# A Comparison Study on Criteria to Select the Most Adequate Weighting Matrix

Marcos Herrera (University of Salta, Argentina); mherreragomez@gmail.com

Jesús Mur (University of Zaragoza); jmur@unizar.es[1]

Manuel Ruiz (University of Murcia, Spain); manuelruiz.spain@gmail.com

## Introduction

The weighting matrix issue is an ever-present topic in spatial econometrics, which reflects the great influence of the time series literature on this field. It is clear the vital importance of attaining a proper specification of this matrix in order to avoid biases and inconsistencies in applied work. Considering its importance, it is a bit surprising the treatment, unfocused, arbitrary to high extent, that the literature has so far given to it. Typically, the researcher provides only one matrix, or a few candidates, very similar among them, which are hardly questioned. Theoretical justification for the chosen matrix tends to be rather vague, and the selection or comparison problems are seldom considered. Fortunately, things have begun to change in recent years where they have appeared several proposal that advocate for a more data-driven approach. Indeed, it is easy to critizise the traditional mode of working in relation to W but this is not our purpose. Rather we prefer to try to improve the basis for choosing a given matrix, from among a finite number of different alternatives; in order to do that, we are going to look at the performance of several simple criteria already present in the spatial econometrics literature. Furthermore, we introduce a new non-parametric criterion, based on symbolic entropy, which offers some advantages to the previous criteria in conditions that deviate somewhat from the standard case.

Roughly, we may distinguish two approaches to the building of W (Harris et al., 2011): (i) specifying the matrix exogenously; (ii) estimating the matrix from data. The exogenous approach is by far the most common and includes, for example, use of a binary contiguity criterion, k-nearest neighbours, kernel functions based on distance, etc. The second approach uses the topology of the space and the nature of the data, and can take many forms. Most are ad-hoc procedures in which the researcher selects an objective which guides the search for the best W. Also included is the most recent work

[1]Corresponding author: Department of Economic Analysis, University of Zaragoza, Spain.

in which this matrix is directly estimated from the data as, for example, Benjanuvatra and Burridge (2015). More flexible approaches to W are possible if repeated information in a panel dataset is available as suggested initially by Meen (1996). Battacharjee and Jensen-Butler (2013) consider a panel data model with SEM errors, whereas Beenstock and Felsenstein (2012) pose the case of a SLM model with unobserved random effects. Another strand of the literature follows the approach of Case (1991), where the weights in the W matrix are endogeneous and changing over time. The recent contributions of Kelejian and Piras (2014), Qu and Lee (2015) or Kuersteiner and Prucha (2015) are in this line.

Against this background, our contribution focuses on the specific problem of choosing a matrix from a finite number of alternatives. It is assumed that the researcher has discussed the case under study and has defined a set of weighting matrices, which are compatible with the expected interaction channels in the model. Then there is the problem of selecting the best alternative among them or, in other words, the necessity of substantiate on objective premises the preference for a particular one.

## Comparing Weighting Matrices

There are two main principles that guide our approach to the W issue:

(i) The weighting matrix can be constructed in different ways using different interaction hypothesis. Each hypothesis results in a different weighting matrix and in different spatial lags, containing different information. In sum, different weighting matrices means different models.

(ii) There are general guidelines in relation to how a weighting matrix should be built (nearness, accessibility, influence, etc). However, a priori is difficult to discern which of them is preferable. This is a topic clearly dominated by uncertainty.

This is a well-known problem in the literature on spatial econometrics where we can find several criteria. Let us mention some of the more relevant.

First is the J test, inititally adated to this literature by Kelejian (2008) in a $SARAR$ framework, requiring of $GMM$ estimators. The test can be formulated as a Wald statistic whose asymptotic distribution, under usual condition, is a Chi-square. Burridge and Fingleton (2010) advocate for a bootstrap resampling procedure that appears to improve the poor behaviour of the Chi-square aproximation, in terms both of size and power. Burridge (2012) suggests a mixture between GMM and likelihood-based moment conditions in order to more effectively control the size of the test whereas Piras and Lozano (2010) present new evidence on the use of the test that relates its power to a wise selection of the instruments. Recently Hageman (2012) introduced a variant of the J test, called MJ (minimum J), that avoids sequential testing and thus the situation of having to conclude that both specifications explain the data equally well. Moreover, this variant does not require the correct model (or weight matrix in our case) to be among the considered specifications. On the negative side, we should mention the necessity to bootstrap the test which complicates its evaluation.

The problem of model selection has also been treated, very successfully, from a Bayesian perspective. Lesage and Pace (2009) show that the spatial framework fits quite will into a Bayesian approach. The same as the J-test, the starting point is a finite set of alternative models, $M = \{M_1; M_2; \ldots; M_R\}$,

whose specification coincides except for the spatial weighting matrix. Then, the posterior probability summarizes, for each model, the support that the data offers to the corresponding weight matrix. As is generally recognized, the Bayesian approach is very powerful to calibrate the suitability of a list of rival alternatives but is also highly demanding in terms of information needed; moreover, there remain doubts with respect to the robustness of the procedure under non-standard conditions.

Model selection techniques also have a role in this problem, specially if we do not have a clear preference for any weighting matrix, which tends to blur the meaning of null hypothesis. There is a huge literature on model selection for nested and non-nested models, with different purposes and combining different approaches. In our case, we join the mainstream view in favor of the Akaike information criterion as a simple and reasonable approximation to the Kullback-Leibler measure.

From a slightly different perspective, Hansen (2007) introduced the concept of model averaging. The purpose is to produce a linear combination of a finite set of existing models with the purpose of minimizing the mean square estimation error. The optimal decision, the same as the Bayesian posterior probability or other selection criteria, is to select the estimator with the lowest risk. This discussion can be adapted to the case of selecting the most adequate matrix from among a finite set of alternatives, $W = \{W_1; W_2; \ldots; W_Q\}$. The solution is a new matrix, $W_n$, which minimizes the mean square error of the estimates. Then the weights of the linear combination of $W_n$ can be used to solve the selection problem.

Finally, we also introduce a new non-parametric procedure for tackling the problem, based on the principle that the most adequate matrix should produce more relevant information with respect to the variables involved in the model. This information measure appears as a reformulation of the traditional entropy index in terms of what is called symbolic entropy (Matilla and Ruiz, 2008), and it does not requires prior information from the researcher.

## The Monte Carlo Experiment

The comparison between the five approaches to the problem of selecting the most adequate weighting matrix is made through a comprehensive Monte Carlo experiment, still under process, and whose design is described below.

For simplicity, in a first step, we are going to use a simple panel equation, containing only one regressors and its spatial lag in the right hand side, like the following:

$$y_{it} = f\left(\beta x_{it}; \theta \sum_{j=1}^{N} \omega_{ij} x_{jt}; \mu_{it}; \varepsilon_{it}\right) \; i = 1, ..., N; \; t = 1, ..., T. \tag{1}$$

where $\beta$ and $\theta$ are parameters; $\varepsilon_{it}$ is the error term of unit $i$ and period $t$ and $\mu_{it}$ its corresponding unobserved effects. Each experiment begins by producing a random map in a two-dimensional space which is reflected in the corresponding $W_0$ matrix, the true weighting matrix. Global parameters of the DGP are:

- The cross-sectional dimension, from small to large: $N \in \{50, 100, 500\}$
- The time dimension, reflecting small panels: $T \in \{5, 10, 20\}$

- The strength of the relation between $y$ and $x$, from weak to strong: $\beta \in \{0.5, 1.0, 2.0\}$
- The strength of spatial interaction from weak to strong: $\theta \in \{0.3, 0.5, 0.8\}$
- The form of the true weighting matrix which is built using different criteria: (i)- k-neighbors; (ii)- A continuous distance-decay function, and (iii)- Random assignment of contacts between the spatial units.
- The list and form of the rival weighting matrices. Four matrices are considered in the list of possible candidates, two of discrete type and two of a continuous type. The list of candidates may contain, or not, the true weighting matrix.
- Functional form. Function $f$ in the DGP may be linear o nonlinear but it is assumed that, as usual, the researcher works under the assumption of linearity in (1)
- Error terms and unobserved effects. Different combinations for both terms will be simulated, including standard assumptions and increasing departures from them.
- Nature of the relation, which can be static or dynamic.

# References


[1] Battacharjee, A. and C. Jensen-Butler (2013): Estimation of the Spatial Weights Matrix under Structural Constraints. *Regional Science and Urban Economics*, 43 617-634.

[2] Beenstock, M. and D. Felsenstein (2012): Nonparametric estimation of the spatial connectivitymatrix using spatial panel data. *Geographical Analysis*, 44 386-397.

[3] Benjanuvatra, S. and P. Burridge (2015): QML estimation of the spatial weight matrix in the MR-SAR model. York, DERS University of York Working Paper

[4] Burridge, P. (2012): Improving the J test in the SARAR model by likelihood-based estimation. *Spatial Economic Analysis* 7 75-107.

[5] Burridge, P. and Fingleton, B. (2010): Bootstrap inference in spatial econometrics: the J-test. *Spatial Economic Analysis* 5 93-119.

[6] Case, A. (1991): Spatial Patterns in Household Demand. *Econometrica*, 59 953-965.

[7] Hansen, B. (2007): Least Squares Model Averaging. *Econometrica,* 75, 1175-1189.

[8] Hageman, A. (2012): A simple test for regression specification with non-nested alternatives. *Journal of Econometrics* 166 247-254.

[9] Harris, R., J. Moffat and V. Kravtsova (2011): In search of W. *Spatial Economic Analysis*, 6 249-270.

[10] Kelejian, H (2008): A spatial J-test for Model Specification Against a Single or a Set of Non-Nested Alternatives. *Letters in Spatial and Resource Sciences*, 1 3-11.

[11] Kelejian, H. and G. Piras (2014): Estimation of spatial models with endoge- nous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics*, 46 140-149.

[12] Kuersteiner, G. and I. Prucha. (2015): Dynamic Spatial Panel Models: Networks, Common Shocks and Sequential Exogeneity. CESIFO Working Paper no 5445.

[13] Lesage, J. and K. Pace (2009): *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

[14] Meen, G. (1996): Spatial aggregation, spatial dependence and predictability in the UK housing market. Housing Studies, 11 345-372.

[15] Matilla, M. and M. Ruiz (2008): A non-parametric independence test using permutation entropy. *Journal of Econometrics*, 144, 139-155.

[16] Piras, G and N Lozano (2010): Spatial J-test: some Monte Carlo evidence. *Statistics and Computing,* DOI: 10.1007/s11222-010-9215-y.

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