20, 21, 22 · November 2019 · Castelló
XLV Reunión de Estudios Regionales · VI Jornades Valencianes d'Estudis Regionals
International Conference on Regional Science
Tackling with societal, technological and climate challenges
in peripheral territories
Universitat Jaume I

PAPER

Title: Moran's I revisited. Small sample case and the factor of scale.

Authors and e-mails: Jesús Mur (jmur@unizar.es**)**

Department: Economic Analysis
University: University of Zaragoza

Subject area: Métodos para el análisis espacial I: econometría espacial

Abstract:

A general purpose of the analysis of spatial data is the detection of cross-sectional dependencies in a given series, as a previous step to model the data. These mechanisms may combine different elements such as autoregressive or moving average structures among the most popular. In any case, we should test for the assumption of randomness using some of the tests proposed in the spatial econometrics toolbox. Especially outstanding among these is the Moran I test (Moran, 1950) whose characteristics (simplicity, efficiency) gave it a prominent role in applied literature. However, this test has also some weaknesses such its sensitivity to the scale of the data, which affects the power of the test, and its (unknown) distribution in a situation of small sample sizes.

In this paper we present evidence in relation to the first question and a new test of spatial autocorrelation, called I_{GQ} (in reference to the Goldfeld-Quandt test variant), whose distribution function is known for all sample sizes, it is not affected by scale factor problems of the data and appears to have better properties than the Moran's I, especially for the case of small sample sizes.

Keywords: Spatial Autocorrelation; Moran's I, Small Samples, Scale factor.

JEL codes: C12, C13, C15, C52; R15

1.- Introduction

In 1950 Moran introduced one of the most popular statistics in the analysis of spatial data, the Moran's I test, which addresses a simple question: are there cross-sectional dependencies in the data? The original framework of Moran was a single cross-section, where the connections were coded by a simple binary scheme. Since then, the situation has evolved quite a lot. For example, the weighting matrix can be symmetric or non-symmetric, and be based in any measure of distance. The test has been extended to panel data sets (Elhorst, 2010), to spatio-temporal dependencies (Lopez et al, 2011), and also to simultaneous (Kelejian and Prucha, 2001) and SUR (López et al., 2019) systems. The distribution of the statistic has been the subject of frequent research, with good results for the asymptotic case including several suggestions for finite sample sizes (Tiefelsdorf and Boots, 1995, Tiefelsdorf, 2000). In case of no normality, the randomization approach constitutes an elegant and efficient way to solve the inference (Cliff and Ord, 1981). Moreover, the available evidence points to Moran's I being one of the most efficient test to detect cross-sectiona dependencies (Anselin and Florax, 1995). Indeed, this is a very intuitive statistic, with a well-defined range of values, easy to obtain and useful for the user. In the end, it has laid aside other contemporary tests, such as Dacey’s d (Dacey, 1965) or Geary’s c (Geary, 1954), and remains competitive with other tests appeared later in the literature, such as the Lagrange Multipliers. It is not surprising that the Moran's I has almost 2 millions of items in Google.

The purpose of this paper is to contribute to the Moran’s I literature by highlighting some weaknesses that is worth to consider in applied work. In particular, we point to two issues which are (i) the problems of the Moran's I in situations of small samples (wrong size, low power) and (ii) the impact of the scale of the data on the behavior of the test. The last point is not very well known by practitioners but we are going to show that, under some circumstances, the power of the test is zero when this scale increases.

In the following section we present some basic results related to the test, while the third discusses the issues related to the sample size and the scale. In the fourth section we present some alternatives to deal with these difficulties, which are tested in the simulation solved in the Fifth section. The work finishes with the Sixth section dedicated to main findings and conclusions.

2. – Brief presentation of the Moran's I

The Moran statistics can be seen as an approximation to the coefficient of correlation between the original series, y, and its spatial lag, Wy, obtained using the so-called weighting matrix, W. It is expression has become popular:

$$I = \frac{R}{S_0} \frac{\sum_{r,s}^{R}\left(y_r - \bar{y}\right) w_{rs} \left(y_s - \bar{y}\right)}{\sum_{r,s}^{R}\left(y_r - \bar{y}\right)^2} = \frac{n}{S_0} \frac{y' DWDy}{y' BDy} \qquad (2.1)$$

with $\{y_r, r = 1, 2, \ldots, n\}$ the observations of the variable in n different spatial points and $\bar{y}$ its corresponding sampling mean; w_{rs} (r,s =1, 2, ... ,n) are the elements of the weighting matrix W, with $S_0 = \sum_{r,s}^{n} w_{rs}$. This statistic takes values in the interval $\left[\frac{n}{S_0}\gamma_{MIN}; \frac{n}{S_0}\gamma_{MAX}\right]$ with γ_{MIN} and γ_{MAX} the lowest and highest eigenvalues of the matrix DWD with $D = \left[I - \frac{\tau\tau'}{R}\right]$ the demeaning matrix (De Jong et al, 1984); τ is a (nxl) vector of ones and I is the identity matrix of order n.

The moments of the Moran's I, even under the assumption of independence, are not immediate to obtain as can be seen in Moran (1950). The discussion simplifies if it is possible to assume normality; if this is not the case, Cliff and Ord (1981) suggest a randomization approach which is very effective. Moreover, Cliff and Ord (1972) show that the asymptotic distribution of the test approaches normality as the sample size increases. The conditions that support this approximation are relatively weak, as can be seen in Sen (1976). Its distribution is unknown for small samples, although Sen (1990) and Tiefelsdorf and Boots (1995), develop procedures to obtain the exact distribution function using numerical integration methods. Furthermore, King (1981) demonstrates that the statistic of Moran is a Locally Best Invariant (LBI) test in the neighborhood of the null hypothesis and, under certain conditions, also a Uniformly Most Powerful Invariant (UMPI) test. Burridge (1980) obtains its formal equivalence with the Lagrange Multipliers (LM), while Anselin and Rey (1991) and Anselin and Florax (1995) demonstrate, using Monte Carlo methods, its good performance and superiority as a test of spatial autocorrelation in relation to other competitors. These results confer the Moran's I test a central position in the field of spatial analysis.

There are also flaws and weaknesses that affects the Moran's I such as, for example, that the alternative hypothesis, in case of rejecting the null of no correlation, remains unspecific. Moreover, it is a conditional test in the sense that the conclusion is restricted to the W matrix used in the analysis. Finally, we need to assume that, under the hypothesis of independence, the first and second order moments of the variable are constant across space. As will be clear below, this point is the nexus between the issues of autocorrelation and spatial heterogeneity, and one of the causes of false rejections of the null hypothesis when using the Moran's I.

3. – The scale factor and the sample size issues

As indicated, one of the limitations with the Moran's I is the lack of a well-defined alternative hypothesis. In case of rejecting the null hypothesis, it is frequent to specify a spatial autoregressive process, SAR, for the series. However, there are other options such as spatial moving average, SMA, or a spatial error component, SEC (Kelejian and Robinson, 1993). The first involves a structure of global dependencies which is of local nature for the other two. Furthermore, a SAR process means, in general, that the first and second order moments of the variable changes for each point in space which results in heterogeneity, a property shared by most spatial series. A typical SMA or SEC series will show up more regularity across space, with a constant first order moment.

These observations have an impact in the Moran's I through the sample mean because this is an unbiased and consistent estimator only for the SMA and SEC cases:

$$y = \mu\tau + (I - \delta W)u = \mu\tau + Bu$$

$$\bar{y} = \frac{\tau' y}{R} \rightarrow \begin{cases} E[\bar{y}] = \mu & \rightarrow \lim\limits_{n\to\infty} E[\bar{y}] = \mu \\ V[\bar{y}] = \dfrac{\sigma^2}{n}\left(1 + \delta\dfrac{S_0}{n}(\delta - 2)\right) & \rightarrow \lim\limits_{n\to\infty} V[\bar{y}] = 0 \end{cases} \qquad (1)$$

with B=I-δW. As said, the first order moment for the SAR case, assuming that μ ≠ 0, changes from point to point:

$$y = \mu\tau + \delta W y + u \Rightarrow y = B^{-1}(\mu\tau + u) \Rightarrow E[y] = \mu B^{-1}\tau \qquad (2)$$

The sampling mean is centered on the mean of these expected values, $E[\bar{y}] = \mu\frac{\tau' B^{-1}\tau}{n}$, but it is a statistic lacking any meaning[1]. The same as with SMA case and under general conditions, the variance of the sampling mean asymptotically goes to zero[2] $\lim_{n\to\infty} V[\bar{y}] = \sigma^2\frac{\tau' B^{-2}\tau}{n^2} = 0$.

The use of the sampling mean in the Moran's I is justified in order to remove a common factor in the data, which is only present in SMA or SEC processes. Thus the question is the incidence in the behavior of the test. To simplify, let us assume that the W matrix is symmetric. We can write for the SMA, I_{SMA}, and SAR, I_{SAR}, cases that:

$$I_{SMA} = \frac{n}{S_0}\frac{(\mu\tau+Bu)^{'}DWD(\mu\tau+Bu)}{(\mu\tau+Bu)'D(\mu\tau+Bu)} \qquad I_{SAR} = \frac{n}{S_0}\frac{(\mu\tau+u)'B^{-1}DWDB^{-1}(\mu\tau+u)}{(\mu\tau+u)'B^{-1}DB^{-1}(\mu\tau+u)} \quad (3)$$

The factor of scale disappears from the SMA case, $I_{SMA} = \frac{n}{S_0}\frac{u'BDWDBu}{u'BDBu}$, which becomes a quotient of two quadratic forms of n normal N(0,1) variates, u/σ, on two symmetric but singular matrices where $rg(BDWDB) = Min\{n-1, rg(W)\}$ and $rg(BDB) = n-1$.

The distribution of both quadratic forms will pertain to the chi-squared family, although they will not be independent. Using the results of Yule and Kendall (1950), we can approximate their expected value though:

Assume that the weighting matrix is symmetric, then the statistic in (3.3) is a quotient of two quadratic forms of a vector of N(0, σ^2) variates on two symmetric and singular matrices. The two quadratic forms are not independent but we can use the result of Yule and Kendall (1950) to approach the expected value of the quotient

$$E\left[\frac{N}{D}\right] = \frac{E(N)}{E(D)}\left[1 + \frac{V(N)}{E(D)^2} - \frac{Cov(N,D)}{E(N)E(D)}\right] + o(R^{-2}) \qquad (3.4)$$

[1] Except when the weighting matrix is row-standardized in which case the mean value of the variable remains constant, $E[y] = \mu\frac{\tau}{1-\delta}$ and $E[\bar{y}] = \frac{\mu}{1-\delta}$

[2] If the matrix is row-standardized, $\lim_{n\to\infty} V[\bar{y}] = \frac{\sigma^2}{(1-\delta)^2}\frac{1}{n} = 0$).

where o(-) means "of smaller order than". Solving the above expression, the expected value of Moran's I is:

$$E[I] \simeq \frac{R}{S_0}\frac{\mathrm{tr}BDWDB}{\mathrm{tr}BDB}\left[1+\frac{\mathrm{tr}(BDB)(BDB)}{(\mathrm{tr}BDB)^2}-\frac{\mathrm{tr}(BDWDB)(BDB)}{\mathrm{tr}(BDWDB)\,\mathrm{tr}(BDB)}\right] \qquad (3.5)$$

which can be approximated by:

$$E[I] \simeq \frac{R}{S_0}\frac{\sum_{r=1}^{R}\lambda_r(1-\delta\lambda_r)^2}{\sum_{r=1}^{R}(1-\delta\lambda_r)^2}\left[1+\frac{\sum_{r=1}^{R}(1-\delta\lambda_r)^4}{\left(\sum_{r=1}^{R}(1-\delta\lambda_r)^2\right)^2}+\frac{\sum_{r=1}^{R}\lambda_r(1-\delta\lambda_r)^4}{\sum_{r=1}^{R}\lambda_r(1-\delta\lambda_r)^2\sum_{r=1}^{R}(1-\delta\lambda_r)^2}\right] \qquad (3.6)$$

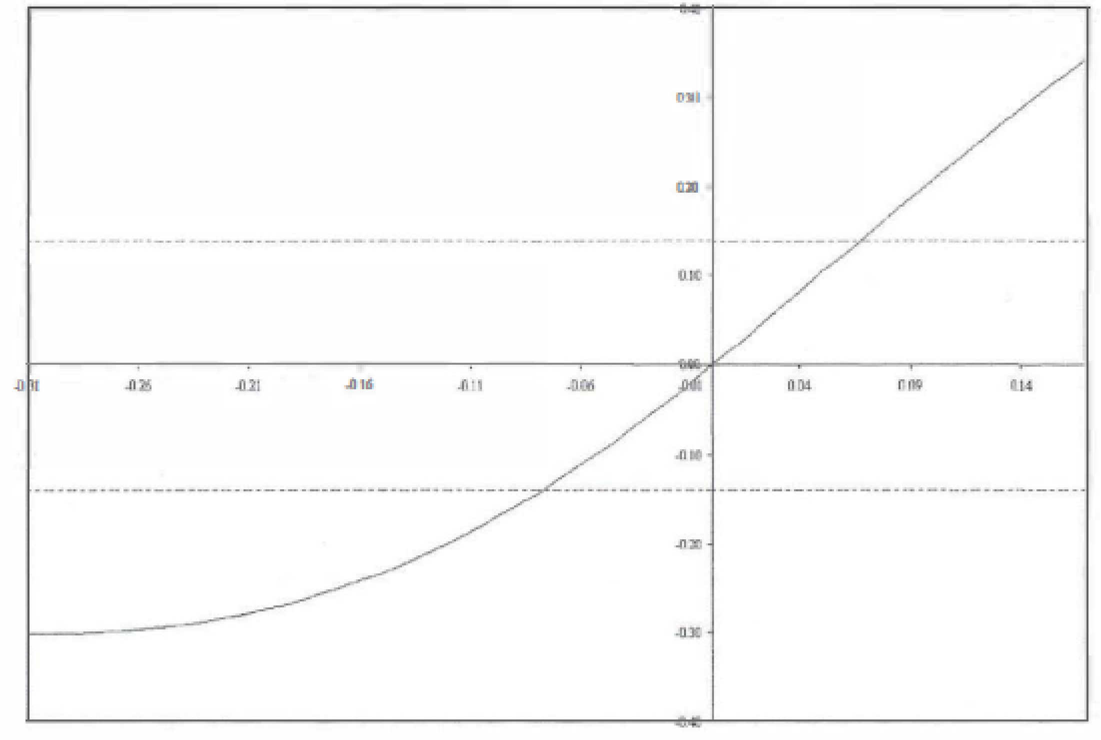
with λ_r. the r-th characteristic root of W. For positive values of the parameter δ, the expected value of I will be negative and positive for negative values of the former. The two quotients that appear in the square brackets are positive and less than one (using the Cauchy-Swartz inequality). Their contribution will become less significant as the sample size increases, so that the above expression can be reduced, for large sample sizes, to:

$$E[I] \rightarrow \frac{R}{S_0}\frac{\sum_{r=1}^{R}\lambda_r(1-\delta\lambda_r)^2}{\sum_{r=1}^{R}(1-\delta\lambda_r)^2} \qquad (3.7)$$

Independently of the accuracy of these approximations, the relevant aspect of all of them is that at no moment is the distribution of Moran's I affected by the scale of the process, being efficiently neutralized by the sampling mean.

The above result allows us to present graphs such as that of Figure 3.1. With the continuous line we represent the expected value of Moran's 1 according to (3.7) and with the dotted lines the acceptance limits of the null hypothesis of independence at a significance level of 5% (that is 1.96xDT (I), where DT(I) is the standard deviation under the null hypothesis). The reference matrix, of the order (74x74), corresponds to the NUTS Il European regional system of 12 member states. The stability interval (if by such we understand that in which it is true that $|\delta\lambda_r| < 1; \forall r$ associated with this matrix is (-0.31, 0.17).

Figure 3.1: Expected value of Moran's I for SMA processes.



The situation in the SAR case is different, given that, as we mentioned before, the first order moment of the series is not constant, which means that the sampling mean will be a biased estimator of the scale factor. Taking (3.2) as a point of reference, it follows that:

$$E[I] \to \frac{R}{S_0} \frac{\sum_{r=1}^{R} \lambda_r (1-\delta\lambda_r)^2}{\sum_{r=1}^{R} (1-\delta\lambda_r)^2}$$

$$E[\bar{y}] = \mu \frac{\tau' B^{-1} \tau}{R} \Rightarrow \begin{cases} \text{if} \quad \delta < 0 \Rightarrow 0 < \dfrac{\tau' B^{-1} \tau}{R} < 1 \Rightarrow E[\bar{y}] < \mu \\ \text{if} \quad \delta > 0 \Rightarrow \dfrac{\tau' B^{-1} \tau}{R} > 1 \Rightarrow E[\bar{y}] > \mu \end{cases} \tag{3.8}$$

In this type of process, Moran's I can be developed as:

$$I = \frac{R}{S_0} \frac{(u+\mu\tau)' B^{-1} DWD B^{-1} (u+\mu\tau)}{(u+\mu\tau)' B^{-1} D B^{-1} (u+\mu\tau)} \tag{3.9}$$

Using the approximation of (3.4) again, its expected value can be expressed as:

$$E[I] \simeq \frac{R}{S_0}\frac{trB^{-2}DWD + c^2\tau' B^{-1}DWDB^{-1}\tau}{trB^{-2}D + c^2\tau' B^{-1}DB^{-1}\tau}\left[1 + \frac{trB^{-2}DB^{-2}D + 4c^2\tau' B^{-1}DB^{-2}DB^{-1}\tau}{\left(trB^{-2}D + c^2\tau' B^{-1}DB^{-1}\tau\right)^2}\right. \tag{3.10}$$

$$\left. - \frac{trB^{-1}DB^{-2}DWDB^{-1} + 4c^2\tau' B^{-1}DWDB^{-2}DB^{-1}\tau}{\left(trB^{-2}DWD + c^2\tau' B^{-1}DWDB^{-1}\tau\right)\left(trB^{-2}D + c^2\tau' B^{-1}DB^{-1}\tau\right)}\right]$$

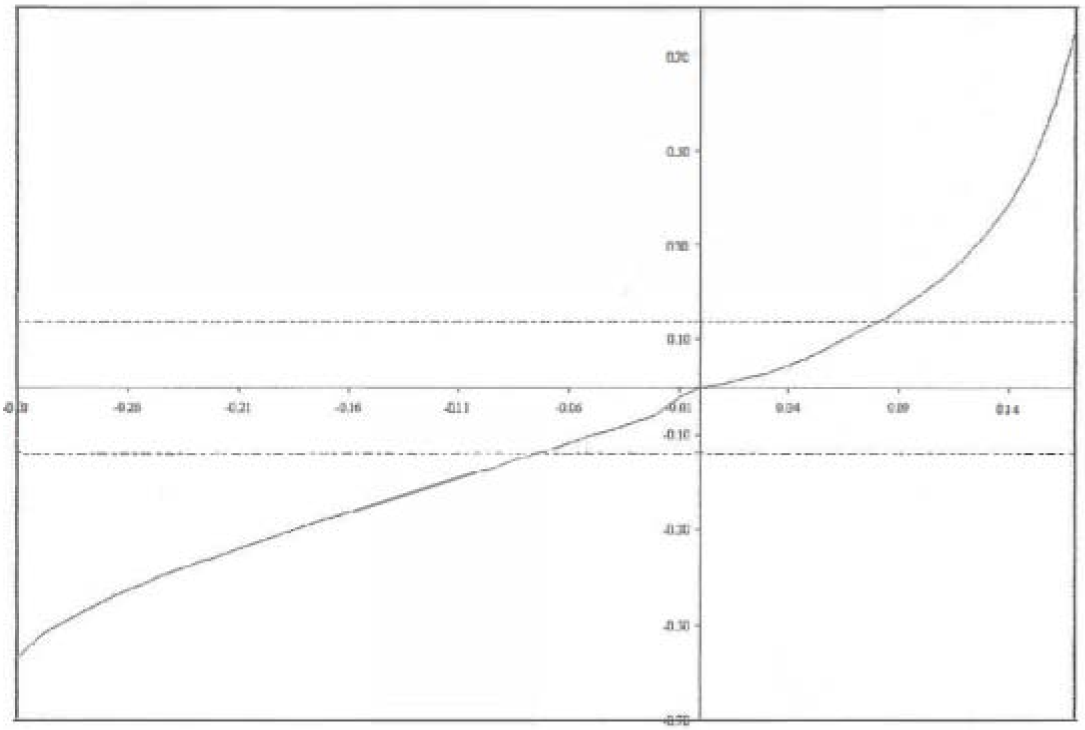
with $c=\mu\backslash\sigma^2$ the coefficient of variation of the process. The above result is intractable, although the probability limit of (3.9) can be considered as an approximation:

$$p\lim I \simeq \frac{\lim\limits_{R\to\infty}\left[trB^{-1}DWDB^{-1}/S_0\right] + c^2\lim\limits_{R\to\infty}\left[\tau' B^{-1}DWDB^{-1}\tau/S_0\right]}{\lim\limits_{R\to\infty}\left[trB^{-1}DB^{-1}/R\right] + c^2\lim\limits_{R\to\infty}\left[\tau' B^{-1}DB^{-1}\tau/R\right]} = \frac{n_1 + c^2 n_2}{d_1 + c^2 d_2} \tag{3.11}$$

The terms d_1 and $_{d2}$ are positive for any δ, while n_1 and n_2 will be negative for $\delta < 0$ and positive if the opposite is true. Given that c^2 will al so be positive, the presence of a factor of scale in the DGP of the series will not affect the sign of the test. However, as the scale increases the I statistic will tend towards the quotient (n_2/d_2), terms strictly associated with the scale (and with the error in the estimation of the first order moment).

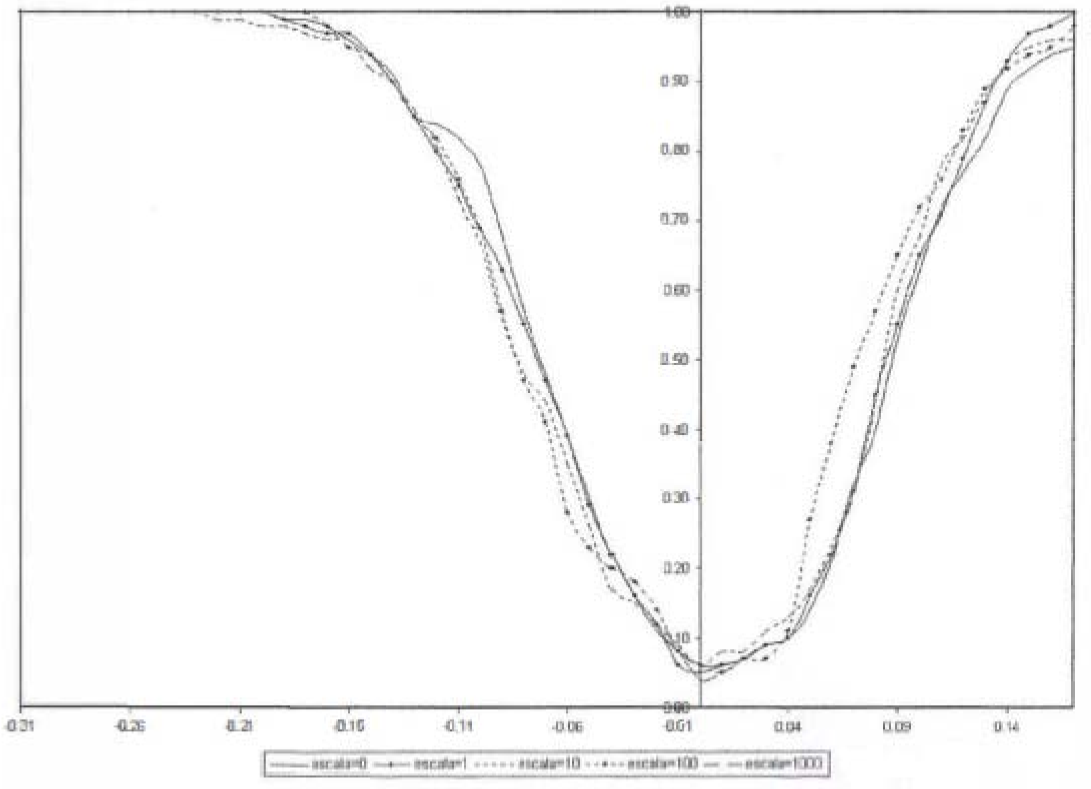
This situation can be represented on a graph as it appears in Figure 3.2. The continuous line represents the quotient (n_2/d_2) obtained for the same contiguity matrix used in Figure 3.1. This is the limit of Moran’s I when we increase the factor of scale of the series indefinitely. The dotted lines represent the acceptance limits of the null hypothesis of independence at a significance level of 5%.

Figure 3.2: Limit of the expected value of Moran's 1 for SAR processes.



Figures 3.1 and 3.2 are apparently similar, though in the first we are representing an expected value around which the finally observed value of the I statistic will fluctuate (SMA case), while in the second we represent the convergence limit of the same statistic (SAR case). That is to say that, in the first case, Moran's I will have a certain capacity to detect relationships of spatial autocorrelation even when the coefficient of the process is close to zero, while in the second there is a zone of values in which the test will not have any power at all (if the scale is high).

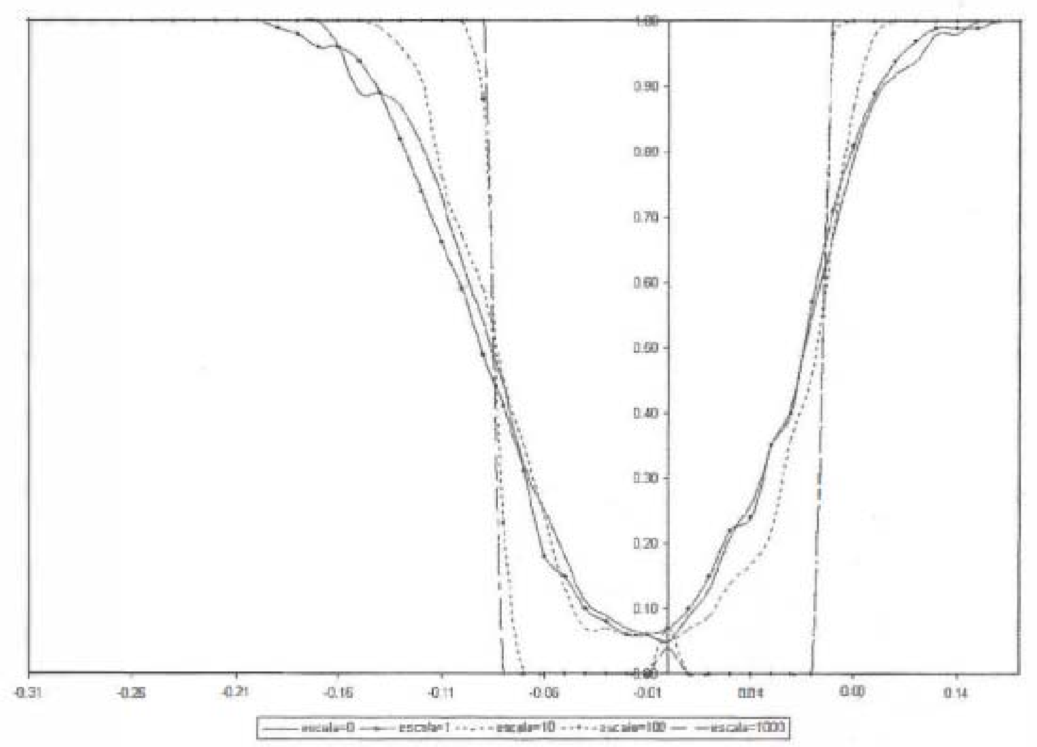
Figure 3.3: Estimated power function of Moran's I. SMA case.



In Figures 3.3 and 3.4 we reproduce the results of a small simulation (100 replications in each experiment) carried out to confirm this effect. The contiguity matrix used is the same (that corresponding to the European regional system of 12 member states, binary and of the order 74x74). Random series have been obtained from an N(0,1) distribution, and then transformed in SMA or SAR processes with a factor of scale ranging between 0 and 1000. In the graphs the percentage of rejections of the null hypothesis of no correlation is reflected, obtained in each case using Moran's test.

These graphs tend to corroborate previous comments. Moran's I does not show any special sensitivity to the factor of scale when it is applied to SMA processes, while its impact is evident in SAR structures. In this case, and when the scale of the series is 1000, the power of the test is zero for values of the coefficient of autocorrelation between -0.08 and 0.07. ln the rest of the parametric space the power is one.

Figure 3.4: Estimated power function of Moran's I. SAR case.



4. – Some Proposals

The problem noted in the previous section affects a restricted zone of the parametric space (dependent on the contiguity matrix) and occurs only when the series is of SAR type. It is not a critical problem but does create certain inconveniences. Some solutions appear obvious, such as using a logarithmic transformation on the series before resolving the test. In this paper we present another proposal that may be useful when the sample size is not very large, given that it implies using the eigenvectors and eigenvalues of the contiguity matrix W.

If the series is a SAR, its DGP is given by:

$$y = \mu\tau + \delta W y + u \rightarrow y = \left[I - \delta W\right]^{-1}\left(\mu\tau + u\right) = B^{-1}\left(\mu\tau + u\right) \qquad (4.1)$$

where τ is a vector of ones of (Rx1) order, μ the factor of scale, δ the parameter of autocorrelation and u a white noise vector $N(0, \sigma^2 I)$. If it is a SMA, the associated DGP will be:

$$y = \mu\tau + \left[I - \delta W\right]u = \mu\tau + Bu \qquad (4.2)$$

ln both cases, the series y and u are referenced to the canonical base (e), although we can also use other bases such as that composed by the eigenvectors of B (q), coincident with those of W and independent of the parameters of the process. If we order these vectors in the columns of matrix Q, the mapping matrix from one base to the other will be Q' which, applied to (4.2) allows us to write:

$$Q'y = \mu Q'\tau + Q'Q[I - \delta\Lambda]Q'u \to y^* = \mu q^* + \Delta u^* = \mu q^* + v^* \qquad (4.3)$$

where y*=Q'y and u*=Q'u are the co-ordinates of the original vectors y and u in the new base q and q*=Q'l. Δ is the matrix of characteristic roots of Δ which depends on the corresponding matrix of roots of W (Λ). The transformed noise vector maintains the characteristics of the original: u* ~ N(0, σ^2I). Given that it is not observable, its incidence is resumed in the random term v*, different to the previous one in that it is heteroskedastic: v* ~ N(0, $\sigma^2\Delta^2$).

$$Q'y = Q'Q[I - \delta\Lambda]^{-1}(\mu Q'\tau + Q'u) \to y^* = \Delta^{-1}\left(\mu q^* + u^*\right) = \mu\Delta^{-1}q^* + w^* \quad (4.4)$$

If we resolve a similar transformation in the SAR process of (4.1) we obtain:

$$Q'y = \mu Q'\tau + Q'u \to y^* = \mu q^* + u^* \qquad (4.5)$$

With w = Δ^{-1}u*~ N(0, σ^2 Δ^{-2}). Lastly, when the series is composed of a factor of scale and a white noise without spatial structure (δ=0), the filter above leads to:

The last three equations (4.3), (4.4) and (4.5) allow us to design a strategy of analysis of spatial series based on the following considerations:

(i)- Given that the matrix of eigenvectors Q is not singular and known (it is directly associated with W), the filtering of the original series should not have any effect on the quality or the quantity of information contained in the sample.

(ii)- If a factor of scale intervenes in the DGP of the original series, a term q*=Q'l should appear in the DGP of the filtered series.

(iii)- When there exists some spatial structure (of SAR or SMA type) in the DGP of the original series, the error term linked to the filtered series will be heteroskedastic. Also, the heteroskedastic function will respond exclusively to the series of eigenvalues of the contiguity matrix.

(iv)- The systematic part of the equation that describes the DGP of the filtered series will be linear in the variable q* (proportional in accordance with 4.3) in the SMA case, and non-linear in the SAR case (equation 4.4).

This strategy can be made operative using, as a testing equation:

$$y^* = \alpha + \beta q^* + n^* \qquad (4.6)$$

in which it will be true that α=0, β=μ and n*=u*·when the original series is a white noise. If the DGP is of SMA type, it wi1l be verified that α=0, β=μ and n*=v*. Lastly, when the DGP is of SAR type, the expansion of (4.4) leads to:

$$y^* = \mu q^* + (\mu\delta)\left[\Lambda q^*\right] + \left(\mu\delta^2\right)\left[\Lambda^2 q^*\right] + \cdots + w^* \qquad (4.7)$$

so that, in (4.6), α=0, β=μ and $w^* = v^* + \left(\mu q^*\right)\sum_{j=1}^{\infty}\delta^j\Lambda^j$. The LS estimation of (4.6) will produce unbiased and consistent estimators of α and β when the DGP is of the first or second type (white noise with scale or SMA), but they will be biased and inconsistent in the SAR case. This bias can be corrected, at least partially, using a testing equation such as:

$$y^* = \alpha + \beta_1 q_1^* + \beta_2 q_2^* + \cdots + \beta_p q_p^* + w^* \qquad (4.8)$$

with $q_j^* = \Delta^j q^*$; j = 0, 1, ... , p . For a sufficiently high value of p we can expect that the impact of the specification error (which will still exist in the SAR case) on the LS estimation will be moderate. In any case, the relevant aspect is that if heteroscedasticity is detected in the error term of (4.8), associated explicitly with the structure of the contiguity matrix, this will be an unequivocal sign of spatial autocorrelation in the original series.

Among the various possibilities that exist, the Goldfeld-Quandt test appears to bring together the principal requirements:

- lt is easy to obtain.
- Its distribution is known for all samp1e sizes.
- Spatial structure can be introduced explicitly in the test process.

In respect to the latter, it must be taken into account that in the disturbance of the SMA process of (4.2) it will be true that:

$$V\left[v_r^*\right] = \sigma^2(1-\delta\lambda_r)^2 \Rightarrow \frac{\partial V\left[w_r^*\right]}{\partial\lambda_r} = -2\delta\sigma^2(1-\delta\lambda_r)\begin{cases} <0 \leftrightarrow \delta>0 \\ >0 \leftrightarrow \delta<0 \end{cases} \forall r \quad (4.9)$$

while in SAR type processes:

$$V\left[w_r^*\right] = \frac{\sigma^2}{(1-\delta\lambda_r)^2} \Rightarrow \frac{\partial V\left[w_r^*\right]}{\partial\lambda_r} = \frac{2\delta\sigma^2}{(1-\delta\lambda_r)^3}\begin{cases} >0 \leftrightarrow \delta>0 \\ <0 \leftrightarrow \delta<0 \end{cases} \forall r \qquad (4.10)$$

In both cases, for a concrete value of 8, the variance evolves systematically with λ_r. This is all the information that we need for resolving the Goldfeld-Quandt test. The values of y* will be ordered highest to lowest (or lowest to highest) with λ; the central c observations will be excluded (our experience corroborates the normal practice of excluding a third of the sample); the equation (4.8) will be estimated by LS in the first and last subsamples and the test statistic will be obtained as (more details in the Appendix):

$$GQ = \frac{SR_{MAX}}{SR_{MIN}} \sim F\left(\frac{R-c}{2}-k;\frac{R-c}{2}-k\right) \qquad (4.11)$$

where SR_{MAX} and SR_{MIN} are the highest and lowest residual sums respectively and k=p+l.

In the context in which we have set the discussion, another attractive alternative is the Breusch-Pagan (BP) LM test. In this case, the concrete functional form associated with the heteroskedastic variance is unknown (it could be 4.9 or 4.1 0), although its arguments are identified (the eigenvalues of the contiguity matrix W). Other possibilities (the tests of Szroeter, 1978 and White, 1982) have been considered but without much success.

One result derived from the above discussion is that, if heteroskedastic relationships have been detected in the filtered series, this information can be used to try to identify the DGP of the series. In accordance with (4.3), if the series is of SMA type, the existence of a scale different to zero in the DGP will result in $\beta \neq 0$. Also, if the variance of the error term n* responds to the sequence of values $\{\lambda_r$, r=l, 2, ... , R$\}$, the conclusion is that the original series presents a structure of spatial correlation. On the other hand, if the DGP is of SAR type, the equation of reference is (4.4) expanded in (4.7). That is to say that what differentiates both types of process is that the autoregressive requires a wider structure of regressors (q^*, Λq^*, $\Lambda^2 q^*$, ...) than that of the moving average (only q*).

This reasoning can be developed in different ways. In the first place, (4.3) and (4.4) are different functional forms, which could give rise to a model selection strategy based on examining the suitability of each functional form or on some other more specific criterion. Another possibility could be to use (4.8) as a nesting equation of both processes to contrast the restrictions that lead to an SMA structure ($\beta_2=\beta_3= \ldots = \beta_p=0$) and to an SAR structure (which will be non-linear). Another less rigorous option, although of simpler resolution, could be to accept an SMA structure initially when signs of autocorrelation have been found. The adoption of SAR structures instead of SMA ones would be carried out only when there is strong evidence in their favor. This strategy could be carried out by means of a simple testing equation:

$$y^{*} = \alpha + \beta_1 q_1^{*} + \beta_2 q_2^{*} + w^{*} \qquad (4.12)$$

estimated by LS. The acceptance of β_2=0 implies the adoption of an SMA structure while its rejection leads to the adoption of an SAR structure. The test statistic could be the t-ratio associated with this parameter with the peculiarity that, given that the error term of that equation is heteroskedastic, a consistent estimation of the covariance matrix of the LS estimators must be used. In this sense, the proposal of White (1982) is very useful because it generates consistent estimations of the matrix even under certain misspecifications in the model (White, 1980), among which we can include errors in the functional form as in the case of (4.12) in relation to the SAR structure presented in (4.4).

5. – A small simulation

In the third section we have commented that Moran's I is sensitive to the scale of the series when this has been generated by an SAR type process. This limitation becomes a problem when it occurs in zones of the parametric space in which the test has zero power. In the fourth section some solutions have been put forward: taking logarithms on the series before obtaining the autocorrelation test or resorting to one of the heteroskedasticity tests on the filtered series. The usefulness of these proposals will be checked below by means of a small Monte Carlo exercise whose most relevant characteristics are the following:

- For the moment we have replicated just one sample size (R=74).
- We have used only one contiguity matrix of binary type and of order (74x74) corresponding to the system of European regions (NUTS II level) of 12 member states.
- Series of random numbers of order (7 4 x 1) have been obtained from a N(0,1) distribution, which have later been transformed in SAR or SMA processes using (4.1) or (4.2). In each case 100 replications have been resolved.
- Different factors of scale have been used in order to analyze their impact on each test. The replicated values where t=0, 1, 10, l00 and 1000.
- The stability interval associated with the contiguity matrix is (-0.31, 0.17).

The most relevant results appear in Figures 5.1 to 5.6 in terms of the percentage of rejections of the null hypothesis of independence. Figures 5.1 and 5.2 refer to the GQ test. In the first an SMA has been used as the DGP and in the second an SAR. Figures 5.3 and 5.4 show the performance of the BP test, In the last two figures, 5.5 and 5.6, Moran's I has been used on the logarithms of the original series (to guarantee their existence only scales 10, 100 and 1000 have been replicated.

The testing equation used both for GQ and for BP has been 4.8, fixing p as 3. This value seems to maintain a certain equilibrium between the over-specification that exists when the series has been generated by an SMA (which results in a loss of power of the tests) and the sub-specification characteristic of the SAR case (of which the result is estimated power curves that seem anomalous). Lastly, the heteroskedastic hypothesis of the BP test has been specified using λ_r, λ_r^2; and λ_r^3; as regressors, with the aim of buffering the effects of the error in functional form that exists in the SAR case.

The results collected in these figures allow us to highlight some provisional conclusions:

As was foreseeable, there is no scale effect either in the GQ test or in the BP test when they have been applied to SMA series. However, signs are still appreciable in the SAR case, more clearly with the BP test. Their incidence can be diluted by increasing the order of p in the testing equation (4.8), at the cost of a progressive worsening of the power of both tests.

The GQ test tends to overestimate the size of the test while BP tends to underestimate it. In the first case, the percentage of rejections observed for a zero value of the parameter of autocorrelation is systematically above 5%, in a range comprised between 6% and 10%. The size estimated in the BP test is closer to the theoretical significance level of 5%, although with a tendency to fluctuate between 3% and 4%. We believe this is due to a lack of precision in obtaining the eigenvectors of the contiguity matrix.

The estimated power for the GQ test is clearly superior to that of the BP test in all cases.

The logarithmic transformation of the series does not prevent the scale effect characteristic of Moran's I in SAR processes. The appearance of Figures 5.6 and 3.4 is similar (the same is true for 3.3 and 5.5 in the SMA case), with a slight reduction in the range of zero power (now it is -0.06 to 0.06). However, other problems arise such as the increase in the size of the test as the scale of the process grows (it is 12% with a scale of

1000), or some misleading covariances in the numerator of the test which lead to the change of the sign of the sampling Moran's

Figure 5.1: The Goldfeld-Quandt test. SMA case.

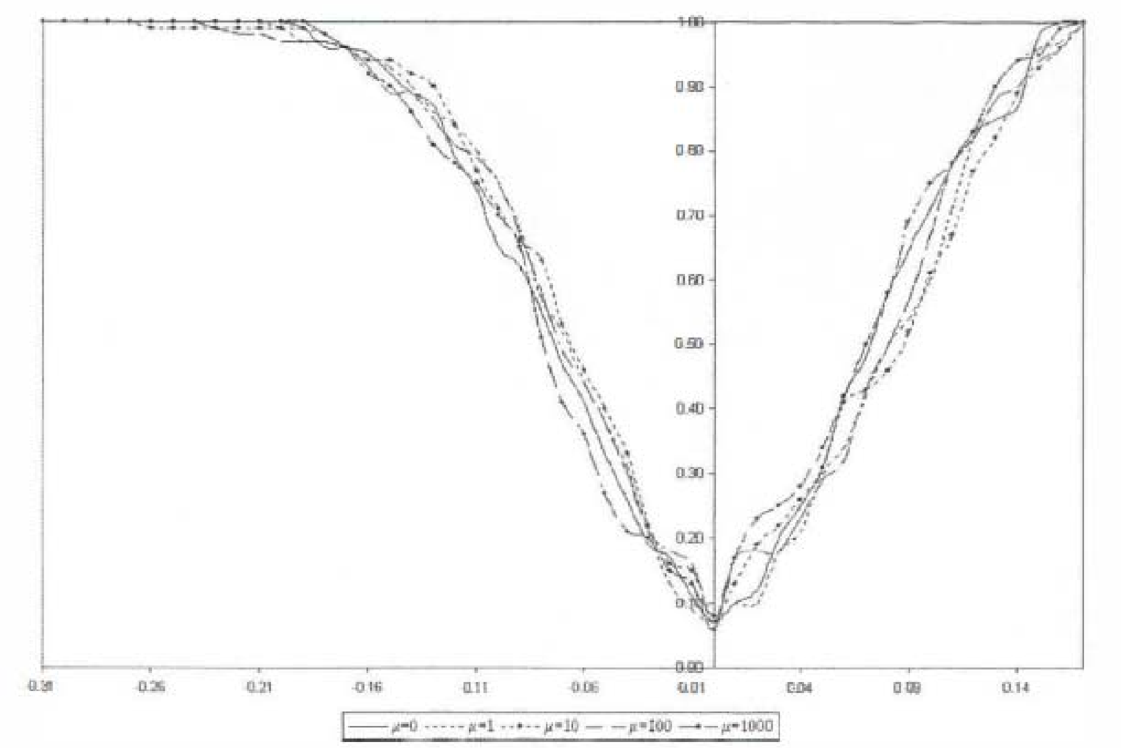


Figure 5.2: The Goldfeld-Quandt test. SAR case.

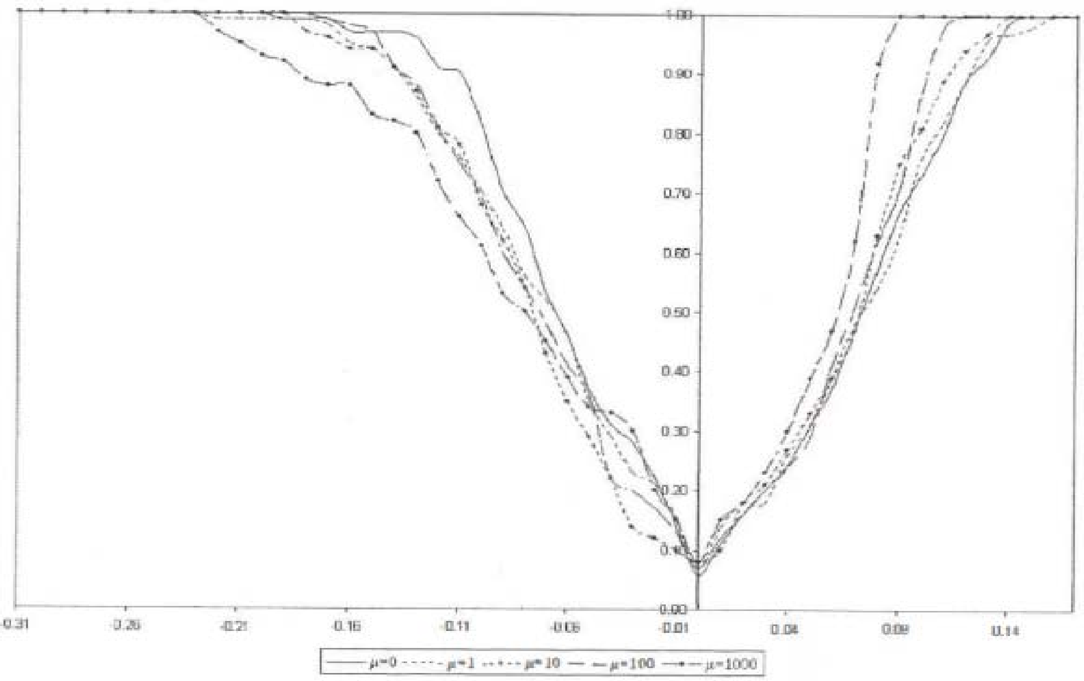


Figure 5.3: The Breusch-Pagan test. SMA case

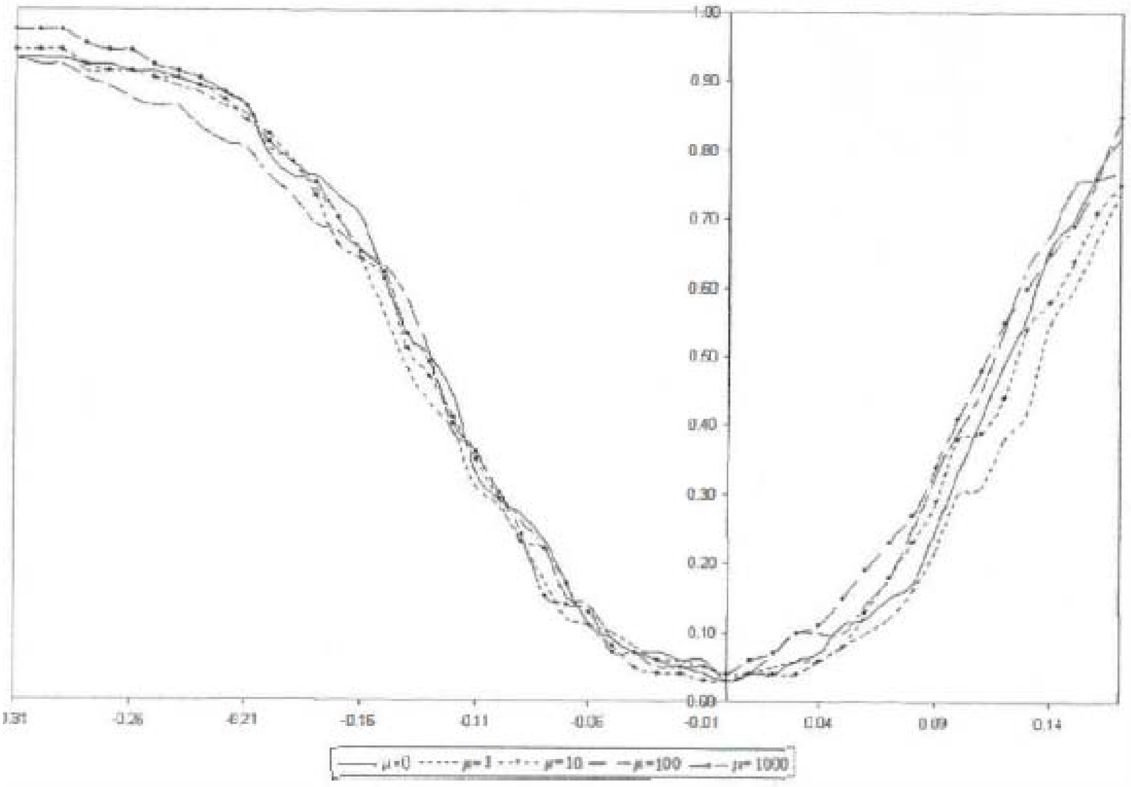


Figure 5.4: The Breusch-Pagan test. SAR case

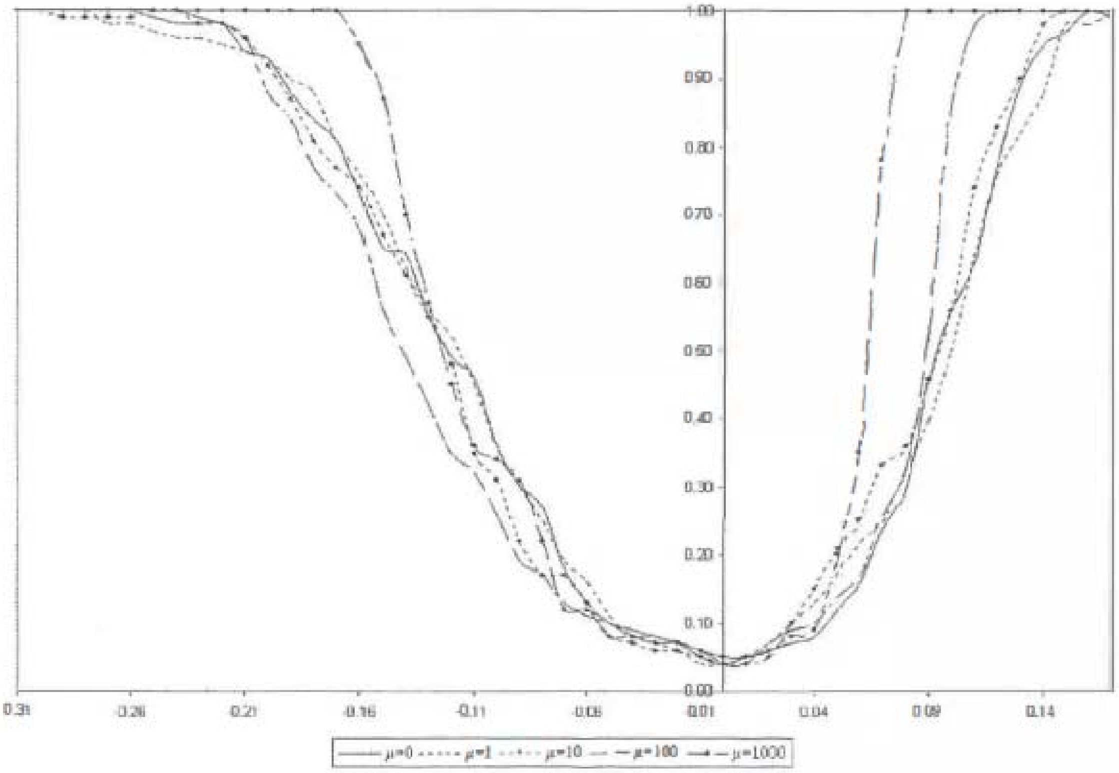


Figure 5.5: Moran's 1 test. Logarithmic Transformation. SMA case.

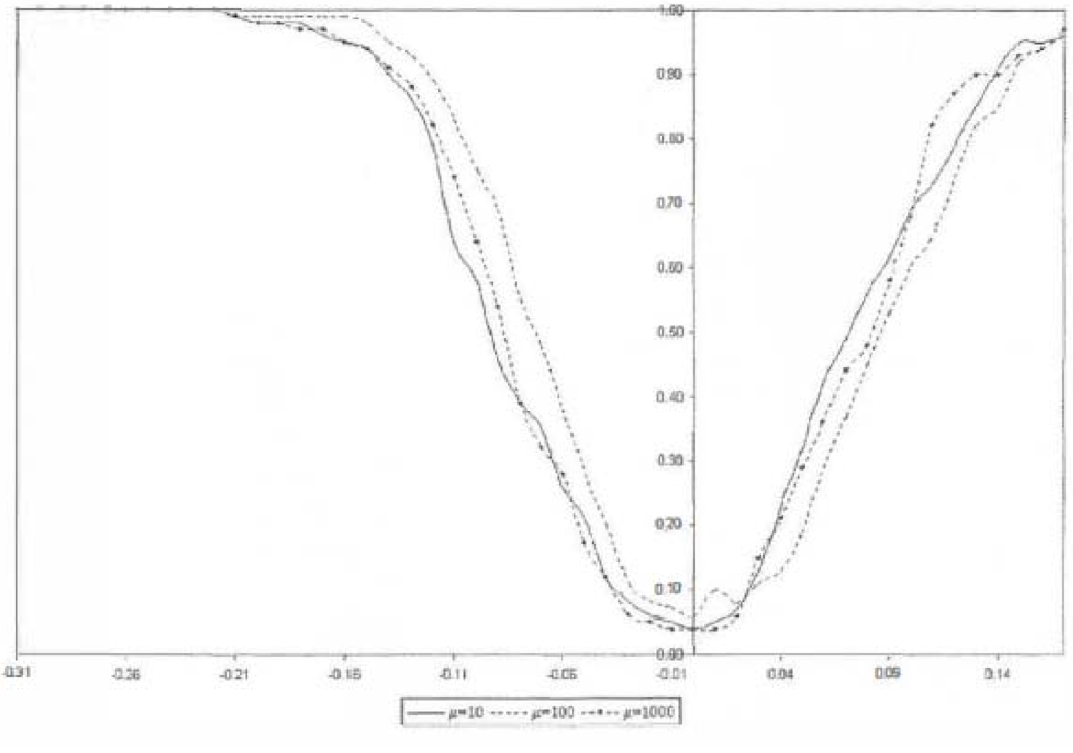
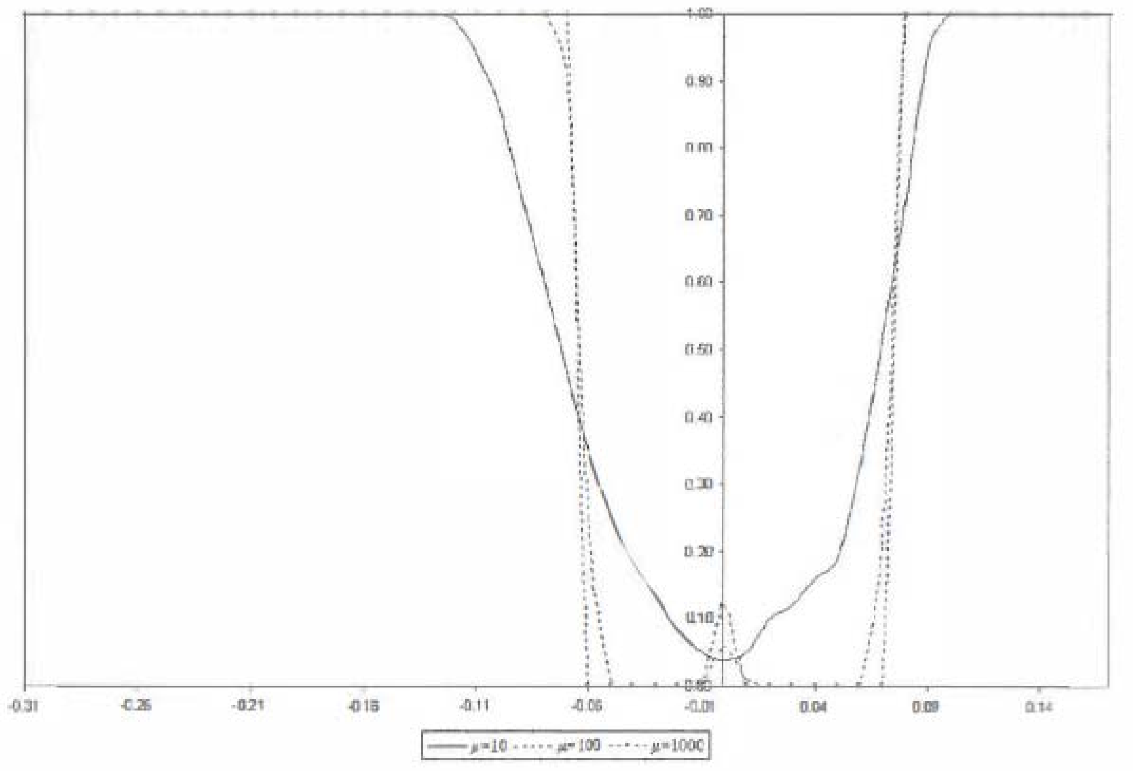


Figure 5.6: Moran's 1 test. Logarithmic Transformation. SAR case.



6. - Conclusions

Moran' s I is an efficient test for detecting relationships of cross-sectional dependence in spatial series. However, its behavior is sensitive to the scale of the process in series of SAR type. When the coefficient of variation of the series is high, the power of the test is zero for a not unimportant range of values of the coefficient of autocorrelation.

In this paper we have noted some solutions and discarded others. Among the latter the re-scaling of the series before resolving Moran's I stands out. In the list of proposals, the GQ and BP tests offer certain guarantees though the solution doesn't seem to be final. It is necessary to elaborate a more structured and consistent analysis framework where the impact of scale in SAR processes can be absorbed. It is also necessary to extend the cases analyzed in order to contemplate different types of scales and of samples sizes. Lastly, another aspect to consider is what happens, with regard to the same question of the factor of scale, with the set of LM tests used in the specification of the spatial dynamics in a causal econometric model.

Appendix: Heteroskedasticiy and Spatial Autocorrelation

In contrast to what occurs in a general context, a series with a structure of spatial autocorrelation will show features of cross-correlation dependencies and also heteroskedasticity. Nevertheless, until now, the analysis has been focussed on the first aspect (Moran's test checks if the covariance of the series is statiscally null), not using the information existing in the variance of the series. This may be due to the fact that it is difficult to isolate the skedastic functions although this is possible when we *filter* the series (King, 1983, suggests a similar approach although in a different context).

Matrix W comes from a decision of the user and its elements are known (ones and zeros) so that the spectral decomposition can be obtained: W= QΛQ' where Q is the matrix of eigenvectors and Λ that of eigenvalues both of order (RxR). If we use the matrix Q to filter the series y* = Q'y the result is that we eliminate the structure of spatial dependencies highlighting the skedastic dimension. This means in the SAR case of (A1):

$$y = (I-\delta W)^{-1}u \Rightarrow Q'y = Q'(I-\delta W)^{-1}u = (I-\delta\Lambda)^{-1}Q'u \Rightarrow y^{*} = \Delta^{-1}u^{*} \qquad \text{(A1)}$$

with $u^{*} \sim N(0,\sigma^{2}I)$. The covariance matrix of the y* series, even supposing spatial correlation, will be diagonal with characteristic elements: $\sigma_r^2 = (1-\delta\lambda_r)^{-2}$, with λr as the r-th eigenvalue of W. If the original series is a moving average the result is:

$$y = (I-\delta W)u \Rightarrow Q'y = Q'(I-\delta W)u = (I-\delta\Lambda)Q'u \Rightarrow y^{*} = \Delta u^{*} \qquad \text{(A2)}$$

The covariance matrix will also be diagonal with characteristic elements $\sigma_r^2 = (1-\delta\lambda_r)^2$. In both cases the skedastic variance is a regular function of eigenvalues of the matrix W, which is a consequence of the structure of transversal correlation (through the same matrix W) which exists in the original series. What we now propose is to use this information to develop a test of spatial autocorrelation to exploit the skedastic structure of the series.

This proposition is possible given that, under the null hypothesis of incorrelation, the variance of the series should not answer to the sequence of eigenvalues of W (this would be mere coincidence), whilst when there is an SAR or SMA structure the answer must be regular. The instrument of analysis could consist of a simple test of heteroskedasticity on the filtered y* series, and that of Goldfeld-Quandt (GQ) seems to offer the maximum guarantees. For this it is only necessary to order the values of the series in ascending order with the associated roots, eliminate the central m to increase its power and obtain the squared sum of the remaining subsamples. If by SR_{MAX} and SR_{MIN} we refer respectively to the highest and lowest value of both sums, the GQ contrast of spatial autocorrelation is simple:

$$\left.\begin{array}{l} H_0 : \delta = 0 \\ H_A : \delta \neq 0 \end{array}\right\} \Rightarrow GQ = \frac{SR_{MAX}}{SR_{MIN}} \sim F(k;k) \qquad (A3)$$

with k=(R-m)/2. If the null hypothesis is true, the expected value of both sums will be k and the GQ statistic will take a value close to one. If there is spatial correlation in the original series ($\delta \neq 0$), the degrees of freedom of each sum will differ so that:

$$SR_j = \sum_{r \in j} y_r^{*2} = \sum_{r \in j} \sigma_r^2 \left(\frac{y_r^*}{\sigma_r} \right)^2 \sim \chi^2(k_j) \qquad (A4)$$

with $k_j = \sum_{r \in j} \sigma_r^2$ and j = MAX, MIN. In this case, the statistic of (A3) will still be a centered F but with k_{MAX} and k_{MIN} degrees of freedom in the numerator and denominator respectively. If, before obtaining the GQ statistic we center the data of the two subsamples (in relation to the respective sample average), the distribution of (A3) will be an F(k-1;k-1) under the null. The degrees of freedom of the χ^2 of (A4) will be $k_j - \bar{k}_j$ with $\bar{k}_j = (\sum_{r \in j} \sigma_r^2)/k$ under the alternative hypothesis. In any case, given that the elements of the distribution function are known, we can obtain the power function of the test:

$$\Pr\left[GQ \geq q_\alpha / \delta \neq 0\right] \qquad (A5)$$

where α is the significance level and q□ their corresponding abscissa, obtained from the distribution F(k,k) of the null hypothesis. For example, if the original series is an SMA, the probability of (A5) will be:

$$\Pr\left[GQ \geq q_\alpha / \delta \neq 0\right] = \Pr\left[\frac{SR_{MAX}}{SR_{MIN}} \geq q_\alpha / \delta \neq 0\right] = \Pr\left[\frac{(SR_{MAX})/k_{MAX}}{(SR_{MIN})/k_{MIN}} \geq \left(\frac{k_{MIN}}{k_{MAX}}\right) q_\alpha / \delta \neq 0\right] =$$

$$\Pr\left[F(k_{MAX}, k_{MIN}) \geq q_\alpha^* / \delta \neq 0\right] \qquad (A6)$$

In (A6), the value of q_α is known once the significance level is fixed and k_j (j=MAX or MIN) depends on δ as is indicated in (A4), which permits the resolution of (A6) and the obtaining of the power function for the GQ statistic of spatial autocorrelation.

References

Anselin, L. (1988): *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Anselin, L. and S. Rey (1991): Properties of Tests of Spatial Dependence in Linear Regression Models. *Geographical Analysis*, 23, 112-131.

Anselin, L., and R. Florax (1995b). Small sample properties of tests for spatial dependence in regression models: some further results. In L. Anselin and R. Florax (eds.) *New Directions in Spatial Econometrics*, pp. 21–74. Berlin: Springer-Verlag.

Arbia, G. (1989): *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problem*s. Dordrecht: Kluwer.

Breusch, T. and A. Pagan (1979): A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47, 1287-1294.

Burridge, P. (1980): On the Cliff-Ord Test for Spatial Correlation. *Journal of the Royal Statistical Society B*, 42, 107-108.

Cliff, A. and K. Ord (1972): Testing for Spatial Autocorrelation Among Regression Residuals. *Geographical Analysis*, 7, 267-284.

Cliff, A. and K. Ord (1981): *Spatial Processes. Models and Applications*. London: Pion.

Dacey, M. (1965): A Review of Measures of Contiguity for Two and k-colour Maps, in *Spatial Analyses: A Reader in Statistical Geography* (pp. 479-495). Ed. B. Berry and D. Marble. Englewood Cliff: Prentice Hall.

De Jong, P. C. Springer and F. van Veen (1984): On Extreme Values of Moran's I and Geary's c. *Geographical Analysis*, 16, 17-24.

Elhorst J. (2010): Spatial Panel Data Models. In: Fischer M., Getis A. (eds) Handbook of Applied Spatial Analysis. Springer: Berlin,

Geary, R. C. (1954): The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5, 115–145.

Godfrey, L. (1988): *Misspecification Tests in Econometrics. The Lagrange Multiplier Principle and other Approaches*. Cambridge: Cambridge University Press.

Goldfeld, S. and R. Quandt (1964): Some Tests for Homoscedasticity. *Journal of the American Statistical Association*, 60, 539-547.

Harvey, A. and G. Phillips (1974): A Comparison of the Power of Some Tests for Heteroskedasticity in the General Linear Model. *Journal of Econometrics*, 2: 307-316.

Kelejian, H., and D. Robinson (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* 72, 297–312.

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

King, M. (1981): A Small-Sample Property of the Cliff-Ord Test for Spatial Correlation. .*Journal of the Royal Statistical Society B*, 43,263-264.

King, M. 1983. Testing for Autoregressive Against Moving Average Errors in the Linear Regression Model. *Journal of Econometrics*, 21: 35-51.

López, F., M. Matilla-García, J, Mur and M. Ruiz (2011): Four tests of independence in spatiotemporal data. Papers in Regional Science, 90: 663-685.

López, F., R. Mínguez, and J. Mur (2019): ML versus IV estimates of spatial SUR models: evidence from the case of Airbnb in Madrid urban area. The Annals of Regional Science, https://doi.org/10.1007/s00168-019-00914-1

Moran, P. (1950): Notes on Continuous Stochastic Phenomena. *Biometrika*, 37, 178-181.

Sen, A. (1976): Large Sample Size Distribution of Statistics Used in Testing for Spatial Correlation. *Geographical Analysis*, 9, 175-184.

Sen, A. (1990): Distribution of Spatial Correlation Statistics, in *Proceeding from the Symposium Spatial Statistics: Past, Present and Future* (pp. 257-272). Ed. D. Griffith. Ann Arbor, Institute of Mathematical.

Szroeter, J. (1978): A Class of Parametric Tests for Heteroskedasticity m Linear Econometric Models. *Econometrica*, 46, 1311-1328.

Tiefelsdorf, M. (2000): *Modelling Spatial Processes. The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I.* Berlin: Springer-Verlag

Tiefelsdorf, M. and B. Boots (1995): The Exact Distribution of Moran's I. *Environment and Planning A*, 27, 985-999.

Yule, U. and M. Kendall (1950): *An Introduction to the Theory of Statistics. London*: Charles Griffin.

White, H. (1980): A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, 48, 817-83 8.

White, H. (1982): Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50, 1-25.