Report 99-002

On spatial regression models: implementation in Matlab

Cheikh A.T.Diack

ISSN: 1389-2355

On spatial regression models: implementation in Matlab

Cheikh A. T. DIACK
EURANDOM, P.O. Box 513, 5600 Eindhoven,
The Netherlands. E-mail:diack@eurandom.tue.nl

January 13, 1999

Abstract

The aim of this paper is to present a class of regression models adapted for spatial data and to discuss its implementation in Matlab.

1 INTRODUCTION

Data coming from a geographical background often exhibit spatial dependence. For example, the evaluated measures of economic activities between two neighbouring districts are often linked. Indeed, the presence of some economic activity in one of these districts may explain the absence of this activity or the presence of a complementary activity in the other. Moreover, rich districts might have attractive effects over their neighbours while poor districts might have repulsive effects. Thus, there are possible interactions between two neighbouring districts: this is the contiguity notion.

Besides, the presence in the physical environment of some fertile lands may result in an over-intensification of the agriculture over these areas. Hence, the position in geographical space is an important characteristic for the spatial objects which are analysed.

One of the main objectives in spatial analysis is to identify the nature of relationships that exist between variables taking into account their covariance structure. For example, in the presence of heteroscedasticity -that is often the case for geographical data- the frequently applied ordinary least squares estimator is inefficient since the estimator of the residual variance is biased and the values of the estimated R^2 are inflated (see Anselin et al. 1995). Hence the tests of spatial dependence are affected. To overcome this problem, one may use the weighted least squares estimator (see Haughton and Haughton 1997).
However, one of the characteristics of models for spatial data is that the errors may be spatially autocorrelated. In this case, the weighted least squares estimator may be inefficient since this one does not incorporate spatial effects.

There is a wide class of possible models which incorporate spatial effects of this nature, and the challenge is to choose the appropriate one.
A conventional approach is to assume that the model has an autoregressive-error structure as in Florax and Folmer (1992). In this class of models, the most general has the following form:

$$\begin{aligned} y &= \psi W_1 y + X\beta + W_2 X^* \rho + \varepsilon, \\ \varepsilon &= \lambda W_3 \varepsilon + \mu, \end{aligned}$$

where y is the $(n \times 1)$ vector of observations on the dependent variable. W_1, W_2 and W_3 are priori specified $(n \times n)$ spatial weights matrices. X is an $(n \times p)$ matrix of observation on the independent variables, X^* the $(n \times (p-1))$ matrix of explanatory variables with constant term deleted. ψ is the autocorrelation coefficient, β the $(p \times 1)$ vector of coefficients of the non weighted independent variables, ρ the $((p-1) \times 1)$ vector of crosscorrelation coefficients, λ the coefficient of the autoregressive error term. μ is a vector of random errors with $E(\mu) = 0$, $E(\mu\mu') = \sigma_\mu^2 \Omega$ and Ω a positive diagonal matrix.

Even though the underlying economic theory provides little guidance on the choice of model, it is very important to incorporate an appropriate specification of spatial effects. This is a real challenge in spatial regression.

The structure of the paper is as follows. In the next section, we present the different regression models which can be derived from the above model by introducing constraints. After, we will briefly discuss the choice between these models. Section 3 describes the implementation of these models in Matlab.

2 Spatial Regression Models

As in times series, several standard models may be considered. Under the assumptions that $(I - \lambda W_3)$ is invertible and $\mid \lambda \mid < 1$ for reasons of stationarity, we can rewrite the above model in the following way:

$$y = \psi W_1 y + X\beta + W_2 X^* \rho + (I - \lambda W_3)^{-1} \mu \qquad (1)$$

In the sequel, we denote model (1) by (GSR): General Spatial Regression.

Let us point out that the (GSR) model admits $(2p + 2)$ unknown parameters $(\psi, \beta', \rho', \lambda, \sigma^2)'$. Hence, this model requires at least two different weights matrices ($W_1 \neq W_2$ or $W_1 \neq W_3$). Otherwise, the unknown parameters are not identified (see Anselin 1988). However, for the below submodels, one may suppose that the a priori specified spatial weights matrices W_1, W_2 and W_3 are equal (say to W). W is the spatial version of the lag operator in times series and is a contiguity matrix. Given a geographical area with n locations, a contiguity matrix is a matrix of size $n \times n$ with element $W(i,j)$ defining the intensity of the dependence between two regions. One often assumes that $\sum_j W(i,j) = 1$. This latter constraint implies that each region is influenced by at least one neighbour.

Let us note that W is not necessarily symmetrical and must receive serious attention. Indeed, the contiguity matrix is a spatial weight matrix which captures the effects of spatial autocorrelation and there are many ways to construct it. For example, one may define a contiguity matrix considering a function of the distance or the time which seperate two regions. Therefore, the problem is to choose an appropriate specification of W. A misspecification of the contiguity matrix has an impact on hypothesis testing with respect to spatial dependence among residuals. Its effect is evaluated by Monte Carlo simulations on the power of the Moran's index and the Lagrange Multiplier tests for spatial errors and|or the spatial lag (see Anselin et al. 1995).

To explain the presence of autocorrelation and crosscorrelation in the model (GSR), one may take the following example from Florax and Folmer (1992):

Consider an aggregate regional production function where regional production is treated as a function of inter alia the availability of labor. Autocorrelation then implies that regional production in region r is also influenced by regional production in regional r', whereas *cross-correlation indicates that the regional production in region r is also influenced by the availability of labor in region $r'(r \neq r')$ (Florax and Folmer, 1992, pp 410).*

Of course, in this model, ψ, β, ρ and λ are unknown and must be estimated. In order to derive the maximum likelihood estimator, let us write the log-likelihood.

Consider the following notations:

$$A = I - \lambda W_3, \quad B = I - \psi W_1, \quad \widetilde{X} = [X|W_2X^*], \quad \gamma = (\beta', \rho')'.$$

When the errors μ are normally distributed, by a straightforward manipulation, one can show that the log-likelihood function is given by:

$$\begin{aligned} L &= -(n/2)\log\pi - (n/2)\log\sigma^2 + \log\det A + \log\det B \\ &\quad +(1/2)\log\det\Omega^{-1} - (1/2\sigma^2)(By - \widetilde{X}\gamma)'(A'\Omega^{-1}A)(By - \widetilde{X}\gamma) \end{aligned}$$

However, in general -that is to say when $\lambda \neq 0$ and|or $\psi \neq 0$– the maximum likelihood estimates for β, ρ and σ^2 can be found through maximisation of the concentrated log-likelihood as suggested in Anselin (1980). That is to say:

first, we assume that ψ and λ are known to derive the maximum likelihood estimator for β, ρ and σ^2. After, we inject these estimators in the log-likelihood function to obtain the concentrated log-likelihood function in ψ and λ. We then get the estimators of the nuisance parameters ψ and λ by maximizing the concentrated log-likelihood function. Via an iterative procedure, maximum likelihood estimates for ψ, β, ρ and λ are obtained.

In what follows, we assume that Ω^{-1} is the diagonal matrix with diagonal term $\{\omega_i\}_{i=1}^n$ and that the errors μ are normally distributed.

2.1 The Weighted Least Squares Model (WLS):

This is the most simple case corresponding to $\psi = \lambda = 0$ and $\rho = 0$.
The model is defined as follows:

$$y = X\beta + \mu$$

The maximum likelihood estimators (MLE) of the coefficient β and the variance σ^2 are given by

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, \quad \hat{\sigma^2} = \frac{\| y - X\hat{\beta} \|^2_{\Omega^{-1}}}{n}.$$

A predicted value of y is given by: $\hat{y} = X\hat{\beta}$.

2.2 The Spatial Autoregressive Error Model (SARE):

This is the case when $\psi = 0$ and $\rho = 0$:

$$y = X\beta + (I - \lambda W_3)^{-1}\mu.$$

For a given λ we have:

$$\hat{\beta}(\lambda) = (X'A'\Omega^{-1}AX)^{-1}X'(A'\Omega^{-1}A)y, \quad \hat{\sigma^2}(\lambda) = \frac{\| y - X\hat{\beta}(\lambda) \|^2_{(A'\Omega^{-1}A)}}{n}.$$

Hence, it is easy to see that the concentrated log-likelihood in λ is given by :

$$\begin{aligned} L(\lambda) &= -(n/2)\log 2\pi - (n/2)\log \hat{\sigma^2}(\lambda) + (1/2)\sum_{i=1}^{n} \log \omega_i \\ &\quad + \sum_{i=1}^{n} \log |1 - \lambda\delta_i| - (n/2) \end{aligned}$$

where $\{\delta_i\}_{i=1}^n$ are the eigenvalues of the weight matrix W_3.
A predicted value of y is given by

$$\hat{y} = X\hat{\beta} + \lambda W_3\hat{\varepsilon} = X\hat{\beta} + \lambda(W_3 y - W_3 X\hat{\beta}).$$

2.3 The first-order Spatial Autoregressive (SLY):

This is the case when $\lambda = 0$ and $\rho = 0$:

$$y = \psi W_1 y + X\beta + \mu$$

For a given ψ we have:

$$\hat{\beta}(\psi) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Ay, \quad \hat{\sigma^2}(\psi) = \frac{\| Ay - X\hat{\beta}(\psi) \|^2_{\Omega^{-1}}}{n}.$$

The concentrated log-likelihood in ψ is given by

$$\begin{aligned} L(\psi) &= -(n/2)\log 2\pi - (n/2)\log\hat{\sigma^2}(\psi) + (1/2)\sum_{i=1}^{n}\log\omega_i \\ &\quad + \sum_{i=1}^{n}\log|1-\psi\delta_i| - (n/2) \end{aligned}$$

where $\{\delta_i\}_{i=1}^n$, as in above, are the eigenvalues of the matrix W_1.
A predicted values of y is given by : $\widehat{y} = \psi W_1 y + X\hat{\beta}$.

2.4 Spatial Weighted Crossregressive Model (SLX):

This is the case when $\psi = \lambda = 0$:

$$y = X\beta + W_2X^*\rho + \mu \tag{2}$$

Model (2) can be rewritten in the following way:

$$y = \widetilde{X}\gamma + \mu. \tag{3}$$

Therefore, the (SLX) model is equivalent from the computational point of view to the (WLS) model with independent variables defined by $\widetilde{X}$ and regression coefficient by γ.

2.5 Mixed Crossregressive and Spatial Regressive Model (SLYLX):

This is the case when $\lambda = 0$:

$$y = \psi W_1 y + X\beta + W_2X^*\rho + \mu \tag{4}$$

One can rewrite model (4) in the following form:

$$y = \psi W_1 + \widetilde{X}\gamma + \mu. \tag{5}$$

We see that the (SLYLX) model is equivalent from computational point of view to the (SLY) model with independent variables defined by $\widetilde{X}$ and regression coefficient by γ.

2.6 Mixed Crossregressive and Spatial Autoregressive Error model (SLXARE):

This is the case when $\psi = 0$:

$$y = X\beta + W_2X^*\rho + (I - \lambda W_3)^{-1}\mu. \tag{6}$$

This model is equivalent from computational point of view to the following (SARE) model:

$$y = \widetilde{X}\gamma + (I - \lambda W_3)^{-1}\mu. \tag{7}$$

2.7 First-order regressive and Spatial Autoregressive Error Model (SLYARE):

This is the case when $\rho = 0$:

$$y = \psi W_1 y + X\beta + (I - \lambda W_3)^{-1}\mu.$$

For given ψ and λ we have:

$$\hat{\beta}(\psi, \lambda) = (X'A'\Omega^{-1}AX)^{-1}X'(A'\Omega^{-1}A)By$$

and

$$\hat{\sigma^2}(\psi, \lambda) = \frac{\| By - X\hat{\beta}(\psi, \lambda) \|^2_{(A'\Omega^{-1}A)}}{n}.$$

The concentrated log-likelihood in ψ and λ is given by:

$$\begin{aligned} L(\psi, \lambda) &= -(n/2)\log 2\pi - (n/2)\log \hat{\sigma^2}(\psi, \lambda) + (1/2)\sum_{i=1}^{n}\log\omega_i \\ &\quad + \sum_{i=1}^{n}\log|1 - \psi\alpha_i| + \sum_{i=1}^{n}\log|1 - \lambda\delta_i| - (n/2) \end{aligned}$$

where $\{\alpha_i\}_{i=1}^n$ and $\{\delta_i\}_{i=1}^n$ are respectively the eigenvalues of the weights matrices W_1 and W_3.
A predicted value of y is given by

$$\widehat{y} = \psi W_1 y + X\hat{\beta} + \lambda W_3(y - X\hat{\beta} - \psi W_1 y).$$

2.8 How can we choose a model?

Once again, a real challenge in applied spatial regression is to choose the appropriate model. Indeed, it is often awkward to estimate directly the general model (GSR). The problem is then: is it appropriate to include autoregressive disturbance and|or a spatially lagged dependent variable? Do we have to introduce spatially lagged independent variables?
To handle this problem, various attemps have been made. There are various test statistics for spatial correlation among the residuals such as the Moran index, Geary's index, Lagrange multiplier tests, the Cliff and Ord statistic, etc (Anselin et al.1995). Moreover, to handle the problem of the misspecification of spatial regression models, Florax and Folmer (1992) propose an algorithm using Lagrange multiplier to choose between these models. We have already mentioned that these tests are crucially dependent on the contiguity matrix.
Let us discuss in more detail the Moran index. For a standardized sequence $y_1 \cdots y_n$ of measures with mean zero and for a given contiguity matrix W such that $\sum_j W_{ij} = 1$ and $W_{ii} = 0$, the Moran I index is defined as follows:

$$I = \frac{(n/S_0)\sum_i \sum_j W_{ij} y_i y_j}{\sum_i y_i^2},$$

where $S_0 = \sum_i \sum_j W_{ij}$. Hence, the Moran index is the ratio of the covariance between neighbouring measures over the variance of these measures. Therefore I has a definition similar to the autocorrelation coefficient. When the y_i are normally distributed and under the hypothesis that there is no spatial autocorrelation, the first two moments of I are given by:

$$E(I) = -(1/n-1), \quad E(I^2) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2}$$

where

$$S_1 = (1/2)\sum_{i,j}(W_{ij} + W_{ji})^2$$

and

$$S_2 = \sum_{i,j}(W_{i.} + W_{.j})^2 \text{ with } W_{i.} = \sum_j W_{ij} \text{ and } W_{.j} = \sum_i W_{ij}.$$

After normalization, one may use the Moran index to test the autocorrelation of the disturbance. It is easily applicable and efficient enough. However, the Moran index points out whether there is or not a spatial autocorrelation but does not tell us what kind of autocorrelation there is. For two appropriate models with spatial autocorrelation (that is to say when models are not rejected by the test using the Moran index), one can use the Akaike information criterium (AIC) or the Schwarz bayesian criterium (SBC) for model selection.

3 Implementation in Matlab

We have implemented all these models in Matlab and we will describe in this section the spatial regression function library. However, one can find in the following web adress http://www.econ.utoledo.edu/matlab_gallery/index.html, some of these packages in a simplified version(corresponding to models SLX, SLY and SLYARE). One will find in this adress other interessing spatial econometrics functions.

Our library contains sevaral packages and each package corresponds to one of these models. Because all these packages are made in the same way, we will describe just one of them: say the package corresponding to model (SLYARE).

This package is formed by three Matlab functions: p_slyare, slyare and l_slyare. First of all, let us recall that the model (SLYARE) is define as follows:

$$y = \psi W_1 y + X\beta + (I - \lambda W_3)^{-1}\mu.$$

- $l = l\_slyare([\psi, \lambda], y, X, W_1, W_3, \Omega)$ returns the reverse of the concentrated log-likelihood function of model (SLYARE). Ω should be a column matrix equivalent to the diagonal covariance matrix of errors. We assume that $X(:,1)$ contains a constant term.

- $slyare(y, X, W_1, W_3, \Omega)$ computes the first-order Regressive and Spatial Autoregressive Error model (SLYARE) y should be a $(n \times 1)$, X is the $(n \times (p-1))$ matrix of explanatory variables with constant term deleted. $results = slyare(y, X, W_1, W_3, \Omega)$ returns several items (see annex) the more interesting are printed out by the Matlab function p_slyare.

- $p\_slyare(y, X, W_1, W_3, \Omega, vnames)$ prints output from slyare regression. Here, vnames should be an optional $(p \times 1)$ vector of variable names ordered with $y, x1, x2$, etc. Results is returned by slyare().

  $results(1:p, 1) = \widehat{\beta}$

  $results(p+1, 1) = \psi$

  $results(p+2, 1) = \lambda$

  $results(1:p, 2) = t$ (asymptotic student statistic of coefficients $\widehat{\beta}$)

  $results(1, 5) = ORSS$ (ordinary residual sum of squares)

  $results(2, 5) = R^2$ (Warning: there is no precise counterpart to R^2 in the generalized regression model but here we will take $R^2 = Fish/((n-p)/(p-1) + Fish)$ where $Fish$ is the Fisher's statistic)

  $results(3, 5) = R^2_{adj}$(adjusted rsaquared)

  $results(4, 5) = l$ (log likelihood)

  $results(5, 5) = n$ (sample size)

  $results(6, 5) = p$ ($p-1$ is the number of independent variables)

$results(8,5) = AIC = l - p$ (Akaike information criterium)

$results(9,5) = SBC = l - .5p\log(n)$ (Schwarz Bayesian criterium)

$results(10,5) = Fish$ (Fisher's statistic)

$results(11,5) = IR$ (Moran's residuals coefficient)

$results(12,5) = stIR$(Standardized IR)

Let us point out the fact that all these functions require to run first *eigv*(). This latter compute the eigenvalues of contiguity matrices W_1, W_3.

eigv(W_1, W_3) returns the $(n \times 2)$ matrix of the global vectors *delta* (size $(n \times 1)$) and *gamma* (size $(n \times 1)$) the eigenvalues of contiguity matrices W_1, W_3 respectively. *eigv*() run fast and of course we need just to turn it once only whatever they may be the following functions.

Finally, our procedures are useful. In application with a real data set, all these procedures with $n = 162$ and more than 5 explanatory variables took about 5 secondes of real time on a 333 MHz machine using a PC version of Matlab v. One may download a self executable file containing all these functions from the following Web adress:http://www.eurandom.tue.nl/diack

Acknowledgments:*I wish to thank Christine Thomas for helpful comments. I am indebted to her.*

References

[1] L. Anselin, *Estimation methods for spatial autoregressive structures,* Regional Science Dissertation and Monograph Series 8 (Cornell University,Ithaca, NY) 1980.

[2] L. Anselin and R.J., Florax *New directions in spatial econometrics: Introduction,* in Luc Anselin and Raymond Florax 9eds.), New Directions in Spatial Econometrics, Springer 1995.

[3] Y. Aragon, D. Haughton, J. Haughton, E. Leconte, E. Malin, A. Ruiz-Gazen and C. Thomas-Agnan, *Female labor force participation in the Midi-Pyrénées.* Université de Toulouse I 1997.

[4] R. Florax and H. Folmer, *Specification and estimation of spatial linear regression models: Monte carlo evaluation and pre-test estimator,* Regional Science and Urban Economics, 1992, 22, 405-432.

[5] D. Haughton and J. Haughton *Expalining child nutrition in Viet-Nam,* Economic Development and Cultural Change 1997.