Spatial statistics in the analysis of county budget incomes in Poland with the R CRAN

Alicja Wolny-Dominiak[1], Katarzyna Zeug-Żebro[2]

Abstract. Since Waldo Tobler [13] formulated the first law of geography, which says that everything is related, but near objects are more related than distant ones, spatial modeling has become an important research area. The methods which were developed proved to be excellent tools which can also be used in regional analysis. The most common are measures of spatial autocorrelations, which show the dependence of variables in respect of spatial localization. Spatial correlation allows to determine that intensification of a given phenomenon is more perceivable in neighboring units than in units distant from each other. The main objective of this paper is to present spatial dependences analysis using measures of global and local spatial autocorrelation with a free software environment R CRAN. The analysis is carried out using the real data set of budget incomes of counties in Poland.

Keywords: spatial autocorrelation, global and local statistics, R CRAN

JEL Classification: C44
AMS Classification: 65C60

1 Introduction

Methods of spatial statistics are used to identify spatial patterns and spatial dependency. Testing occurrence of spatial dependency boils down to verify the hypothesis of the existence of spatial autocorrelation in the data spatially localized. The evaluation of spatial autocorrelation requires the knowledge of the extent and specificity of spatial diversity, i.e. diversity of characteristics of individual sites and geographic regions.

Until recently, the rare use of spatial autocorrelation measures in practice resulted from complex and time-consuming calculation procedures. For some time, however, there has been a rapid development in computer software that allows to carry out research (often very complex) in the field of spatial statistics and econometrics. One of such programs is the R CRAN, within which packages `{spdep}` [4] and `{maptools}`, used to analyze regional and spatial data dependences, are developed. R CRAN can successfully replace the familiar, expensive software because it is multifunctional and available for free.

The objective of this paper is to study spatial dependency using of global and local spatial autocorrelation measures. All calculations and maps were made in the statistical program R CRAN based on the data relating to the budget incomes of counties in Poland in 2010. The data was obtained from the Local Data Bank of the Central Statistical Office (www.stat.gov.pl).

2 Spatial statistics

There are two types of indicators of spatial associations (ISA): global and local measures of autocorrelation. The global autocorrelation follows from the existence of correlations across the spatial unit test. The local measure shows a spatial dependency the variable with neighboring units in a particular location. The most commonly used global and local measures are: the Moran statistics I [11] and the Geary statistics C [6], [1]. The spatial autocorrelation occurs when a certain phenomenon in a single spatial unit alters the probability of occurrence of this phenomenon in the neighboring units [3]. In general, the positive spatial autocorrelation occurs when we observe the accumulation, in terms of the location, high or low values of observed variables. In the case of negative autocorrelation, high values adjacent to low, and low to high, creating a kind of checkerboard [12]. The lack of spatial autocorrelation means the spatial randomness, i.e. the high and low values of observed variables are distributed independently.

[1] University of Economics in Katowice, Department of Statistical and Mathematical Methods in Economics, alicja.wolny-dominiak@ue.katowice.pl.
[2] University of Economics in Katowice, Department of Mathematics, katarzyna.zeug-zebro@ue.katowice.pl.

2.1 Selected global statistics

The Moran statistics is one of the most widely used measures in the study of spatial autocorrelation. The Global Moran's I is defined as follows:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \cdot \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{n}{S_0} \cdot \frac{z^{\mathrm{T}} W z}{z^{\mathrm{T}} z} \tag{1}$$

where: x_i, x_j are the values of variables in spatial unit *i* and *j*, $\bar{x}$ is the mean of variable for all units, *n* is the total number of spatial units that are included in the study, S_0 is the sum of all elements of a spatial weight matrix, *z* is a column vector of elements $z_i = x_i - \bar{x}$, *W* is the spatial weight matrix degree *n*, defining the structure of the neighborhood, w_{ij} is an element of weights matrix *W* [10]. This statistic takes values ranging from $[-1,1]$: positive, when tested objects are similar, negative, when there is no similarity between them and approximately equal to 0 for a random distribution of objects.

Cliff and Ord [5] have shown that the distribution of Moran statistics is asymptotically normal. Thus, the statistical significance of spatial autocorrelation can be verified using normalized statistics : $I_S \sim N(0,1)$:

$$I^S = \frac{I - \mathrm{E}(I)}{\sqrt{\mathrm{Var}(I)}} \tag{2}$$

where: $\mathrm{E}(I)$ is the expected value of Moran's and $\mathrm{Var}(I)$ is its variance:

$$\mathrm{E}(I) = -\frac{1}{n-1},\ \mathrm{Var}(I) = \frac{n^2 S_1 - n S_2 + 3S_0^2}{\left(n^2-1\right)S_0^2} - \frac{1}{(n-1)^2} \tag{3}$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ ,\ S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2\ ,\ S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2 \tag{4}$$

If the Moran statistic has a value $I \approx -(n-1)^{-1}$, $I^S \approx 0$ it indicates a random spatial pattern. However, when $I > -(n-1)^{-1}$, $I^S > 0$ the spatial autocorrelations is positive, and if $I < -(n-1)^{-1}$, $I^S < 0$, the spatial autocorrelations is negative.

Another global measure of spatial autocorrelation is Global Geary's *C*. This statistic, is given by

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - x_j\right)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \cdot \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{n}{(n-1)}\left[\frac{n}{S_0} \cdot \frac{z^{\mathrm{T}}\mathrm{diag}(w_{i.})z}{z^{\mathrm{T}} z} - I\right] \tag{5}$$

where all elements of the formula are defined as in statistic *I*. The above formula shows that the Geary measure can be expressed by the Moran statistic [8]. Although Moran and Geary measures give similar results, the Moran statistic is more effective. This is due to greater sensitivity of the variance of the Geary statistic to the distribution of sample. Values of this statistic can be impaired when the matrix of weights is asymmetrical. In order to verify the hypothesis of no spatial correlation, the Geary statistic can be standardized:

$$C^S = \frac{C - \mathrm{E}(C)}{\sqrt{\mathrm{Var}(C)}} \sim N(0,1) \tag{6}$$

where: $\mathrm{E}(C)$ is the expected value of Geary's and $\mathrm{Var}(C)$ is its variance:

$$\mathrm{E}(C) = 1,\ \mathrm{Var}(C) = \frac{(n-1)(2S_1 + S_2) - 4S_0^2}{2(n+1)S_0^2} \tag{7}$$

The value of Global Geary's *C* is always positive and takes values ranging from $[0, 2]$. In the case, of: $1 < C < 2$, $C^S > 0$, the spatial autocorrelation is negative; when $0 < C < 1$, $C^S < 0$, the spatial autocorrelation is positive; finally, when $C \approx 1$, $C^S \approx 0$, there is no spatial autocorrelation.

2.2 Selected local statistics

We can use local indicators of spatial association (LISA), a Local Moran statistics and a Local Geary statistics, to identify spatial systems. The Local Moran determines clusters of spatial units and studies whether the unit is surrounded by neighboring units with similar or different values of the variable studied in relation to the random distribution of these values in the studied space [10].

In the case of non-standardized values of the variable and row-standardized spatial weight matrix [2] ($\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = n$), the local Moran is given by:

$$I_i = \left[(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij} (x_j - \bar{x}) \right] \Bigg/ \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \right] \tag{8}$$

where all elements of the formula are defined as in the Global Moran's *I*. The standardized Local Moran's I_i^S is used to test the statistical significance of local spatial autocorrelation [1]:

$$I_i^S = \frac{I_i - \mathrm{E}(I_i)}{\sqrt{\mathrm{Var}(I_i)}} \sim N(0,1) \tag{9}$$

where: $\mathrm{E}(I_i)$ is the expected value of the Local Moran and $\mathrm{Var}(I_i)$ is its variance

$$\mathrm{E}(I_i) = -\frac{\sum_{j=1}^{n} w_{ij}}{n-1} \text{ i } \mathrm{Var}(I_i) = \frac{(n-k)\sum_{i\neq j} w_{ij}^2}{n-1} + \frac{2(2k-n)\sum_{l\neq i}\sum_{hi\neq i} w_{il} w_{ih}}{(n-1)(n-2)} - \left(\frac{-\sum_{i\neq j} w_{ij}}{n-1} \right)^2 \tag{10}$$

where $k = \left(\frac{1}{n}\sum_i (x_i - \bar{x})^4 \right) \Big/ \left(\frac{1}{n}\sum_i (x_i - \bar{x})^2 \right)^2$.

When I_i^S is negative, the spatial autocorrelation is negative too, i.e. when the object is surrounded by spatial units with significantly different values of the studied variable. The spatial autocorrelation is positive when $I_i^S > 0$, the object is surrounded by similar neighboring units.

According to Anselin [1] a Local Geary statistics for an observation *i* may be defined as

$$C_i = \sum_{j\neq i}^{n} w_{ij} (z_i - z_j)^2 \tag{11}$$

where $z_i = x_i - \bar{x}$, $z_j = x_j - \bar{x}$ and w_{ij} are the elements of the row-standardized binary symmetric spatial weight matrix **W**. The test statistic for C_i^S is

$$C_i^S = \frac{C_i - \mathrm{E}(C_i)}{\sqrt{\mathrm{Var}(C_i)}} \sim N(0,1) \tag{12}$$

where: $\mathrm{E}(I_i)$ is the expected value of the Local Moran and $\mathrm{Var}(I_i)$ is its variance

$$\mathrm{E}(C_i) = \frac{n\sum_{j=1}^{n} w_{ij} \cdot \sum_{j=1}^{n} (z_i - z_j)^2}{(n-1)^2} \text{ i } \mathrm{Var}(C_i) = \frac{\left[(n-1)\sum_{i=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij} \right)^2 \right] \cdot \left[(n-1)\sum_{j=1}^{n} (z_i - z_j)^4 - \left[\sum_{j=1}^{n} (z_i - z_j)^2 \right]^2 \right]}{(n-1)^2 (n-2)} \tag{13}$$

the significant testing on local spatial association can be conducted based on the calculated test statistics above. The C_i statistic is interpreted in the same way as the Local Moran.

3 ISA for incomes of counties in Poland with R

In the empirical example, we analyzed global and local indicators of spatial associations (ISA) for incomes of counties in Poland in 2010 year. For all computations and maps we used the free software environment R. We started with the calculation of the spatial weight matrix for 376 counties in Poland which measures spatial links between objects. This matrix is necessary to analyze the neighborhood. Based on the weight matrix we computed the neighborhood matrix according to adjacency criteria. The neighborhood map of counties is as follows:

number of links	number of counties	% of counties
1	31	8.24%
2	12	3.19%
3	21	5.59%
4	41	10.90%
5	90	23.94%
6	77	20.48%
7	57	15.16%
8	29	7.71%
9	11	2.93%
10	5	1.33%
11	2	0.53%

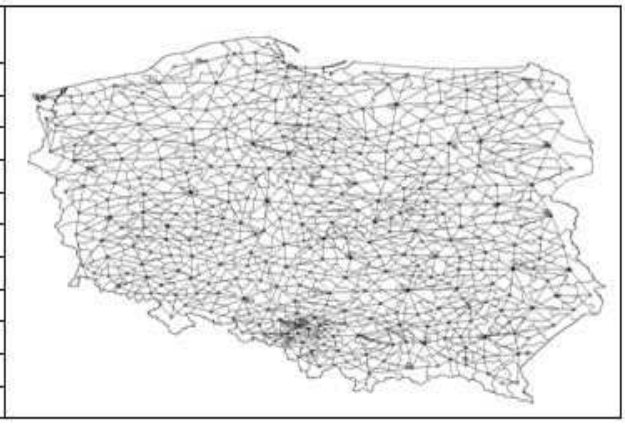

Figure 1 Distribution of the number of links for the counties and the neighborhood matrix for Poland

The number of nonzero links in all counties is equal to 1996 and the average number of links for one county is 5.31. As we can see there are 31 least connected counties with 1 link and 2 most connected counties with 11 links. Over 23% of counties have five neighbors.

R code:
```
> map<-readShapePoly("C:/DANE/POL/POL_adm2.shp")
> map.nb<- poly2nb(as(map,"SpatialPolygons"))
> map.listw<-nb2listw(map.nb, style="W")
> coord=coordinates(map)
> plot(map.nb,coord, add=TRUE)
```

Then we calculated Moran's I global statistics using the test under randomization and $I = 0.0615$ with expectation $E(I) = -0.0027$, variance $Var(I) = 0.0005$. The small p-value at 0.0025 shows significance of the statistics. The value of Moran I is close to zero, which indicates no spatial autocorrelation. This means that there is no similarity between neighboring counties in terms of incomes.

R code:
```
> moran<-moran.test(data$Income, map.listw)
> moran.plot((data$Income-mean(data$Income))/sd(data$Income),map.listw, xlab="Income budget of counties in Poland ", ylab="Spatial lags for Income")
```

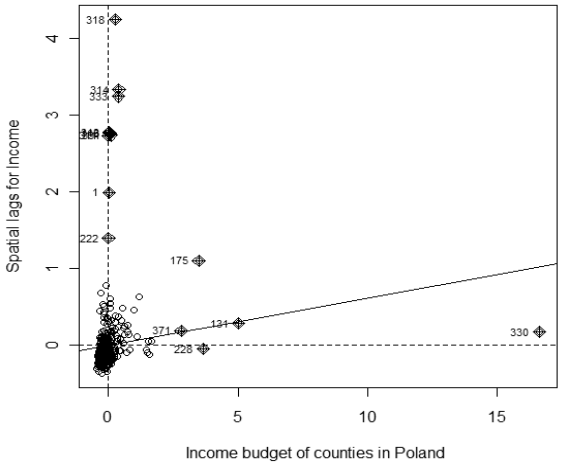

Figure 2 Scatter plot for the Moran global statistic

We also computed global Geary's statistics using the test under randomization. The results are similar to Global Moran statistics except for a p-value, which is equal to 0.5062. The value $C = 1.0031$ is near one which indicates no spatial autocorrelation, but p-value proves that Global Geary's statistics is insignificant.

In order to analyze the spatial autocorrelation in every county we calculated Local Moran statistics and Local Geary statistics. First of all, we tested significance of both statistics. The results are shown on the following maps:

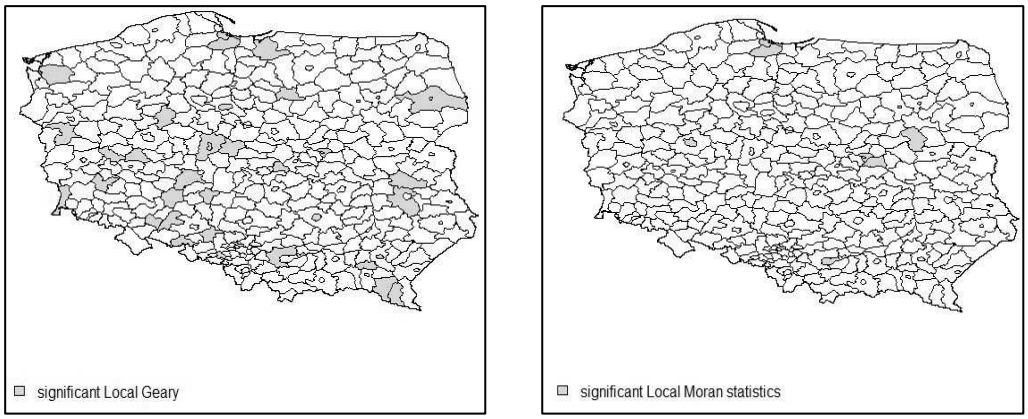

Figure 3 Counties with significant Local Geary and Local Moran statistics

The Local Moran is significant only for 8 counties: “Powiat m. Krakow”, “Powiat m. Poznań”, Powiat piaseczynski”, “Powiat pruszkowski”, “Powiat m. Warszawa”, „Powiat wołomiński”, „Powiat m. Gdańsk”, „Powiat m. Gdynia”. For all these counties Local Moran is significantly positive with the p-value below 0.05 which means that those counties are surrounded by objects with similar value of incomes, but we can-not say which counties are rich or poor. The Local Geary is significant for a greater number of counties against Local Moran. The interpretation is similar.

R code:

```
> moran.local<-localmoran(data$Income, map.listw)
> sig<-ifelse(moran.local[,5]<=0.05,"*"," insig ")
> break=c(0.0000000000000001, 0.05, 0.95, 0.999999999999999999)
> colors=cm.colors(1:3, alpha=1)
> moran.local.df=as.data.frame(moran.local)
> plot(map,col=colors[findInterval(moran.local.df[,5],break)])
> legend("bottomleft", legend=c("significant Local Moran"), fill=colors, bty="n")
> Gi.local<-localG(data$Income, map.listw)
> Gi.local.df<-as.data.frame(as.vector(Gi.local))
> sig<-ifelse(as.data.frame(as.vector(Gi.local))<=-3.083|as.data.frame(as.vector(Gi.local))<=3.083,"*"," insig ")
> colors=cm.colors(1:3, alpha=1)
> plot(map,col=colors[findInterval(Gi.local.df[,5],break)])
> legend("bottomleft", legend=c("significant Local Geary"), fill=colors, bty="n")
```

Based on Local Moran, we identified spatial regimes which show counties and neighbors with high and low values of incomes.

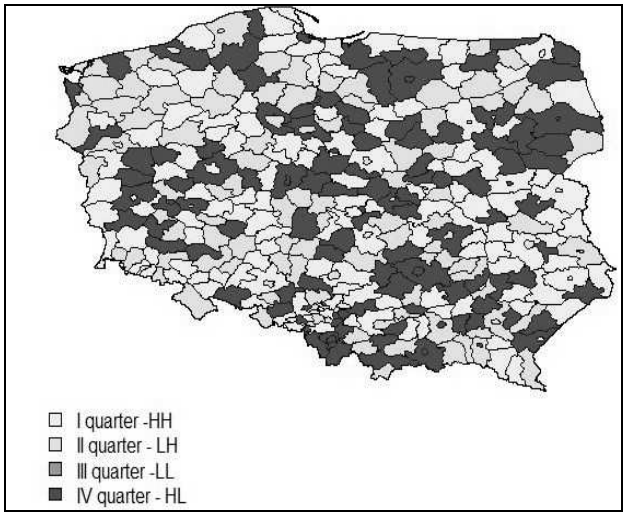

Figure 4 Spatial regimes

The classification of spatial regimes is as follows:

(i) I quarter HH – Cluster " I have high budget incomes and so do my neighbors"
(ii) II quarter LH – Outlier " I have low budget incomes among neighbors with high incomes"
(iii) III quarter LL – Cluster " I have low budget incomes and so do my neighbors"
(iv) IV quarter HL – Outlier " I have high budget incomes among neighbors with low incomes"

R code

```
> Z<-(dane$X15-mean(dane$X15))/sd(dane$X15)
> lag.Z<-lag.listw(map.listw, Z)
> q1<-ifelse(Z>0&lag.Z>0,1,0)
> q2<-ifelse(Z>0&lag.Z<0,2,0)
> q3<-ifelse(Z<0&lag.Z>0,3,0)
> q4<-ifelse(Z<0&lag.Z>0,4,0)
> q<-q1+q2+q3+q4
> q.data<-as.data.frame(q)
> break=c(1,2,3,4)
> colors=rev(heat.colors(4))
> plot(map,col=colors[findInterval(q.data$q,break)], forcefill=FALSE)
> legend("bottomleft", legend=c("I quarter -HH", "II quarter - LH", "III quarter -LL", "IV quarter - HL"),
fill=colors, bty="n")
```

4 Conclusions

Spatial methods are used increasingly frequently in the analysis of economic processes. One of the reason is the fact that spatial autocorrelation local and global measures, informing about the type and strength of spatial dependency, allow on: fuller use of the measure; to determine the relationship between reference entities; to define spatial structures [9]. Additionally, there are rapid developments in software that offers computational procedures in the field of spatial statistics and econometrics. Their effects can be observed, inter alia, in the R CRAN, which is useful for all professionals and scientists dealing with the analysis of spatial data.

References

[1] Anselin, L.: *Local Indicators of Spatial Association*-LISA. Geographical Analysis, 27, (1995), 93-115.
[2] Arbia, G.: *Spatial Econometrics: Statistical Foundations and Applications to Regional Growth Convergence*. Springer, New York, (2006).
[3] Bivand, R.: Autokorelacja przestrzenna a metody analizy statystycznej w geografii. In: *Analiza regresji geografii* (Chojnicki, Z., ed), PWN, Poznań, (1980), 23-38.
[4] Bivand, R.: *Spatial Econometrics Functions in R: Classes and Methods*. Journal of Geographical System, (2003)
[5] Cliff, A.D., Ord, J.K.: *Spatial Autocorrelation*. Pion, London, (1973).
[6] Geary, R.: *The Contiguity Ratio and Statistical Mapping*. The Incorporated Statistician, 5, (1954), 115–145.
[7] Getis, A., Ord, J.K.: *The Analysis of Spatial Association by Use of Distance Statistics*. Geographical Analysis, 24, (1992), 189–206.
[8] Griffith, D.A.: *Spatial Autocorrelations and Spatial Filtering*. Springer, Berlin-Heidelberg, (2003).
[9] Janc, K.: Zjawisko autokorelacji przestrzennej na przykładzie statystyki I Morana oraz lokalnych wskaźników zależności przestrzennej (LISA) – wybrane zagadnienia metodyczne. In: *Idee i praktyczny uniwersalizm geografii. Dokumentacja Geograficzna* (Komornicki, T., Podgórski, T., eds), 33, (2006) 76–83.
[10] Kopczewska, K.: Ekonometria *i statystyka przestrzenna z wykorzystaniem programu* R CRAN. Cedewu.pl, Warszawa, (2006).
[11] Moran, P. A. P.: *Notes on Continuous Stochastic Phenomena*. Biometrika 37 (1), (1950), 17–23.
[12] Suchecki, B. (red.): *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*. Wydawnictwo C.H. Beck, Warszawa, (2010).
[13] Tobler, W.: *A computer model simulating urban growth in Detroit region*. Economic Geography, 46(2), (1970), 234–240.