3. Spatial autocorrelation indices

BOUAYAD AGHA SALIMA
GAINS (TEPP) et CREST
Le Mans Université

MARIE-PIERRE DE BELLEFON
INSEE

Abstract

Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in space. The more the observation values are influenced by observation values that are geographically close to them, the greater the spatial correlation.
This chapter defines spatial autocorrelation, then describes spatial autocorrelation indices at the global and local levels — principles, properties, practical implementation with R and interpretation of their significance.

R Prior reading of Chapters 1 "Descriptive spatial analysis" and 2 "Codifying the neighbourhood structure" is recommended.

Very often, the variables for which geolocated information is available are characterised by spatial dependencies, which are all the stronger as the locations are closer. Thus, increasingly frequent access to spatial data makes it possible to better take into account interactions and spatial externalities in analysing the economic decisions made by agents. Analysis of spatial structures included in the data is essential for addressing, if necessary, any violation of the hypothesis of spatial independence of variables. Secondly, when it comes to interpretation, the analysis of spatial autocorrelation enables quantified analysis of the spatial structure of the phenomenon in question. Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in space.

3.1 What is spatial autocorrelation?

Autocorrelation measures the correlation of a variable with itself, when the observations are considered with a time lag (temporal autocorrelation) or in space (spatial autocorrelation). Spatial autocorrelation is defined as the positive or negative correlation of a variable with itself due to the spatial location of the observations. This spatial autocorrelation can first be the result of unobservable or difficult-to-quantify processes that combine different locations and, as a result, give rise to a spatial structuring of activities: interaction phenomena – between agents' decisions, for example – or dissemination – such as phenomena of technological diffusion – in space are each phenomena that can produce spatial autocorrelation. Secondly, in the context of the specification of econometric models, measuring spatial autocorrelation can be considered a tool for diagnosing and detecting an incorrect specification – omission of spatial variables that are spatially correlated, errors on the choice of scale on which the spatial phenomenon is analysed, etc.

From a statistical point of view, many analyses – analysis of correlations, linear regressions, etc. – are based on the hypothesis of independence of variables. When a variable is spatially auto-correlated, the independence hypothesis is no longer respected, thus challenging the validity of the hypotheses on the basis of which these analyses are carried out. Secondly, analysis of spatial autocorrelation enables quantified analysis of the spatial structure of the studied phenomenon.

It should be emphasised that spatial structure and spatial autocorrelation cannot exist independently of one another (Tiefelsdorf 1998):

— The term spatial structure refers to all the links with which the autocorrelated phenomenon will spread;
— without the presence of a significant autocorrelated process, the spatial structure cannot be empirically observed.

The spatial distribution observed is then considered the manifestation of the underlying spatial process.

3.1.1 Empirical observation of spatial autocorrelation

In the presence of spatial autocorrelation, it is observed that the value of a variable for an observation is linked to the values of the same variable for the neighbouring observations.

— Spatial autocorrelation is positive when similar values of the variable to be studied are grouped geographically.
— Spatial autocorrelation is negative when the dissimilar values of the variable to be studied come together geographically — nearby locations are more different than remote locations. This type of situation is usually found in the presence of spatial competition.

— In the absence of spatial autocorrelation, it can be considered that the spatial allocation of the observations is random.

Spatial autocorrelation indices make it possible to assess spatial dependence between values of the same variable in different places in space and test the significance of the identified spatial structure. To show this, the indices take into account two criteria:

— spatial proximity;
— the similarity or dissimilarity of the values of the variable for the spatial units considered.

Beware: if the data are aggregated following a breakdown that does not respect the underlying phenomenon, the strength of the spatial link will be overestimated or underestimated.

The measurement of a global spatial autocorrelation is distinguished from that in a given space and local autocorrelation in each unit of this space. This measures the intensity and significance of local dependence between the value of a variable in a spatial unit and the values of the same variable in neighbouring units (more or less close, depending on the neighbourhood criterion used).

3.1.2 Moran's diagram

Moran's diagram allows a rapid reading of the spatial structure. This is a scatter graph with the values of variable y centred on the x-axis and the average values of the variable for the neighbouring observations Wy in the y-axis, where W is the normalized weight matrix. The two properties, y *centred* and W *normalized* imply that empirical average Wy is equal to that of y and therefore 0. The straight regression line of Wy is also drawn depending on y and equation lines $y = 0$ and $Wy = 0$ that delineate the quadrants.

If the observations are randomly distributed in space, there is no particular relationship between y and Wy. The slope of the linear regression line is zero, and the observations are evenly allocated in each quadrant. If, on the contrary, observations have a particular spatial structure, the linear regression slope is non-null since there is a correlation between y and Wy. Each of the quadrants defined by $y = 0$ and $Wy = 0$ matches up with a type of specific space association (Figures 3.1 and 3.2).

— The observations in the upper right – quadrant 1 – show values of the variable that are higher than average, in a neighbourhood similar to it — positive spatial autocorrelation and high index value; high-high structure.
— At the bottom left – quadrant 3 – the observations show lower variable values than average, in a neighbourhood similar to it — positive space autocorrelation and low index value; low-low structure.
— Observations located at the bottom right – quadrant 2 – have higher values of the variable than average in a neighbourhood not similar to it — negative spatial autocorrelation and high index value; high-low structure.
— At the top left – quadrant 4 – the observations show values for the variable that are lower than the average in a neighbourhood not similar to it — negative spatial autocorrelation and low index value; low-high structure.

The density of points in each of the quadrants is used to visualise the dominant spatial structure. Moran's diagram also makes it possible to see the atypical points that move away from this spatial structure.

To understand how spatial autocorrelation can be seen on Moran's diagram, we simulated a growing spatial autocorrelation of incomes by IRIS (Figures 3.3 and 3.4). Parameter ρ that defines spatial autocorrelation is the slope in Moran's chart. Apart from extreme values, it is difficult to identify the sign and the strength of spatial autocorrelation by simply looking at the maps of the various values. On the other hand, Moran's diagrams make it possible to clearly identify the various scenarios.

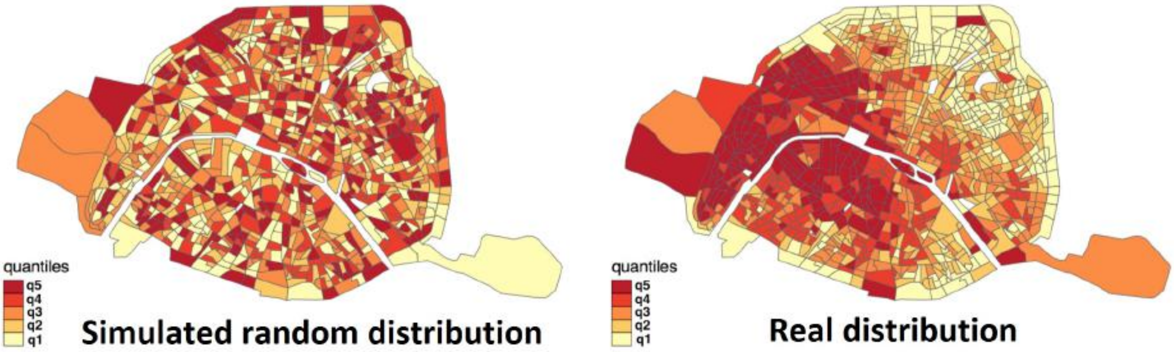


Figure 3.1 – Illustration, on Parisian census districts (IRIS), of the gap between random distribution and spatially autocorrelated distribution
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

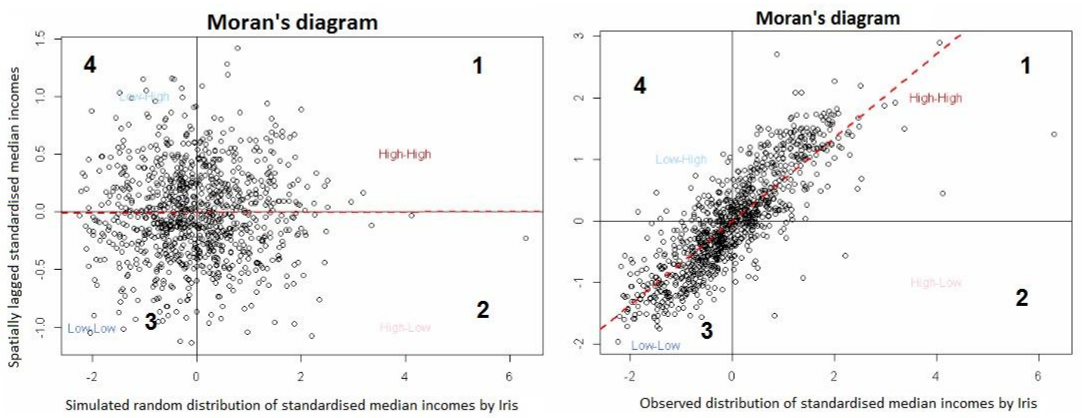


Figure 3.2 – Moran's diagram of a simulated random distribution of standardised median incomes by IRIS and standardised median incomes by IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

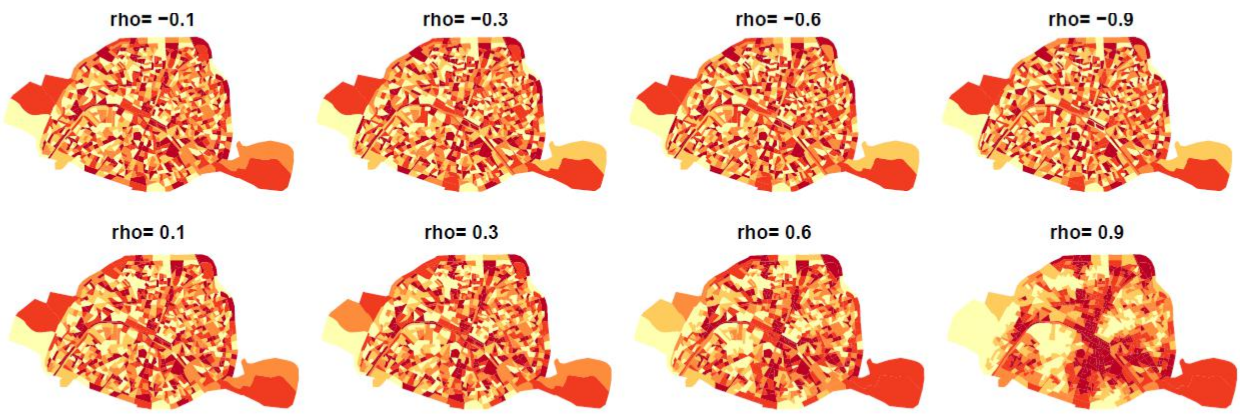


Figure 3.3 – Simulation of increasing spatial autocorrelation of incomes by IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

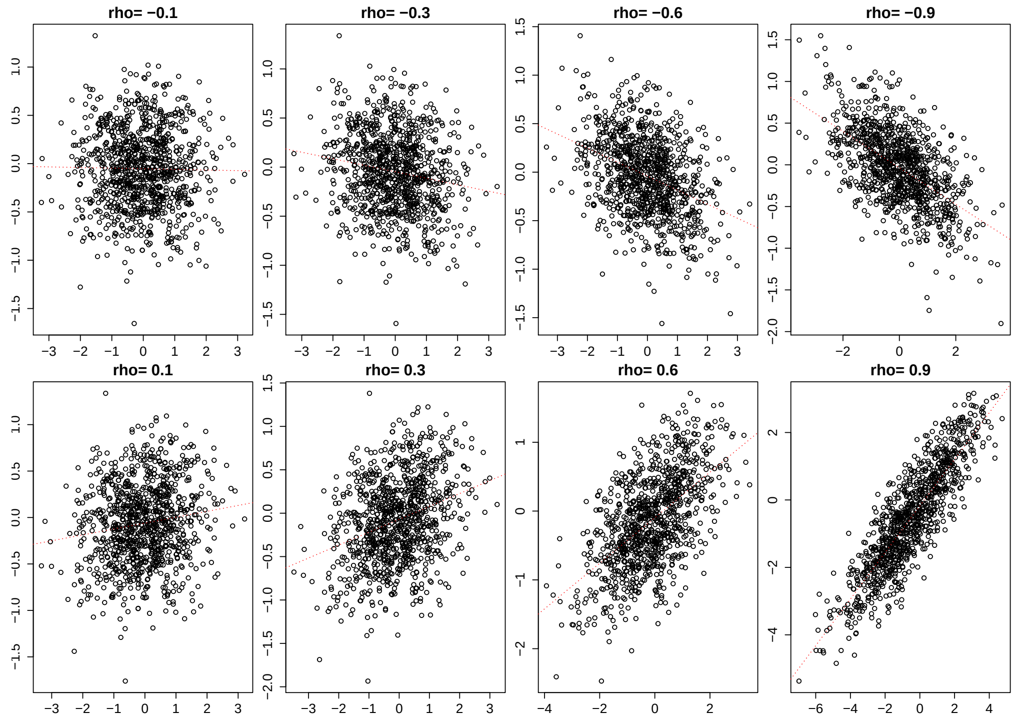


Figure 3.4 – Moran diagrams according to simulated autocorrelated income, for Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

3.2 Global measures of spatial autocorrelation

3.2.1 Spatial autocorrelation indices

When Moran's diagram highlights a particular spatial structure, calculating the spatial autocorrelation indices is used to answer two questions:

— Could the values taken by the neighbouring observations have been comparable (or also dissimilar) by mere chance?
— If not, then we are dealing with a case of spatial autocorrelation. How is this denoted and what is the strength of the said autocorrelation?

To answer the first question, we must test the hypothesis of absence of spatial autocorrelation for a gross variable y.

— H_0: no spatial autocorrelation
— H_1: spatial autocorrelation

To carry out this test, it is necessary to specify the distribution of the variable of interest y, in the absence of spatial autocorrelation (under H_0). In this context, statistical inference is generally conducted considering either of the following assumptions:

Normality hypothesis: each of the values of the variable, or y_i, is the result of an independent draw in the **normal distribution specific to each geographical area i on which this variable is measured**.

Randomisation hypothesis: The inference over Moran's I is usually conducted under the randomisation hypothesis. The estimated statistic calculated from data is compared with **the distribution of the data derived by randomly re-ordering the data – permutations**. The idea is simply that if the null hypothesis is true, then all possible combinations of data are equiprobable. The data observed are then only one of the many outcomes possible. In the case of spatial autocorrelation, the null hypothesis is always that there is no spatial association and the values of the variable are randomly assigned to the spatial units in order to calculate the test statistic. If the null hypothesis is rejected, *i.e.* if spatial autocorrelation is found, we can then calculate the range of values that governs the spatial autocorrelation index and thus answer the question as to the signals and strength of the spatial autocorrelation: the closest this index is to 1 in absolute value, the greater is the correlation. This interval depends on the weight matrix and can sometimes vary outside the interval $[-1;1]$, hence the importance of calculating the limits of this interval.

Very generally speaking, spatial autocorrelation indices are used to characterise the correlation between measures that are geographically similar to a measured phenomenon. If WY is the vector of means of variable Y (where W is the spatial weights matrix) in the neighbourhood of each spatial unit, spatial autocorrelation indices occur as:

$$Corr(Y,WY) = \frac{Cov(Y,WY)}{\sqrt{Var(Y).Var(WY)}} \tag{3.1}$$

Based on this very general formulation, for quantitative variables, two main indices are used to test for spatial autocorrelation — the Moran index and the Geary index. The former considers the variances and covariances taking into account the difference between each observation and the average of all observations. The Geary index takes into account the difference between the respective observations. In the literature, Moran's index is often preferred to that of Geary due to greater general stability (see in particular Upton et al. 1985).

Moran index

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.2}$$

— H_0: The neighbours do not **co-vary** in any particular way.
— $I_W > 0$ => positive spatial autocorrelation.

Geary index

$$c_W = \frac{n-1}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.3}$$

— H_0: The **differences** between neighbours have no particular structure.

— $c_W < 1$ => positive spatial autocorrelation.

Depending on the distribution chosen for the variable in the absence of spatial autocorrelation, the calculation of the variance of the indices is modified. In contrast, the equations that yield the expression of the expectancy of the indices (3.4) and the test statistic (3.5) remain the same. These relationships thus make it possible to assess the significance of spatial autocorrelation.

$$E(I_W) = E(c_w) = -\frac{1}{n-1} \tag{3.4}$$

$$\frac{I_W - E(I_W)}{\sqrt{Var(I_W)}} \sim \frac{c_W - E(c_W)}{\sqrt{Var(c_W)}} \sim \mathcal{N}(0,1) \tag{3.5}$$

As spatial autocorrelation is measured based on a comparison of the value of an individual variable with that of its neighbours, the definition of the neighbourhood will have a significant impact on the measurement of spatial autocorrelation. As explained in Chapter 2 "Codifying the neighbourhood structure", the larger the planned neighbourhood, the greater the number of neighbours considered, and the greater the probability that their average will be closer to the population's average, which may lead to a relatively low value for spatial autocorrelation.

A change in scale can also have implications when measuring spatial autocorrelation. The term MAUP (Modifiable Areal Unit Problem; Openshaw et al. 1979b) is used to describe the influence of spatial breakdown on the results of statistical processing or modelling.

More precisely, the irregular forms and limits of the administrative levels that do not necessarily reflect the reality of the spatial distributions studied are an obstacle to the comparability of the irregularly distributed spatial units. According to Openshaw 1984, MAUP is a combination of two distinct but similar problems:

— The scale problem stems from a change in the information generated when a set of spatial units is aggregated to form smaller and larger units for the needs of an analysis or due to data availability issues;

— The aggregation problem – or zoning – stems from a change in the diversity of information generated by the various aggregation schemes possible at a same scale. This effect is characteristic of administrative partitioning – particularly electoral – and adds to the scale effect.

■ **Example 3.1 — Spatial autocorrelation of median income in Paris.** What is the intensity of spatial autocorrelation in the income of Parisians? Is it significant? To what extend does it depend on the specification of spatial relations – type of neighbourhood, aggregation scale –?

Measuring spatial autocorrelation

Source	I_W	c_w	p value	H0	limits of I_W
Income: breakdown observed	0.68	0.281	3.10^{-6}	rejected	[-1.06.1.06]
Income: simulated random distribution	0.0027	1.0056	> 0.5	accepted	[-1.06,1.06]

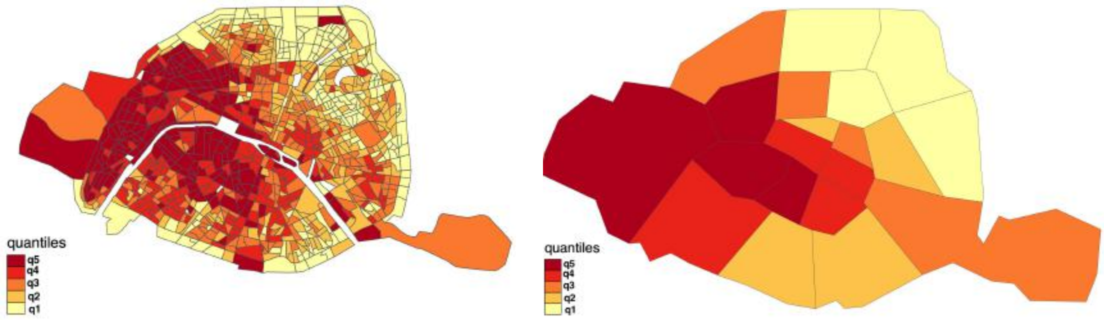
Table 3.1 – Moran and Geary indices of the median income earned by Parisians by IRIS: observed and simulated distribution
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Influence of neighbourhood choice

Type of neighbourhood	I_W	p value	H0
QUEEN	0.68	3.10^{-6}	rejected
ROOK	0.57	2.10^{-6}	rejected
1NN	0.30	0.07	rejected
3NN	0.58	9.10^{-6}	rejected
Delaunay	0.57	6.10^{-7}	rejected

Table 3.2 – Moran and Geary indices of the median income earned by Parisians by IRIS according to the neighbourhood defined
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Influence of the aggregation scale



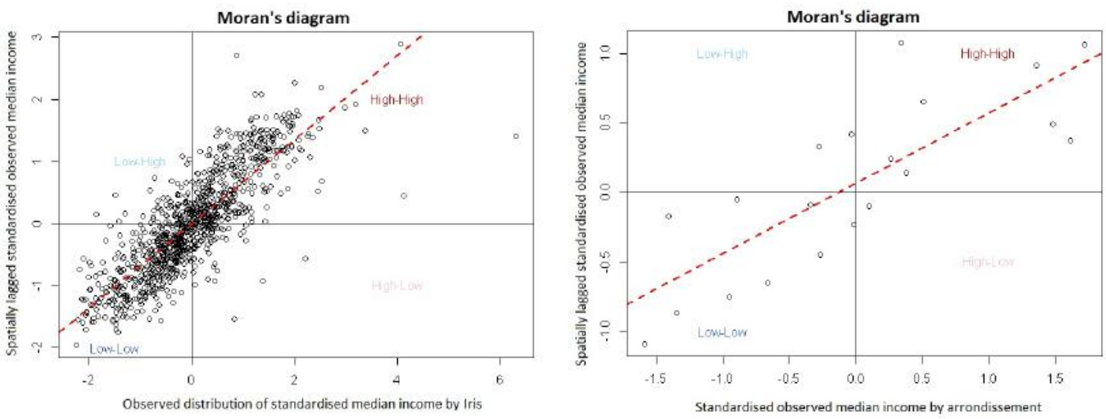
(a) By IRIS

(b) By arrondissement

Figure 3.5 – Aggregation of income in Paris
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Aggregation scale	I_W	p value	H0	Boundaries of I_W
IRIS	0.68	3.10^{-6}	rejected	[-1.06,1.06]
Arrondissement	0.51	$<9.10^{-9}$	rejected	[-0.53,1.01]

Table 3.3 – Value and significance of the Moran's I as a function of the chosen aggregation scale
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*



(a) By IRIS

(b) By arrondissement

Figure 3.6 – Moran's scatterplots for income distribution in Paris
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

In this example, whatever the definition of the neighbourhood or the aggregation scale, the spatial autocorrelation of the income earned by Parisians is positive and significant. The strength of spatial autocorrelation varies slightly depending on the type of neighbourhood used. In particular, looking only at the nearest neighbours slightly decreases the strength of spatial autocorrelation in this example.

■

Application with R

The *spdep* package is used to calculate spatial autocorrelation indices and their significance using functions `moran.test` and `geary.test`.
By default, the distribution of the variable of interest under the null hypothesis is derived by randomisation. The `randomisation = FALSE` argument makes it possible to assume that this is a normal distribution.

Box 3.2.1 — If certain entities do not have neighbours. In order for the package functions *spdep* to accept spatial weight matrices in which certain units do not have neighbours, it is necessary to specify the option: `zero.policy=TRUE`. By default, the size of the matrix is reduced to exclude observations without neighbours. The opposite can be specified with the option: `adjust.n=FALSE`. In this case, the absolute value of the test statistic increases, and the absolute value of its expected maturity and variance decreases (Bivand et al. 2013a). Generally speaking, spatial autocorrelation indices were developed assuming that all units had neighbours, and there are different opinions on what to do when this is not the case.

As seen before, two approaches are used to estimate the significance of these indices — an analytical solution based on the normality hypothesis and a Monte Carlo solution based on the randomisation hypothesis. The analytical solution, used by the `moran.test` function, is based on the assumption that the test statistic asymptotically follows a normal distribution with mean 0 and variance 1. This is not always the most accurate measure of significance as convergence towards this distribution may depend on the arrangement of the polygons. Instead, the `moran.mc` function can be used, allowing to choose the number of permutations to calculate the simulated distribution of Moran's I. Comparing the significance levels calculated from functions `moran.mc` and `moran.test` makes it possible to ensure the robustness of the conclusions.

```
library(spdep)

###############################
# Data preparation  ###########
###############################

#Extraction of list of neighbours (defined by default with Queen contiguity
    )
iris75.nb <- poly2nb(iris75)
#Creation of weight matrix
iris75.lw <- nb2listw(iris75.nb,zero.policy=TRUE)
#Calculation of standardised median income
iris75.data <- as.data.frame(iris75)
iris75.data$med_revenu_std <- scale(iris75.data$med_revenu)

#####################
# Moran's diagram
#####################
```

```
moran.plot(iris75.data$med_revenu_std, iris75.lw, labels=FALSE,
xlab='observed distribution of standardised median income by IRIS', ylab='
    Spatially lagged standardised median incomes')

######################
# Moran's I test
#####################

moran.test(iris75.data$med_revenu_std,iris75.lw, zero.policy=TRUE,
    randomisation=FALSE)

#Calculation of the range of Moran's I
moran.range <- function(lw) {
    wmat <- listw2mat(lw)
    return(range(eigen((wmat+t(wmat))/2)$values))
}

moran.range(iris75.lw)
```

3.2.2 Spatial autocorrelation of categorical variables

When the variable of interest is not continuous but categorical, the degree of local association is measured by analysing the statistics of the *join count* (Zhukov 2010).

To illustrate the calculation of these statistics, we consider a binary variable representing two colours, White (B) and Black (N) so that a relation can be called White-White, Black-Black or White-Black. It can be seen that:

- positive spatial autocorrelation occurs if the number of White-Black relations is significantly **lower than** what would have occurred with random spatial distribution;
- negative spatial autocorrelation occurs if the number of White-Black relations is significant **greater than** what would have occurred with random spatial distribution;
- no positive spatial autocorrelation occurs if the number of White-Black links is approximately **identical to** what would have occurred with random spatial distribution;

If there are n observations, n_b white observations and $n_n = n - n_b$ black observations, the probability of a white observation occurring is: $P_b = \frac{n_b}{n}$ and the likelihood of a black observation occurring is: $P_n = 1 - P_b$.

In the absence of spatial autocorrelation, the probabilities of observations of the same colour occurring in two neighbouring cells are: $P_{bb} = P_b * P_b = P_b^2$ and $P_{nn} = P_n * P_n = (1 - P_b)^2$.

The probability of obtaining different colour observation occurring in two neighbouring cells is: $P_{bn} = P_b * (1 - P_b) + (1 - P_b) * P_b = 2P_b * (1 - P_b)$.

As $\frac{1}{2} \sum_i \sum_j w_{ij}$ measures the number of existing relations, assuming random spatial distribution of the observations, it can be asserted that:

$$
\begin{aligned}
E[bb] &= \frac{1}{2}\sum_i \sum_j w_{ij} P_b^2 \\
E[nn] &= \frac{1}{2}\sum_i \sum_j w_{ij}(1-P_b)^2 \\
E[bn] &= \frac{1}{2}\sum_i \sum_j w_{ij} 2P_b * (1-P_b)
\end{aligned}
\tag{3.6}
$$

Assuming $y_i = 1$ when the observation is black and $y_i = 0$ in the opposite case (white colour), the empirical counterparts (observed values) of these mathematical expectations can be written:

$$
\begin{aligned}
nn &= \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \\
bb &= \frac{1}{2}\sum_i \sum_j w_{ij}(1-y_i)(1-y_j) \\
bn &= \frac{1}{2}\sum_i \sum_j w_{ij}(y_i - y_j)^2
\end{aligned}
\tag{3.7}
$$

In this case, the test statistic to assess the significance of spatial autocorrelation is based on the assumption that in the absence of spatial autocorrelation, the statistics of *joincount* (*bb, nn* and *bn*) follow a normal distribution. It can be written that:

$$
\frac{bn - E(bn)}{\sqrt{Var(bn)}} \sim \mathcal{N}(0,1) \quad \frac{bb - E(bb)}{\sqrt{Var(bb)}} \sim \mathcal{N}(0,1) \quad \frac{nn - E(nn)}{\sqrt{Var(nn)}} \sim \mathcal{N}(0,1) \tag{3.8}
$$

■ **Example 3.2 — *Joincount* statistics on the employment of individuals in Paris.** [1]

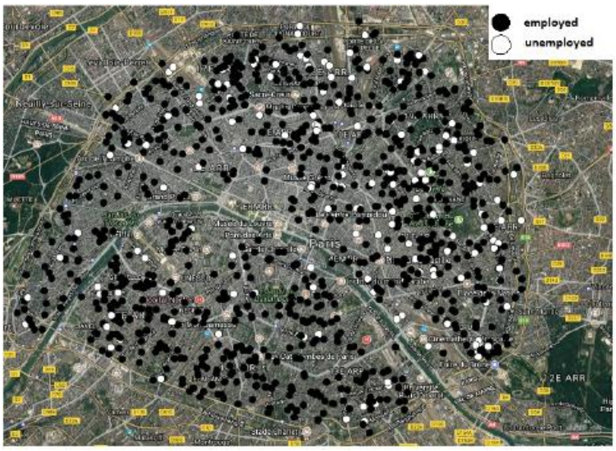


Figure 3.7 – Employment of a sample of 1,000 individuals in Paris
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

The binary variable is considered to be 1 if individual i is unemployed and 0 otherwise. The aim is to determine whether unemployed Parisians are more grouped into space than if they were randomly distributed. *Join count* statistics make it possible to answer this question. From the table 3.4, the location of the unemployed can be observed to be significantly correlated, as is the location of the employed.

1. The purpose of this example is not to detail the results of an economic study, but to illustrate the techniques implemented. There is no interpretation to be derived from this.

Variable	p-value of the join count statistic of spatial association	H0
Unemployed	$5.439.10^{-}3$	rejected
Active workers	$9.085.10^{-}5$	rejected

Table 3.4 – Significance of the join count statistic of Parisian unemployed
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Application with R

Join count statistic is reached by implementing `joincount.test` function of the *spdep* package in R.

```
library(spdep)

# Conversion as factor
menir10_d75_subset$unemployment <- ifelse(menir10_d75_subset$ZCHOM>0, 3, 1)
unemployment <- as.factor(menir10_d75_subset$unemployment, levels=c("
    employed", "unemployed"))

# Neighbours list and spatial weight matrices
coordinates(menir10_d75_subset) <- c("PLG_X", "PLG_Y")
proj4string(menir10_d75_subset) <- CRS("+init=epsg:27572 +proj=lcc +lat_
    1=46.8 +lat_0=46.8 +lon_0=0 +k_0=0.99987742 +x_0=600000 +y_0=2200000 +a
    =6378249.2 +b=6356515 +towgs84=-168,-60,320,0,0,0,0 +pm=paris +units=m
    +no_defs")
menir10_d75_subset <- spTransform (menir10_d75_subset, CRS ("+init=epsg
    :2154") )

menir75.nb <- knn2nb(knearneigh(menir10_d75_subset,k=2))

# Implementation of the test
joincount.test(unemployment, listw2U(nb2listw(menir75.nb)))
```

Where dealing with several categories, the `joincount.multi` function of package *spdep* tests the significance, in accordance with the same principle, namely of spatial association of different variables.

■

3.3 Local measures of spatial autocorrelation

Global statistics are based on **the assumption of a spatial stationary process**: spatial autocorrelation would be the same throughout space. However, this assumption is all the less realistic as the number of observations is high.

3.3.1 Getis and Ord index

Getis and Ord (Getis et al. 1992) offer an indicator for identifying local spatial dependencies that do not appear in the global analysis.

Getis and Ord indicator

$$G_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \tag{3.9}$$

$G_i > 0$ indicates a grouping of values higher than average.
$G_i < 0$ indicates a grouping of values lower than average.

The significance of the Getis and Ord indicator can be tested by making the assumption, in the absence of local spatial dependency, of a normal distribution.

$$z(G_i) = \frac{G_i - E(G_i)}{\sqrt{Var(G_i)}} \sim \mathcal{N}(0,1) \tag{3.10}$$

Application with R

The `localG` function of package *spdep* makes it possible to use this indicator.

3.3.2 Local spatial autocorrelation indicators

Anselin (Anselin 1995) develops the concepts introduced by Getis and Ord by defining *local spatial autocorrelation indicators*. These must measure the intensity and significance of local autocorrelation between the value of a variable in a spatial unit and the value of the same variable in the surrounding spatial units. More specifically, these indicators make it possible to:

- detect significant groupings of identical values around a particular location (clusters);
- identify spatial non-stationarity zones, which do not follow the global process.

The Getis and Ord indicators serve only the first of these two objectives. To be considered as local spatial association measures – (LISA; *Local Indicators of Spatial Association*) – as defined by Anselin, these indicators must verify the following two properties:

- for each observation, they indicate the intensity of the grouping of similar – or opposite in trend – values around this observation;
- the sum of local indices on all observations is proportional to the corresponding global index.

One of the most used LISA is the local Moran's I.

Local Moran's I

$$I_i = (y_i - \bar{y}) \sum_j w_{ij}(y_j - \bar{y}) \tag{3.11}$$

$$I_W = constante * \sum_i I_i \tag{3.12}$$

$I_i > 0$ indicates a grouping of similar values (higher or lower than average).
$I_i < 0$ indicates a combination of dissimilar values (*e.g.* high values surrounded by low values).

3.3.3 Significance of the local Moran's I

Significant LISAs are combinations of similar or dissimilar values more marked than what might have been observed based on random spatial distribution. These groupings can match up with the four types of spatial groupings described in 3.1 and identifiable on Moran's diagram (high-high, low-low, high-low or low-low). The significance test of each local association indicator is based on

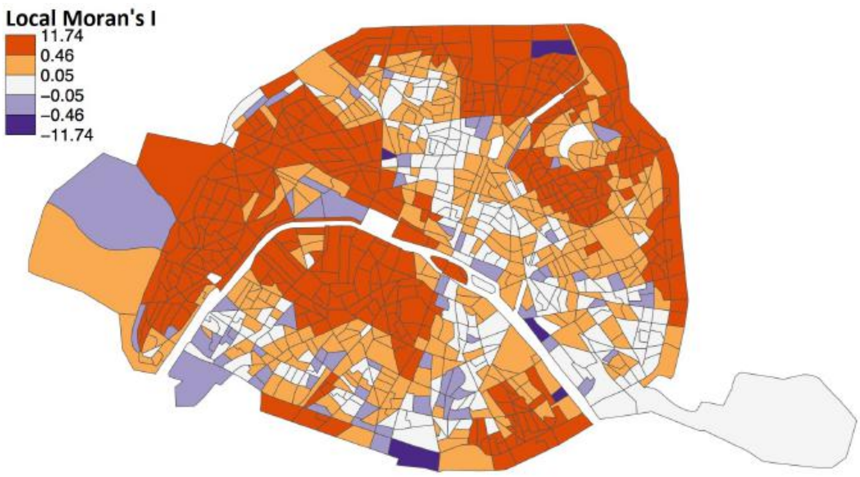


Figure 3.8 – Values of local Moran's I, on Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

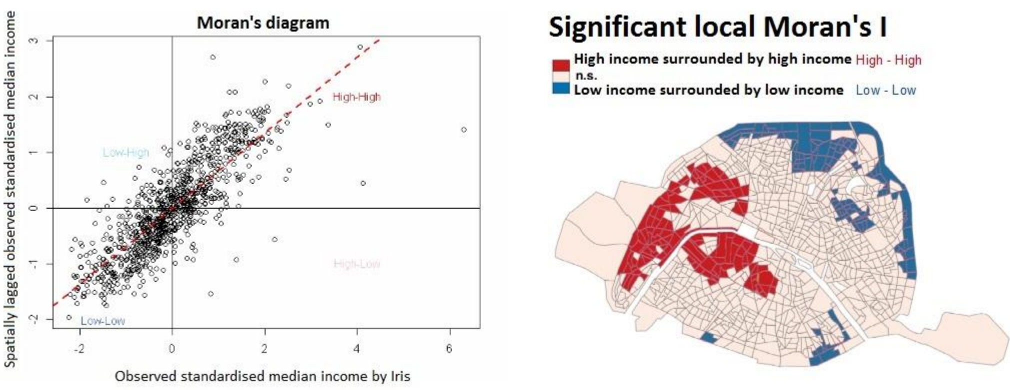


Figure 3.9 – Significant local Moran's I, on Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

a statistic assumed to asymptotically follow a normal distribution under the null hypothesis. If the assumption of normality holds, $z(I_i) = \frac{I - E(I_i)}{\sqrt{Var(I_i)}} \sim \mathscr{N}(0,1)$.

To test the validity of the normality assumption of the LISAs under the null hypothesis, several random distributions are simulated in the space of the variable of interest and the local indicators associated with these simulations are calculated.

Taking up the example of Parisians' income once again, we can see that (Figure 3.10) the extreme quantiles of the distribution of the local Is are higher than those of a normal distribution. The *p-values* calculated under the normality assumption are therefore to be used with caution. This is because, as Anselin (Anselin 1995) shows, based on simulations (Figure 3.11), **in the presence of global spatial autocorrelation, the normality assumption of I_i no longer holds.**

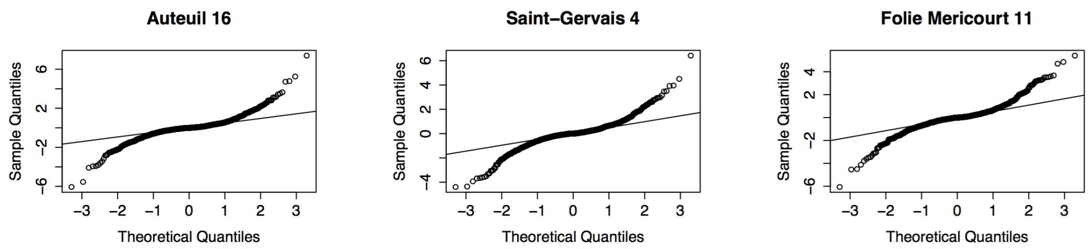


Figure 3.10 – Testing the normality assumption of of local Moran's Is distribution on three Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

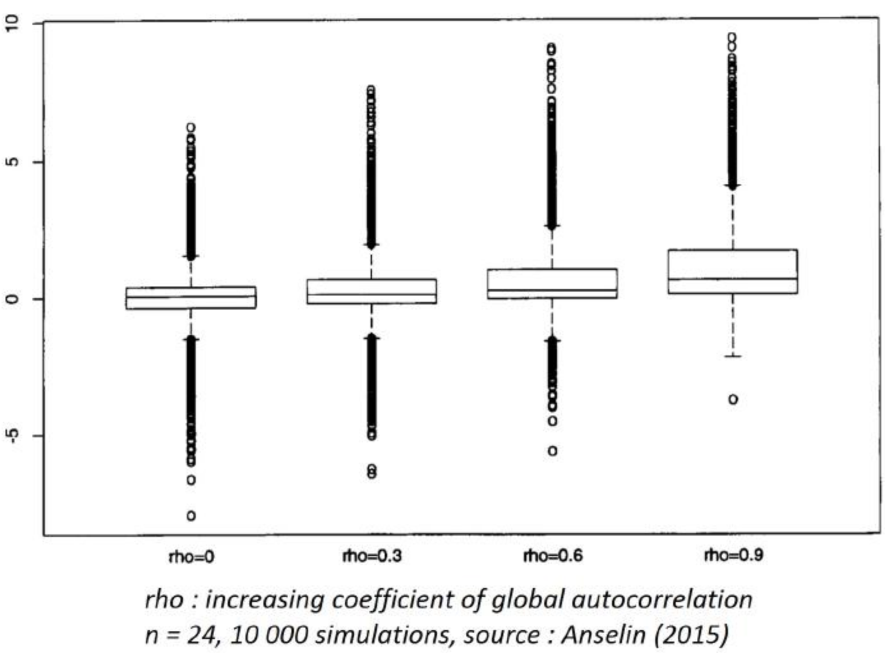


Figure 3.11 – Distribution of local Moran's Is where global spatial autocorrelation exists
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Moreover, the LISAs significance test raises a problem encountered every time multiple comparisons are made. Indeed, when several statistical tests are carried out simultaneously using the same dataset, the global risk of errors in decision of the first kind – probability of wrongly rejecting the null hypothesis – increases. At each test, the risk of randomly having a significant result is repeated. This increases the global risk to wrongly accept the significance of local index. Thus, in our case, we will conclude positively as to the **existence** of local spatial autocorrelation if **at least one** local spatial autocorrelation index is significant out of all indices in the study area. If there are 100 local spatial autocorrelation indices, there is a 100-fold increase in the risk of incorrectly detecting at least one significant one (exact formula in box 3.3.1) . Risk inflation α (type I error) is the risk of concluding incorrectly that local spatial autocorrelation has occurred is increased (Anselin 1995, Ord et al. 1995).

Different methods have been developed to prevent risk inflation α when multiple statistical comparisons are needed. Some of them are described below. Let α be the significance level selected for each local index.

Box 3.3.1 — The Bonferroni method – the historical method. The probability of not wrongly rejecting H_0 is $1-\alpha$ by polygon, therefore $(1-\alpha)^n$ for the whole zone, with n the number of polygons.
The probability of rejecting H_0 wrongly at least once is $\alpha^* = 1-(1-\alpha)^n \approx n\alpha$.
If the overall risk is to be maintained at approximately α, it is thus possible to choose $\alpha' \approx \frac{\alpha^*}{n}$ as a level for each individual test. For example, for $\alpha = 0.05$, a grouping is significant if its p-value is $\frac{0.05}{n}$.
The R software allows this method to be applied with the `method='bonferroni'` option of the `p.adjust` function.
It is considered that this method only yields good results when the number of tests carried out is small. In the case of the LISAs, it is a bit too restrictive and can lead to risk, given the number of comparisons, not to detect certain significant LISAs.

Box 3.3.2 — The Holm Adjustment Method makes it possible to detect a spatial cluster. The Holm adjustment method (Holm 1979) takes into account the fact that out of n polygons, k are truly significant spatial clusters, thus the probability of incorrectly rejecting H_0 on the whole area is not $(1-\alpha)^n$ but $(1-\alpha)^{n-k}$, where α is the desired significance level.

The Holm method classifies the p-values from α_1 lowest to α_n highest. If $\alpha_1' \sim n\alpha_1 < \alpha$, *i.e.* $\alpha_1 < \frac{\alpha}{n}$, it is considered that this local index is indeed significant since it meets the most restrictive criterion. Attention is then turned to whether $\alpha_2 < \frac{\alpha}{n-1}$, and so on until testing whether $\alpha_k < \frac{\alpha}{n-k+1}$.

The R software makes it possible to apply this method with the `method='holm'` option of the `p.adjust` function.

The Holm adjustment method leads to more significant clusters than the Bonferroni method. It is therefore the most often preferred. However, this method also focuses on detecting **at least one cluster throughout the zone**.

Box 3.3.3 — The False Discovery Rate Method – locating spatial clusters. The False Discovery Rate (FDR) method was introduced by Benjamini et al. 1995. With this method, the risk of judging – incorrectly – a cluster as significant is higher, but conversely the risk of judging – incorrectly – a cluster as non-significant is lower. Caldas de Castro et al. 2006 prove the interest of this method to **locate** significant spatial clusters.

The FDR method classifies the p-values from α_1 lowest to $\alpha_n$highest.
Let k be the largest whole number such that $\alpha_k \leq \frac{k}{n}\alpha$. Benjamini and Hochberg explain that the null hypothesis of absence of local spatial autocorrelation for all clusters whose p-values are less than or equal to α_k can be rejected. The R software makes it possible to apply this method with

the `method='fdr'` option of the `p.adjust` function.

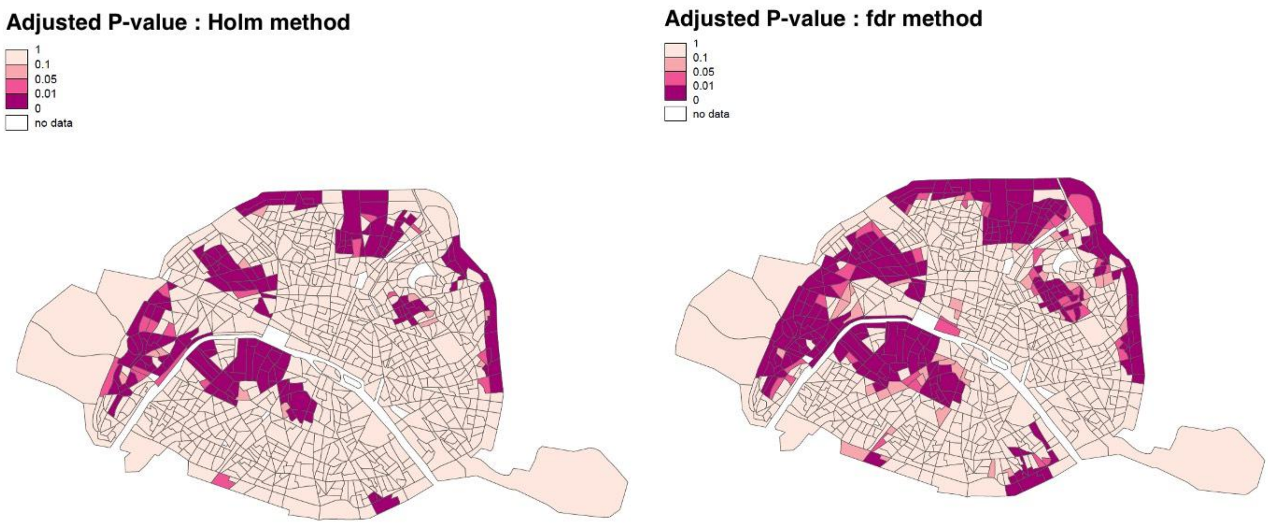


Figure 3.12 – Testing the significance of the local Moran's I, on Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

In the example of Parisians income (Figure 3.12), it is clear that the adjustment of the p-values using the Holm method leads to less significant p-values than the adjustment by the FDR method. The Holm method reduces the risk of making incorrect positive conclusions as to the **existence** of local spatial autocorrelation. On the other hand, this method increases the risk of overlooking a local cluster. The choice of the adjustment method will therefore depend on the objectives of the study and the risks that are favoured.

Application with R

```
lisa_revenus<-  localmoran(iris75.data$med_revenu, iris75.lw, zero.policy=
   TRUE)

# Calculation of adjusted p-values
iris75.data.LISA$pvalue_ajuste<-
p.adjust(iris75.data.LISA$pvalue_LISA, method='bonferroni')
```

3.3.4 Interpretation of local indices

In the absence of global spatial autocorrelation

The LISAs make it possible to **identify areas where similar values are significantly grouped**. These are areas where the local spatial structure is such that the relations between neighbours are particularly strong.

In the presence of global spatial autocorrelation

LISAs **indicate areas that have a particular impact on the global process** (local autocorrelation more pronounced than global autocorrelation), **or, on the contrary, which stand out from it** (lower autocorrelation). Thus, using the median income of Parisians as an example, it can be seen that the distribution of local Moran's I is not centred on the global Moran's I (Figure 3.13). Some zones have a significantly different spatial association structure than the global process.

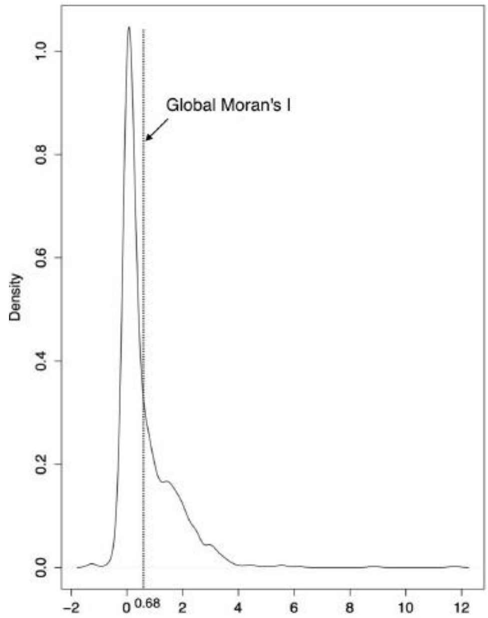


Figure 3.13 – Distribution of local Moran's I of median incomes by Parisian IRIS
Source: *INSEE, Localised Tax Revenues System (RFL) 2010*

Even adjusted, the p-values are at risk of being too low, as the distribution of I_i moves away from the norm. The more the global autocorrelation increases, the higher the number of extreme values. High LISAs can therefore hardly be interpreted as significant groupings of similar value. In this case, LISAs are interpreted as indicators of a certain type of **local instability**.

3.4 Spatio-temporal indices

It is not unusual for a geolocalised database to have observations raised at different points in time, as is the case with databases that list real estate transactions. It may be interesting to understand how a localised phenomenon has spread and evolved in space and time and how it can be linked to the conditions of the environment surrounding it. In this case, it is important to be able to assess how the underlying spatial structures change over different periods of time. On spatio-temporal data, prior graphical exploration of cross-section data (standard Moran's I) can be used to study the existence and change in grouping or dispersion trends that are statistically significantly different from random models. Many recent developments show a growing interest in analysing spatio-temporal data in many areas of research such as physics, meteorology, economics and environmental studies. By extending the Moran index to include time attributes, it becomes possible to calculate global and localized indices that concurrently take into account spatial and temporal auto-correlations. This can also be done on the basis of spatial-temporal risk weighting matrices. The work of Martin et al. 1975, Wang et al. n.d., López-Hernández et al. 2007 suggests extensions of Moran's I, traditionally used to measure spatial dependence, to calculate a spatio-temporal Moran's I. Chen et al. 2013 develop an enhanced analytical approach based on the traditional Moran's I, based on stationary data over time. As Lee et al. 2017 note, geolocated time series are usually non-stationary. When this assumption is not respected, Moran's spatiotemporal index suggested by Chen et al. 2013 can be fallacious. Lee et al. 2017 suggest to bypass this difficulty by applying a correction of fluctuations around the trend – detrended fluctuation analysis, DFA – and suggest a new method for calculating this index.

Conclusion

Spatial autocorrelation indices are exploratory statistical tools that make it possible to bring out the existence of a significant spatial phenomenon. Sections 2 and 3 present different methods of taking this spatial phenomenon into account, at global or local level, for quantitative or qualitative variables. It is important to know whether the autocorrelation is insignificant, but also to measure the extent to which the autocorrelation is significant in order to determine the scale of spatial dependence. The study of spatial autocorrelation is an essential step before considering any specification of spatial interactions in an appropriate model.

References - Chapter 3

Anselin, Luc (1995). « Local indicators of spatial association—LISA ». *Geographical analysis* 27.2, pp. 93–115.

Benjamini, Yoav and Yosef Hochberg (1995). « Controlling the false discovery rate: a practical and powerful approach to multiple testing ». *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Bivand, Roger S, Edzer Pebesma, and Virgilio Gomez-Rubio (2013a). *Applied spatial data analysis with R*. Vol. 10. Springer Science & Business Media.

Caldas de Castro, Marcia and Burton H Singer (2006). « Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association ». *Geographical Analysis* 38.2, pp. 180–208.

Chen, Shao-Kuan et al. (2013). « Analysis on Urban Traffic Status Based on Improved Spatiotemporal Moran's I ». *Acta Physica Sinica* 62.14.

Getis, Arthur and J Keith Ord (1992). « The analysis of spatial association by use of distance statistics ». *Geographical analysis* 24.3, pp. 189–206.

Holm, Sture (1979). « A simple sequentially rejective multiple test procedure ». *Scandinavian journal of statistics*, pp. 65–70.

Lee, Jay and Shengwen Li (2017). « Extending moran's index for measuring spatiotemporal clustering of geographic events ». *Geographical Analysis* 49.1, pp. 36–57.

López-Hernández, Fernando A and Coro Chasco-Yrigoyen (2007). « Time-trend in spatial dependence: Specification strategy in the first-order spatial autoregressive model ». *Estudios de Economia Aplicada* 25.2.

Martin, Russell L and JE Oeppen (1975). « The identification of regional forecasting models using space: time correlation functions ». *Transactions of the Institute of British Geographers*, pp. 95–118.

Openshaw, Stan (1984). *The modifiable areal unit problem*. Vol. CATMOG 38. GeoBooks, Norwich, England.

Openshaw, Stan and Peter Taylor (1979b). « A million or so correlation coefficients ». *Statistical methods in the spatial sciences*, pp. 127–144.

Ord, J Keith and Arthur Getis (1995). « Local spatial autocorrelation statistics: distributional issues and an application ». *Geographical analysis* 27.4, pp. 286–306.

Tiefelsdorf, Michael (1998). « Modelling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». PhD thesis. Université Wilfrid Laurier.

Upton, Graham, Bernard Fingleton, et al. (1985). *Spatial data analysis by example. Volume 1: Point pattern and quantitative data.* John W & Sons Ltd.

Wang, Y. F. and H. L. He. « Spatial Data Analysis Method ». *Science Press, Beijing, China*.

Zhukov, Yuri M (2010). « Applied spatial statistics in R, Section 2 ». *Geostatistics.[Online] Available: http://www. people. fas. harvard. edu/~ zhukov/Spatial5. pdf.*