24 - 26 | November 2021 | Madrid
XLVI Reunión de Estudios Regionales
International Conference on Regional Science
Full cities, empty territories
Universidad Autónoma de Madrid

Searching the correct specification in Spatial Probit Model

Abstract:

The correct selection of a model with spatial effects has been one of the topics that has received more attention in the literature in spatial econometric. The wide variety of specifications (SEM, SAR, SDM, SLX, SARAR,…) and the difficulty to identify correctly the true model has been the warhorse in any modelling exercise. The Florax, Folmer and Rey (2003) paper was a guide that marked a turning point in the spatial econometric literature giving clear rules to select selection to the researchers. Other papers (Mur and Angulo, 2009, Mur, López and Angulo, 2014) continue the discussion giving more information regarding the properties of various specification strategies for spatial econometric models. Unfortunately, for models with limited variable there is no guides to select the correct specification. This paper proposes the first keys to the model section in spatial probit. A Monte Carlo experiment in small samples size compare two well-known model selection strategies, the so-called Specific-to-General, Stge, and General-to-Specific, Gets. The two strategies are based on a battery of misspecification tests obtained in a maximum likelihood framework.

Keywords: probit model, correct specification, selection model, spatial econometrics.

Introduction

When the making of a decision or the occurrence of an event follows a Bernoulli process and there is also a certain spatial dependence between observations, the standard probit regression fails to model it. The classical econometric models assume, among other key hypotheses, independence of the observations. If such independence is breached, then econometric models would provide inconsistent and inefficient estimations (McMillen, 1992).

Spatial probit models arise to extend the classical probit model by introducing some kind of spatial dependence. The outcome of the model not only depends on the characteristics of each observation, such as in the classical model, but it also relates observations closed to each other. The spatial structure can arise both by spatial dependence of the dependent variable itself or spatial dependence immersed in the errors of the model (LeSage and Pace, 2009).

Throughout the last decades the interest of researchers has been directed specially to developing algorithms that would produce consistent and efficient estimators to solve binary regression with spatial interconnections. In addition, there are numerous papers with test developments indicative of the existence of spatial dependence.

However, as far as we know, within the spatial probit environment, no researcher has designed a flow whereby the correct specification of the models is revealed. Without a standard process to verify the specification of the spatial probit models, the validity of the Axiom of Correct Specification (Leamer, 1978) could not be assured. Therefore, it could not be guaranteed that there is no correlation between the researcher's beliefs and the final model (Mur and Angulo, 2009).

The objective of this research is to propose a strategy to be able to define the type of specification among the best-known models within the spatial probit field. There are two classic methods that look for the correct specification of a model. The backwards method (General to Specific Method - GtS) where you start with a complete model and many times with existing multicollinearity and through different statistical tests, a routine is established to clean and reduce it. The other alternative is the forward method (Specific to General Method -StG) in which we start with a basic model and through tests on its residuals, elements are incorporated.

There is currently an open debate on the best strategy to follow to find the bests spatial econometric model. Florax et al, (2003) and years later Mur and Angulo (2009) stated the good properties of the GtS strategy. Until that moment, there was more or less econometric consensus that the strategy to follow was better than StG. . To carry out the strategy and to be able to discern at each stage of the algorithm which is the best model, the most reliable econometric tests available in standard spatial econometrics are used. Our proposal in this paper is to extend this discussion to probit models. Given the idiosyncrasy of the spatial probit, it is not possible to extrapolate the strategies followed so far and it is necessary to go to the specific tests for binary responses to establish a model selection process.

Taxonomy and Estimation of Spatial Probit Models

Sim Probit Model

Probit models arise to solve regression models where the dependent variable (Y^*) is continuous and unobservable called latent variable and to which a binary and observable response (Y) is associated such that:

$$Y^* = X\beta + e$$
$$e \sim N(0, \sigma^2 I)$$

and

$$Y := \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{if } otherwise \end{cases}]$$

Then, let (Y) be a binary Nx1 vector that reflects information that can be summarized in 0 when some event is failure and 1 for success. (Y) will be assumed to be explained linearly by a set of explicative variables (X). And the function that links the explicative variables (X) to the dependent variable (Y) is the cumulated density function of a standardized normal.

Spatial Probit Models

Spatial probit models arise to extend the classical probit model by introducing some kind of spatial dependence. The outcome of the model not only depends on the characteristics of each observation, such as in the classical model, but it also relates observations closed to each other. The spatial structure can arise both by spatial

dependence of the dependent variable itself or spatial dependence immersed in the errors of the model (LeSage and Pace, 2009).

As in the classic spatial probit, let (Y*) be a binary Nx1 continious stochastic vector of the latent variable. The broader specification would cover all types of spatial dependency. This model is the general nonlinear nesting model (GNNM)

$$Y^* = X\beta + \rho W_1 Y^* + Z\theta + u$$
$$u = \lambda W_2 u + e$$
$$e \sim N(0, \sigma^2 I)$$

where X is a matrix containing the initial exogenous variables and Z the same variables spatially lagged

$$Z = W_1 X$$

.Then, W{1,2} are the adjacency matrices by which the observations are linked to each other. Therefore,

$$w_{ij}^* := \begin{cases} 1 & \text{If j represents one of the neighbours to i,} \\ 0 & \text{If i = j or j is not one of the neighbours to i,} \end{cases}]$$

The final W matrices are commonly row-standarized, so that

$$w_{ij} = \frac{w_{ij}^*}{\sum_{j=1}^{n} w_{ij}^*}$$

which means that W matrices become frequently asymmetric. However, there are many other specifications for weighted matrices. See Yrigoyen, page 59 (2003).

The set of coefficients to be estimated are

$$\beta$$

for each of the independent variables and their spatially lagged and also

$$(\rho, \lambda, \theta)$$

which refer to the autorregresive parameters. And finally, Y is the observable variable defined by

$$Y := \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{if } otherwise \end{cases}]$$

The disturbance of the model (e) is a multivariate normal variable with mean equals to 0 and finite variance, although in many theoretical research is established as 1.

Before starting with the most popular specific spatial models, it is important to highlight the assumptions made in Kelejian and Prucha (2010) in the spatial contiguity matrices and in the autoregressive parameters. The first assumption is related to the main

diagonal of W{1,2} which is supposed to be 0 in all the instances. It means that no observation is a neighbor of itself. Second assumption states that the autorregresive parameters should be in the interval $\left(-\frac{1}{\tau},\frac{1}{\tau}\right)$. Where τ is the spectral radius of the adjacency matrices (W). The third assumption reflects that the aggregation of rows and columns of $W_1, W_2, (I-\rho W_1)^{-1}$ and $(I-\lambda W_2)^{-1}$ are bounded uniformly in absolute value. Finally, the fourth hypothesis establishes X with full rank and (X'X) nonsingular. By setting any of the spatial parameters to zero (ρ, λ, θ), we arrive at the specific models that have been studied in the academic literature. Next, we will see the specification of these models.

Probit-SAR Model.

If θ and λ are set to 0, then we obtain the so called spatial autoregregressive probit model (SAR). In this model initially presented by Anselin (1988), the endogenous variable is influenced by independent variables (X) and also depends on the lagged variable (WY*), that is, on the value of the dependent variable in neighboring locations.

$$Y^* = X\beta + \rho W_1 Y^* + u$$
$$u \sim N(0, \sigma^2 I)$$

So that, the dependent variable (Y*) located in (i), can be explained by exogenous variables located in (i) and located in other locations through the spatial multiplier

$$(I-\rho W_1)^{-1}$$

, as it is reflected in next equation

$$Y^* = (I-\rho W_1)^{-1} X\beta + (I-\rho W_1)^{-1} u$$
$$u \sim N(0, \sigma^2 I)$$

In this model, a spatial process is obtained by which in the case that Y *> 0 and therefore Y = 1 then the probability that neighboring observations have value 1 would increase.

Probit-SEM Model.

In this case, if θ and ρ are set to 0, then we obtain the so called spatial error probit model (SEM). This model has been widely used in research studies in the case of Gaussian regression, unlike for probit regression. In this type of model, they assume that there are variables with a certain spatial autocorrelation and correlated with the

dependent variable, and yet they are not contemplated in the model specification. Therefore, this omission of variables in the specification goes to model rediduals.

$$Y^* = X\beta + u$$
$$u = \lambda W_2 u + e$$
$$e \sim N(0, \sigma^2 I)$$

So that, the dependent variable (Y*) located in (i), can be explained by exogenous variables located in (i) and the residuals of adjacent observations through the multiplier $(I - \lambda W_2)^{-1}$.

.

$$Y^* = X\beta + (I - \lambda W_2)^{-1} e$$
$$e \sim N(0, \sigma^2 I)$$

Rest of spatial probit models.

The previous models presented (SAR and SEM) are the ones that first emerged in the spatial econometrics literature and the most widely used in research. Both models present global spillovers, as the autoregressive parameters affect the model globally. However, by starting from the general model we can arrive at many more different models, which are those revealed by Florax and Folmer (1992).

The first specification worth addressing is the so-called Spatial Lag of X probit model (SLX). It is based on setting λ and ρ to zero. According to Elhorst (2017), these types of models have received very little attention over the years. However, they do deserve it since if a good specification is established with a good W adjacency matrix, it can produce very significant flexible spillovers. These spillovers in this case are considered local, given that since the diffusion effect of the spatial multiplier is less than in global cases (Yrigoyen, 2003).

$$Y^* = X\beta + Z\theta + u$$
$$u \sim N(0, \sigma^2 I)$$

Where Z is the spatial lag of variable X. We would find this type of model in a probit environment when the probability that Y = 1 in the locality (i) not only depends on the characteristics of said locality but also depends on the characteristics of the neighboring localities.

The next model to present is the one that only sets $\lambda = 0$. This would be the so-called *spatial Durbin probit model* (SDM). This model would be a combination of the

previously presented SAR and SLX models, which means that combines global and local spillovers. This model introduced in LeSage and Pace (2009) takes the following form.

$$Y^* = X\beta + \rho W_1 Y^* + Z\theta + u$$
$$u \sim N(0, \sigma^2 I)$$

which can be expressed as,

$$Y^* = (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1} Z\theta + (I - \rho W_1)^{-1} u$$
$$u \sim N(0, \sigma^2 I)$$

In SDM the probability that Y*>0 in location (i) depends on features at location (i), also depends on features located in other locations through the spatial multiplier $(I - \rho W_1)^{-1}$, and spatial lagged features on the points surrounding (i) with the same spatial multiplier.

In case SAR and SEM models are combined, then we obtain the *spatial autoregressive model with autoregressive disturbances* (SARAR). In this model, the parameter θ is equal to zero.

$$Y^* = X\beta + \rho W_1 Y^* + u$$
$$u = \lambda W_2 u + e$$
$$e \sim N(0, \sigma^2 I)$$

This is a very interesting model because of the nuances it contains. A considerable aspect to note is that in this type of model the adjacency matrices W_1 and W_2 should not be exactly the same to avoid parameter estimation problems (Anselin, 1988). The model can be writen also as

$$Y^* = (I - \rho W_1)^{-1}(X\beta + (I - \lambda W_2)^{-1} e)$$
$$e \sim N(0, \sigma^2 I)$$

The last remarkable combination of space models is bringing together the SLX and SEM models. This would lead us to set the rho parameter to zero and give rise to the so-called modelSpatial Lag of X probit model with autoregressive disturbances.

$$Y^* = X\beta + Z\theta + u$$
$$u = \lambda W_2 u + e$$
$$e \sim N(0, \sigma^2 I)$$

which can also be written as

$$Y^* = X\beta + Z\theta + (I - \lambda W_2)^{-1}e$$
$$e \sim N(0, \sigma^2 I)$$

To finalize the taxonomy of the main spatial models, it can be observed how the introduction of different second-order spatial adjacency matrices would change the specification of the model. However, in this paper, we will be focused on the specification of (SIM, SAR,SEM,SARAR,SLX-SAR,SLX-SEM).

Estimation.

Spatial Probit Models have not received as much attention as classic regression models. One of the reasons why this fact occurs is due to the added complexity to solve these models and the high computational cost which increases as the data grows.

For instance, the specification of spatial autoregressive models (SAR) is

$$Y^* = (I - \rho W_1)^{-1}(X\beta + u)$$

Therefore, the whole error term is

$$\psi = (I - \rho W_1)^{-1}u$$

and it's variance can be expressed as

$$\Omega = E(\psi\psi') = \sigma^2((I - \rho W_1)^{-1})((I - \rho W_1)^{-1})^t)$$

As we are in a probit environment, then we assume a multivariate normal distribution for the likelihood function with a covariate structure as the observations are interconnected.

$$L(\beta, \rho) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_n}^{b_n} e^{-\frac{t^t \Omega^{-1} t}{2}} dt$$

The intervals of integrations depend on the value of the observable Y.

$$\text{if } y_i = 1 \text{ then: } a = -\infty; b = ((I - \rho W_1)^{-1})X\beta$$
$$\text{if } y_i = 0 \text{ then: } a = ((I - \rho W_1)^{-1})X\beta; b = \infty$$

Over the last few years, algorithms capable of dealing with the estimation of the parameters of these spatial probit models have emerged. The first algorithm that emerged was the Expectation-Maximization (EM) proposed by McMillen (1992). Then, Pinkse and Slade (1998) proposed the solution by using the Generalized method of moments (GMM). LeSage (2000) uses Gibbs Bayesian Sampling to estimate a heteroskedastic spatial autoregressive problem. Beron and Vijverberg (2004) proposed

to use the Recursive Importance Sampling (RIS). In Klier and McMillen (2008) a modification of GMM algorithm is proposed by linearizing it. Calabrese and Elkink (2014) made an in-depth review of these methods, EM, Gibbs and RIS are very accurate in the estimates, however for samples greater than 1500 observations in practice are intractable. In contrast, GMM is very fast even when the number of observations is large, but it fails to achieve unbiased estimates. Then, Martinetti and Geniaux (2017) proposed a different estimation proposing a different estimate from what has been available so far. They proposed an approximation to maximize the likelihood function. Focusing on the optimization in time of the algorithm using different mathematical and computational techniques, the authors achieve consistent and efficient estimates in a reasonable amount of time. That reason is why we have chosen this method to compute our MonteCarlo simulations. The algorithm is in contained in ProbitSpatial R Package developed by same authors.

Probit Specification Selection Strategy

The problem lies in finding the underlying spatial structure given some observed data. Given a spatially georeferenced database, the only thing that is known is the structure of the adjacency matrix. However, the specification of the model is unknown. The most widespread methods for comparing models are the Akaike Information Criteria (Akaike, 1973) and the Bayesian Information Criteria (Schwarz, 1978). These methods provide simplicity and speed when choosing between models. At the moment there is no evidence on the power for discriminating capacity of these criteria to choose the real model behind some data.

The tests that have been most used to undertake the selection of the ideal model are the Lagrange Multiplier tests for error dependence and for spatially lagged dependence. These tests provide very good results especially when the autoregressive parameters are greater than 0.3. However, they are only prepared for linear Gaussian spatial models. The fact of not having tests specifically prepared for the selection of models is a clear disadvantage for the purpose of our research.

The proposed strategy is StG, where the correct selection of six very analyzed spatial models is attempted. (SIM, SAR, SEM, SARAR, SLX-SAR, SLX-SEM). Since the correct specification of our model is not known, nor whether the data comes from a standard or spatial theoretical model. It is vitally necessary to have tests that at least

diagnose whether the errors evaluated in the standard model (SIM) are spatially correlated or not. Although, what these tests would not tell us is what kind of spatial specification we would be talking about. In this sense, the first test used is the one proposed by Kelejian and Prucha (2001). They generalize de famous I-Moran test (Moran, 1950) in order to be able to apply it to those problems in which the dependent variable is binary. This test outperforms in size and power to others for probit model. (Amaral et al. (2013)).

H0: Disturbances are Spatially Independent
H1: Disturbances are Spatially Dependent of each other

$$KP_{test} = \frac{(e^t We)^2}{\sqrt{tr(W\Sigma W\Sigma + W^t \Sigma W\Sigma)}} \sim N(0,1)$$

where

$$\Sigma = matrix(diag((\Phi(\hat{y}_i^*)(1 - \Phi(\hat{y}_i^*)))))$$

And

$$e_i = y_i - \Phi(\hat{y}_i^*)$$

The other test that will be used in the selection strategy is the Join-Count test (Cliff and Ord, 1973). It is used to test global spatial autocorrelation pattern of both binomial and multinomial variables. Join-Count statistic tests the null of the random co-localized pattern by counting the number of each possible joins between neighbors. In this paper, the possible join 1-1 is the one used. It counts the observed number of joins and compare them with the expected number under the null.

H0: Response Variable is Spatially Independent
H1: Response Variable is are Spatially Dependent of each other

$$J11 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} 11_{ij} \sim N(0,1)$$

The rest of the tests used to investigate the origin of the data are directly the t-test of the coefficients of the exogenous variables and of the autoregressive parameters. The

Vuong's closeness test is also used based on a likelihood ratio and which compares two models that can be nested or not nested.

H0: Both models are equally close to true data
H1: One model is closer to true data

The flow presented consists of seven stages. In each of these stages, one decision is made. The combination of tests within each stage might identify the model and then the process would stop or might not identify the model and the process would continue.

In figure 1, the strategy to select the proper model is depicted. In step 1, in case null hypothesis for Generalized Moran Test over the residuals of SIM model is not rejected and Join Count null hypothesis is not rejected either, then process would stop identifying the model as a SIM model. In step 2, in case step 1 has not finished the process, then if null hypothesis for Generalized Moran Test over the residuals of SIM and SLX models are not rejected and Join Count null hypothesis is now rejected, then process will finish identifying a SLX. In step 3, when rejecting the null hypothesis for Generalized Moran Test over the residuals of SIM and rejecting the Join Count test and not rejecting the t-test for the spatial lagged exogenous variable of the SLX model, then process will finish identifying a SEM. In step 4, after solving the SLXSAR, in case of not rejecting the t-test for the spatial lagged exogenous variable of the SLXSAR model and rejecting the null of Vuong test comparing SLXSAR and SLXSEM in favor of SLXSAR, then process will end identifying a SLXSAR. In step 5, is similar to step 4, but rejecting the null of Vuong test in favor of SLXSEM, and adding another condition which is rejecting of Vuong test comparing SEM and SLX in favor of SLX, which brings us to a SLXSEM. In step 5, Vuong test is also rejected comparing SAR and SARAR in favor of SAR, which leads to SAR model. If none of the previous conditions are fulfilled, then the model will be a SARAR.

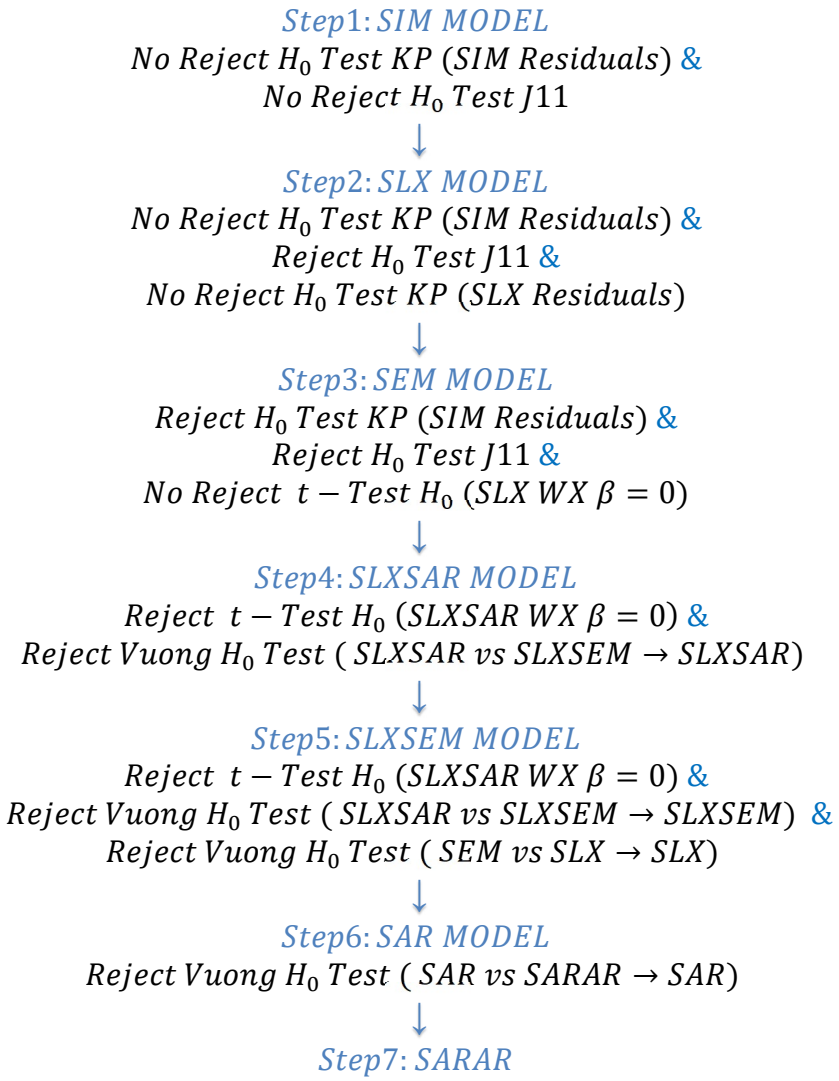


Figure 1

Monte Carlo Experiment

In order to propose the simulation work, data is generated according to one of the seven models to be identified. The number of observations is defined as an element to be studied, so N = (400,900,1600,2500) is proposed. The adjacency matrices W1 and W2 are assumed to be known. A grid division of space has been considered. Both contiguity matrices are not the same, since otherwise the model could produce inconsistent estimates in the case of SARAR (Billé and Arbia, 2013). Therefore, W1 type of contiguity is the Rook criteria and W2 is a Queen criteria.

The theoretical coefficients of the model have been previously studied in such a way that appropriate characteristics are fulfilled for the formulation of MonteCarlo's work. The coefficients chosen are c (1, -0.5, -0.4) = c (b0, b1, b2). b0 refers to the intercept, b1 to the exogenous variable and b2 to the spatially lagged exogenous variable if

applicable. These coefficients guarantee the balancing of the data (0,1). Furthermore, these coefficients keep the AUC (ROC) of all models always greater than 0.8. The spatial autocorrelation coefficients also vary according to each iteration in order to see the effect they have on the predictive capacity of the proposed algorithm. Both Rho and Lambda will take the value of c (0.4, 0.6). For each of these combinations and models, 48 samples have been run. All these assumptions are briefly summarized in table 1. The combination of all the parameters and executions makes a total of 5379 samples to see the accuracy of our strategy.

48 simulations

Models	N Parameters	Parameters
SIM	2	(1, -0.5)
SLX	3	(1, -0.5, -0.4)
SEM	3	(1, -0.5, Lambda)
SAR	3	(1, -0.5, Rho)
SARAR	4	(1, -0.5, Rho, Lambda)
SLXSAR	4	(1, -0.5, -0.4, Rho)
SLXSEM	4	(1, -0.5, -0.4, Lambda)

n Observations	400,900,1600,2500
Rho	0.4,0.6
Lambda	0.4,0.6

table 1

Before reviewing the results of the model, an important point is to note that the models have been built under ideal conditions. That is, there is no endogeneity in the residuals of the generated models, there are no outliers, nor is there heteroscedasticity in the residuals of the models. The theoretical residuals are generated to follow a normal distribution and there is no local spatial dependence at any point in space. Obviously, they are very strong hypotheses that do not accurately reflect what happens in real problems. However, as it is a pioneering study in discovering the pattern to follow to achieve the correct specification of probit spatial models, we leave that point open for future research.

Monte Carlo Results

The following table2 shows the main results of the strategy followed. In each of the steps the process ends for a number of observations of which a percentage of times is correct. The algorithm loses precision as it takes more steps to finish the process. In other words, as we enter more complex specifications or whose spatial structure can sometimes be confused, it is more difficult to identify them with the available tests.

Step	***Model if Process Ends***	***Process Ends***	***% Correct***
1	SIM	684	89%
2	SLX	804	84%
3	SEM	708	85%
4	SLXSAR	864	77%
5	SLXSEM	678	81%
6	SAR	840	68%
7	SARAR	798	55%

table 2

An important aspect to note when studying the error rates, in which the model misidentifies the specification, is the basis of the error. When the algorithm is wrong, if we look at table3, we can see how in most cases the mistake points to another specific model, instead of being distributed evenly among the rest of the models. For example, in step 1, there are a lot of SEM models in which the join of the KP and Join Count Test are not able to identify. Another example in the case of step 2 and 3, the algorithm grants SLX, SEM when in fact an underlying model SLXSEM, SARAR respectively.

		Estimated Model						
		SIM	**SLX**	**SAR**	**SEM**	**SARAR**	**SLXSAR**	**SLXSEM**
Real Model	**SIM**	606	60	48	36	18	-	-
	SLX	12	678	6	-	-	36	36
	SAR	6	12	570	12	150	12	6

	SEM	60	-	-	600	108	-	-
	SARAR	-	-	144	60	438	126	-
	SLXSAR	-	6	6	-	-	666	90
	SLXSEM	-	48	66	-	84	24	546
		Table 3						

The next part of the study consists of the analysis of those factors that intervene in the correct specification of the model. As previously mentioned, simulations have been produced with different number of observations, different coefficients of the autoregressive parameter, and therefore, the data sets that are generated contain different percentages of 1s, which vary between 40% and 70%.

A logistic regression model is proposed, simply with the objective of seeing how the commented variables affect. The interest lies in seeing if the slope of the beta is positive or negative and in the case of categorical variables to be able to see the comparison between their categories. Table4 shows the regression model. It can be clearly seen how the number of observations plays a fundamental role in being able to correctly discriminate the model. In fact, you can even see how the positive slope increases when going from 1600 to 2500 observations. When talking about polygons it is a difficult number to arrive at, however in the point pattern it is really easy to have samples of these sizes. Regarding the percentage of 1s within the model, it is also noteworthy that at normal balancing levels it does not affect the discriminant power. However, when the sample has many 1s, specifically more than 80 percent, then the predictions are not as accurate. Regarding the value taken by the autoregressive parameters, it is clear that the greatest influence is in using or not using the parameter, as can be seen in table2. The precision drops substantially in those models that use the Rho parameter (SAR, SLXSAR and SARAR). Although it is quite interesting that when the parameter takes a high value (0.6) the predictions improve. This last consideration is equivalent for the case of the autoregressive parameter lambda. When the parameters match and both have the same effect on the final model, then the model loses precision. This is probably a side effect of parameter estimation. In the case in which rho and lambda are equal, the

approximation of the likelihood function is more expensive and, in some cases, it is not accurate with the estimation of the true parameters.

Variable	Estimate	Std. Error	z value	
(Intercept)	0.6735	0.2867	2.3490	*
Observations=900	0.5722	0.0534	10.7220	***
Observations=1600	0.6301	0.0540	11.6670	***
Observations=2500	0.7824	0.0558	14.0180	***
Rho=0.4 (When applicable)	-0.1456	0.0602	-2.4200	*
Rho=0.6 (When applicable)	0.2086	0.0672	3.1040	**
Lambda=0.4 (When applicable)	-0.4417	0.0523	-8.4410	***
Lambda=0.6 (When applicable)	-0.3331	0.0525	-6.3490	***
Rho>0 & Lambda>0 and Rho=Lambda	-0.2603	0.0828	-3.1430	**
% 1s Percentaje = 50%	-0.0401	0.2886	-0.1390	
% 1s Percentaje = 60%	-0.1194	0.2886	-0.4140	
% 1s Percentaje = 70%	-0.3977	0.2865	-1.3880	
% 1s Percentaje = 80%	-0.8223	0.2928	-2.8080	*

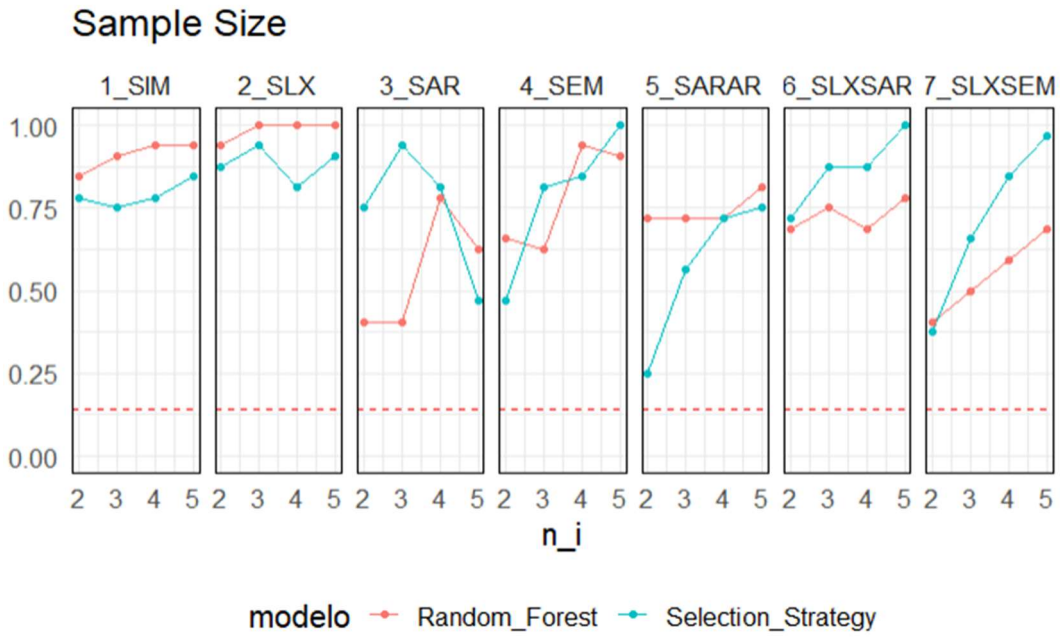
Comparative Results

In a last point of the investigation, we would like to think about the degree of precision of the presented decision flow. Four reference tests are involved in the decision process. The Join Count test and the generalized Moran test, which detect autocorrelation patterns. This Moran test has been tested in SAR models by (Kelejian and Prucha (2001) and in SEM models by (Amaral et al., 2013). However, other papers such as Pinkse (1999) warned that this test could have difficulties to detect patterns when the explanatory variables are spatially correlated Therefore, since there is no single test that solves the problem, our proposal is the combination of tests to reach the solution. Another test used so far is the student's t-test, to assess the significance of certain specific parameters of models to discard or accept them. And finally, the Vuong test that compares two models and performs the test of whether both models are equally represented by the data collected or if a model is closer. This Vuong test varies depending on whether the models are nested or not. The test varies slightly from one to another.

La cuestión ahora que nos planteamos es si con todo el elenco de test que tenemos en nuestra mano hemos elegido el mejor camino para seleccionar el mejor modelo. De alguna manera, para nosotros tiene sentido dicho flujo de decisión. Los test utilizados para cada una de las decisiones hacen referencia de alguna manera al modelo que estamos apuntando. SIn embargo, no hemos hecho todas las combinaciones de test posibles para ver el mejor diagnóstico. Adicionalmente, tenemos en nuestra disposición

los test que hablábamos al comienzo tan utilizados en el mundo de los datos como el AIC y el BIC. Por último, también tenemos el test de verosimilitud (Neyman and Pearson, 1933) para las las combinaciones de modelos. Therefore, we are left with a large database with which we can make some type of independent model that guides us on the precision of our strategy. We would compare the output of an algorithm whose objective is the minimization of the entropy within the outgoing nodes against the result of the decision algorithm presented. Without much intention of carrying out a grid search that provides us with a global optimum of the problem, we propose a random forest with a multinomial output. The maximum depth that we are going to put in this random forest is 2, so that in each iteration, it gets reliable estimates step by step. Imitating the behavior that we have done in our flow. The number of trees established to make the model is 1000 and the rest of the parameters have been left as the default. The package that has been used to undertake the model has been randomForest (version 4.6-14) in Cran R. We consider 60% training database and 40% test data to avoid overfitting and that the output misleads us.

Figure 2 shall present these results in the form of a comparison between the model selection strategy and the output from the random forest. The comparison is shown based on the most influential variable in the model, which in this case is the sample size.



Where n_i=(2,3,4,5) belongs to the sample size (400,900,1600,2500) and the dotted red line is equal to 1/7, which is the probability of picking one model randomly.

Given that the random forest is a non-interpretable algorithm, it is difficult to know what decisions it has taken in each of the 1000 iterations that have been proposed. It can be seen how in SIM and SLX it is higher in the random forest. Probably by adding complexity to the strategy we could reach levels of 100% precision. In the case of the SEM, the precision is similar between both models. For specifications that combine SLX with SAR or SEM, the strategy followed provides better predictions, especially as the sample size increases. And the algorithms ultimately deliver opposite results in the accuracy of SAR and SARAR. This is probably because we have put the SARAR in the last step. By changing the order, the results would probably line up.

Conclusions

The search for the correct specification is one of the topics that has received less attention within spatial econometrics. Over the last 40 years, different taxonomies have emerged that allow a better approach to reality. At the same time, algorithms and computational techniques allow us to solve these problems in reasonable times. The identification of the ideal specification given some data should be very relevant, otherwise we leave everything to the discretion of the researcher.

In this paper we collect for the first time a series of rules that could be followed in order to find the correct specification of spatial probit. Currently there are numerous tests to detect spatial autocorrelation, detect significance of parameters or compare models, but as can be seen in the development of the topic, only the combination of them can lead us to correctly discriminate the underlying specification given some data. In a MonteCarlo simulation exercise, we have developed the test combination that leads to a reasonable precision to detect the ideal specification to follow.

Obviously, not all the spatial specifications are ruled, nor have all the existing tests related to the spatial probit been used. However, we believe that it is a good beginning of analysis. At the same time, we must know what happens when we do not incorporate all the information into the model, or when there is initial heteroscedasticity in the model that is not detected. We leave these matters for further investigations in the field.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory (pp. 267-281). Akadémiai Kiadó Location Budapest, Hungary.

Amaral, P. V., Anselin, L., & Arribas-Bel, D. (2013). Testing for spatial error dependence in probit models. Letters in Spatial and Resource Sciences, 6(2), 91-101.

Anselin, L. (1988). Spatial Econometrics: Methods and Models (Vol. 4). Springer Science & Business Media.

Beron, K. J., & Vijverberg, W. P. (2004). Probit in a spatial context: a Monte Carlo analysis. In Advances in spatial econometrics (pp. 169-195). Springer, Berlin, Heidelberg.

Billé, A. G., & Arbia, G. (2013). Spatial discrete choice and spatial limited dependent variable models: a review with an emphasis on the use in regional health economics. arXiv preprint arXiv:1302.2267.

Calabrese, R., & Elkink, J. A. (2014). Estimators of binary spatial autoregressive models: A Monte Carlo study. Journal of Regional Science, 54(4), 664-687.

Clif A, Ord JK (1973) Spatial autocorrelation. Pion, London

Elhorst, J. P., & Halleck Vega, S. (2017). The SLX model: extensions and the sensitivity of spatial spillovers to W. Papeles de Economía Española, 152, 34-50.

Florax, R., & Folmer, H. (1992). Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. Regional science and urban economics, 22(3), 405-432.

Florax, R. J., Folmer, H., & Rey, S. J. (2003). Specification searches in spatial econometrics: the relevance of Hendry's methodology. Regional Science and Urban Economics, 33(5), 557-579.

Kelejian, H. H., & Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. Journal of Econometrics, 104(2), 219-257.

Kelejian, H. H., & Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of econometrics, 157(1), 53-67.

Klier, T., & McMillen, D. P. (2008). Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples. Journal of Business & Economic Statistics, 26(4), 460-471.

Leamer, E. E., & Leamer, E. E. (1978). Specification searches: Ad hoc inference with nonexperimental data (Vol. 53). John Wiley & Sons Incorporated.

LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. Geographical Analysis, 32(1), 19-35.

LeSage, J., & Pace, R. K. (2009). Introduction to spatial econometrics. Chapman and Hall/CRC.

López, F. A., Mur, J., & Angulo, A. (2014). Spatial model selection strategies in a SUR framework. The case of regional productivity in EU. The Annals of regional science, 53(1), 197-220.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706), 289-337.

Martinetti, D., & Geniaux, G. (2017). Approximate likelihood estimation of spatial probit models. Regional Science and Urban Economics, 64, 30-45.

McMillen, D. P. (1992). Probit with spatial autocorrelation. Journal of Regional Science, 32(3), 335-348.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Mur, J., & Angulo, A. (2009). Model selection strategies in a spatial setting: Some additional results. Regional Science and Urban Economics, 39(2), 200-213.

Pinkse, J., & Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. Journal of Econometrics, 85(1), 125-154.

Pinkse, J. (1999). Asymptotic properties of Moran and related tests and testing for spatial correlation in probit models. Department of Economics, University of British Columbia and University College London.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Yrigoyen, C. C. (2003). Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales. Dirección General de Economía y Planificación.