A comparison of spatial outliers using different algorithms for constructing of weights system

Massimo Mucciardi,
Dip. D.E.S.Ma.S, University of Messina
massimo.mucciardi@unime.it

Pietro Bertuccelli,
Dip. S.E.F.I.S.A.T., University of Messina
pbertuccelli@unime.it

Abstract: *The aim of this work is to highlight the effects of the application of different algorithms for determinates spatial interactions and local Moran index to an economic variable regarding the value added per inhabitant.*

Keywords: Local Moran, Spatial outliers, Weight matrix

1. Introduction

Statistical data analysis generally studies phenomena by examining the chance of the existence of relations between variables and/or their temporal evolution. However, in the last couple of years there has been the need of considering phenomena, especially in the economic field, from the spatial point of view. This fact implies taking into account possible spatial effects due to a particular territorial configuration, which condition in some way the values kept by variables. For this reason, the aspect concerning spatial autocorrelation has great relevance, because eventual interactions between territorial units can heavily influence the phenomenon we are studying hence leading the researcher to draw conclusions which could be potentially not true. The so-called local statistics, among which Getis and Ord's G_i (Getis et Ord, 1995) and local Moran (Anselin, 1995) stand out, and the Moran scatterplot (Anselin, 1996) appear to be of specific interest, because they allow to highlight in a more precise way how each datum relates to the data observed in the near territorial units. Although Moran scatterplot is a very good tool to visually find spatial clustering phenomena and outliers, it doesn't offer any analytical tool to strictly identify them. In order to achieve the above mentioned purpose, in our work we'll use local Moran to perform an analysis on spatial data. As a matter of fact Anselin (1995) gives two interpretations in defining local indicators of spatial association. On the one hand he argues that they can be used as indicators of local spatial clusters; on the other that they can put second, into evidence local instability or spatial outliers. As it is known, the local Moran index is given by:

$$I_i = z_i \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} z_j \ , \qquad (1)$$

where z_i are standardized and weights w_{ij} are row standardized. If statistically meaningful, a positive value of I_i means a clustering phenomenon of similar values, whereas a negative value means a clustering of dissimilar values. In this contest, the choice of which weight matrix W to apply in a study becomes crucial because different spatial effects often correspond to different spatial patterns. For this reason we'll apply local Moran using three weight matrix generation methods: "classical spatial adjacency" (Cliff et Ord, 1973), "MaxMin" (Mucciardi, 1998) and "k-

nearest neighbors" (Fix et Hodges, 1951); the aim is to highlight differences or similarities existing between these methods on the effects of their application to some real economic variable.

2. Adjacency, k-nn and MaxMin: a comparison

The W matrices we'll apply are of binary type (1-0), in other words the generic element will be 1 when some kind of spatial relationship exists between two generic units and 0 when it does not exist any relationship. Matrices will be then row standardized so that the total sum of weights by row is equal to 1. The three proposed methods considerably differ about the way they generate the weight matrix, and for this reason we'll shortly describe the characteristics of each of them. The first method, known as spatial adjacency (traditional criteria) , essentially consists of considering adjacent two spatial units, obviously of polygonal type, which have in common one or more points of its surface. This method, maybe the most popular for its simplicity, uses the same terminology of the game of chess in order to define the model of spatial adjacency. In facts, since territorial subdivisions initially used into researches were regular quad grids, the adjacency was defined in terms of queen, rook and bishop case, according to whether two spatial units share almost one point (queen), a side (rook), one and only one point (bishop). In this paper, we'll consider the queen case, although here the terminology is some way unsuitable because the territorial partition we'll use to determine the weight matrix, composed by 103 Italian provinces, is highly irregular. The adjacency method generates, without applying any generalized weight system, symmetrical matrices. If we indicate with s_i the generic territorial unit belonging to territorial partition S and with R a relation which establishes a boundary condition between two units, the W matrix is given by:

$$\begin{cases} w_{ij} = 1 & \text{if} \quad s_i \; R \; s_j \\ w_{ij} = 0 & \text{if} \quad s_i \; \overline{R} \; s_j \end{cases} \qquad \forall\, i \neq j \tag{2}$$

Among the disadvantages of this method there is the fact that it never considers, during the matrix creation process, spatial units physically non linked to the others (isles). In the above mentioned situations/context this may cause potentially unbiased estimations of the autocorrelation indices. As an example, if we consider a territorial partition composed by Italian regions. According to this method Sicily and Sardinia are never linked to other regions; this is a matter we must keep into account if we are to draw inferential conclusions. Centroid based procedures are used to avoid these problems, because they permit to consider the map as if it was composed only by points. K-nn and MaxMin are part of these methods. K-nn, with exception of the others two used methods, generates, unless it is applied to a particular territorial configuration, asymmetric matrices. Such characteristic is due to the fact that for each spatial unit the k nearest units are considered linked, and so a biunique relation between two units doesn't necessarily exist (let u_i and u_j be two spatial units, if u_j is among the k nearest units to u_i, it isn't necessary true that u_i belongs to the set of the k nearest units to u_j). Another relevant characteristic of this procedure is that the number of neighbors for each unit is constant, whereas in the other methods we are considering the neighboring elements have variable numerousness. Formally the matrix is given by:

$$\begin{cases} w_{ij} = 1 & \text{if} \quad d_{ij} \leq d_{ik} \\ w_{ij} = 0 & \text{if} \quad d_{ij} > d_{ik} \end{cases} \quad \forall\, i \neq j \quad \text{with} \quad k = 1,..,K \tag{3}$$

where d_k stands for the k-th the smallest distance between two spatial units i and j so that the number of neighbors is K. One of the troubles of this method is represented by the subjective choice of the number of neighbors. Thus it is a researcher's duty to evaluate the solution to take according to the territorial partition taken into consideration. MaxMin procedure, opposed to k-nn,

generates interconnection rays h^k without causing any constriction on the number of neighbors and so arbitrariness in the choice of them. In this case distance h^k is chosen in such way to satisfy the relation:

$$h^k = \max\left(\mathbf{e}^k{}_1, \mathbf{e}^k{}_2, ...\mathbf{e}^k{}_j .....\mathbf{e}^k{}_n\right), \quad (4)$$

where $\mathbf{e}^k{}_j$ represents minimum distance between unit j and the other units (for details, see Mucciardi, 1996).

An application on value added per capita

For our research we used as study variable (expressed in standardized form) the value added per capita[1] in the year 2005 concerning 103 italian provinces (Istat, 2008). The comparison between the three procedures early mentioned, is carried out by the calculation of the classic Moran I in order to perform local analysis, of the local Moran and of the Moran scatterplot, due to their characteristic to bring out spatial outliers or spatial clusters. The calculations are performed using S-Joint software (Mucciardi-Bertuccelli, 2007). From the results of the analysis, we can observe that the adjacency and the MaxMin method are substantially equivalent from the point of view of the global autocorrelation. The values of the Moran I differ very little: 0.7878 versus 0.8077 (Tab 1). The k-nn produces values of the index I which are very similar to those previously obtained until we consider a number of neighbours for each unit comparable to that generated by the other two methods (obviously this depends on the territorial partition). On the contrary, the application of local Moran and of the Moran scatterplot, seems to highlight some differences between the two methods used. Specifically, adjacency and MaxMin detect the same outlier (Massa Carrara) with highest residue between the theoretical lagged standardized variable and the observed lagged standardized variable. The k-nn instead, generates results which change according to the number of neighbors. As we can observe from the table 1, this method identifies a different spatial outlier compared to the other two procedures (Verb-Cusio -Ossola), considering only one neighbor, but if we consider a number of higher neighbors (four spatial units), then we can observe not only that the territorial unit with higher residual value is the same for all the three methods, but also that adjacency and k-nn share four out of five spatial units in the first positions, although their ranks are different. This can mean that when applied to this map the k-nn with four neighbors determines a spatial configuration very similar to the other two procedures. Such interpretation seems to be strengthened by the analysis of the lowest residual values, where the outliers identified are the same (Roma, Milano) independently of the matrix generation method used. Moreover, as we have pointed out for the highest residual values, there is a very high similarity between adjacency and k-nn with four neighbors, having in this case the same five spatial units in the same rank position (the explanation is rather simple because we obtain very near values of the Moran I and evidently the territorial partition is such that the two methods generate matrices with many units in the same positions). The value of the local Moran, can help us to understand in a better way the relations existing between near spatial units. An interesting result emerging from the analysis, which seem to confirm the goodness of spatial autocorrelation indices as tools to explain possible relations between values of a certain variable, is that concerning the two outliers referring to Roma and Milano. In facts we can observe a negative value of the local index for the first outlier, which means that the Roma's province is negatively correlated with its neighbors. This result puts into evidence the fact the great "economic influence" of the city seems to depress near provinces, which lack of a developed productive structure. Completely different is the case of Milano's province, where its industrialized neighbors seem to positively amplify the effect on it.

[1] Estimates

Moran I	Rank of residue		Province	**Residue	Zi	*WZI Theor.	WZI Obs.	Pos.***	Local Moran
Adjacency	Highest	1°	Massa Carrara	1.1228	-0.3948	-0.3110	0.8118	L-H	-0.3205
		2°	Verb-Cus-Ossola	1.0001	-0.4177	-0.3291	0.6710	L-H	-0.2803
0,7878		3°	Pistoia	0.9138	0.2658	0.2094	1.1232	H-H	0.2985
		4°	Ferrara	0.9038	0.2974	0.2343	1.1381	H-H	0.3385
		5°	Como	0.8843	0.3484	0.2745	1.1588	H-H	0.4038
	Lowest	1°	Roma	-1.8090	1.7613	1.3877	-0.4214	H_L	-0.7422
		2°	Milano	-1.4376	2.5837	2.0355	0.5979	H_H	1.5449
		3°	Cagliari	-0.7082	-0.6032	-0.4753	-1.1835	L_L	0.7139
		4°	Bologna	-0.6475	1.8358	1.4464	0.7988	H_H	1.4666
		5°	Catanzaro	-0.6082	-1.0107	-0.7963	-1.4045	L_L	1.4196
MaxMin	Highest	1°	Massa Carrara	1.0796	-0.3948	-0.3188	0.7607	L-H	-0.3003
		2°	Latina	0.8925	-0.1662	-0.1342	0.7582	L-H	-0.1260
0,8077		3°	Ferrara	0.8133	0.2974	0.2402	1.0535	H-H	0.3133
		4°	Imperia	0.8104	-0.2148	-0.1735	0.6369	L-H	-0.1368
		5°	Lodi	0.7996	0.3166	0.2557	1.0553	H-H	0.3341
	Lowest	1°	Roma	-1.8132	1.7613	1.4226	-0.3906	H_L	-0.6880
		2°	Milano	-1.4851	2.5837	2.0868	0.6017	H_H	1.5546
		3°	Palermo	-0.7836	-1.2339	-0.9966	-1.7802	L_L	2.1966
		4°	Cagliari	-0.7718	-0.6032	-0.4872	-1.2590	L_L	0.7595
		5°	Bologna	-0.6752	1.8358	1.4828	0.8076	H_H	1.4827
Knn1	Highest	1°	Verb-Cus-Ossola	1.0567	-0.4177	-0.3324	0.7243	L-H	-0.3026
		2°	Asti	0.9694	-0.0654	-0.0521	0.9173	L-H	-0.0600
0,7956		3°	Gorizia	0.8445	0.4291	0.3414	1.1860	H-H	0.5090
		4°	Pesaro e Urbino	0.7646	0.0491	0.0390	0.8037	H-H	0.0394
		5°	Piacenza	0.7569	0.6920	0.5505	1.3074	H-H	0.9047
	Lowest	1°	Roma	-1.9830	1.7613	1.4013	-0.5817	H_L	-1.0245
		2°	Milano	-1.7649	2.5837	2.0556	0.2907	H_H	0.7511
		3°	Aosta	-0.8578	0.9959	0.7924	-0.0654	H_L	-0.0652
		4°	Palermo	-0.7985	-1.2339	-0.9817	-1.7802	L_L	2.1966
		5°	Alessandria	-0.7953	0.9173	0.7298	-0.0654	H_L	-0.0600
Knn4	Highest	1°	Massa Carrara	1.1219	-0.3948	-0.3101	0.8118	L-H	-0.3205
		2°	Como	1.0201	0.3484	0.2738	1.2938	H-H	0.4508
0,7857		3°	Pavia	0.8990	0.2907	0.2284	1.1274	H-H	0.3278
		4°	Ferrara	0.8198	0.2974	0.2337	1.0535	H-H	0.3133
		5°	Verb-Cus-Ossola	0.8151	-0.4177	-0.3282	0.4869	L-H	-0.2034
	Lowest	1°	Roma	-1.7658	1.7613	1.3838	-0.3820	H_L	-0.6728
		2°	Milano	-1.5102	2.5837	2.0299	0.5197	H_H	1.3427
		3°	Cagliari	-0.6745	-0.6032	-0.4740	-1.1484	L_L	0.6928
		4°	Bologna	-0.6316	1.8358	1.4424	0.8107	H_H	1.4884
		5°	Catanzaro	-0.6104	-1.0107	-0.7941	-1.4045	L_L	1.4196

*WZI Theor = I*Zi where "I" is global Moran's index
**Residue = (WZ_I_Theor- WZ_I_Obs)
***Pos = position of the spatial unit in Moran's scatterplot
H-H= High-High; H-L=High-Low; L-H=Low-High;: L-L=Low-Low

Table 1: comparison of Adjacency, K-nn and MaxMin output

Bibliography

Anselin L. (1995), Local indicators of spatial association-LISA. Geographical Analysis 27: 93–115.

Anselin L. (1996), The Moran scatterplot as an ESDA tool to assess local instability in spatial association in: Fisher M, Scholten HJ, Unwin D (eds) Spatial analytical perspectives on GIS, Taylor&Francis, London.

Cliff J., Ord K. (1973), Spatial Autocorrelation, Pion Limited, London.

Fix E., Hodges J.L. (1951), Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Istat, (2008), Occupazione e valore aggiunto nelle province. Statistiche in breve, on-line document

Mucciardi M. Bertuccelli P. (2007) S-Joint: a new software for the analysis of spatial data, in: Atti della Riunione Scientifica Intermedia della Società Italiana di Statistica, Venezia.

Mucciardi M. (1998), La procedura D.S.M.A. per la misura dell'intensità dei legami tra unità spaziali di tipo puntuale, *SIS XXXIX Riunione Scientifica, Sorrento*, 773-779.

Ord J.K., Getis A. (1995), Local spatial autocorrelation statistics: distributional issues and an application, Geographical Analysis 27: 286–305.

