Investigaciones Regionales – Journal of Regional Research
https://doi.org/10.38191/iirr-jorr.24.060

Journal of Regional Research
Investigaciones Regionales

Accepted Article

Spatial Trends and Spatial Econometric Structures: practical application to a different context data

Maryna Makeienko[*], *Mariano Matilla-García*[**]

Received: 26 October 2023
Accepted: 17 September 2024

Abstract:

Spatial trend concept was proved to be useful to depict the systematic variations of the phenomenon concerned over a region based on geographical locations. We use three different geographical datasets to check if there exist potential leading deterministic spatial components and whether we can econometrically model spatial economic relations that might contain unobserved spatial structure of unknown form. Hypothesis testing is conducted with a symbolic-entropy based non-parametric statistical procedure, proposed in Garcia-Cordoba et al. (2019), which does not rely on prior weight matrices assumptions. Geographically restricted semiparametric spatial models are taken to perform a modeling strategy for cross-sectional data sets. The main question to be responded is whether the models that merely incorporate space coordinates might be sufficient to capture space dependence when applied to different types of data. Moreover, it is important to study what intrinsic characteristics of the economic problem or the dependent variable itself make feasible (and optimal) to use the specific methodological approach.

Keywords: Symbolic entropy; spatial trends; applied analysis.
JEL Classification: C01; C51; C21.

Tendencias espaciales y estructuras econométricas espaciales: aplicación práctica a un contexto de datos diferente

Resumen:

El concepto de tendencia espacial ha demostrado su utilidad para describir las variaciones sistemáticas del fenómeno en cuestión en una región basada en ubicaciones geográficas. Utilizamos tres conjuntos de datos geográficos diferentes para comprobar si existen posibles componentes espaciales deterministas principales y si podemos modelizar econométricamente las relaciones económicas espaciales que podrían contener una estructura espacial no observada de forma desconocida. La comprobación de hipótesis se realiza con un procedimiento estadístico no paramétrico basado en la entropía simbólica, propuesto en García-Córdoba et al. (2019), que no se basa en supuestos de matrices de pesos previos. Se toman modelos espaciales semiparamétricos geográficamente restringidos para realizar una estrategia de modelización de conjuntos de datos transversales. El problema principal a resolver es si los modelos que simplemente incorporan coordenadas espaciales podrían ser suficientes para capturar la dependencia espacial cuando se aplican a diferentes tipos de datos. Además, es importante estudiar qué características intrínsecas del problema económico o de la propia variable dependiente hacen factible (y óptimo) utilizar el enfoque metodológico específico.

Palabras clave: Entropía simbólica; tendencias espaciales; análisis aplicado.
Clasificación JEL: C01; C51; C21.

[*] Universidad Villanueva. España. maryna.makeienko@villanueva.edu
[**] Universidad Nacional de Educación a Distancia. UNED. España. mmatilla@cee.uned.es
Corresponding author: mmatilla@cee.uned.es

1. Introduction

Historically, spatial trends received less attention when it comes to understanding the association of outcomes across different geographical locations. This discrepancy is particularly noticeable within the field of spatial econometrics. Unlike the emphasis on time trends in time series econometrics for explaining proximate economic outcomes, spatial trends have not been as thoroughly explored. One potential reason for this disparity is the lack of sufficient statistical tools to assess the presence of spatial trends in the data. However, in the last few years, some statistical tools were developed. This paper seeks to address this gap by exploring the utilization of possible instruments to analyze spatial trends in modeling and examining economic relationships, especially in scenarios where spatially correlated variables are involved, either directly or indirectly.

From the methodological point of view, this paper elaborates on the work of Garcia-Cordoba et al. (2019), which develops a spatial econometric test for linear and nonlinear spatial structures. In this seminal paper, the authors delineate a statistical procedure based on the entropy concept to determine whether a cross-sectional data set contains a leading deterministic component in the form of either a trend or a chaotic non-linear process, building on the previous studies of these authors, but then within a time-series context. The work by Garcia-Cordoba et al. (2019) could generate research interest in testing for weak spatial dependence in the presence of a leading deterministic component, like time-series tests for unit roots in the presence of drift and/or trend. This is especially relevant as it has been recently shown (Müller and Watson, 2023) that using spatial data will easily lead researchers to spurious results and therefore to bad quality inference. In this regard, this paper is related with and can be used to avoid the perils of an invalid spatial model specification.

This paper applied the methodological approach given in Makeienko (2020) by applying it to a wider practical analysis of different types of datasets with different characteristics and economic backgrounds where spatial structures might not be so easy model them. This type of analysis allows us to capture the main characteristics of the datasets, where we can control both deterministic structure and spatial structure of the data. The analysis performed is aimed to answer the important question of whether there might be a simple model that, taking into consideration only the geographical position of the unit, might help us control the spatial dependence better than currently existing procedures and models. One would like to find the deterministic part of different datasets, that might be useful to develop a generalized method of using each model specifications.

The structure of the paper goes as follows. Part 2 clarifies theoretical points on Spatial Trends and Spatial Econometric structures, Part 3 describes in detail the datasets, specific characteristics of each dataset and general conclusions on the analysis of each of it. Finally, Part 4 concludes and opens the path of the possible further research.

2. Spatial trends and spatial econometric structures

Spatial trends have not been widely emphasized as a factor for describing and comprehending the interconnection between outcomes in a particular geographical area and those in its proximate regions, countries, or spatial points. It becomes a challenge when one tries to include the dependence lag in the spatial data. Difficulties arise from the way the data generating process is formed.

The data generating process (DGP) for a conventional cross-sectional non-spatial sample of n independent observations $\{y_i;\ i = 1, \dots, n\}$ is introduced as:

$$y_i = X_i\beta + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

where X_i is a 1 × k vector of explanatory variables and β is a k × 1 vector of parameters. It suits for linear regression models with mean $X_i\beta$ and a random component ε_i.

Spatial dependence has a dependence model similar to that of time series, as values observed in one location depend on the values of the neighboring observations in the nearby locations.

Suppose that i and $j \in \{1, \ldots, n\}$ with i≠j are two neighborhoods, then a DGP is given by:

$$y_i = \alpha_i y_j + X_i\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$y_j = \alpha_j y_i + X_j\beta + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2)$$

assumes that there is a simultaneous spatial dependence between y_i and y_j. Under standard econometric modeling, it is impossible to model spatial dependency.

A simple way to introduce a spatial dependency and spatial structure is to define a $W = (w_1, \ldots, w_n)$ (nxn) matrix to reflect spatial connectivity among neighbors, a so-called spatial weight matrix, which has served as a basis for different econometric model specifications that explicitly incorporate spatial lags. It imposes a structure in terms of what the neighbors are for each location and assigns weights that measure the intensity of the relationship among pairs of spatial units.

However, the use of the weight matrix W has been a controversial issue over the past few years. The two main and most severe critiques are McMillen (2012) and Gibbons and Overman (2012).

The problem of selecting a weight matrix among the different possibilities is a problem of model selection. In fact, different weight matrices result in different spatial lags of the endogenous or the exogenous variables included in the model. Different equations with different regressors amount to model selection problems, even when the weighting matrix appears in the equation of the errors. Moreover, these different specifications are generally impossible to distinguish without assuming prior knowledge about the true data generating process that we often do not possess in practice. This decision is extremely important because if matrix W is misspecified in some way, parameter estimates are likely to be biased and they will be inconsistent in models that contain some spatial lag, as stated in Mur et al. (2011). Furthermore, the consequences for evaluating effects of policy decisions can be serious if model specification is not conducted properly.

Serious problems also arise if there is spatial correlation in the unobserved components of the model u_i. This may happen because of sorting when unobservable similar agents tend to be co-located. They might have common unobserved shocks or causal linkages between neighbors (unobserved characteristics). For simplicity, assume that neighborhood exogenous characteristics (Xw_i) do not directly affect outcomes:

$$y_i = \rho w'_i y + x'_i\beta + w'_i X\gamma + u_i.$$

The estimation of this model provides two coefficients which identify β but do not allow separate identification of ρ and γ. In this case it is impossible to distinguish between the ways spatial correlation is driven. In traditional spatial econometric models, it is the assumption that most standard W matrices are not idempotent, which allows identification (Gibbons and Overman, 2012).

2.1. Spatial models, parametric approach

Ideally, spatial economic theories should provide the researcher with sufficient prior information to enable the construction of fully specified spatial econometric models. In such a situation, the researcher can make an unambiguous choice from a wide range of possible model specifications and appropriate econometric/statistical methods in accordance with various criteria such as unbiasedness, consistency, efficiency, etc. Unfortunately, this is not the common situation in Spatial Econometrics. Therefore, researchers from the social sciences are confronted with substantial specification uncertainty.

Let us consider the OLS model[1] given by:

[1] We refer as OLS because it is commonly estimated by ordinary least squares (OLS).

$$Y = \alpha \iota_N + X\beta + \varepsilon, \tag{1}$$

where Y represents an $N \times 1$ vector consisting of one observation on the dependent variable for every unit in the sample ($i = 1, \ldots, N$), ι_N is an $N \times 1$ vector of ones associated with the constant term parameter, X denotes an $N \times K$ matrix of explanatory variables associated with the $K \times 1$ parameter vector, and $\varepsilon = (\varepsilon 1, \ldots, \varepsilon N)^T$ is a vector of independently and identically distributed disturbance terms with zero mean and variance σ^2.

Spatial econometrics literature has developed models that treat three different types of interaction effects among units:

- endogenous interaction effects among the dependent variables,
- exogenous interaction effects among the explanatory variables, and
- interaction effects among the error terms.

As we mentioned before, a number of models exist, where space enters into the equation through *W*, such as SAR, SEM, SAC, etc.[2] Taking into account the number of existing models, economists propose different approaches when it comes to choosing the best-fitting spatial model. Two of the most used are the *top-down* and the *bottom-up*.

The *top-down* approach consists of starting from the General Nesting Spatial (GNS) model that includes all types of interaction effects and is given by:

$$Y = \rho WY + \alpha \iota_N + X\beta + WX\theta + u, \; u = \lambda Wu + \varepsilon \tag{2}$$

where $W(N \times N)$ is the spatial weights matrix, which is assumed known and describes the structure of dependence between units in the sample. The variable WY denotes the endogenous interaction effects among the dependent variables, WX the exogenous interaction effects among the explanatory variables, and Wu the interaction effects among the disturbance terms of the different observations. The scalar parameters ρ and λ measure the strength of dependence between units, while θ, like β, is a $K \times 1$ vector of response parameters. The other variables and parameters are defined as in the OLS model (1). Since the GNS model incorporates all interaction effects, models that contain less interaction effects can be obtained by imposing restrictions on one or more of the parameters. Various methods can be applied to estimate spatial econometric models such as Maximum Likelihood (ML), Instrumental Variables or Generalized Method of Moments (IV/GMM), and Bayesian methods (Rossi (2018), Baum, et al. (2002), van de Schoot, et al. (2021)).

The *bottom-up* approach consists of starting with the non-spatial model (see Le Gallo, 2002 for a summary). The Lagrange multiplier tests (Anselin, 1988) for the SAR and SEM model specification tests, robust to the presence of other types of spatial interactions, are used to choose between SAR, SEM or non-spatial models. This approach was widely favored until the 2000s because the tests developed by Anselin et al. (1996) are based on the residuals of the non-spatial model. They are therefore inexpensive from a computational point of view. Florax et al. (2003) have also shown, using simulations, that this procedure was the most effective when the real model is a SAR or SEM model.

There is extensive literature on how the coefficients of each of the interaction effects can be estimated. However, considerably less attention has been paid to the interpretation of these coefficients. Many empirical studies use the point estimates of the interaction effects to test the hypothesis as to whether spillovers exist. Only recently, thanks to the work of LeSage and Pace (2009), researchers started to realize that this may lead to erroneous conclusions, and that a partial derivative interpretation of the impact from changes to the variables of different model specifications represents a more valid basis for testing this hypothesis.

Parametric methods are helpful in a lot of cases. However, they become unfeasible in the simultaneous presence of different sources of model misspecification, such as substantial spatial dependence, nonlinear

[2] We refer to these models as “classic spatial models”.

relationship of spatially correlated independent variables, unobserved spatial heterogeneity, spatially varying relationships, and common factors (Basile and Minguez, 2018). That leads to the impossibility of obtaining consistent and efficient estimates. Thus, a number of non-parametric and semiparametric frameworks, that are more flexible to be able to deal with the problem of spatial dependence, have been developed.

2.2. Semiparametric approach and Splines

Spatial econometric frameworks that include parametric methods appear to be unfeasible when another source of model misspecification appears. The latter can include substantial spatial dependence, nonlinear relationship of spatially correlated independent variables, unobserved spatial heterogeneity, spatially varying relationships, and common factors. Though non-parametric methods have already gained a great popularity in time series analysis, their usage in spatial econometrics is still scarce. Some contributions (Basile and Minguez, 2018 and Montero et al., 2012) attempt to promote a more flexible estimation framework to address this problem.

Nonparametric and semiparametric models are attractive alternatives to parametric variations because they admit at the start that the structure of a true model is unknown. This type of models can be used to carry out hypothesis testing, and they can be easily implemented.

Recently, Geniaux and Martinetti (2018) have introduced a new class of models, called MGWR-SAR (Mixed Geographically Weighted Regression Simultaneous Auto Regressive models), where the regression parameters and the spatial dependence coefficient can vary over space. In its most general form, the MGWR-SAR is specified as:

$$y \;=\; \rho(x_{s1}, x_{s2}; h)Wy + X^{*}\beta^{*} \; + \beta(x_{s1}, x_{s2}; h)X + \varepsilon$$

where y is the N-vector of the continuous dependent variable, X^{*} is a matrix of k_1 exogenous explanatory variables entering the model linearly (i.e. with spatially stationary coefficients β^{*}), while X is a matrix of k_2 exogenous explanatory variables with non-stationary coefficients $\beta(x_{s1}, x_{s2}; h)$, x_{s1} , x_{s2} are spatial coordinates, W is the spatial weights matrix, ρ the spatial spillover parameter and ε is an i.i.d. error vector.

In this way, they relax the hypothesis that the spatial parameter ρ and the regression parameter β are constant over the coordinate space. The value of these parameters, in fact, depends on the coordinates. The parameters $\rho(x_{s1}, x_{s2})$ and $\beta(x_{s1}, x_{s2})$, are only required to be spatially smoothed. The use of the Spatial Two-Stage Least Squares (S2SLS) technique is proposed for the estimation of these types of models. These authors propose a 5-step approach that uses, a local linear estimator (a variant of the GWR) and Cross Validation for the selection of the bandwidth parameter.

A characteristic of this approach is that it only considers spatial parameter heterogeneity (i.e. parameter heterogeneity over the space of coordinates), while neglecting the possibility of pure nonlinearities (i.e. parameter heterogeneity over the domain of the explanatory variable). However, it remains very important to assess the existence of pure nonlinearities in the relationship between the response variable and the covariates. Moreover, keeping the spatial autocorrelation parameter (ρ) constant over space can be a valid option: in that case, the feedback effects of spatial autocorrelation have a clearer definition and the interpretation of direct and indirect effects is easier.

Next, we will discuss the types of splines we are going to use in our analysis. We consider the following configurations of the nonparametric part:

$$\textit{Spline}\text{: } f(z) \;=\; f(a, b) \tag{3}$$

where $f(z)$ is fully nonparametric and is limited to longitude (a) and latitude (b) variables.

$$\begin{aligned} C - spline\text{: } f(x) &= \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 \\ &+ \beta_3(x - x_0)^3 + \textstyle\sum_{s=1}^{S} \delta_s(x - x_s)^3 D_s, \end{aligned} \tag{4}$$

where the spline simply adds a set of interaction terms between dummy variables and cubic terms to a standard cubic function, and where S is the number of equal length intervals ranging from $x_0 = min(x)$ to $x_S = max(x)$ and the dummy variable D_S indicates whether x is greater than x_s. C-spline is used as an analogy approximation to one of delta models (G-model) that is introduced in next section. This allows us to make better comparison of the models that are in the same analysis line. Lastly, a Fourier based spline of the form:

$$F - spline{:}\, f(z) = \beta_0 + \beta_1 z + \beta_2 \mathrm{z}^2 + \sum_{j=1}^{J} \gamma_s \sin(\mathrm{jz}) + \lambda_j \sin(\mathrm{jz}),$$

where $z = 2\pi(x - min(x))/(max(x) - min(x))$

It should be recalled that splines and series regression are based on the mathematical theory of the approximation of functions. Particularly, spatial-econometricians that are concerned with approximating the conditional expectation function, find the Weierstrass-Stone Theorem, which states that, under mild regularity conditions, any continuous function can be uniformly well approximated by a polynomial of sufficiently high order, very useful (Stone, 1948). There are mathematical results that point out that, when the true conditional expectation function is smoother, it is possible to approximate it with a fewer number of series terms. This explains why other spline methods like B-splines or P-Splines can be used instead of (or together with) the ones we have selected. The central point is the same one as in the delta-models that are introduced in the following section, which consider that basic coordinates can be a first step to control for spatial relationships. One or more of these simple structures can approximate a spatial trend even in the case of a nonlinear spatial trend.

3. Methodology

This part proposes statistically study other ways of incorporating space to control for unknown sources of spatial dependence before relying on W. We firstly focus on testing for (weak) spatial dependence in the presence of leading deterministic components, similar to time-series tests for unit roots in the presence of drift and/or trend. To do so, we rely on a recent statistical procedure based on symbolic entropy developed in Garcia-Cordoba et al. (2019) to determine whether a cross-sectional dataset is statistically compatible with a leading deterministic component in the form of a spatial trend. The possibility of some spatial trend capturing the spatial dependence is studied. Thirdly, for those cases that were found statistically compatible with a spatial trend, a geographically restricted semiparametric approach is proposed to specify a model avoiding the critical points on W.

3.1. Delta test

The delta-test, that we briefly describe below, tests for the null of the existence of a non- stochastic leading term in a spatial dataset $\{X_S\}s\in_S$ where S is a set of coordinates. To do so the spatial realization $\{X_S\}s\in_S$ is embedded in an m-dimensional space:

$$X_m(s_0) \; = \; (X_{s_0}, X_{s_1}, \dots, X_{s_{m-1}}) \text{ for } s_0 \in S$$

Where $\{s_1, \dots s_{m-1}\}$ are the $m - 1$ nearest neighbors to s_0. A symbolization map is then defined $f{:}\{X\}_{s\in S} \rightarrow \Gamma \subseteq \{0,1\} \mathrm{x}^{\text{(m-1 times)}} \mathrm{x}\{0,1\}$ as:

$$f(X_s) \; = \; (I_{s\, s_1}, I_{s\, s_2}, \dots, I_{s\, s_{m-1}}) \tag{6}$$

where $I_{s\, s_j}$ is an agreement indicator function of being above or below the median at locations s and s_j, Γ is the subset of $2^{\mathrm{m-1}}$ different vectors of dimension $m - 1$ with entries in the set {0, 1}, where we refer to each symbol by σ_{i} (see Garcia-Cordoba et al. (2019) for more details). Obviously, it is required that the spatial process $\{X_s;\ s\in S\}$ has a finite median, otherwise the test cannot be applied, which is not a very strict limitation. Then the relative frequency, p_σ, of each symbol is computed from the data, and the associated entropy of the dataset is calculated: $h(\Gamma) = -\sum_{\sigma\epsilon\Gamma} \rho_\sigma \ln\left(\rho_\sigma\right)$ The delta-test consists of estimating the behavior of a function of the difference between entropies $h^{W_{j+1}}(\Gamma) - h^{W_j}(\Gamma)$ where $\{\mathcal{W}\ j\ ;j=1,\ldots,k\}$ are sets of symbols chosen at random from Γ. Under the null of a non-stochastic spatial structure, that difference does not increase with the number of symbols considered.

Particularly, the delta-test is implemented by testing if $\alpha_1 = 0$ in the following regression:

$$dh^{W_j}(\Gamma) = \alpha_0 + \alpha_1 j + \varepsilon_j\,, \quad for\, j = 1,2,\ldots,k-1, \tag{7}$$

where

$$dh^{W_j}(\Gamma) = \frac{h^{W_{j+1}}(\Gamma) - h^{W_j}(\Gamma)}{log\left(\frac{j+1}{j}\right)}$$

As shown in Garcia-Cordoba et al. (2019), the delta statistic is a test well-suited to detecting simple and complex spatial trends. Provided with the delta- test, (*dh – test*), we can supplement the spatial analysis by applying the test to the spatial raw data. In case of an acceptance of the null hypothesis of the non-stochastic spatial leading term, the possibility of specification of a scenario with spatial deterministic trends opens up for the econometric modeler. A natural way for modeling this situation from an econometric point of view is by using what we call restricted semiparametric regression:

$$Y = \alpha\iota_n + X\beta + f(a, b) + \varepsilon, \tag{8}$$

where each element on vector Y is a continuous output variable in a given location and $X\beta$ contains all explanatory variables (i.e., a set of explanatory variables that can include categorical variables and where vector β collects fixed parameters) and the nonparametric part $f(a, b)$ is restricted to geographic functions of longitude and latitude, *a, b*, respectively. At this point, according to the acceptance of the null hypothesis of the non-stochastic spatial leading term, there is no evidence for introducing a weight matrix (W) into the model, neither in the parametric part $X\beta$ nor in the nonparametric one.

Several comments are important in this respect. The previous family of models aims to ascertain whether a specification of space via latitude and longitude might serve to control for spatial heterogeneity once the researcher has had statistical evidence of a spatial trend. At this stage, prior to the use of a given W weight matrix, we wonder if considering some form of geographical variables in the model is enough to correctly estimate vector β. This will avoid the severe consequences in estimation and inference (about β) of not considering spatial heterogeneity when it really exists, as occurs in many fields. Notice also that the family of models (8) will not be the object of the main critiques that spatial econometrics has received by scholars, upon which we have commented in the previous section.

The delta-test can be used as a diagnostic tool helping in the model selection procedure. Consider a model that erroneously omits some form of spatial dependence:

$$Y = \alpha\iota_n + X\beta + u,$$

we understand that the omission can be in form of a linear spatial dependence or in the form of a spatial trend. An example of the former is

$$u := WX\theta + \varepsilon,$$

while the latter can be of the form

$$u := \mu f(a, b) + \varepsilon.$$

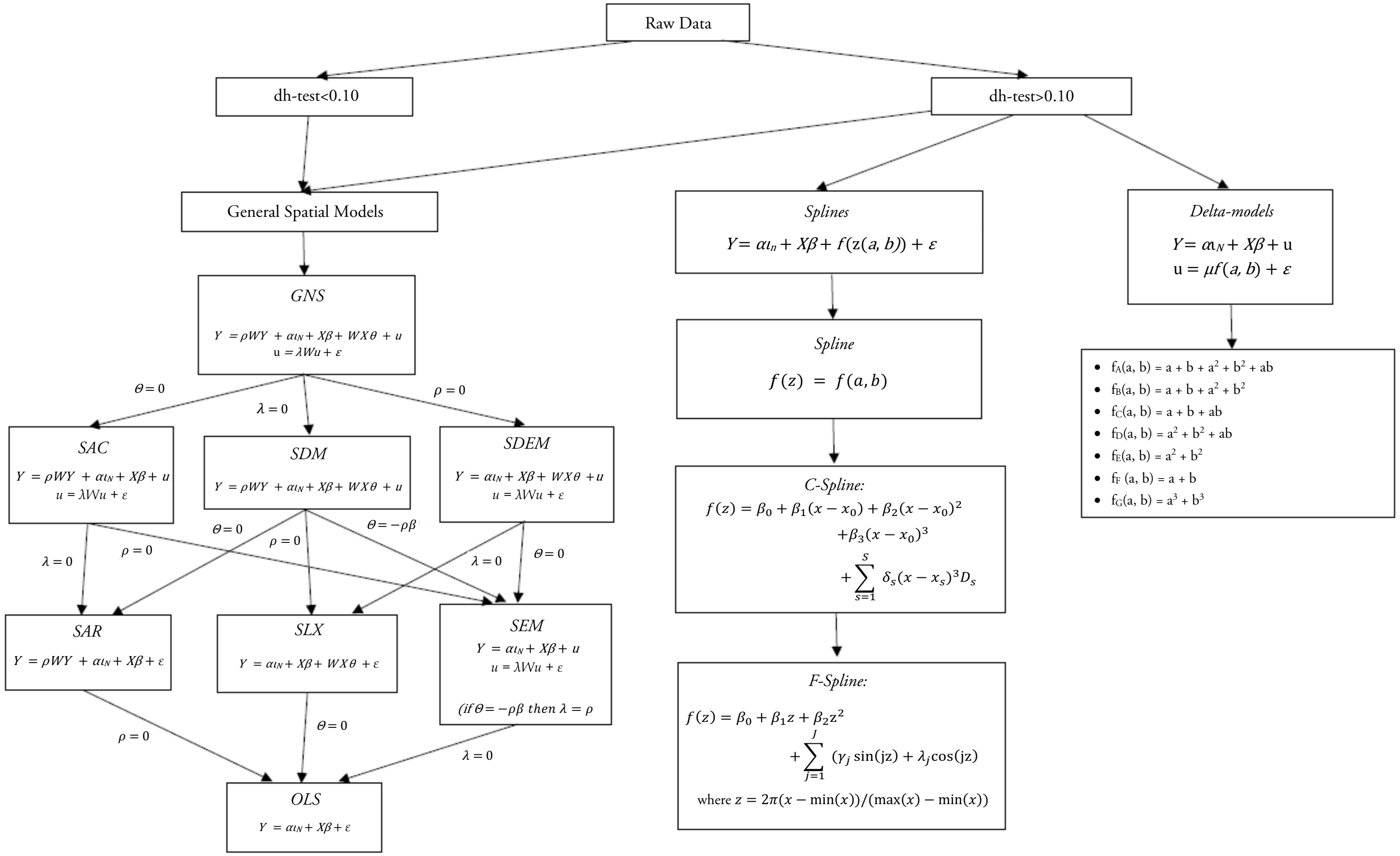
How to choose between these specifications is far from straightforward. As shown in Garcia-Cordoba et al. (2019), the delta-test can be used to distinguish between them if the test is applied to the residuals of the mis specified model, that is, if it is applied to $\hat{u}$. In the case of a true spatial dependence via *W*, the delta-test will tend to point out that no spatial trend is found in the residuals, and therefore the researcher will have to deal with a statistically correct specification of the model (this will probably be done through well-known models in the spatial econometrics literature, as we indicate later in this paper). In this regard, we will expect that Moran's *I* test will correctly indicate spatial autocorrelation in the residuals. On the contrary, the delta-test will highlight that a spatial trend is omitted if the true spatial dependence comes in the form of a non-stochastic geographic spatial structure (spatial trend). Obviously, the researcher should now take a different modeling strategy, as he/she has put forward a statistically compatible spatial trend. In other words, the proposal of some form of $f(a, b)$ should be required.

Our procedure consists of specifying the model using the previous diagnostics' tools. Particularly we firstly run delta-test on the raw data to check for the existence of a deterministic structure and Moran's test to check if there is a spatial autocorrelation in the data we use. If delta-test cannot reject the statistical existence of a spatial trend, we introduce a geographical additive model of the form given in (8). In particular, we consider and study two forms for the restricted nonparametric part, $f(a, b)$. The first way (that we will refer to as delta-model strategy) is to restrict $f(a, b)$ to be low-degree polynomials of geographical coordinates, which is inspired by the practice of including powers of *t*-time in time-series modeling:

- $f_A(a, b) = a + b + a^2 + b^2 + ab$
- $f_B(a, b) = a + b + a^2 + b^2$
- $f_C(a, b) = a + b + ab$
- $f_D(a, b) = a^2 + b^2 + ab$
- $f_E(a, b) = a^2 + b^2$
- $f_F(a, b) = a + b$
- $f_G(a, b) = a^3 + b^3$

We will use letters *A, B, ..., G* to indicate the model specification we refer to. For example, by Model B we will mean $Y = \alpha\iota_n + X\beta + f_B(a, b) + \varepsilon$. The choice of the best delta model is done based on the results of Moran test and delta test. The procedure we follow to make our analysis is presented in Figure 1 (Makeienko, 2020).

FIGURE 1.
Procedure of choosing a model[3]


Raw Data
dh-test<0.10
dh-test>0.10
General Spatial Models
Splines
$Y = \alpha\iota_n + X\beta + f(z(a, b)) + \varepsilon$
Delta-models
$Y = \alpha\iota_N + X\beta + u$
$u = \mu f(a, b) + \varepsilon$
GNS
$Y = \rho WY + \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
$\Theta = 0$
$\lambda = 0$
$\rho = 0$
SAC
$Y = \rho WY + \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
SDM
$Y = \rho WY + \alpha\iota_N + X\beta + WX\theta + u$
SDEM
$Y = \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
$\lambda = 0$
$\rho = 0$
$\Theta = 0$
$\rho = 0$
$\Theta = -\rho\beta$
$\lambda = 0$
$\Theta = 0$
SAR
$Y = \rho WY + \alpha\iota_N + X\beta + \varepsilon$
SLX
$Y = \alpha\iota_N + X\beta + WX\theta + \varepsilon$
SEM
$Y = \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
$(if\ \Theta = -\rho\beta\ then\ \lambda = \rho$
$\rho = 0$
$\Theta = 0$
$\lambda = 0$
OLS
$Y = \alpha\iota_N + X\beta + \varepsilon$
Spline
$f(z) = f(a, b)$
C-Spline:
$f(z) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \beta_3(x - x_0)^3 + \sum_{s=1}^{S} \delta_s(x - x_s)^3 D_s$
F-Spline:
$f(z) = \beta_0 + \beta_1 z + \beta_2 z^2 + \sum_{j=1}^{J} (\gamma_j \sin(jz) + \lambda_j \cos(jz))$
where $z = 2\pi(x - \min(x))/(\max(x) - \min(x))$
$f_A(a, b) = a + b + a^2 + b^2 + ab$
$f_B(a, b) = a + b + a^2 + b^2$
$f_C(a, b) = a + b + ab$
$f_D(a, b) = a^2 + b^2 + ab$
$f_E(a, b) = a^2 + b^2$
$f_F(a, b) = a + b$
$f_G(a, b) = a^3 + b^3$

[3] dh-test represents the p-value of the delta test.

3.2. Software

In order to apply the methodology, the first software program used in the analysis performed is the toolbox for spatial econometric models written by LeSage and Pace (2009) in MATLAB (MATLAB 2017). Some functions, also in MATLAB, are used to estimate static and dynamic spatial panel data models developed by Elhorst (2013). Moreover, the R packages *spdep* and *sp* (Pebesma and Bivand (2005), Bivand et al. (2013)) are used. They facilitate the creation, transformation and manipulation of spatial objects, neighborhood matrices and the computation of descriptive measures of spatial autocorrelation. Focusing on semiparametric spatial data models, *McSpatial* package from McMillen (2015), McMillen (2012) includes routines to estimate nonparametric and conditionally parametric versions of spatial linear regression and spatial models with a binary dependent variable. It mainly uses kernel techniques to perform the nonparametric estimations.

4. Empirical Illustrations

This section of analysis is based on three different datasets. Each of the dataset includes the full information on the object of the analysis, where a certain relation can be found. Apart from the special characteristics of the units, every dataset includes the information of the geographical position (longitude and latitude) of the units described. Thus, we have a possibility to compare the general characteristics of data analyzed to produce a better methodology of specifying a trend methodology (including a delta test usage). The process of choosing the best model and main steps of the analysis are based on the scheme presented in the Figure 1[4]. As mentioned before, we present the results of only 3 datasets in total, however, more than 15 different datasets with similar characteristics were previously analyzed. Taking into account, that the first step of our analysis reveals the existence (or absence) of the spatial dependence in the raw dataset, using Moran's *I* test, we have found that only 6 out of 15 datasets presented the existence of spatial autocorrelation in it. The second step is to check the existence of deterministic component in the data, using the dh-test. It might be the case, that the data presents the existence of spatial dependence, but not the deterministic part.5 Nevertheless, we present all the datasets where the existence of spatial dependence was confirmed. Two of them resulted having no deterministic component. Still, we use these datasets to additionally analyze probable common characteristics to be taken into account for our further research.

4.1. NUTS2

This dataset is analyzed by estimating a number of growth regression models on a sample of 249 NUTS 2 regions belonging to the enlarged Europe (EU 27). We start from the linear specification of the neoclassical growth model proposed by Mankiw et al. (1992). The dependent variable is the per-worker income growth rate, $gy = lny_T - lny_0$, computed for the 1990-2004 period. The model predicts that gy is higher in the economies with higher rates of investment in physical and human capital (s_K and s_H, respectively), lower initial conditions, $ln\ y_0$, and lower effective depreciation rates ($n + g + d$), with n the working-age population growth rate, g the common exogenous technology growth rate and d the rate of depreciation of physical capital assumed identical in all economies. Basic data to measure these variables come from the EUROSTAT Regio and Cambridge Econometrics databases, which include information on real gross value added, employment, investment and tertiary education. We measure per worker income levels, y, as the ratio between total real value added and total employment; the physical capital accumulation rate, s_{K}, as the average share of gross investments on real gross value added; the human capital accumulation rate, s_H, as the percentage of a region's working population that is in the tertiary level of the education process; n is the average growth of total employment; and we also use the information on GDP. Finally, we assume, as it is usual, that ($n + g + d$) is equal to 0,05 (see Mankiw et al., 1992). Furthermore, as mentioned before, we include the information on longitude and latitude as well to control for the geographical position of the unit analyzed.

[4] The explicit explanation of the steps and models results for each dataset can be seen in detail in Appendix.

[5] Case of Chicago Airbnb and Earthquake datasets, more information can be found in Appendix.

TABLE 1.
Results for NUTS2 Dataset

	OLS	SAC	GNS	C-spline	F-spline	Model E	Model G
Constant	0.093	0.096***	0.080***	0.055	-0.018	0.098***	0.088**
Human capital	0.000	-0.003	-0.003	-0.001	-0.001	0.001	0.001
GDP	-0.001	0.002**	0.002*	0.001	0.001	-0.001	-0.001
Population Growth	0.000	0.019**	0.015*	-0.004	-0.004	0.001	-0.002
Physical Capital	0.035	0.022***	0.021***	0.033***	0.033***	0.036	0.036***
Moran(p-value)	0	0.68	0.72	0.01	0.01	0	0
dh-test(p-value)	0.07	0.01	0.02	0.13	0.12	0.07	0.09
	OLS	SAC	GNS	C-spline	F-spline	Model E	Model G

#of embedding dimensions m=6, dh-test p-value on raw data 0.12, Moran test p-value on raw data 0.001
***, **, * =coefficient estimates that are significant at the 0.01, 0.05 and 0.1 level respectively.

Moran's test on the data on the NUTS2 dataset, gives a clear evidence of the spatial autocorrelation (Table 1). The delta-test on the raw data confirms the presence of deterministic structure, that gives evidence in favor of running restricted semiparametric analysis, including spatial trend. Following the modeling proposal of the paper, we firstly model the deterministic part by using the so-called delta-models. Results for both models G and E are clearly in favor, as controlling for spatial trend is concerned (p-value = 0.09, 0.07, respectively). However, based on the Moran test results we cannot be sure that the estimated model controls for the spatial heterogeneity of the data. The same conclusion is reached if we opt by some spline-based methods.

If instead we model according to the classic spatial models, we find that the best spatial models for our data are SAC and GNS models, based on test results and AIC criteria (Appendix). One interesting conclusion of the results found, is that neither SEM nor the SDEM models are able, according to delta-test results, to remove the previously found spatial trend (see Table A.1 in the Appendix). In other words, the residuals of these models are compatible with a deterministic structure that have not been yet removed. For this reason, results seem to point that restricted semiparametric models work better in this case, as they let us get rid of the spatial structure of the model and thus get more credible results on the estimates. The practical implications for NUTS2 dataset are mainly relative to the partial effects of several explanatory variables, but not to the list of significant variables, nor to the signs in general if different approaches are analyzed. However, considering only SAC and GNS models, one can observe that both the significance and the sign of the significant variables coincide in both models.

4.2. GECON

Another dataset is based on the G-Econ data. The G-Econ research project is devoted to developing a geophysical based data set on economic activity for the world. Current dataset used in the performed analysis (GEcon 4.0) is now publicly available and covers "gross cell product" for all regions for 1990, 1995, 2000, and 2005 and includes 27,500 terrestrial observations. The basic metric is the regional equivalent of gross domestic product. Gross cell product (GCP) is measured at a 1-degree longitude by 1-degree latitude resolution at a global scale. This dataset includes such characteristics as:

- Gross cell product, 2005 US $ at market exchange rates, 2000
- Distance to coast (km)
- Elevation (km)
- Distance to major navigable lake (km)
- Distance to major navigable river (km)

- Distance to ice-free ocean (km)
- Distance to navigable river (km)
- Vegetation category
- Grid cell population, 2000
- Average precipitation, prior data
- Soil category
- Average temperature, prior data
- Geographical position (1ongitude and 1atitude)

This dataset is interesting mainly because of the complete information on the geographical characteristics, that might be important when analyzing data with spatial components.

As in the previous case, Moran's test on the data on the GEcon dataset, gives clear evidence of the spatial autocorrelation (Table 2). Moreover, the presence of deterministic structure is confirmed by running a delta-test (p-value=0.39). We apply restricted semiparametric analysis, including spatial trend and classical standard models. In this case, no model (either classical or delta) is able to control for the deterministic component (see Table A.2 in the Appendix). In this scenario, our reasonable choice would be GNS and SAC models that at least can correct for spatial structure in the sense that Moran's test statistically indicates that spatial structure has been controlled, see Table 2. As happened with the previous data set, the list of relevant explanatory variables is common to all the models. Variations are again on the partial effects. However, this situation is compatible with potential spatial units roots in the variables, and therefore it should be advisable to test for it as soon as there are available tests.

4.3. California housing prices

Next dataset is the most common dataset on housing prices. This is the dataset used in Geron (2017), that contains information from the 1990 California census and pertains to the houses found in a given California district and some summary statistics about them based on the 1990 census data. The variables we use are as follows:

- Housing median age
- Total room number
- Total bedrooms number
- Population
- Households
- Median income
- Median house value
- Proximity to the ocean(km)
- Geographical position (1ongitude and 1atitude)

We got the information on the variables in using all the block groups in California from the 1990 Census. In this sample, a block group on average includes 1425.5 individuals living in a geographically compact area. Naturally, the geographical area included varies inversely with the population density. We computed distances among the centroids of each block group as measured in latitude and longitude. We excluded all the block groups reporting zero entries for the independent and dependent variables. The final data contained 20640 observations on 9 characteristics. Table 3 present the results of the analysis performed.

TABLE 2.
Results for G-Econ Dataset

	OLS	**SAC**	**GNS**	**C-spline**	**F-spline**	**Model A**	**Model G**
Constant	3.805***	4.198***	3.252***	2.597***	10.73**	2.664	3.494***
Distance to coast (km)	-471.761	107.520***	139.548***	-567.8	-549.8	-516.009	-471.904
Distance to coast (km)	-0.001	-0.004**	-0.004**	0.001	0.001	-0.001	-0.001
Elevation (km)	-0.001*	0.001	0.001	-0.002***	-0.002***	-0.002***	-0.001*
Dist. to mn lake (km)	0.001***	0.001	0.01***	-0.001***	-0.001***	-0.001	-0.001**
Dist. to mn river (km)	0.001	-0.001	-0.01*	-0.001	0.001	0.001	0.001
Dist. to ice-free ocean (km)	0.472	-0.108***	-0.139***	0.568	0.055	0.516	0.472
Dist. to navigable river (km)	-0.001***	-0.001***	-0.001***	0.001	0.001	-0.001**	-0.001***
Veg. category	0.056***	0.018	0.019	0.047*	0.047*	0.063***	0.056***
Grid cell population, 2000	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
Avg precipitation, prior data	0.001**	-0.001**	-0.001**	0.001	0.001	0.001	0.001**
Soil category	0.004**	0.003	0.003	0.003	0.004*	0.002	0.003
Avg temperature, prior data	-0.189***	-0.204***	-0.199***	-0.097***	-0.089***	-0.168***	-0.191***
Moran(p-value)	0	0.82	0.87	0	0	0	0
dh-test(p-value)	0.42	0.36	0.35	0.42	0.43	0.29	0.42
	OLS	SAC	GNS	C-spline	F-spline	Model A	Model G

#of embedding dimensions m=12 dh-test p-value on raw data 0.39 Moran test p-value on raw data 0.01
***, **, * =coefficient estimates that are significant at the 0.01, 0.05 and 0.1 level respectively.

TABLE 3.
Results for California housing prices Dataset

	OLS	SLX	SAC	C-spline	F-spline	Model D	Model H
Constant	-46139.647***	32398.866***	30653.921***	-517600000*	1704000***	-1975883.03***	-660179826.1***
Housing median age	1882.121***	1149.584***	1212.721***	1145.00***	1127.00***	1172.236***	1141.883***
Total room number	-19.733***	-9.949***	-7.288***	-7.97***	-7.85***	-8.067***	-7.071***
Bedroom number	100.944***	75.520***	53.326***	115.80***	116.1***	117.242***	89.731***
Population	-35.319***	-30.679***	-30.274***	-38.76***	-38.79***	-37.416***	-38.025***
Households	124.803***	76.439***	84.744***	45.19***	44.04***	40.836***	67.839***
Median income	47748.381***	36562.699***	36141.689***	4025.00***	4016.00***	40327.951***	39785.866***
Moran(p-value)	0	0	0	0	0	0	0
dh-test(p-value)	0.10	0.09	0.05	0.09	0.11	0.09	0.09
	OLS	SLX	SAC	C-spline	F-spline	Model D	Model H

#of embedding dimensions m=12 dh-test p-value on raw data 0.40 Moran test p-value on raw data 0.001
***, **, * =coefficient estimates that are significant at the 0.01, 0.05 and 0.1 level respectively.

Running the dh-test for the data, we get that there is a spatial trend in the data (dh-test (p-value) =0.40). Following the same steps as before, we can conclude, that neither delta models, nor classic spatial models can control both spatial heterogeneity and deterministic part. However, even we cannot control for spatial heterogeneity and given the statistical importance of controlling for spatial trends, it is worth mentioning that there are several models that perform well controlling the deterministic part of the data (see Table A.3 in the Appendix). Thus, SLX and SAC classic models and Models D and H are models of different nature that can control the spatial trend of the data, although none of them capture the spatial dependence found in the residuals. In this case, it seems that the perils of not modelling a spatial deterministic trend and/or a spatial unit root have been under control, and therefore other forms of weak dependence (different from the ones used in this section) might try to solve the specification problem.

5. Conclusions

The general contribution of this work is on methodological aspects regarding with challenge of dealing with different forms of spatial dependence: we study several standard tools and other newer ones to see how they perform in data sets and models of very different nature. Once we recognized the limited ability to accurately model spatial data, it is important to explore how different analysis techniques perform once applied to different types of data. This allows us to make the process of analysis more efficient and precise, when trying to overcome the problems we might face when processing spatial data. We add to the importance of specifications tests usage in order to validate general results. In particular, this paper has studied different types of spatial data to be able to highlight some common characteristics, both for datasets where the spatial part is controlled by means of what has been labeled as delta-models, and other datasets where classic models perform better when addressing spatial dependence.

The general results of the analysis performed allows us to draw some conclusions and to open new questions related with modelling different forms of spatial dependence. First, neither delta models, nor classic spatial models can control the spatial component of the data in all the types of data we have chosen for our analysis. Among other things, this means that it might be other techniques to deal with spatial dependence, particularly it would be worth studying whether a spatial unit root can help in the modelling procedure. The literature on spatial unit root is quite limited, although promising steps are currently being given (see Baltagi and Shu, 2024). Similar as it happens within the context of time series analysis, a wrong distinction between spatial trend and a spatial unit root might easily drive researchers to statistically invalid conclusions. This explains why it is relevant to deal with spatial trends, in the sense of modelling them properly. Second, given the importance of treating spatial trends as better as possible, we have observed that delta models seem to perform better with the data that have some specific theoretical model behind, as in the case of NUTS2 data, where we found that the spatial trend is controlled with some of these delta models, while this does not happen for the classical ones. On the other hand, we also observed that classic spatial models perform better with the data that have some detailed geographic information, as in the case of GEcon dataset, despite the fact that in this dataset classical models are also unable of dealing with spatial trends. In this scenario, it would be advisable to use some spatial unit root test. Finally, considering that the California dataset has a lot in common with other dataset that are based on the hedonic models, we find that both classical and delta models can deal with the spatial trend, although there still is room for modelling spatial heterogeneity (in a form of weak dependence). This could be due to the lack of data in this dataset, as adding some more characteristics might help delta models in controlling for the trend. Other datasets have not presented any clear evidence in favor of classic or delta models.

The next steps of our research might include the application of the methodology developed to datasets with more detailed characteristics. Moreover, a step-by-step analysis might be considered, repeating the same analysis when adding characteristics one by one. This is one of the ways to detect crucial characteristics of the observations, that can help us to control both the spatial heterogeneity and spatial deterministic part.

References

Anselin, L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical analysis*, *20*(1), 1-17.

Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, *26*(1), 77-104.

Baltagi, B.H., & Shu, J. (2024). A Survey of Spatial Unit Roots. *Mathematics 2024*, *12*, 1052. https://doi.org/10.3390/math12071052

Basile, R., and Minguez, R. (2018). *Advances in Spatial Econometrics: Parametric vs. Semiparametric Spatial Autoregressive Models*, (pp. 81-106). Springer.

Bivand, R., Pebesma, E., and Gomez Rubio, V. (2013). *Applied Spatial Data Analysis with R* (pp. 59-82). Springer.

Elhorst, J. P. (2013). *Spatial Econometrics*. Springer-Verlag Berlin Heidelberg.

Florax, R., Folmer, H., and Rey, S. (2003). Specification searches in spatial econometrics: The relevance of hendry's methodology. *Regional Science and Urban Economics*, *33*, 557-579.

Garcia-Cordoba, J., Matilla-Garcia, M., and Ruiz Marin, M. (2019). A test for deterministic dynamics in spatial processes. *Spatial Economic Analysis*, *14*(3), 361-377.

Geniaux, G., and Martinetti, D. (2018). A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models. *Regional Science and Urban Economics*, *72*, 74-85.

Geron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gibbons, S., and Overman, H. G. (2012). Mostly pointless spatial econometrics? *Journal of Regional Science*, *52*(2), 172-191.

Le Gallo, J. (2002). *Econometrie spatiale: l autocorrelation spatiale dans les modeles de regression lineaire*.

LeSage, J., and Pace, R. (2009). *lntroduction to Spatial Econometrics. Statistics: A Series of Textbooks and Monographs.* CRC Press.

Makeienko, M. (2020). Symbolic Analysis Applied to the Specification of Spatial Trends and Spatial Dependence. *Entropy*, *22*(4), 466. https://doi.org/10.3390/e22040466

Mankiw, N., Romer, D., and Weil, D. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, *107*(May), 407-437.

The MathWorks Inc. (2017). MATLAB version: (R2017a). The MathWorks Inc. https://www.mathworks.com

McMillen, D. (2012). Perspectives on spatial econometrics: linear smoothing with structured models. *Journal of Regional Science*, *52*(2), 192-209.

McMillen, D. (2015). *Package McSpatial.*

Montero, J., Minguez Salido, R., and Durban, M. (2012). Sar models with nonparametric spatial trends. a p-spline approach. *Estadistica Española*, *54*, 89-111.

Mur, J., Herrera, M., and Ruiz, M. (2011). Selecting the W matrix. Parametric vs Non Parametric Approaches. MPRA Paper No. 71181, posted 11 May 2016 15_26 UTC.

Müller, U., & Watson, M. (2023). *Spatial Unit Roots, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"*. ZBW - Leibniz Information Centre for Economics.

Pebesma, E., and Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, *5*.

Rossi, R. J. (2018). *Mathematical Statistics: An Introduction to Likelihood Based Inference* (pp. 227). John Wiley & Sons.

van de Schoot, R., Depaoli, S., & King, R. et al. (2021) Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1. https://doi.org/10.1038/s43586-020-00001-2

ORCID

Maryna Makeienko	https://orcid.org/0000-0002-5949-4856
Mariano Matilla-García	https://orcid.org/0000-0002-6007-3522

Appendix

Selection of embedding dimensions(m), number of neighbors and weight matrix W

The user must choose the parameter m, which determines the number of symbols used to analyze the dependence structure in spatial processes. In particular, the number of symbols increases according to 2^{m-1}. A larger number of symbols results in a larger number of spatial structures or number of potentially recognizable spatial patterns. Therefore, the greater the *m*, the better or finer will the test be at detecting deterministic spatial dependences. However, it is impossible to increase *m* if the sample size does not increase as well. Thus, given the number of observations, we selected *m* that is as large as possible to obtain a larger range of recognizable spatial structures. A simple rule is proposed to select *m*: use the largest *m* subject to the restriction that, given the number of spatial observations, the ratio number of observations on number of symbols is ≥5. The selection of the neighbors is done based on the value of *m* as well. The weight matrix *W* is constructed with geographical coordinates to produce spatial contiguity weight matrices with Delaunay routine. It can be changed based on the users´ selection.

The process of choosing the best model

After running the spatial models' regressions, one of the criteria used to choose the best model are the ones that control for spatial structure or the deterministic component, based on Moran and delta-test. Another criterion is the likelihood ratio (LR) test based on the log-likelihood function values of the different models. The LR test is based on minus two times the difference between the value of the log-likelihood function in the restricted model and the value of the log-likelihood function of the unrestricted model: −2 × (logLrestricted − logLunrestricted). This test statistic has a Chi squared distribution χ_n^2 with n degrees of freedom equal to the number of restrictions imposed. The election rule states that if LRtest > $\chi_{n,\alpha}^2$ where α is the signification level and $\chi_{n,\alpha}^2$ is the (1-α)-quantile of the Chi squared distribution χ_n^2, then the unrestricted model performs better than the restricted one. Using this criterion we can make a comparison of the models, as detailed below.

- OLS vs SLX
- OLS vs SAR
- OLS vs SEM
- SAR vs SAC
- SEM vs SAC
- SLX vs SDM
- SEM vs SDM

- SAR vs SDM
- SLX vs SDEM
- SAR vs SDEM
- SEM vs SDEM

Other models cannot be compared among themselves with LR test, as they are not nested. The only two models that are not nested but can be compared are SAR and SEM models. The criteria used to make a comparison are the Lagrange multiplier tests that make it possible to choose between spatial or non-spatial model according to the mechanism we explain below. The rest of the models can only be subjectively compared, checking the significance of the estimated coefficients and the structure of the model.

The last criteria when working with the rest of the models, as a mix of classic models, delta models and splines is an AIC criterion.

Additional results

The analysis is based on the datasets that include the full information on the object of the analysis, where a certain relation can be found. Apart from the special characteristics of the units, every dataset includes the information of the geographical position (longitude and latitude) of the units described. Thus, we have a possibility to compare the general characteristics of data analyzed to produce a better methodology of specifying a trend methodology (including a delta test usage). We present the results of only 5 datasets in total, however, more than 15 different datasets with similar characteristics were previously analyzed. Taking into account, that the first step of our analysis reveals the existence (or absence) of the spatial dependence in the raw dataset, using Moran's *I* test, we have found that only 6 out of 15 datasets presented the existence of spatial autocorrelation in it. The second step is to check the existence of deterministic component in the data, using the dh-test. It might be the case, that the data presents the existence of spatial dependence, but not the deterministic part. Nevertheless, we present all the datasets where the existence of spatial dependence was confirmed. Two of them resulted having no deterministic component. We use these datasets to additionally analyze probable common characteristics to be taken into account for our further research.

This first dataset includes data on the Airbnb prices in Chicago. The data were collected on October 3rd, 2015 and includes 77 observations from 2008 to 2015. It includes response rate, acceptance rate, review rating, price per included guest, room type (1 is entire home/apartment, 2 is private room, and 3 shared room), number of Airbnb spots. The socioeconomic indicators are percentages by community area: households below poverty, housing crowed, under 18 or over 64 years old (dependency), aged 25+ without high school diploma, and unemployed above 16 years old. Also per capita income and hardship index are included. These indicators were built for the period 2008 – 2012. The crime data include the number of crimes (battery, burglary, gambling, homicide, kidnapping, robbery, stalking, homicide, and theft, among others; murders with data for each victim are not included) and thefts from October 2014 to September 2015 (one year before the Airbnb data). Population by community area based on Census 2010 data.

The first step taken was to check if there exists any deterministic component in the data. In this case, the result is negative, there is no deterministic part that could be controlled. Thus, we do not proceed with the whole analysis and pass to the next dataset.

Another dataset contains data about the earthquakes that hit the center of Italy between August and November 2016. The data was taken from the National Earthquake Information Center (NEIC), that determines the location and size of all significant earthquakes that occur worldwide and disseminates this information immediately to national and international agencies, scientists, critical facilities, and the general public.

The NEIC compiles and provides to scientists and to the public an extensive seismic database that serves as a foundation for scientific research through the operation of modern digital national and global seismograph networks and cooperative international agreements. The NEIC is the national data center and archive for earthquake information. This dataset includes a record of the date, time, location, depth, magnitude, and source of every earthquake with a reported magnitude 5.5 or higher since 1965 with 8087 observations.

Same as in the case of the previous dataset, the results of the delta-test show no evidence of deterministic component in the data structure. Taking this into account, we proceed with the next datasets, where we will be able to take all the necessary steps of the analysis.

Table A.1.
Results for NUTS2 Dataset

	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH	SAC	SAR	SDEM	SDM	SEM	SLX
Constant	0,080”’	0,093	0,007”	0,08	0,000”	0,09”’	-0,09”’	0,077”	0.088”	0.049	0,096”’	0,058”	0,094”’	0,074”	0,096”’	0,090”’
Human capital	-0,003	0,000	-0,001	-0,001	0,000	-0,001	0,000	0,000	0,001	-0,002	-0,003	-0,002	-0,002	-0,003	-0,003	-0,002
GDP	0,002’	-0,001	0,000	0,000	0,000	0,000	-0,001	0,000	-0,001	0,001	0,002”	0,001	0,001	0,002	0,002”	0,001
Population Growth	0,015’	0,000	0,008	0,006	-0,002	0,004	0,001	-0,003	-0,002	0,015	0,019”	0,009	0,015’	0,014	0,019	0,015
Physical Capital	0,021”’	0,035	0,033”’	0,033”’	0,033”’	0,033”’	0,036”’	0,034”’	0,036”’	0,030”’	0,022”’	0,020”’	0,023”’	0,021”’	0,023”’	0,021”’
WHuman capital	0,010’												0,009	0,010		0,010
WGDP	-0,007”’												-0,007”	-0,007”		-0,009”’
WPopulation Growth	-0,019”												-0,029”’	-0,016”		-0,042”’
WPhysical Capital	-0,002												0,010	-0,004		0,023”’
rho	0,591”’										-0,115	0,660”’		0,662”’		
lambda	0,123										0,842”’		0,696”’		0,794”’	
Moran(p-value)	0,72	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,68	0,00	0,00	0,00	0,00	0,00
dh-test(p-value)	0,02	0,07	0,29	0,31	0,13	0,14	0,10	0,13	0,10	0,11	0,02	0,10	0,10	0,10	0,15	0,07
	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH	SAC	SAR	SDEM	SDM	SEM	SLX

Table A.2.
Results for G-Econ Dataset

	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH	SAC	SAR	SDEM	SEM	SLX
Constant	3,252”’	3,805”’	0,000’	0,000	0,000”	0,000”’	0,001”	0,019”’	0,000	0,000	4,199”’	1,156”	3,084”’	3,144”’	3,350”’
Distance to coast (km)	139,548”’	-471,761	-516,009	-423,227	-507,959	-590,273	-462,758	-462,758	-471,904	-1211,060	107,525”’	-237,028”’	-100,677”’	55,290”’	-546,409
Distance to coast (km)	-0,004”	-0,001	0,000	-0,001	-0,001	0,000	-0,001	-0,001	-0,001	0,001	-0,004”	-0,001”’	-0,005”	-0,003”’	-0,004
Elevation (km)	0,001	-0,001’	-0,002”’	-0,002”’	-0,002”’	-0,002”’	-0,001	-0,002”’	-0,001’	-0,003”’	0,001	0,000	0,002	0,000	0,002
Dist. to mn lake (km)	0,000”’	0,000”’	0,000	0,000”	0,000	0,000	0,000”’	0,000”	0,000”	0,000”	0,000	0,000	0,000”’	0,000	0,000’
Dist. to mn river (km)	0,000’	0,000	0,000	0,000”’	0,000	0,000”’	0,000	0,000”’	0,000	0,000	0,000	0,000	0,000	0,000	0,000’
Dist. to ice-free ocean (km)	-0,140”’	0,472	0,516	0,423	0,508	0,590	0,463	0,463	0,472	1,211	-0,108”’	0,237”’	0,101”’	-0,055”’	0,546
Dist. to navigable river (km)	-0,001”	-0,001”’	-0,001”	-0,001”	-0,001”’	-0,001”	0,000	-0,001”’	-0,001”’	0,000	-0,001”’	-0,001”’	-0,001”	-0,001”’	-0,001’
Veg. category	0,019	0,056”’	0,063”’	0,062”’	0,061”’	0,061”’	0,056”’	0,060”’	0,056”’	0,062”’	0,018	0,060”’	0,016	0,034	0,019
Grid cell population, 2000	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’	0,000”’
Avg precipitation, prior data	-0,001”	0,001”	0,000	0,000	0,000’	0,000	0,001”’	0,000’	0,001”	0,000	-0,001”	0,000”	-0,001”	0,000	-0,001”
Soil category	0,003	0,004”	0,002	0,002	0,002	0,002	0,004’	0,002	0,003	0,002	0,003	0,004”	0,002	0,003	0,003
Avg temperature, prior data	-0,200”’	-0,189”’	-0,168”’	-0,163”’	-0,181”’	-0,158”’	-0,086	-0,174”’	-0,191”’	0,006	-0,204”’	-0,144”’	-0,201”’	-0,196”’	-0,189”’
WDistance to coast (km)	405,584”’												-899,864”’		-5019,592
WDistance to coast (km)	0,005												0,005		0,004
WElevation (km)	-0,004												-0,003’		-0,004”
WDist. to mn lake (km)	0,000”’												0,000”’		0,000”’
WDist. to mn river (km)	0,000												0,000		0,000
WDist. to ice-free ocean (km)	-0,406”’												0,900”’		5,020

Table A.2. cont.
Results for G-Econ Dataset

	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH	SAC	SAR	SDEM	SEM	SLX
WDist. to navigable river (km)	0,000												0,000		0,000
WVeg. category	0,045												0,020		0,018
WGrid cell population, 2000	0,000”’												0,000”’		0,000”’
WAvg precipitation, prior data	0,002”												0,002”’		0,002”’
WSoil category	0,003												-0,001		-0,002
WAvg temperature, prior data	-0,036												0,053		0,040
rho	-0,500”’										-0,586”’	0,436”’			
lambda	0,790”’										0,816”’		0,542”’	0,579”’	
Moran(p-value)	0,874	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,818	0,000	0,000	0,000	0,000
dh-test(p-value)	0,351	0,420	0,288	0,313	0,421	0,416	0,420	0,423	0,420	0,426	0,362	0,421	0,400	0,421	0,398
	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH	SAC	SAR	SDEM	SEM	SLX

TABLE A.3.
Results for California Prices Dataset

	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH
Constant	30949,965’’’	-46139,647’’’	22230,990’’’	-1869,001’’’	679,201’’’	2130,442’’’	-569,229’’’	-42509,737’’’	-10,066’’’	2,587
Housing age	1159,251’’’	1882,121’’’	1170,281’’’	1162,020’’’	1152,106’’’	1172,236’’’	1161,584’’’	1157,900’’’	1170,883’’’	1141,883’’’
Total room	-10,042’’’	-19,733’’’	-7,251’’’	-7,839’’’	-8,126’’’	-8,067’’’	-8,403’’’	-8,250’’’	-8,642’’’	-7,071’’’
Bed. number	75,731’’’	100,944’’’	94,186’’’	122,837’’’	113,182’’’	117,242’’’	111,883’’’	113,821’’’	110,074’’’	89,731’’’
Population	-31,378’’’	-35,319’’’	-37,755’’’	-37,988’’’	-38,705’’’	-37,416’’’	-38,641’’’	-38,386’’’	-38,840’’’	-38,025’’’
Households	79,585’’’	124,803’’’	63,447’’’	35,087’’’	48,605’’’	40,836’’’	51,321’’’	47,701’’’	55,186’’’	67,840’’’
Median inc	36734,281’’’	47748,381’’’	39869,022’’’	40269,322’’’	40230,435’’’	40327,951’’’	40354,844’’’	40297,522’’’	40465,702’’’	39785,866’’’
WHousing age	1461,697’’’									
WTot. room	-40,181’’’									
WBed. number	211,435’’’									
WPop	-27,032’’’									
WHouseholds	100,585’’’									
WMed.inc	32553,167’’’									
rho	-0,041									
lambda	0,653’’’									
Moran(p-value)	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
dh-test(p-value)	0,038	0,104	0,392	0,391	0,106	0,099	0,105	0,105	0,105	0,093
	GNS	OLS	MA	MB	MC	MD	ME	MF	MG	MH