Spatial Weighting Matrix Estimation through Statistical Learning: Analyzing Argentinean Salary Dynamics under Structural Breaks

Pablo Quintana[*], Marcos Herrera-Gómez[†]

Abstract

The spatial weighting matrix plays a pivotal role in spatial econometrics and remains an active area of research. In this study, we apply recent advancements in machine learning for estimating the spatial weights matrix in econometric models. By employing LASSO strategies and incorporating geographical restrictions, we directly derive the spatial weighting matrix from the available data. This approach removes the necessity for arbitrary criteria set by researchers. As an empirical example, we explore the relationship among the salary of registered salary workers of Argentine provinces. Using monthly information between 2014 and 2022, we identify breakpoints in the wage time series and determine whether the breaks occur due to the movements within each province or due to neighboring provinces.

JEL Classification: C21, C23, C53.

Keywords: Spatial Weighting Matrix, Spatial Econometrics, Statistical Learning, LASSO.

[*] Doctoral Program in Economics, Faculty of Economics, National University of Cuyo, Centro Universitario Parque General San Martin, M5502 JMA, Mendoza, Argentina, email: pabanib@hotmail.com.

[†] CONICET - Department of Economics, National University of Rio Cuarto, Ruta Nac. 36-Km 601 (X5804BYA), Argentina; email: mherreragomez@conicet.gov.ar.

1 Introduction

Spatial econometrics is a specialized field within econometrics that focuses on spatial dependence and spatial heterogeneity in regression models for cross-sectional and panel data. As Griffith et al. (2013) mention, the name of this field was initially introduced by Jean Paelinck in 1974, a Belgian economist considered the founder of this modern discipline with the seminal introductory monograph on this topic, Spatial Econometrics (Paelinck and Klaassen, 1979). Initially, the term was coined to encompass an expanding range of models that specifically integrated the concept of "space" (referring to geography) and aimed to address estimation and testing challenges prevalent in fields like regional science, urban studies, and real estate economics. However, nowadays, spatial econometrics is applied in models that incorporate the "space" in a broad sense, no only geography, defining a general interaction between different units or accounting for non-independent data observations. In economics, the used of spatial econometric models is present in fields such as international economics, labor economics, public finance, education economics, environmental economics, among others.

In spatial econometric models, the spatial interactions are typically considered using what is known as a "spatial weighting matrix", a concept introduced by Moran (1948). This matrix is commonly represented by the letter $\mathbf{W}$, a $(n \times n)$ positive symmetric and non-stochastic matrix with the element at position (i, j), w_{ij}, representing the degree of "closeness" between unit i and unit j based on a distance measure. If w_{ij} is not equal to 0, unit j is considered a neighbor of unit i; conversely, if w_{ij} equals 0, unit j is not considered a neighbor of unit i. Also, by definition, the unit i cannot be a neighbor to itself, i.e. $w_{ii} = 0$. Units that are considered neighbors to a particular unit have a meaningful impact on that unit. This interaction could relate to spillovers, copy-cat strategies, geographic proximity issues, similarity of markets, welfare benefits, tax issues, just to mention a few.

Spatial weighting matrices provide a convenient form to implement Tobler's first law of geography which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970) which applies whether the space is geographic, social or economic. In this form, the models discussed in the literature include cross-unit interactions and correlation in terms of spatial lags, which may involve the dependent variable, the exogenous variables, and the disturbances. A spatial lag of a variable is defined as a weighted average of observations on the variable over neighboring units. To illustrate, the spatial lag of a variable $\mathbf{y}$ is expressed as $\mathbf{Wy}$, representing the values of $\mathbf{y}$ in the neighbourhood. Therefore, the determination of this matrix is of utmost importance for the study of the data.

Although in most cases the $\mathbf{W}$ is unknown, traditionally it is presumed known from different criteria specifying exogenously. This approach is by far the most popular in the empirical literature and includes, for example, use of a contiguity criteria (queen, rook, block contiguity) and distance criteria (e.g., distance band, k-nearest neighbors, kernel functions) or a combination of these.

In recent times, the literature has made progress in detecting the weight matrix using endogenous approaches, thus avoiding the influence of the arbitrary criteria of the researcher. This new literature uses the topology and the nature of the data such as the local statistical model of Getis and Aldstadt (2004) or the algorithm under partial knowledge about the form of the weights of Benjanuvatra and Burridge (2015). Under panel data information, there are approaches more flexible such as in Beenstock and Felsenstein (2012); Bhattacharjee and Jensen-Butler (2013) or more recent advances like the proposed by Angulo et al. (2017, 2018).

Our contribution pertains to the endogenous approach but under recent developments based on the revolution of statistical learning. Particularly, the present work focuses on finding a matrix that reflects the true relationship in a spatial variable. The primary objective is to detect breakpoints within the variable and ascertain whether these discontinuities result from internal shifts or are influenced by neighboring entities. This process facilitates the creation of a spatial weights matrix derived directly from the data.

The paper is organized as follows: Section 2 gives a brief description of the literature on this subject. Section 3 presents the model followed by the method of estimation that has been chosen presented by Section 4. Section 5 presents the main results obtained from simulations. Section 6 presents an empirical example using information of Argentina. Finally, Section 7 includes the conclusions and discussions on the subject.

2 Literature Review

Spatial econometrics is a discipline that analyzes the interactions between entities or units in a given space. How these interactions manifest can vary, and it is necessary to be able to detect and represent them through a mathematical object, such as a matrix of spatial weights $\mathbf{W}$. This matrix is a key element in the econometric model, and its specification probably reflects the great influence of time series analysis, where the common temporal lags are needed to specify the dynamic models. The natural ordering of temporal precedence allows the researcher to specify parsimonious time series models. However, the situation is more complex in space, where the relations are multidirectional.

The necessity of the spatial weighting matrix, $\mathbf{W}$, is an emerging solution for a spatial cross-sectional data. As is mentioned by Paelinck and Klaassen (1979), there is an identification problem which is evident in the following simple interdependent specifications with three different spatial units, 1, 2 and 3:

$$\begin{gathered}
y_1 = \alpha_{12} y_2 + \alpha_{13} y_3 + \iota + \varepsilon_1, \\
y_2 = \alpha_{21} y_1 + \alpha_{23} y_3 + \iota + \varepsilon_2, \\
y_3 = \alpha_{31} y_1 + \alpha_{32} y_2 + \iota + \varepsilon_3, \\
\varepsilon_j ; \varepsilon_k ; \varepsilon_l \sim i.i.d.\left(0, \sigma^2\right)
\end{gathered} \tag{1}$$

where y_i is the response variable in region i $(i = 1, 2, 3)$, ι is the constant term and α_{ij} $(i, j = 1, 2, 3)$ are unknowns parameters that capture the essential interaction. In terms of LeSage and Pace (2009), this unrestricted spatial autoregressive process: *would be of little practical usefulness since it would result in a system with many more parameters than observations. The solution to the over-parametrization problem that arises when we allow each dependence relation to have relation-specific parameters is to impose structure on the spatial dependence parameters.* The parametrization procedure is, in fact, the way preferred in the empirical applications.

Formally, rewritten the main part of the system of (1), we introduce a new term ρ that captures the spatial dependence of the system and $\mathbf{W}$, a square $(n \times n)$ matrix (in this case, 3×3), whose diagonal elements are all zero and the off-diagonal elements are, usually, nonnegative:

$$\mathbf{\Gamma} = \begin{bmatrix} 0 & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & 0 & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & 0 \end{bmatrix} \Longrightarrow \rho \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} = \rho\mathbf{W}. \tag{2}$$

In general terms, for a sample of n observations:

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \cdots & w_{1j} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2j} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \cdots & \cdots & \cdots \\ w_{i1} & w_{i2} & \vdots & 0 & \cdots & w_{in} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ w_{n1} & w_{n2} & \vdots & w_{nj} & \vdots & 0 \end{bmatrix}. \tag{3}$$

The terms w_{ij} are called the spatial weights. According to Kelejian and Prucha (1998), the spatial matrix should be uniformly bounded in absolute value (that is, $max_{1\leq j\leq N}\sum_{i=1}^{n}|w_{ij}| < c < \infty$ and $max_{1\leq i\leq N}\sum_{j=1}^{n}|w_{ij}| < c < \infty$, where c is a constant). In order to avoid cases of isolation, an additional restriction is add: the row sums are uniformly bounded away from zero $min_{1\leq j\leq n}\sum_{i=1}^{n}|w_{ij}| > 0$. It is typical to row-standardize the matrix before being used in the model.

As said previously, the weight matrix is associated to the interaction among the spatial units in a spatial model. If the agents interact between themselves, if they are dispersed over the space and if the space is divided in regions, then is reasonable to find spatial interaction between the regions capture for a ρ different to zero. $\mathbf{W}$ is a simple and intuitive way to create spatial lags of the variables: spatial dependence results from lags of the endogenous variable.

Based on the preceding discussion, the importance of matrix $\mathbf{W}$ within a spatial model becomes evident. The subsequent task of quantifying spatial dependence within a dataset involves the precise specification of this matrix. Broadly speaking, existing literature identifies three distinct approaches (one exogenous and two endogenous) for constructing the spatial weighting matrix:

1. Specifying the matrix exogenously;

2. Specifying the matrix from the data;

3. Estimating the matrix from the data.

The exogenous approach is widely favored by researchers. It relies on making an initial judgment about how the spatial units are arranged geographically or on other important aspects of the data. Typical examples of this approach include the binary contiguity criterion, the k-nearest neighbors, the great circle distance criterion, as well as kernel functions that depend on the distance between centroids, among others. There are important advances to make this approach more flexible, for example, Lee and Yu (2012) developed a quasi-maximum likelihood estimator (QML) for dynamic panel data models with spatial weights that vary over time and are assumed to be known and exogenous. They found that if the time-invariance of the spatial weights matrix ($\mathbf{W}$) is assumed incorrectly, significant biases can occur, particularly for the estimates of the marginal direct/indirect effects. All these methods assume prior knowledge about the real process that generates the data. However, in practice, we often lack such understanding, as noted by Gibbons and Overman (2012).

The second approach considers both the structure of the space and the characteristics of the data being modeled. Many examples of this combined approach have emerged following pioneering work by Bodson and Peeters (1975) and Kooijman (1976). More recent contributions include the works of Getis and Aldstadt (2004), the AMOEBA procedure by Aldstadt and Getis (2006), and the complete correlation coefficient (CCC) criteria by Mur and Paelinck (2011). In all these cases, there is a clear connection between existing knowledge and the insights derived from analyzing the data.

The third approach involves directly estimating the weight matrix from both the data and the model being constructed. This task is complex due to the considerable number of parameters that require estimation (potentially around n^2, depending on the assumed constraints). Beenstock and Felsenstein (2012) and Bhattacharjee and Jensen-Butler (2013) address this challenge within a panel data framework, while Benjanuvatra and Burridge (2015) focuses on a pure cross-section, adopting a slightly more parameterized strategy. Additional proposals for constructing weight matrices based on non-geographical criteria are available in works by Maggioni et al. (2007) and Autant-Bernard and LeSage (2011).

In recent times, the third approach has been more relevant, it has been tried to find the spatial matrix that arises exclusively from the data, in this respect Benjanuvatra and Burridge (2015) seek to determine the gamma parameter on which depends the chosen distance function to determine the matrix. Usually two ways are used to determine this distance one is with the exponential formula and the other with inverse distance. In such studies the authors use maximum likelihood and try to find this parameter along with the rest of the parameters of the model like in Angulo et al. (2017, 2018).

Snijders et al. (2007) introduced the concept of co-evolution, where $\mathbf{W}$ and the explained variable evolve jointly over time in network dynamics. Other researchers, such as Qu and Lee (2015), have also explored models where $\mathbf{W}$ may depend on observable or unobservable covariates. Hays et al. (2010) proposed a SpatioTemporal AutoRegressive (STAR) model, which estimates $\mathbf{W}$ using conditional maximum likelihood, including both exogenous and lagged endogenous components.

Given the new tools coming from statistical learning, mainly the LASSO penalty method, originally proposed by Tibshirani (1996), which is simply a regression with penalty of parameters to avoid over-adjustments, some authors have made progress in obtaining the matrix of spatial weights from the data. One of these examples is applied by Ahrens and Bhattacharjee (2015), which based on an SLX panel model, estimate the $\mathbf{W}$ using a two-stage LASSO post estimator. On the other hand you can see the work of Lam and Souza (2014) that estimate a spatial model auto-regressive and lagging in the X also, using adaptive LASSO techniques.

In another more recent work Lam and Souza (2019) incorporate into their model a linear combination of the spatial matrices in which a given expert matrix intervenes with usual techniques of spatial econometrics and on the other hand a matrix that arises from the application of adaptive LASSO methods to determine it directly from the data. Merk and Otto (2022) propose a way to estimate it only with cross-sectional data increasing the statistical power with re-sampling techniques and thus be able to obtain the matrix of spatial weights sought.

It is evident that current practices related to building $\mathbf{W}$ are heterogeneous. These practices span a wide range, our proposal involves directly estimating the weight matrix from both the data and the model under construction, similar to the third approach. However, we impose constraints on the estimation based on data specifications. In this context, our strategy integrates aspects from both the second and third methodologies.

3 The model

We assume that the data is generated by a spatio-temporal process $\{\mathbf{Y}_t(\mathbf{s}) : t \in \mathcal{D}_t, \mathbf{s} \in \mathcal{D}_\mathbf{s}\}$, where $\mathcal{D}_t$ is the temporal domain and $\mathcal{D}_\mathbf{s}$ is the spatial domain of the process. Also, the process is observed for $t = 1, \ldots, T$ at a constant set of spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n\}$. The temporal domain is a discrete-time of points, $\mathcal{D}_t \in \mathbb{Z}$ and the spatial domain is a discrete-irregular space.

The vector of observations of the spatio-temporal process $\{\mathbf{Y}_t(\mathbf{s})\}$ is represented in two alternative forms: $\mathbf{Y}_{\bullet t} = (Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_i), \ldots, Y_t(\mathbf{s}_n)' : t = 1, \ldots, T)$ where the observations are stacked as successive cross-sections for $t = 1, ..., T$. This is the usual notation in spatial panel data (see, for example, Elhorst, 2014, chapter 3). Alternatively, we use $\mathbf{Y}_{\bullet i} = (Y_1(\mathbf{s}_i), \ldots, Y_t(\mathbf{s}_i), \ldots, Y_T(\mathbf{s}_i)' : i = 1, \ldots, n)$ representing the usual structure in panel data. Then, in compact form, we introduce the $(T \times n)$ matrix $\mathcal{Y} = (\mathbf{Y}_{\bullet 1}, \ldots, \mathbf{Y}_{\bullet i}, \ldots, \mathbf{Y}_{\bullet n})$.

We consider that the process has an autoregressive dependence structure across space and that may have unknown variations in different periods of time. Temporal breaks can occur in different time periods for each i location and can be of different magnitudes. The model can be written as follows:

$$\mathbf{Y}_{\bullet t} = \mathbf{W}\mathbf{Y}_{\bullet t} + \boldsymbol{\delta}_t + \boldsymbol{\varepsilon}_{\bullet t}, \tag{4}$$

where $\mathbf{Y}_{\bullet t}$ is a $(n \times 1)$ vector of observations of the spatio-temporal random process in the period t, the $(n \times n)$ matrix $\mathbf{W}$ represents the unknown spatial interaction (the spatial weighting matrix) between the locations, the $(n \times 1)$ vector $\boldsymbol{\varepsilon}_t$ is an independent and identically distributed error term with zero mean and constant variance, $i.i.d.(0, \sigma^2)$, and $\boldsymbol{\delta}_t = (\delta_{1t}, \ldots, \delta_{nt})'$ is a $(n \times 1)$ vector of scale factors that represents the local mean level for location in the period t.

Since it is a spatial autoregressive model, any change in $\boldsymbol{\delta}_t$ will generate movements in the other locations through the spillover effects. At this point, we can distinguish the true process mean given by: $\boldsymbol{\mu}_t = E(\mathbf{Y}_{\bullet t}) = (\mathbf{I} - \mathbf{W})^{-1}\boldsymbol{\delta}_t$, determining that the expected value of the model in (4) only can be changes by the change of the $\boldsymbol{\delta}_t$, a vector of parameters to be estimated. Each parameter of $\boldsymbol{\delta}_t$ represents a local mean that may be determined by different unknown factors to the model or it can be represented common shocks (Pesaran, 2006).

Also, the structure of (4) can be extended with classic exogenous variables, $\mathbf{X}_t$, and the respective parameters, $\boldsymbol{\beta}$:

$$\mathbf{Y}_{\bullet t} = \mathbf{W}\mathbf{Y}_{\bullet t} + \boldsymbol{\delta}_t + \mathbf{X}_{\bullet t}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\bullet t}. \tag{5}$$

The main problem with these models is that all the spatial structure must be estimated simultaneously with local means and beta coefficients. Consequently, the parameter count substantially exceeds the number of observations, presenting a notable issue. Therefore, as we review in the previous section, the main strategy followed in the literature was to introduce a matrix constructed exogenously, using external information like geographical factors. However, the creation of this matrix is defined by the arbitrariness of the researcher and, due to the large number of possible candidates, a search for the matrix that best suits would be costly. That is why this work follows a methodology proposed by Otto and Steinert (2022) that allows detecting a matrix of connections between the spatial units that arises purely from the data. The following section gives an explanation of how the estimation method works.

4 Estimation

The causes for which the dependent variable may change are many, it may be due to changes in the real form or unknown confounding factors. These causes can be identified according to the period in which they occurred without needing to know what caused it. Therefore the strategy would be to try to detect those periods in which one of these breaks occurs. After identifying the breakpoints we can say that we have been able to identify the average level $\boldsymbol{\delta}_t$ referred to in equation (4). With these determined points, we proceed to detect the matrix of spatial weights that allows us to conclude the model. The estimate is made through a two-stage LASSO estimator. The idea of doing it in two stages is to separate in the estimation what is the detection of temporary change points with the spatial dependence. Therefore, in a first stage we will try to detect temporary change points in the data that will allow us to reduce the temporal parameters only in relation to those that have actually caused a change shock in the variable. The second stage consists in estimating the $\mathbf{W}$ that allows relating the different spatial units among themselves.

4.1 First stage: the search of break points

Under a model like the expressed in equation (4), we can define the level in all time periods of each spatial unit $(i = 1, ..., n)$ as follows:

$$\mathbf{Y}_{\bullet i} = \mathbf{K}\tilde{\boldsymbol{\mu}}_i + \boldsymbol{\varepsilon}_i, \tag{6}$$

where $\mathbf{Y}_{\bullet i}$ is a $(T \times 1)$ vector, with $\mathbf{K}$ as a $(T \times T)$ lower triangular matrix, and $\boldsymbol{\varepsilon}_i$ denotes an independent and identically random error term, with mean 0 and constant variance. The vector of coefficients $\tilde{\boldsymbol{\mu}}_i = (\tilde{\mu}_{t,i})_{t=1,...,T}$ represents the changes of the overall mean levels. The vector $\mathbf{K}\tilde{\boldsymbol{\mu}}_i$ coincides with the overall mean levels $\boldsymbol{\mu}_i$.

The equation (6) is a fully identified model where the total of parameters coincides with the total of equations, therefore, solving it will not provide anything relevant to interpret. The estimation strategy proposed by Otto and Steinert (2022) is to eliminate all those changes in the general average that are not significant and only keep the strong structural breaks. To find these change points, the adaptive LASSO method proposed by Zou (2006) is applied. This method allows that the estimated coefficients $\hat{\tilde{\mu}}_{t,i}$ to consistently become zero when T tends to infinity, a property that is not obtained with the original LASSO method. The estimated coefficients, with the appropriate level of penalty, consistently select those periods that represent structural breaks in the variable for each spatial unit.

In all those periods where there is no break point, we will have that the general average of the period will be equal to that of the previous period (this is, $\hat{\tilde{\mu}}_{t,i} = \hat{\tilde{\mu}}_{t-1,i}$). But the points that really interest us are those that represent a significant change. If the spatial units are really connected to each other, a breaking point in one unit will cause the change in the general average of all its neighbors. Therefore, we can determine the set of candidate change points for $\boldsymbol{\delta}_t$ in the next stage. This selection is determine in the following way: $\mathcal{T}_i = \{\tau : \hat{\tilde{\mu}}_{t,i} \neq \hat{\tilde{\mu}}_{t-1,i}\}$.

4.2 Second stage: estimation of the full model

At this stage, we estimate the full model that includes the changes in the mean, $\boldsymbol{\delta}_t$, and the spatial dependence, $\mathbf{W}$, that captures the neighborhood influence. The model (4) can be rewritten in the following way:

$$\mathbf{Y} = \boldsymbol{\Psi}\tilde{\boldsymbol{\delta}} + \mathbf{Z}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{7}$$

where $\mathbf{Y} = vec(\mathcal{Y})$ is a $(nT \times 1)$ vector, where $vec(\cdot)$ is the vectorization operator such as $vec(\mathcal{Y}) = (y_{11}, \ldots, y_{1T}, y_{21}, \ldots, y_{2T}, \ldots, y_{n1}, \ldots, y_{nT})'$; $\boldsymbol{\Psi}$ is a lower triangular, block diagonal matrix, in fact it is the direct sum of n triangular matrices $\mathbf{K}$ of dimension $(T \times T)$. The vector of coefficients $\tilde{\boldsymbol{\delta}}$ is defined as the changes in $\boldsymbol{\delta}$, that is, $\boldsymbol{\delta} = \boldsymbol{\Psi}\tilde{\boldsymbol{\delta}}$ assuming $\tilde{\delta}_{t,i} = 0$ for $t \notin \mathcal{T}_i$ for all $i = 1, \ldots, n$, according the results in the first stage. Also, $\mathbf{Z} = \mathbf{I}_n \otimes \mathcal{Y}$, where $\otimes$ is the Kronecker product, then the dimension of matrix $\mathbf{Z}$ is $nT \times n^2$ $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T)'$, with the coefficients $\boldsymbol{\xi}$ as the vectorized spatial weights ($\boldsymbol{\xi} = vec(\mathbf{W})$). Finally, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T)'$ is a $(nT \times 1)$ vector of stacked errors.

Given the previous conditions, the estimated coefficients are given by the minimization of the next problem:

$$\begin{aligned}
\left(\hat{\tilde{\boldsymbol{\delta}}}, \hat{\boldsymbol{\xi}}\right)' \;&\in\; \underset{(\tilde{\boldsymbol{\delta}},\boldsymbol{\xi})'}{\arg\min} \left\|\boldsymbol{Z}\boldsymbol{\xi} + \boldsymbol{\Psi}\tilde{\boldsymbol{\delta}} - \boldsymbol{Y}\right\|_2^2 + \lambda_b \left\|\boldsymbol{\xi}\right\|_1, \\
\text{subject to:} \qquad & \tilde{\delta}_{t,i} = 0 \text{ for all}, t \notin \mathcal{T}_i,\ 1 \leq i \leq n; \\
& w_{ij} \geq 0 \text{ and } w_{ii} = 0 \text{ for all}\, 1 \leq i, j \leq n; \\
& \sum_{j=1}^{n} w_{ij} \leq 1 \text{ for all}\, i = 1, ..., n.
\end{aligned}$$

For this second stage, the estimation was made with the LASSO restriction. The restriction $\sum_{j=1}^{n} w_{ij} \leq 1$, for all i, ensures the non-singularity of the matrix $(\mathbf{I} - \mathbf{W})$, an important condition to be able to determine the model.

In summary, in this second stage all the change points that were obtained in the first stage, which represent the general average level, are consider whether these are due to changes in the structures of their neighbors or their own changes. These results that will be in charge of calculating this second stage, taking to zero those neighbors that are not those that are included in the general level and seeing if the changes produced are only made by internal changes to the unit.

4.3 Spatial Adjustments

The estimation presented in the previous sections does not take spatial relationships into account. The method only establishes neighbor connections considering their correlations arising from the data but without any theoretical foundation. However, in contrast to what Otto and Steinert (2022) suggest, we modify this by adding a spatial information to the model, considering some kind of relationship to guide the estimation from the theory.

If we consider all units and write it in matrix form, we have a problem very similar to the one previously stated:

$$\left(\hat{\tilde{\boldsymbol{\delta}}}, \hat{\boldsymbol{\xi}}\right) \in \arg \min_{\left(\tilde{\boldsymbol{\delta}}, \boldsymbol{\xi}\right)'} \left\|\boldsymbol{Z}\boldsymbol{\xi} + \boldsymbol{\Psi}\tilde{\boldsymbol{\delta}} - \boldsymbol{Y}\right\|_2^2 + \lambda_b \left\|\tilde{\boldsymbol{\xi}}\right\|_1, \text{ subject to:}$$

$$\tilde{\delta}_{t,i} = 0 \quad \text{for all } t \notin \mathcal{T}_i,\ 1 \leq i \leq n;$$

$$w_{ij} \geq 0 \text{ and } w_{ii} = 0 \quad \text{for all } 2 \leq i, j \leq n;$$

$$\sum_{j=1}^{n} w_{ij} \leq 1 \quad \text{for all } i = 1, \ldots, n,$$

where $\tilde{\boldsymbol{\xi}} = \text{vec}\left(\boldsymbol{W} \odot \boldsymbol{D}\right)$ with $\odot$ being the Hadamard product (also known as the element-wise product), and $\boldsymbol{D}$ being an $n \times n$ matrix representing the matrix of geographical distances according to a chosen function.

5 Monte Carlo simulation

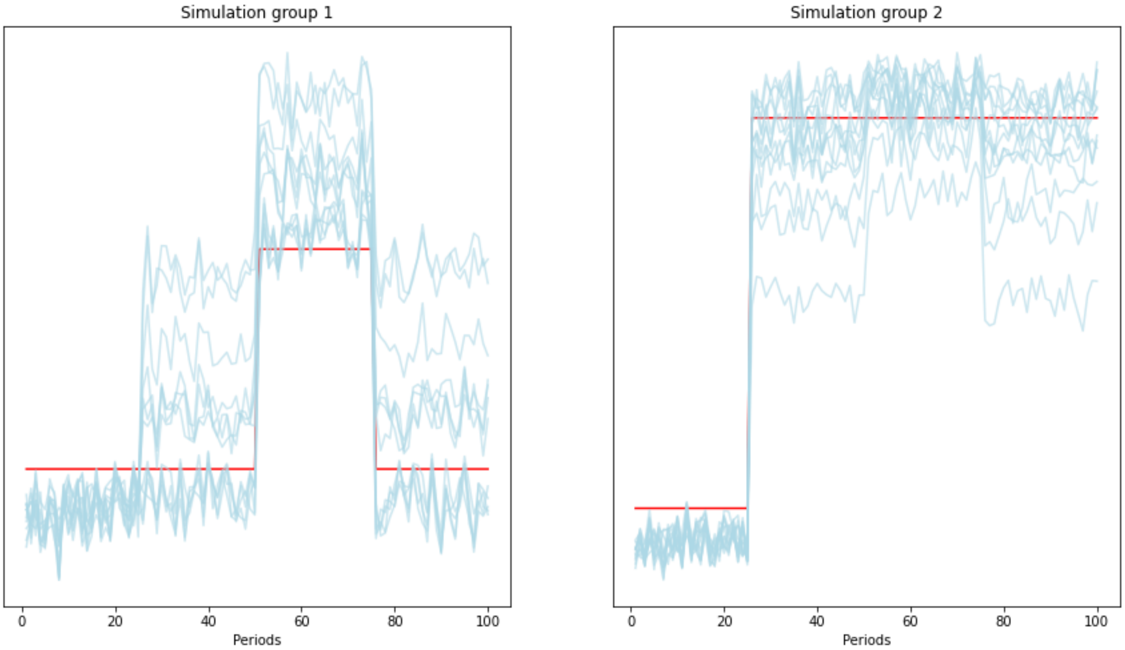
5.1 Setup

In the simulation process, a strategy similar to the one used by Otto and Steinert (2022) is employed. Twenty-five areas are taken, located within a five-by-five square, with a randomness in the distances between each area. These areas are divided into two groups, each having different periods in which the shock occurs and with different impacts. In the first group, breakpoint moments occur in periods 50 and 75, where the first breakpoint is of magnitude 3 and the second is -3, returning the level to its initial state. In the second group, there is a single breakpoint of magnitude 7 at period 25. It should be noted that determining the breakpoints is not innocuous. If their intensity decreases or becomes more random, the algorithm struggles to identify the true connections. Moreover, increasing the number of breakpoint periods negatively affects performance. However, the breakpoints being investigated represent structural changes in the variable, which is why it is more realistic to work with few significant changes.

The criteria for selecting the spatial weights matrix was that of the four nearest neighbors. This matrix is used to determine the relationships between areas, but the intensity of these relationships is established randomly. Therefore, the impact of each neighbor will not be the same for each unit. Subsequently, the matrix is row-normalized to ensure non-singularity. After constructing this initial matrix, it is multiplied by a coefficient ρ, determining the overall intensity of the matrix. This also influences the algorithm's performance, as a low-intensity spatial weights matrix tends to result in fewer identified true relationships.

To the above-described process, an independent random component is added for each area, centered around 0 and taking values between -0.5 and 0.5. This data generation process is repeated 100 times to estimate the matrix for each instance. For selecting the distance weights proposed in equation (8), two criteria are considered: the first involves considering the distances directly, while the other considers an exponential distance $\exp(d)$ to penalize more distant neighbors more strongly. The comparison is also made without considering any weight. The choice of an additional weight for each neighbor, as proposed in this study, is related to the process of choosing the parameter λ, commonly used in Lasso regularization. If the chosen parameter is larger, penalization is increased when adding distances, leading to excessive regularization. In this case, favoring not considering an extra weight for neighbors is more appropriate.

Figure 1: Spatial Simulations



Note: Left: simulations of group 1. Right: simulation of group 2. The red line represents the local mean level δ. The light blue lines represents the individual realizations of each units.

5.2 Results

Various metrics are utilized to evaluate the results, and respective comparisons are conducted. Using the Euclidean distance between the true and estimated matrices is the initial measure, which concentrates on determining the intensity of the relationships

Another group of metrics focuses on determining whether the algorithm correctly identifies the relationships or not. Following the spatial econometrics literature, we know that sparse matrices are more suitable for an autoregressive process. In this study, the focus is on measuring how well the model detects the zeros in the matrix. The first performance metric used is the proportion of true positives.

$$\text{pvp} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(\hat{w}_{ij} > 0, \text{when } \ w^*_{ij} > 0)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(w^*_{ij} > 0)}, \tag{8}$$

where I_{ij} is an indicator function that takes 1 if the condition between parenthesis is satisfied and zero otherwise; $\hat{w}_{ij}$ represents the predicted weight from the algorithm and w^*_{ij} is the real positive weight in the simulation. This measure allows us to determine what proportion of true relationships the algorithm was able to capture. The second metric used is the proportion of true no-relationships:

$$\text{pvn} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(\hat{w}_{ij} = 0, \text{when } \ w^*_{ij} = 0)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(w^*_{ij} = 0) - n}. \tag{9}$$

This measure helps us identify the algorithm's tendency to detect relationships that do not exist.

The other metrics tested are commonly used for classification models and include balanced accuracy (see Guyon et al., 2015).

$$\text{b.accuracy} = \frac{\text{pvp} + \text{pvn}}{2} \tag{10}$$

Goutte and Gaussier (2005) suggest to consider the precision. This metric allows us to evaluate the proportion of true relationships on all estimated positive relationships:

$$\text{precision} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(\hat{w}_{ij} > 0, \text{when } \ w^*_{ij} > 0)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}(\hat{w}_{ij} > 0)} \tag{11}$$

Additionally, the f_1-score is considered, which combines precision and recall or pvp into a single index (see Guyon et al. (2015); Goutte and Gaussier (2005):

$$\text{f}_1 = 2 \times \frac{\text{precision} \times \text{pvp}}{\text{precision} + \text{pvp}}. \tag{12}$$

The measurement of metrics is based on a scale from 0 to 1 and implies that higher values indicate better model performance.

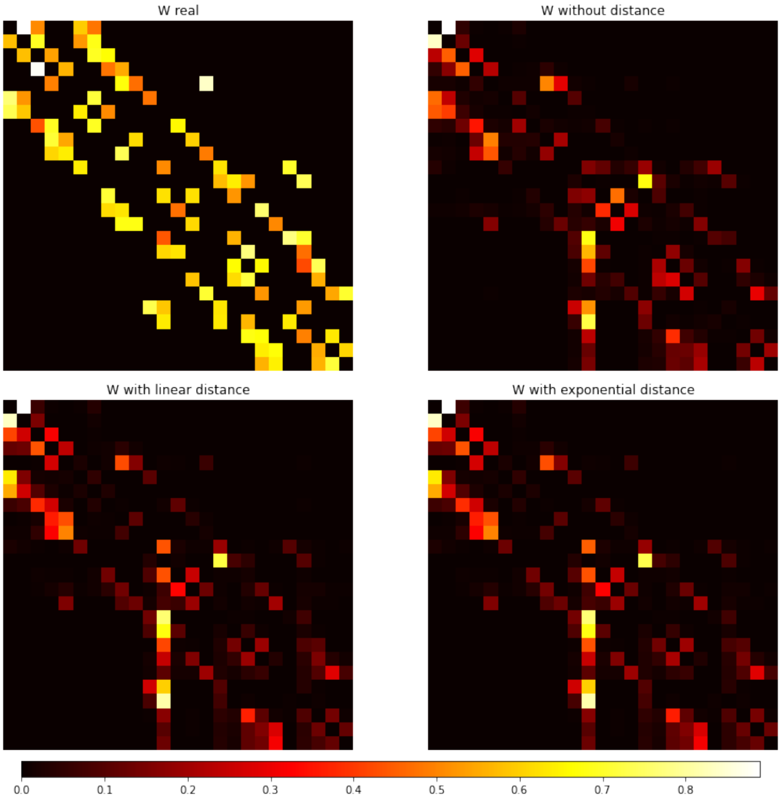
In Table 1, the results obtained by the different proposed methodologies are shown. The model without distance matrix achieves better results in the Euclidean distance measure, although in all cases, the distance is not very high. However, in all other metrics, the proposed models seem to obtain better results, indicating that adding distance weights as an additional regularization parameter is beneficial. Metrics such as pvn and precision yield improved outcomes when a distance matrix is integrated into the penalty framework. Generally, these metrics highlight the instances where our proposed approach demonstrates superior performance. This enhancement is indicative of a reduction in matrix sparsity and the prevention of false spatial relationships being incorporated into the model.

Table 1: Results of the metrics of simulations for different distance penalty matrices

Method	Distance	pvp	pvn	b.accuracy	precision	f1
Without dist	1.2349	0.87	0.82	0.83	0.49	0.63
Linear dist.	1.4358	0.90	0.85	0.85	0.54	0.67
Exponential dist.	1.4438	0.90	0.85	0.85	0.54	0.67

Figure 2 displays a heatmap illustrating the original matrix and each of the simulated matrices. At first glance, the three simulations do not exhibit significant differences. However, it can be observed that the relationships in these matrices appear darker, suggesting that the relationships detected by the simulations are generally weaker.

Figure 2: Matrices of simulation



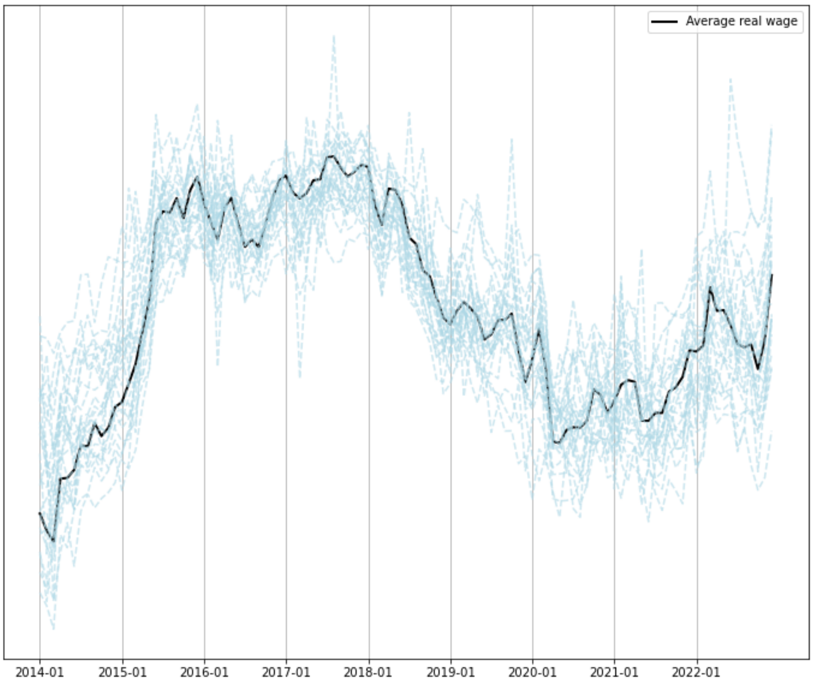
6 Empirical Application

In the labour market, the wages of workers in each region depend to a large extent on the economic activities carried out there. However, it should also be borne in mind that wages in an area may be affected by wages in neighbouring regions. This reflects the fact that shocks that cause real wage increases and decreases in a region do not always come from the same region but may have occurred in related provinces. The reasons why two regions are related can be diverse and exceed the objectives of this work, which mainly seeks to find such relationships using new techniques established in spatial econometrics.

The data is derived from public databases in Argentina and corresponds to a monthly time series from January 2014 to November 2022, encompassing the 23 provinces and the Autonomous City of Buenos Aires (CABA). The main objective is to estimate the real salary by province and identify any spatial or proximity-related relationships among them. It's important to note that salary movements often exhibit seasonal patterns, such as in agricultural work, and another recurring cycle affecting all sectors of workers, which is the annual supplementary wage observed in the months of June and December each year. To mitigate potential distortions caused by these patterns, both cycles were removed, and a deseasonalized variable was employed. Additionally, to enhance comparability and refine the method detailed in section 4, the variable was standardized for each province. Standardization allows us to focus on the patterns exhibited by the variables, disregarding the magnitude of average salaries in each jurisdiction. Thus, it is normal to find relationships between provinces with significant differences in average salaries that have displayed similar trends throughout the entire period.

Argentina undergoes various periods of fluctuations, where real salary not only depends on the labor economic situation but is also impacted by the recurrent issue of inflation in the country. In Figure 3 , one can observe that, in general, the provincial movements have followed the trajectory of the nation as a whole, with some peaks experienced by specific provinces during certain periods. However, overall, all provinces have undergone a growth period in the first two years, followed by a decline phase starting from early 2018, and subsequently, a slight recovery beginning in 2021.

Figure 3: Evolution of real wage



Note: Real wage after normal transformation for each province shows in the above graphic with light blue color. Black line show the all provinces real wage average

6.1 Matrix construction

In the section 4, the addition of a penalty matrix **D** was proposed, which serves to constrain the results derived from the data using a theoretical matrix that may underlie the process. Proximity among provinces influences salary for various reasons, ranging from sharing the same labor sectors to other factors beyond the scope of this study. To account for this, the distance matrix was considered, both directly and exponentially, between the centroids of the most populous departments within each province. Furthermore, a distance matrix proportional to the midpoint between Buenos Aires and the Autonomous City was considered, under the assumption that the influence might primarily emanate from the country's capital. This implies that if a province is close to the capital, nearby provinces lose weight, as the influence would be directly absorbed from the capital.

In the Table 2 you can see the results of the different methods.Evaluating the optimal methodology in a real-world scenario becomes intricate due to the unavailability of the true weight matrix. The mean squared error (MSE) serves as a means to assess the method's performance in prediction and provides an indication of the matrix's efficacy. The MSE reveals that the best-performing method in this metric is the method without distance matrix, achieving a value of 0.1157. However, measuring the MSE using the same dataset on which the model was fitted tends to favor models with less penalization. To address this, the MSE was computed out-of-sample with data from the subsequent four months. In this context, the optimal outcome was 0.479, achieved by the methodology employing exponential distance matrix **D**.

Table 2: Metrics comparison for different distance matrices applied to the method.

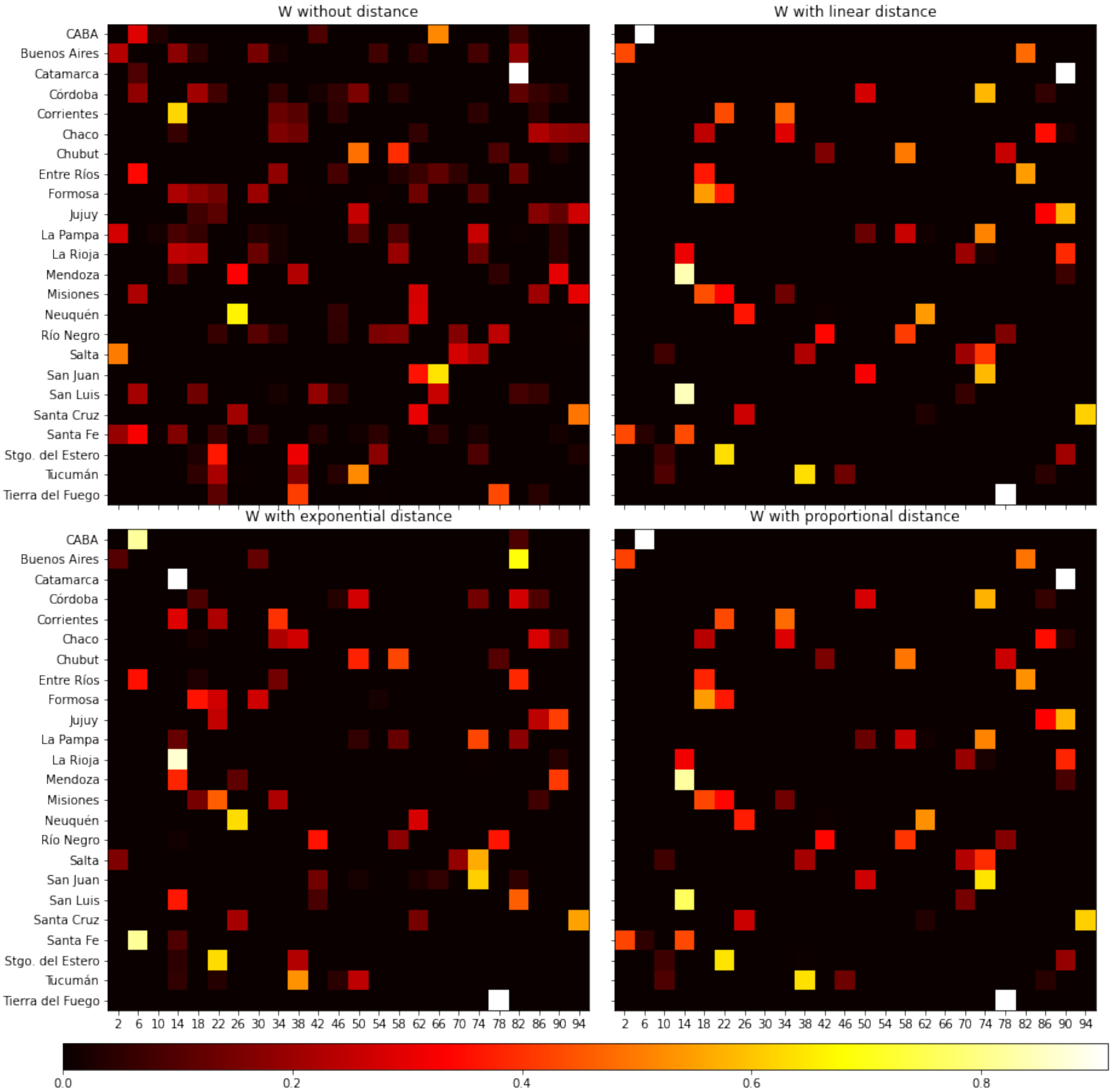
	mse	mse test	no zero prop.	norm
Without dist.	0.1157	0.5241	0.2622	2.6928
Linear dist.	0.1721	0.5347	0.1094	3.6751
Exponential dist.	0.1808	0.4790	0.1406	3.4913
Proportional dist.	0.1671	0.5235	0.1128	3.6504

One proposed objective is to attain a sparser **W** matrix. This aim is grounded in the spatial econometrics theory, which asserts that sparse matrices yield enhanced performance in autoregressive models such as the one presented. The proportion of non-zeros metric effectively elucidates this phenomenon. A noticeable discrepancy is apparent between methods employing penalty matrix **D** and those that do not, particularly in this metric. This disparity underscores that the inclusion of an additional penalty is conducive to the methodology in this regard.

Moreover, to assess the magnitude of relationships, the Euclidean norm is employed. In this regard, it becomes evident that the proposed methods establish more restrained yet potent relationships, as they exhibit higher norms and, consequently, greater intensity.

Figure 4 presents a heatmap of the matrices **W** constructed. On the y-axis, the provinces are listed alongside their respective names, while on the x-axis, in the same order, they are identified by their INDEC province codes. The rows signify the influence that a province receives from all others. Darker hues indicate proximity to zero in terms of relationships, while progressively lighter shades, even approaching white, signify maximized relationships in this case, denoting 1 due to matrix normalization. Examining this depiction, it becomes evident that the darkest points, the predominant blacks, are primarily present within the matrices constructed using our proposed methodology and the matrix derived from the methodology that disregards distance considerations.

Figure 4: Estimated Weighting Matrices

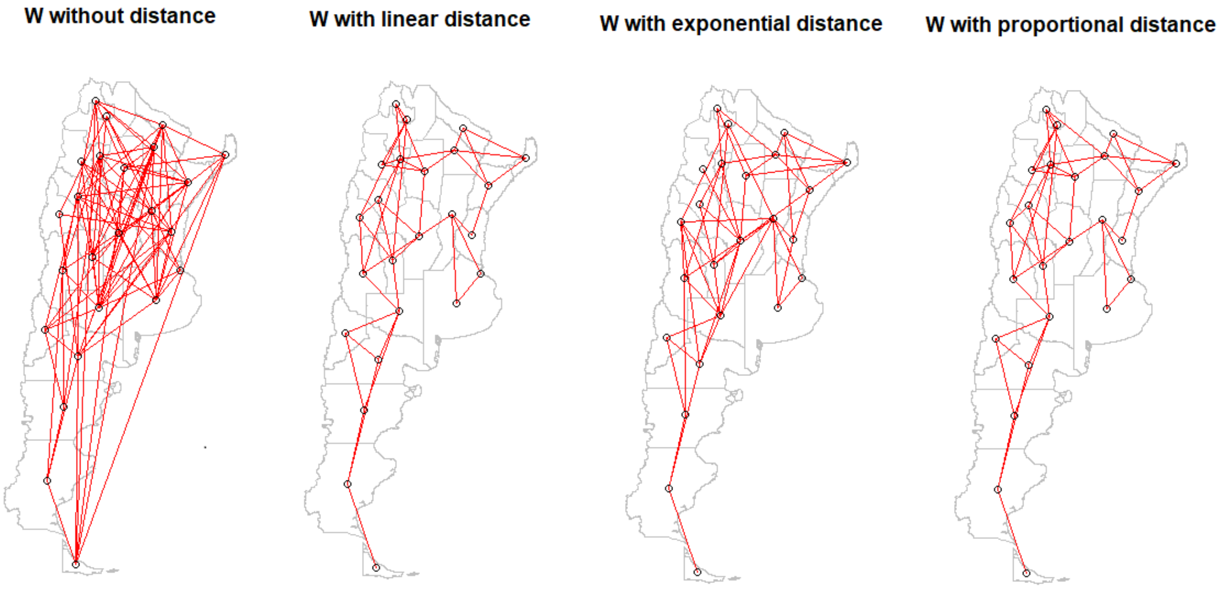


Note: Heatmap reflecting the intensity of relationships among provinces. Darker values represent weaker dependencies.

Conversely, this distinction becomes significantly more apparent perhaps when we binarize matrix **W** and transform it in the connection matrix **C**, where $c_{ij} = 1$ if $(w_{ij} > 0)$ and $c_{ij} = 0$ if $(w_{ij} = 0)$. The matrices and proceed to visualize connection quantities in the subsequent Figure 5 , where the connectivity map is displayed. The connections generated by the method while applying different matrices D are depicted individually in corresponding maps. The left map represents the method without distance penalty matrix **D** show a overabundance of connections. Several of these relationships prove challenging to justify from an economic theory perspective. For instance, this is evident in the southernmost province, Tierra del Fuego.

Despite the significant geographical distance that separates it from the northern provinces, the methodology establishes connections between them. However, such a phenomenon does not occur in the remaining three cases, where a markedly reduced number of relationships per province is observed. Typically, these relationships form among geographically proximate provinces.These instances can be better expounded upon through theoretical lenses, although the outcomes have arisen from data analysis.

Figure 5: Maps of connections



Note: red line represents the connection between province i and province j determined for c_{ij}.

Through methodologies of statistical learning, it becomes feasible to derive a correlation matrix among provinces. However, assessing whether this matrix accurately reflects the true connections between them proves to be a more challenging endeavor. The methodology proposed within this study endeavors to guide relationships within a more economically reasoned framework. Primarily, this is achieved by leveraging temporal breakpoints to identify relationships and subsequently applying a penalty informed by supplementary data pertinent to the specific problem. In this case, geographic information was harnessed for this purpose. Consequently, this approach leads to a reduction in connections, enabling the derivation of a correlation matrix more aligned with the existing literature of spatial econometrics.

7 Conclusions

The identification of interaction effects is crucial for the understanding of how individuals, firms and regions interact. The full estimation of these spatial effects is not possible due to the dimensionality problem, more when the spatial dataset is large.

In spatial econometric models, the spatial interactions are typically considered using what is known as a "spatial weighting matrix", to incorporate these interaction effects. Then, the selection of the spatial weighting matrix is a hot topic in Spatial Econometrics. This key element is usually choice by the researcher and the a priori selection put in doubt the inferential results because it conditioned about that choice.

In this paper, we propose a technique with good performance to detect spatial relationships between units in the geo-databases. Using recent advances from statistical learning, a procedure of two-stap LASSO, with geographical restrictions, is proposed.

The main results of the simulations of Monte Carlo show that the procedure performs well, managing to detect the moments of structural break points in time and when an effect overflows and spreads to the neighbors.

Our empirical example uses monthly time series from January 2014 to November 2022, encompassing the 23 provinces and the Autonomous City of Buenos Aires (CABA). We show how works the proposed technique to estimate the real salary by province and identify any spatial or proximity related relationships among them.

In next advances, we pretend to integrate the estimation methodology in combination of large T panel estimators. Also, we will improve the empirical example, adding control variables and using a more complex model.

References

Ahrens, A. and Bhattacharjee, A. (2015). Two-step Lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128–155.

Aldstadt, J. and Getis, A. (2006). Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4):327–343.

Angulo, A., Burridge, P., and Mur, J. (2017). Testing for a structural break in the weight matrix of the spatial error or spatial lag model. *Spatial Economic Analysis*, 12(2-3):161–181.

Angulo, A., Burridge, P., and Mur, J. (2018). Testing for breaks in the weighting matrix. *Regional Science and Urban Economics*, 68:115–129.

Autant-Bernard, C. and LeSage, J. (2011). Quantifying knowledge spillovers using spatial econometric models. *Journal of Regional Science*, 51(3):471–496.

Beenstock, M. and Felsenstein, D. (2012). Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis*, 44(4):386–397.

Benjanuvatra, S. and Burridge, P. (2015). QML estimation of the spatial weight matrix in the MR-SAR model.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617–634.

Bodson, P. and Peeters, D. (1975). Estimation of the coefficients of a linear regression in the presence of spatial autocorrelation. an application to a belgian labour-demand function. *Environment and Planning A*, 7(4):455–472.

Elhorst, J. P. (2014). *Spatial econometrics: from cross-sectional data to spatial panels.* Springer.

Getis, A. and Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2):90–104.

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

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Griffith, D., Amrhein, C., and Huriot, J.-M. (2013). *Econometric advances in spatial modelling and methodology: essays in honour of Jean Paelinck*, volume 35. Springer Science & Business Media.

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Ho, T. K., Macià, N., Ray, B., Saeed, M., Statnikov, A., and Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hays, J., Kachi, A., and Franzese, R. (2010). A spatial model incorporating dynamic, endogenous network interdependence: A political science application. *Statistical Methodology*, 7(3):406–428.

Kelejian, H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17:99–121.

Kooijman, S. (1976). Some remarks on the statistical analysis of grids especially with respect to ecology. *Annals of Systems Research*, 5:113–132.

Lam, C. and Souza, P. (2014). Regularization for spatial panel time series using the adaptive lasso.

Lam, C. and Souza, P. (2019). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics*, pages 1–41.

Lee, L.-f. and Yu, J. (2012). Qml estimation of spatial dynamic panel data models with time varying spatial weights matrices. *Spatial Economic Analysis*, 7(1):31–74.

LeSage, J. and Pace, R. (2009). *Introduction to spatial econometrics.* Statistics: A Series of Textbooks and Monographs. Chapman and Hall, CRC press.

Maggioni, M. A., Nosvelli, M., and Uberti, T. E. (2007). Space versus networks in the geography of innovation: A european analysis. *Papers in Regional Science*, 86(3):471–493.

Merk, M. and Otto, P. (2022). Estimation of the spatial weighting matrix for regular lattice data. an adaptive Lasso approach with cross-sectional resampling. *Environmetrics.*

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Mur, J. and Paelinck, J. (2011). Deriving the w-matrix via p-median complete correlation analysis of residuals. *The Annals of Regional Science*, 47:253–267.

Otto, P. and Steinert, R. (2022). Estimation of the spatial weighting matrix for spatiotemporal data under the presence of structural breaks. *Journal of Computational and Graphical Statistics*.

Paelinck, J. and Klaassen, L. (1979). *Spatial econometrics*. Saxon House, Farnborough.

Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Qu, X. and Lee, L.-F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, 184:209–232.

Snijders, T., Steglich, C., and Schweinberger, M. (2007). *Longitudinal models in the Behavioural ond Related Sciences*, chapter Modeling the co-evolution of networks and behavior, pages 41–71. Lawrence Erlbaum.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240.

Zou, H. (2006). The Adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.