations of coordinates ($\mathbf{s}_1$, $\mathbf{s}_2$ **.....**$\mathbf{s}_n$) on a continuous or discrete space. We shall assume that we want to build up a model which explains the behaviour of the economic variable Y_i in location $\mathbf{s}_i$ in terms of the behaviour of the other random variables $\mathbf{X}_i$ that constitute the random field. We shall indicate the random field $\mathbf{Z}(\mathbf{s}_i) = \mathbf{Z}_i = (Y_i, \mathbf{X}_i^T)^T$ and the sample observations $\mathbf{z}(\mathbf{s}_i) = \mathbf{z}_i = [y_i, \mathbf{x}_i^T]^T$ **.**

To start with let us summarize the basic assumptions on which is based a linear regression model specified in a conditional form.

The fundamental assumption is that the joint distribution of the random variables involved (both Y_i and the explicative variables $\mathbf{X}_i$) is multivariate Gaussian, that is:

$$\Phi = \left\{ f_{\mathbf{Z}_i}(\mathbf{Z}_i = \mathbf{z}_i, \boldsymbol{\theta}_i); \boldsymbol{\theta} \in \Theta; \Theta \subset \Re^{k+1} \right\} \qquad \text{and} \qquad Z_i \sim \text{MVN} \qquad (63)$$

where $\boldsymbol{\Phi}$ represents a parametric family of density functions, $\boldsymbol{\theta}$ the associated parametrization and Θ the parametric space

All other hypotheses concerning the probability model (PM) are consequences of this basic assumption. In fact from the joint normality it follows the normality of the conditional distributions, the linearity of the expected value (regression function), the constancy of the conditional variance (also called skedasticity function) and, finally, the hypothesis of spatial invariance of the parameters.

The statistical generating mechanism (GM) it is constituted by a systematic (forecastable) component and a non-systematic (non forecastable) component. In the basic linear regression model the two components are combined linearly. In particular, the systematic component is represented by the conditional expectation of y_i given $\mathbf{x}_i$ while the non-systematic component is simply the unexplained part of the model, measured by the difference between the observed value and the systematic component. From the linearity of the mean, assumed in the PM, we then have :

$$Y_i = \boldsymbol{\beta}^{\mathbf{T}} \mathbf{x}_i + u_i \qquad (64)$$

From the normality postulated in the PM, we also have that the only parameters of interest of the model are $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}; \sigma^2)$ $\theta \in \Theta \subset \Re^{k+1}$ and that $\mathbf{X}_i$ is weakly exogenous with respect to $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}; \sigma^2)$. Finally no restrictions are imposed a priori on the range (if deterministic) or on the distribution (if stochastic) of the parameters $\boldsymbol{\theta}$ and we assume that the observed data matrix is of full rank whatever the observed sample

Finally concerning the sampling model (SM) the basic assumption is that the data are drawn with a simple random criterion from the conditional distribution of Y given **X**. The SM assumption of simple random sampling is certainly the most important among those the linear regression model is based on. In practice none of the results related to the estimation and hypothesis testing remain valid if it is rejected on the basis of empirical data. More precisely the implications of the violation of the simple random sampling hypothesis on SM are that the OLS estimates of $\boldsymbol{\beta}$ and σ^2 are inefficient and inconsistent even if still unbiased. Moreover the sampling variances are biased and in most cases significantly underestimated. As a consequence the coefficient of

determination (R^2) as well as the test statistics t and F tend to be inflated leading to acceptance of the model more frequently than it should (Maddala, 2001).

If we retain the PM assumptions, but we consider the violation of the SM hypothesis, we need to re-specify our model as a vector Gaussian random field as described in Section 2 for which we have:

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \\ \\ \mathbf{X}_n \end{pmatrix} \sim MVN \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \\ \\ \boldsymbol{\mu}_n \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 & \mathbf{C}_{12} & & \mathbf{C}_{1n} \\ & \mathbf{V}_2 & & \\ & & & \\ & & & \\ \mathbf{C}_{n1} & & & \mathbf{V}_n \end{pmatrix} \tag{65}$$

where each $\boldsymbol{\mu}_i$ represents a k-by-1 vector of expected values at site *i*, and $\mathbf{V}_i$ and $\mathbf{C}_{ij}$ the matrices of cross-covariance and, respectively, of spatial autocovariance and spatial cross-covariance between pairs of sites defined by:

$$\mathbf{V}_i = \begin{pmatrix} \gamma_{11}(ii) & \gamma_{12}(ii) & & \gamma_{1k}(ii) \\ & & & \\ & & & \\ \gamma_{k1}(ii) & & & \gamma_{kk}(ii) \end{pmatrix}; \text{ and } C_{ij} = \begin{pmatrix} \gamma_{11}(ij) & \gamma_{12}(ij) & & \gamma_{1k}(ij) \\ & & & \\ & & & \\ \gamma_{k1}(ij) & & & \gamma_{kk}(ij) \end{pmatrix} \tag{66}$$

$\gamma_{kl}(ii)$ being the covariance between the random variables X_k and X_l at location i and $\gamma_{kl}(ij)$ the cross- covariance between X_k and X_l at locations i and j.

In practice, in order to obtain an operative sampling model, it is useful to limit ourselves to one of the random fields introduced in Section 2 of which the properties are known. Given the continuous nature of many economic variables, and due to its simplicity, the obvious choice is represented by the auto-normal field. However the framework presented here is general enough to allow the application to any other random fields in those cases when the phenomenon under study requires a different specification. In the present section we shall redefine the basic hypothesis of the linear regression model in the case of a non-independent sampling model by making explicit reference to the auto-normal random field.

5.2 Respecification of the hypotheses

A first way of redefining the linear regression model to keep into account the spatial nature of data is by redefining the probability model in such a way that the vector of random variables involved are assumed to obey an autonormal field, that is:

$$\Phi = \left\{ f_{\mathbf{Z}_i}(\mathbf{Z}_i = \mathbf{z}_i, \boldsymbol{\theta}_i); \boldsymbol{\theta} \in \Theta; \Theta \subset \Re^{2(k+1)} \right\} \qquad \text{and} \qquad \mathrm{Z_i} \sim \mathrm{MVN} \tag{67}$$

From this this, fundamental, assumption we can derive a series of consequences in terms of the probability model. First of all we have the normality of the conditional distributions:

$$f_{Y_i|\mathbf{X}_i Y_j}(y_i|\mathbf{X}_i = \mathbf{x}_i, Y_j = y_j; j \in N(i); \boldsymbol{\theta}_i) \sim \text{N} \qquad (68)$$

As a consequence of PM1 we have the linearity of the expected value that now (recalling the definition of a bivariate autonormal random field) can be expressed as

$$E(Y_i|\ \mathbf{X}_i\ = \mathbf{x}_i;\ Y_j = y_j;\ j \in N(i),\ \boldsymbol{\theta}) = \boldsymbol{\alpha}^T \Sigma_{i \neq j}\ w_{ij}\ \mathbf{x}_j + \boldsymbol{\beta}^T \mathbf{x}_i + \rho\ \Sigma_{i \neq j}\ w_{ij}\ y_j$$

where $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2, \ldots, \alpha_k)$, $\Sigma_{i \neq j}\ w_{ij}\ \mathbf{x}_j \equiv (\Sigma_{i \neq j}\ w_{ij}\ x_{1j}, \Sigma_{i \neq j}\ w_{ij}\ x_{2j}, \ldots, \Sigma_{i \neq j}\ w_{ij}\ x_{kj})^T\ \boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and ρ being the parameters to be estimated and, in particular, $\boldsymbol{\alpha}$ and ρ, are those that regulate the amount of spatial depenedence in the independent variables and, respectively in the lagged dependent variable. A second consequence of PM1 is the constancy of the conditional variance (homoskedasticity) that is:

$$\text{Var}(Y_i|\ \mathbf{X}_i\ = \mathbf{x}_i;\ Y_j = y_j;\ j \in N(i),\ \boldsymbol{\theta}) = \sigma^2 \qquad \forall\ \mathbf{x}_i \qquad (69)$$

We also derive the constancy of all parameters with respect to space, that is:

$$\boldsymbol{\theta}_i\ (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \rho, \sigma^2) = \boldsymbol{\theta} \qquad \forall i. \qquad (70)$$

From the statistical generation model we maintain the main hypothesis that the observations of the random variable Y_i are generated by a linear combination of a systematic component and a non-systematic. The systematic component (say μ_i) is constituted by the expectation of the variable Y in location *i* conditional upon the variables Y in the surrounding locations and on the variable X recorded at the same location *i* and in the surrounding locations that is:

$$\mu_i = E(y_i|\ \mathbf{X}_i = \mathbf{x}_i;\ Y_j = y_j;\ j \in N(i),\ \boldsymbol{\theta}) \qquad (71)$$

From the linearity of the conditional expectation derived from PM2 and substiting into ( ) we obtain

$$Y_i = \boldsymbol{\alpha}^T \Sigma_{i \neq j}\ w_{ij}\ \mathbf{x}_j + \boldsymbol{\beta}^T \mathbf{x}_i + \rho\ \Sigma_{i \neq j}\ w_{ij}\ y_j \qquad (72)$$

From the normality postulated in the PM we also have that the parameters of interst are $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \rho, \sigma^2)$. We also keep the hypothesis of weak exogeneity of $\mathbf{X}_i$ with respect to $\boldsymbol{\theta}$**.** The model considered here imposes some restrictions on the parameters that are connected with the symmetry of the variance-covariance matrix. In particular we have the following restrictions $\alpha_l \sigma_i^2 w_{ij} = \alpha_l \sigma_j^2 w_{ji}$ $\beta_l \sigma_i^2 w_{ij} = \beta_l \sigma_j^2 w_{ji}$ $\rho \sigma_i^2 w_{ij} = \rho \sigma_j^2 w_{ji}$ $\forall$ i, j, l ,that ensure that the variance covariance matrix of the field is symmetrical. In practical terms, due to the assumption of constant variance, these restrictions only requires the choice of a symmetrical **W** matrix, a condition that is almost invariably respected in the generality of geographical applications.

Furthermore in order to avoid singularity of the variance-covariance matrix, we need also to assume that $\mathbf{X} \equiv (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$ is a full rank (n-by-k) matrix for all the observed values of the random variables **X**

Finally in the sampling model we shall assume that $\mathbf{y} \equiv (y_1, y_2, \ldots, y_n)'$ is a sample drawn from a stationary random field characterised by conditional distribution

$$f_{Y_i|\mathbf{X}_i y_j}(y_i|\mathbf{X}_i = \mathbf{x}_i, Y_j = y_j; j \in N(i); \boldsymbol{\theta}_i) \quad (73)$$

is such that $\mathbf{y} \equiv (y_1, y_2, ..., y_n)'$ is independently drawn from $f_{Y_i|\mathbf{X}_i y_j}(y_i|\mathbf{X}_i = \mathbf{x}_i, Y_j = y_j; j \in N(i); \boldsymbol{\theta}_i)$

The model thus specified will be referred to as the *multivariate CAR spatial linear regression model* and notwithstanding its sound probabilistic foundations it was never exploited in the applied spatial econometric literature.

5.3 The likelihood of a bivariate CAR spatial linear regression model

In this section we shall limit ourselves to only bivariate autonormal fields $\mathbf{Z}_i = (Y_i, X_i)'$. The extension to higher dimensional fields is treated in section 7.

If Y and X are jointly distributed as a bivariate random field we know, from Section 2.4.2.8, that we can express the conditional expected value of Y_i as

$$E[Y_i|X_j; Y_j; X_i] = {}_Y\mu_i + \sum_{i \neq j} \alpha_{ij}(X_j - {}_X\mu_j) + \beta(X_i - {}_X\mu_i) + \sum_{i \neq j} \rho_{ij}(Y_j - {}_Y\mu_j) \quad (74)$$

To simplify a notation that runs the risk of becoming too cumbersome, we shall express, without loosing in generality, each random variable as a deviation from the respective expected value. Equation (74) now becomes

$$E[Y_i|X_j; Y_j; X_i] = \sum_{i \neq j} \alpha_{ij} X_j + \beta X_i + \sum_{i \neq j} \rho_{ij} Y_j \quad (75)$$

with $\alpha_{ij} = \alpha w_{ij}$, $\rho_{ij} = \rho w_{ij}$, α, β and ρ parameters and $w_{ij} \in \mathbf{W}$ the elements of a contiguity matrix properly defined.

This expression of the expected conditional value provides an operational form for the systematic component of the model. Therefore if we define the non-systematic component as:

$$u_i = Y_i - E[Y_i|X_j; Y_j; X_i] = Y_i - \sum_{i \neq j} \alpha_{ij} X_j - \beta X_i - \sum_{i \neq j} \rho_{ij} Y_j \quad (76)$$

and we redefine the data generation statistical model as the sum of the systematic and the non-systematic component

$$Y_i = E[Y_i|X_j; Y_j; X_i] + u_i \quad (77)$$

we have

$$Y_i = \alpha \sum_{i \neq j} w_{ij} X_j + \beta X_i + \rho \sum_{i \neq j} w_{ij} Y_j + u_i \quad (78)$$

In this way each observation of the random variable Y at location *i* is expressed as a function of the observation of the variable X in the same location (as in a standard

linear regression model), but also of the spatially lagged values of the variable X and of the variable Y, or, in other words, as a function of the mean of the neighbouring values for both variables.

We know that a bivariate CAR field has a variance-covariance matrix given by Equation (25). Therefore it is immediate to redefine the likelihood function of the sample as a bivariate Gaussian density function with $\mathbf{Q} = \mathbf{Q}(\alpha, \beta, \rho, \sigma_x^2, \sigma_y^2)$ as a variance-covariance matrix, that is:

$$L(\alpha, \beta, \rho, \sigma_x^2, \sigma_y^2; \mathbf{z}) = c(\mathbf{z}) \frac{1}{2\pi} |\mathbf{Q}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{Z}^T \mathbf{Q}^{-1} \mathbf{Z} \right\} \tag{79}$$

and, consequently, the log-likelihood is defined as:

$$l(\alpha, \beta, \rho, \sigma_x^2, \sigma_y^2; \mathbf{z}) = c(\mathbf{z}) - \frac{1}{2} \ln |\mathbf{Q}| - \frac{1}{2} \mathbf{Z}^T \mathbf{Q}^{-1} \mathbf{Z} \tag{80}$$

This expression is highly non linear in the parameters and therefore can only be maximized by using numerical algorithms in order to obtain maximum likelihood estimators. The likelihood thus derived constitute also the basis to build various hypothesis testing procedures as we will show in the next section.

6. Hypothesis testing in the bivariate CAR spatial regression model

Now that we have fully specified the alternative hypothesis to the null hypothesis of spatial independence in the regression model, we are in the position to apply these general procedures to this particular instance. In fact, once the model is respecified as indicated in the preceding section 4.3.3.1, the system of hypotheses can be explicitly obtained by contrasting the null hypothesis $H_0 : \alpha_0 = \rho_0 = 0$, with the alternative hypothesis $H_1 : \alpha \neq 0; \rho \neq 0$

In terms of the likelihood we have, under the null:

$$L(\beta_0, \sigma_{x,0}^2, \sigma_{y,0}^2; \mathbf{z}) = c(\mathbf{z}) |\mathbf{Q}_0|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{Z}^T \mathbf{Q}_0^{-1} \mathbf{Z} \right\} \tag{81}$$

with $\mathbf{Q}_0$ the variance-covariance matrix that in this case is $\mathbf{Q}_0 = \mathbf{Q}(\beta_0, \sigma_{x,0}^2, \sigma_{y,0}^2)$ $= \left\{ \mathbf{I} * - \begin{bmatrix} 0 & \boldsymbol{\beta}\mathbf{I} \\ \boldsymbol{\beta}\mathbf{I} & 0 \end{bmatrix} \right\}^{-1} \boldsymbol{\Sigma}$ implying independence between observations.

As a consequence the log-likelihood can be expressed as:

$$l(\beta_0, \sigma_{x,0}^2, \sigma_{y,0}^2; \mathbf{z}) = c(\mathbf{z}) - \frac{1}{2} \ln |\mathbf{Q}_0|^{-\frac{1}{2}} - \frac{1}{2} \mathbf{Z}^T \mathbf{Q}_0^{-1} \mathbf{Z} \tag{82}$$

In contrast, under the alternative hypothesis of a bivariate CAR random field, the likelihood assumes the expression

$$L(\alpha,\beta,\rho,\sigma_x^2,\sigma_y^2;\mathbf{z}) = c(\mathbf{z})\left|\mathbf{Q}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{Z}^T\mathbf{Q}^{-1}\mathbf{Z}\right\} \tag{83}$$

with $\mathbf{Q} = \mathbf{Q}(\alpha,\beta,\rho,\sigma_x^2,\sigma_y^2)$ provided by Equation (25). Consequently the log-likelihood under the alternative hypothesis becomes:

$$l(\alpha,\beta,\rho,\sigma_x^2,\sigma_y^2;z) = c(\mathbf{z}) - \frac{1}{2}\ln\left|\mathbf{Q}\right| - \frac{1}{2}\mathbf{Z}^T\mathbf{Q}^{-1}\mathbf{Z} \tag{84}$$

A test of spatial independence therefore can be easily derived from the likelihood ratio test criterion:

$$LRT = -2\left[\ln L(\boldsymbol{\theta}_0;X) - \ln L(\boldsymbol{\theta}_1;X)\right] = -2\left[l(\beta_0,\sigma_{x,0}^2,\sigma_{y,0}^2;\mathbf{z}) - l(\alpha,\beta,\rho,\sigma_x^2,\sigma_y^2;\mathbf{z})\right] =$$

$$= \left[\ln\left|\mathbf{Q}_0\right|^{-\frac{1}{2}} + \mathbf{Z}^T\mathbf{Q}_0^{-1}\mathbf{Z} - \ln\left|\mathbf{Q}\right| - \mathbf{Z}^T\mathbf{Q}^{-1}\mathbf{Z}\right] \tag{85}$$

$\boldsymbol{\theta}_0 \equiv \left(\beta_0,\sigma_{x,0}^2,\sigma_{y,0}^2\right)$,and $\boldsymbol{\theta}_1 \equiv (\alpha,\beta,\rho,\sigma_x^2,\sigma_y^2;\mathbf{z})$. Equation (85) represents the formal expression of a likelihood ratio test in the case of a bivariate Gaussian field.

An alternative way of obtaining a test statistics for the hypothesis of independence in a bivariate CAR linear regression model is to consider the fact that the data generation mechanism assumes different formulations if the random sampling hypothesis is veryfied or not. Indeed, when such a condition is respected the data generation model can be expressed as:

$$Y_i = \beta x_i + u_i \tag{86}$$

whereas in the second case, if we postulate a bivariate autonormal random field as an alternative, it is conversely given by:

$$Y_i = \alpha\sum_{i\neq j} w_{ij}x_j + \beta x_i + \rho\sum_{i\neq j} w_{ij}y_j + u_i \tag{87}$$

Consequently, a simple test for the independence hypothesis can be constructed using the general test statistic introduced in Section 3.5:

$$\mathrm{F} = \frac{n-r}{r}\frac{\left[RSS_1 - RSS_0\right]}{RSS_0} \tag{88}$$

with n the number of observations, r the number of constraints under the null hypothesis, RSS_1 the residuals sum of squares of model (87) and RSS_0 the residual sums of squares of model (86). Expression (88) has a distribution of a Student central F-distribution with r and n-r degrees of freedom.

7. Likelihood of a multivariate CAR spatial linear regression model

Let us now consider the more general case of a multivariate linear regression model and the extension of the derivation of the likelihood considered in Equation (80).

From Equation (27), the likelihood of the multivariate CAR model is given by:

$$L(\boldsymbol{\mu},\Omega;\mathbf{z}) = c(\mathbf{z})|\Omega|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \Omega^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\} \tag{89}$$

and, hence, the log-likelihood by

$$l(\boldsymbol{\mu},\Omega;\mathbf{z}) = c(\mathbf{z}) - \frac{1}{2}\ln|\Omega| - \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \Omega^{-1}(\mathbf{z}-\boldsymbol{\mu}), \tag{90}$$

Equation (90) can be expressed in a different way by assuming the following reparametrizations for $\boldsymbol{\mu}_i$, $\mathbf{C}_{ij}$ and
$\mathbf{V}_i : \forall i \ : \ \boldsymbol{\mu}_i = \boldsymbol{\lambda}$ and $\mathbf{V}_i = \mathbf{V} = \text{diag}\left(\upsilon_1^2 \quad \cdots \quad \upsilon_p^2\right), \ \forall i, j \ : \ j \neq i \Rightarrow \mathbf{C}_{ij} = \boldsymbol{\Phi} w_{ij}$,
where $\boldsymbol{\Phi}$ is a k-by-k symmetric matrix and $w_{ij} \in W$ is the generic element of a weights matrix .

Under these assumptions, we can write

$$\Omega^{-1} = \mathbf{I}_n \otimes \mathbf{V}^{-1} - \mathbf{W} \otimes \mathbf{V}^{-1}\boldsymbol{\Phi}, \tag{91}$$

so that the log-likelihood becomes:

$$\begin{aligned} l(\boldsymbol{\lambda},\boldsymbol{\Phi},\mathbf{V};\mathbf{z}) &= c(\mathbf{z}) + \frac{1}{2}\ln\left(\det\left(\mathbf{I}_n \otimes \mathbf{V}^{-1} - \mathbf{W} \otimes \mathbf{V}^{-1}\boldsymbol{\Phi}\right)\right) \\ &- \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\lambda})^T \mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\lambda}) \\ &+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j \ : \ j \neq i}(\mathbf{x}_i - \boldsymbol{\lambda})^T \mathbf{V}^{-1}\boldsymbol{\Phi} w_{ij}(\mathbf{x}_j - \boldsymbol{\lambda}). \end{aligned} \tag{92}$$

and can be used in the estimation and hypothesis testing procedures.

Essential references

Anselin L., 1988. *Spatial Econometrics, Methods and Models*, Kluwer Academic, Boston.

Anselin L.,2001a. Spatial econometrics, in B.H. Baltagi (ed.), *A Companion to Theoretical Econometrics*, Basil Blackwell, Oxford: 310-330.

Anselin L. and Bera A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics, in A. Ullah and D. Giles (eds.), *Handbook of Applied Economic Statistics*, Marcel Dekker, New York: 237-289.

Anselin L. and Florax R., 1995. *New Directions in Spatial Econometrics*, Springer-Verlag, New York.

Anselin L., Florax R.J.G.M. and Rey S., 2004. *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Springer-Verlag, New York.

Anselin L., Le Gallo J. and Jayet H., 2004b. Spatial panel econometrics, in L. Matyas and P. Sevestre (eds), *The Econometrics of Panel Data*, Third edition, Kluwer Academic Publishers, Dordrecht.

Arbia G., 1989. *Spatial Data Configuration in the Statistical Analysis of Regional Economics and Related Problems*, Kluwer Academic Publishers, Dordrecht.

Arbia G., 2006. Spatial Econometrics: Statistical foundations and Applications to Regional Convergence, Springer-Verlag, Heidelberg, (in press).

Arbia G. and Espa G., 1996. *Statistica Economica Territoriale*, CEDAM, Padua.

Baltagi B.H., 2001. *Econometric Analysis of Panel Data*, (second edition), John Wiley and Sons, Chichester, England.

Baltagi B.H. and Li D., 1999. Double-length artificial regressions for testing spatial dependance, *Econometric Review*, 20: 31-40.

LeSage J., 1999. *Spatial Econometrics*: *The Web Book of Regional Science*, Regional Research Institute, West Virginia University, Morgantown, WV.

Paelinck J.H.P. and Klaassen L.H., 1979. *Spatial Econometrics*, Gower, Westmead, Farnborough.

Yaglom A.M. 1962. *An Introduction to the Theory of Stationary Random Functions*, Prenctice-Hall, Englewood Cliffs, New Jersey.