A Tutorial on Spatial Analysis of Areas

Gilberto Câmara[1]*, Marilia Sá Carvalho*[2]

[1]Image Processing Division, National Institute for Space Research (INPE), Av dos Astronautas 1758, São José dos Campos, Brazil

[2]National School for Public Health , Fundacao Oswaldo Cruz

R. Leopoldo Bulhoes, 1480/810, Rio de Janeiro, Brazil

INTRODUCTION

This tutorial discusses methods for the analysis of spatial data whose location is associated to areas delimited by polygons. This situation frequently occurs when we deal with events aggregated by city, districts or census tract, where one doesn't have the exact location of the events, but instead, a value for the whole area. Some of these indicators are counts, as is the case with the majority of the variables collected by the census: for example, the Brazilian Institute of Geography and Statistics (IBGE) provides, for each census tract, the number of family heads for each income range. Various health indicators are also of this type: the Brazilian Ministry of Health and the State Departments of Health organize and distribute data about death and birth rates and contagious diseases by municipality. Using two counts – deaths and population, for instance – density rates of incidence, like death rate or incidence are estimated. Other very useful indicators are: (a) proportions, like the percentage of illiterate adults; (b) averages, like the mean income of the family head by census tract, and (c) medians, like the median age for men.

The usual form for presenting data aggregated by area is using color maps with the spatial pattern of the phenomena. Figure 1 shows the spatial distribution of the social exclusion index[1] for 96 districts of the city of São Paulo, from the data of the 1991 Census. It can be verified that 2/3 of the 96 districts of São Paulo were below the minimum acceptable levels of social inclusion in 1991. A strong polarization downtown-suburbs is clearly perceptible in the map, that presents two great regions of social exclusion, the South and East zones of the city. In the East zone, it is perceptible a gradient in the index of social exclusion/inclusion that worsens as we move away from the center. In the South zone, the index discontinuity is more

[1] The social exclusion/inclusion index is an aggregate measure of the socioeconomic disparities, which vary from -1 to +1, where the value of 0 (zero) indicates a basic level of social inclusion.

relevant, and we notice the existence of districts with high indexes of social exclusion/inclusion close to the excluded areas.

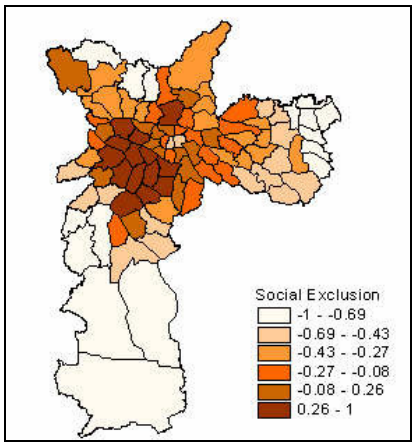


Figure 1 – Index of social Exclusion/Inclusion of the districts in the city of São Paulo from 1991 data, with 96 districts grouped by sextiles.

A large part of the users limits their use of the GIS to these visualization operations, drawing intuitive conclusions. But it is possible to go much further. When we visualize a spatial pattern, it's very useful to translate it into objective considerations: is the observed pattern random or it represents a definite aggregation? Can such distribution be associated to measurable causes? Are the observed values enough to analyze the spatial phenomena to be studied? Is there any group of areas with differentiated patterns within the region of study?

To approach these issues, we presents a set of spatial analysis techniques for data aggregated by areas. The first step is to choose the inference model to be used. The most common hypothesis is to assume that the areas are differentiated and that each one of them has an "identity" of its own. From the statistical point of view, this implies that each area presents a probability distribution different from the others, the so-called *discrete model for spatial data*. The alternative is to assume that the studied phenomenon presents spatial discontinuity, forming a surface, the so-called *continuous model for spatial data*. In this case, the areas are considered just a support for data collecting, and the inference model doesn't take into consideration the limits of each area. The production of surfaces from the areal data will be discussed in the end of this tutorial.

The question of count aggregation in areas poses yet two important conceptual problems: can the individual behavior be estimated from aggregated data? How much does the behavior of the aggregate reflect more than the sum of the individuals? What is the error in the estimation of the

indicators when the counts are very small? In this tutorial, the basic concepts for the spatial analysis of data aggregated by area will be presented after the presentation of the adequate models for the analysis of data aggregated by area.

MODELS FOR DATA DISTRIBUTION IN AREAS

The most frequently used model for areal data is the *model of discrete spatial variation*. Consider the existence of a stochastic process Z_i, $i = 1,\ldots,n$, where Z_i is the realization performance of the spatial process in the area i and n is the total number of areas A_i. The main objective of the analysis is constructing an approximation for the joint distribution of random variables $Z = \{Z_1,\ldots,Z_n\}$, estimating its distribution.

Similar to the model of point events, consider Z_i as the random variable which describes the count, indicator or rate associated to area A_i. We have an observed value z_i, corresponding to the count in the i-th area. The most common hypothesis is to assume that the random variable Z_i, which describes the number of occurrences in each area can be associated to a Poisson Probability Distribution. Such hypothesis is reasonable because this is the statistical distribution most adequate to phenomena involving the counting of events, and such is the case of most of the data aggregated by area. Evidently other distributions could be more adequate, depending on the variable being analyzed. Rates could be modeled using the normal distribution, for even though it admits negative values, which is evidently impossible in this type of indicator, the properties of the normal distribution can be adequate.

The alternative to the hypothesis of the *discrete spatial variation* is to assume that the data presents a *continuous spatial variation*. Let's consider a stochastic process $\{Z(x), x \in A, A \subset \Re^2\}$, whose values can be known in all the points of the study area. In this case, the aggregate counts must be transformed into rates or indicators, because rates vary continuously in space while counts don't. The use of continuous spatial models will be discussed in the end of this tutorial.

THE MODIFIABLE AREA UNIT PROBLEM

One of the basic problems with data aggregated by area is that, for the same population under study, the spatial definition of the frontiers of the areas impacts the results obtained. The estimates obtained within a system of units of area function of the many ways that these units can be grouped;

different results can be obtained by just changing the frontiers of these zones. This problem is known as the "The modifiable area unit problem".

In many studies involving areal data, the aggregate data is the only available source, but the object of study relates to the individual characteristics and relationships. Some of these studies try to establish cause-effect relationship among different measurements, as in the case of regression models; a classic example is the correlation of the years of schooling of the family head and his income, usually strong. Due to the effects of scale and of the aggregation of areas, the coefficients of correlation can be entirely different among the individuals and among the areas.

Consider a set of individuals and two characteristics of each measured according to what is estimated in Figure 2. A regression considering all individuals (black line in the box to the left) results in a positive coefficient of 0.1469. These individuals belong to distinct groups, separating each group by the attribute color, we obtain a negative correlation, varying between –0.5 and –0.8. Using the means of each group (black line in the box to the right), the coefficient gets to 0.99. It's important to observe that each model measures a different aspect and that there's no correct model. In the first case, we can say that without information that allows to differentiate the individuals within the colored groups, the variables interrelate positively. In the last example, the interest of the study is the effect of the variation in the mean of a variable over the mean of another, in the groups. These are different questions and different models.

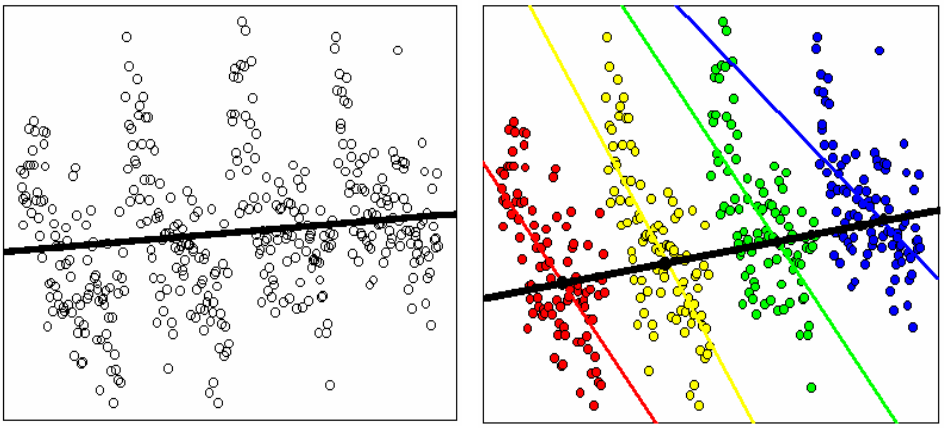


Figure 2 – Regression models: individuals, individuals in different strata and groups.

To illustrate the MAUP, we studied the Belo Horizonte city 1991 census data of 1991 in two scales: the census tracts and the planning units (UP) as shown in Figure 2. The census tracts were used by the Brazilian Census Bureau as the basic collection units, and the planning units correspond to the area aggregations used by the Planning Department of Belo Horizonte.

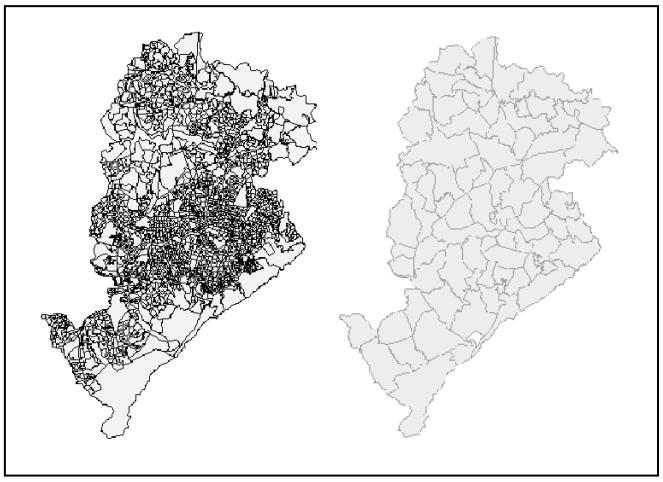


Figure 3. Census Tracts (left) and Planning Units (right) for the city of Belo Horizonte.

We computed 1,000 correlations between pairs of census, for both scales (census tracts and planning units). For example, taking the variables "number of family heads with 1 to 3 years of schooling" and "number of family heads with 1 to 3 minimum wages" results in different correlations in the case of census tracts (0.79) and for planning units (0.96). The results, shown in Table 1, indicate that the correlations in the census tracts are significantly lower than the correlations in the planning units. A total of 773 correlations are lower for the census tracts than for the planning units. Only 40 (4%) show the opposite behavior. In some situations, a change in signal does occur, that is, variables that are negatively correlated in the census tracts

become positively correlated. It is can be verified that the reduction in scale (bigger areas) makes the data more homogeneous, reducing the random fluctuation and reinforcing correlations that, thus, seems to be stronger than in smaller areas.

The above results indicate that one cannot assure that any scale is the "right" one, but which of the models better serve the objective of clarifying the issue: weaker correlations and more random fluctuations, but with more internal homogeneity, or stronger correlations with the bias introduced by not considering the dispersion and the heterogeneity around the mean in the greater areas. As a general rule we have: the more disaggregated the data the greater the flexibility on choosing the models. Aggregating in bigger regions is easy while disaggregating is impossible.

Table 1

Correlation between pairs of variables according to different units of area – Census tracts and Planning Units – for the 1991 Census in the city of Belo Horizonte

		Correlations per Planning Unit							
Correlation per Census Tract		-0,4/-0,2	-0,2/0,0	0,0/0,2	0,2/0,4	0,4/0,6	0,6/0,8	0,8/1,0	Pairs
	-0,8/-0,6	0	0	1	1	1	0	2	**5**
	-0,6/-0,4	2	11	7	4	2	7	0	**33**
	-0,4/-0,2	**3**	23	14	11	10	3	6	**70**
	-0,2/0,0	3	**5**	9	27	34	13	21	**112**
	0,0/0,2	0	1	**2**	42	75	32	55	**207**
	0,2/0,4	0	2	0	**17**	44	50	68	**181**
	0,4/0,6	0	2	3	1	**10**	42	110	**168**
	0,6/0,8	0	0	2	7	8	**9**	75	**101**
	0,8/1,0	0	0	0	4	4	3	**112**	**123**
	Total	**8**	**45**	**38**	**114**	**187**	**159**	**449**	**1000**

In practice, for confidentiality reasons, the individual data are seldom available. What to do then? One possibility is to work with the data in the greatest spatial scale possible, usually denominated micro-areas, for example, census tracts and to use aggregation or combinatorial optimization techniques to obtain more aggregated regions, but that preserve the phenomena under

investigation as much as possible. This way, we must recognize that the problem of scale is an inherent effect of the data aggregated by areas. It cannot be removed and cannot be ignored. To minimize its impact with relation to these studies, one must try to utilize the best available scale of data retrieval using techniques that allow the treatment of the random fluctuation, always looking for criteria for data aggregation that are consistent with the objectives of the study.

EXPLORATORY SPATIAL DATA ANALYSIS

The techniques of exploratory analysis applied to spatial data are essential to the development of the stages of spatial statistics modeling which is, in general, sensitive to the type of distribution, to the presence of extreme values and to the lack of stationarity. The techniques employed are, in general, adaptations of commonplace tools. Thus, if during the investigation of extreme values one utilizes graphic tools like histograms or boxplots, in the spatial analysis it's important to also investigate outliers not only inside the data set but also relative to the neighborhood. Besides, the nonstationarity of the spatial process within the region of study must also be investigated, in its various aspects: mean variation (first order), spatial variance and covariance.

Data visualization

The most simple and intuitive exploratory analysis is the visualization of extreme values in the maps. It's worth pointing out that the use of different cut-points on the variable leads to the visualization of different aspects. The GISs usually make available three methods of variable cut: equal intervals, percentiles, and standard deviation. In the case of *equal intervals* where the maximum and minimum values are divided by the number of classes, if the variable has a distribution that is too concentrated on one side, the cut will leave just a very small number of areas in the classes on the long leg of the distribution; as a result, most of the areas will be allocated to one or two colors. The use of percentiles for the definition of classes forces the allocation of the polygons in equal number of colors; that could masquerade significant differences in extreme values and hinder the identification of critical areas. Finally, the use of standard deviation, where the distribution of the variable is presented in different color gradations for values above and below the mean, supposes the normality of the variable distribution; such hypothesis is unrealistic in the case of census variables in countries of great social inequalities like Brazil. In short, it's an important part of the exploratory analysis to try different cut points in the variable for visualizing maps.

The different techniques of visualization are illustrated in the following example, where we show the spatial distribution of the indicator that measures the proportion of healthy newborns (APGAR Scoring) for the districts of Rio de Janeiro, in 1994.Two visualizations were generated, both with 5 cut points and 5 colors. In Figure 4 we utilized quintiles and in Figure 5, five classes of equal size. Since the variable distribution is not symmetric, when we divide it in classes of amplitude equal to lower (or worse) values, signaled in red, we are reduced to less classes, while in the division by quintiles, one fifth of the classes will be in each area. Then, the question is: what do we want to show? Certainly, the pre-natal care social worker of the region will not be happy seeing one fifth of the districts as being "high" risk areas. On the other side, since the areas where the score is lower have a small population, the reliability of the encountered values can be just an effect of the random fluctuation described before. Is it worth then looking at maps? Yes, of course, in the same way that we look at histograms and boxplots, and always trying to look at the distribution using different cut points. GISs have in general a standard form, but dozens of possibilities can and should be explored.

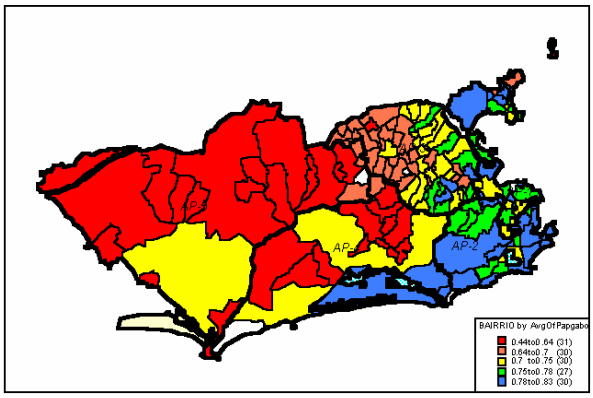


Figure 4– Distribution of the APGAR Score, grouped in quintiles.

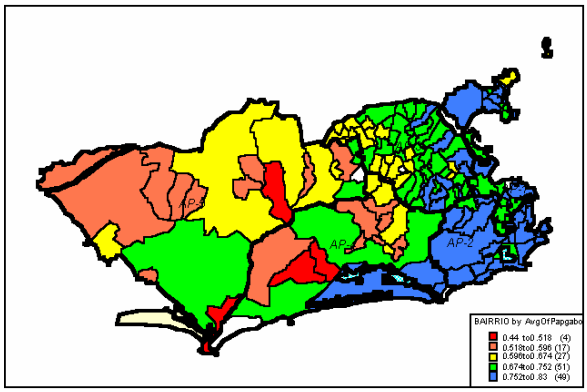


Figure 5 – Distribution of the APGAR Score, grouped in equal amplitude classes

Another interesting question is the comparison of maps. Suppose the spatial distribution of an indicator in different years: how can we visualize the temporal evolution? Certainly the cut points of the variable in the different periods must be the same. Observe in Figure 4 the temporal evolution of the mortality due to homicide for the triennial 79-81 and 90-92, in the state of Rio de Janeiro. The presentation of the quintiles for the joint distribution of the indicators allows a good visualization of the extension of this "disease".

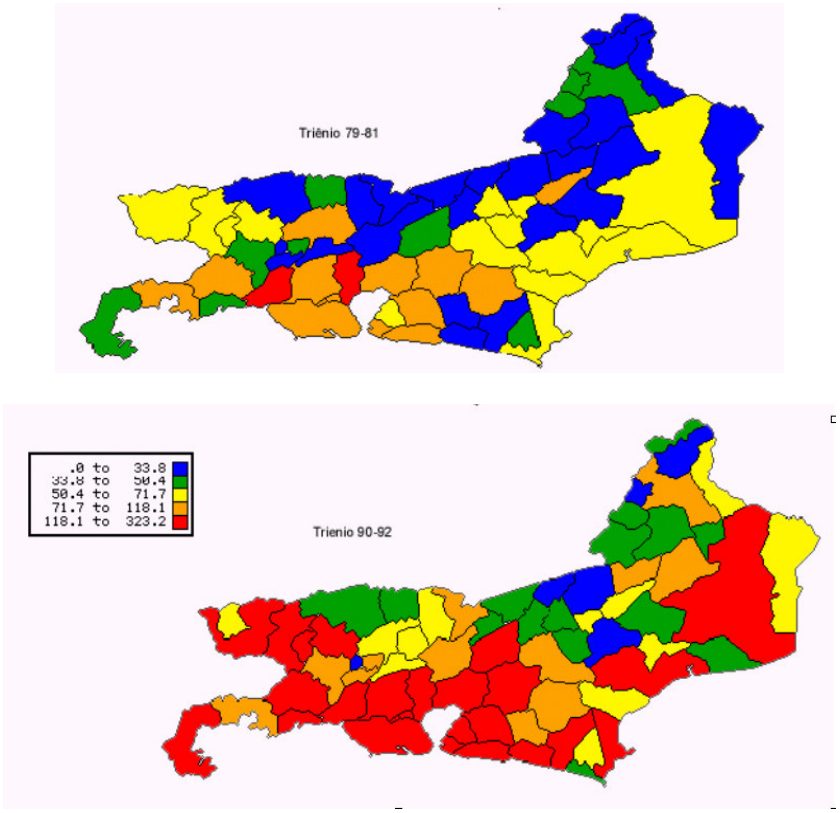


Figure 6 – Mortality due to homicides in Rio de Janeiro, for the triennials 79-81 and 80-92.

Plots of Means and Medians

The plots of means and medians in lines and columns allow the simultaneous exploration of the presence of a trend (first order nonstationarity), and second order nonstationarity, where the variance and covariance between neighbors is not kept constant. To build these plots, we utilize the coordinates of the centroids of the areas, approximating them to a regular spacing in order to mount a matrix. Then we calculate the means and medians of the indicator along the lines (East-West axis) and columns (North-South axis) of this matrix. This technique allows the identification of a fluctuation in the measures along two directions, suggesting the presence of discrepant values when the difference between them is large, and the trend along a direction when the values vary smoothly.

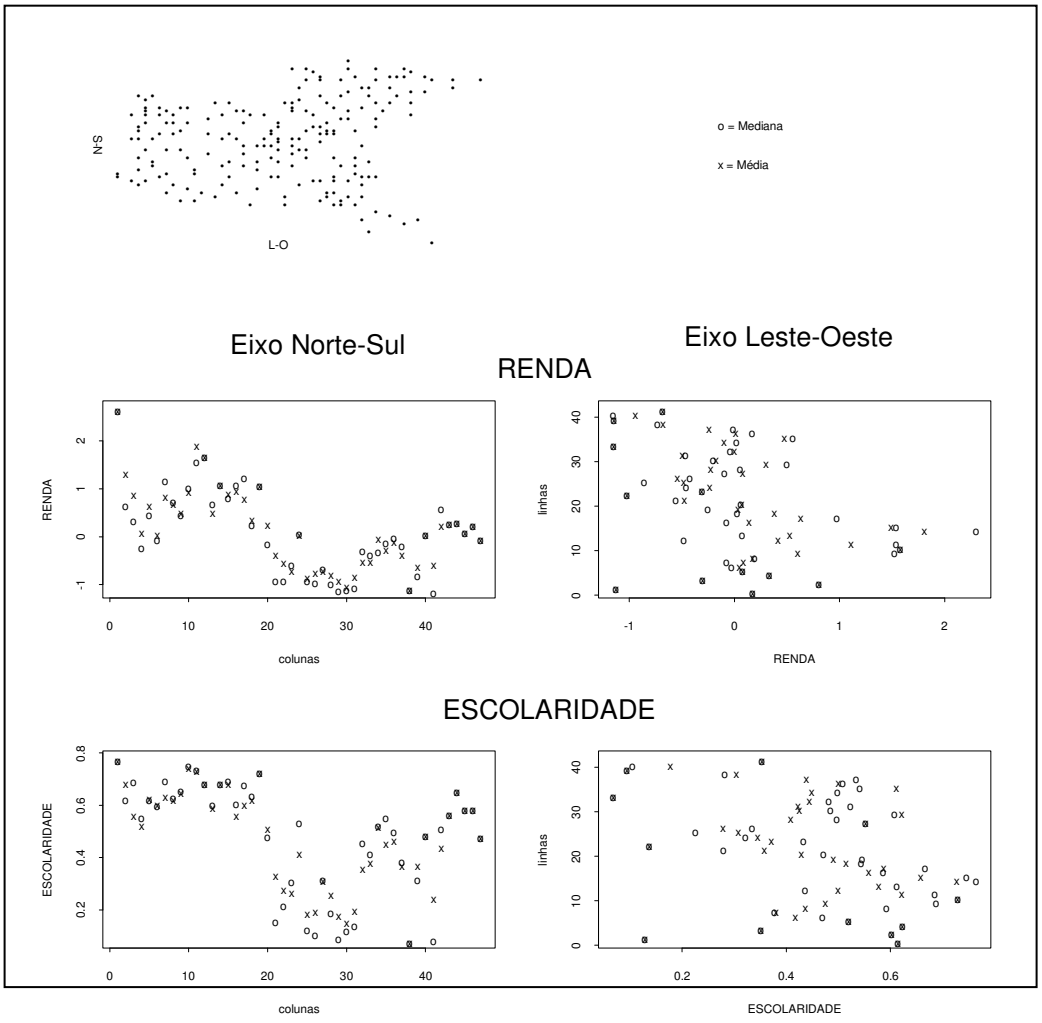


Figure 7 – Means and medians of the schooling and income at the Governador Island. (R.J.)

In Figure 7, we present the results of this technique, applied to two socioeconomic indicators of 1991 Census – the average income of the family head and the proportion of family heads with schooling greater than or equal to secondary school – for the census tracts at Governador Island, in Rio de Janeiro. This is composed of 225 census tracts, of which the centroids are indicated in the first chart of the figure: observe that in the extremities of the "map" the quantity of points is very small, and, consequently, any measures inside this area will not be very robust.

In the North-South axis (columns) we can notice that the average income of the family head presents a variable trend, much smaller in the center of the region. The same occurs with the schooling, although with greater fluctuation. In the East-West (lines) axis, there also seems to be some shifting to higher values towards the east, but the shift of means (x) and medians (o) suggest the presence of extreme values in the indicators. The variation in the mean of the indicators in the region is, apparently, divided between the two directions analyzed, and one can better explore these trends by rotating the reference axis.

Analysis of Spatial Autocorrelation

Another stage of exploratory analysis intends to identify the structure of the spatial correlation that better describes the data. The basic idea is to estimate the magnitude of the spatial autocorrelation between the areas. In this case, the tools utilized are the global Moran index, the Geary index and the variogram. When we have a great number of areas, resulting, for example, from detailed spatial scales, the nature of the processes involved is such that it is highly probable that there are different sorts of spatial correlation in different sub-regions. To illustrate these different sorts of spatial autocorrelation we can use the local indicators of spatial autocorrelation and the Moran scatter plot, also described in this section. All of these statistics depend on the adopted definition of neighborhood, discussed as follows.

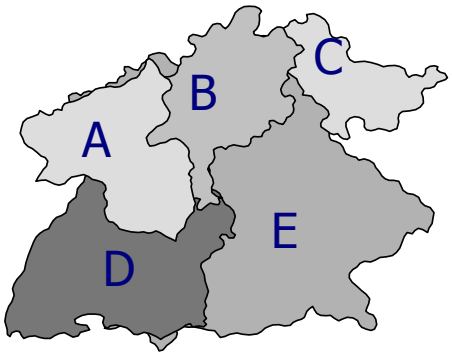
Spatial Proximity Matrices

To estimate the spatial variability of areal data a fundamental tool is the *spatial proximity matrix,* also called the neighborhood matrix. Given a set of n areas $\{A_1,\ldots A_n\}$,we build a matrix $W^{(1)}$ (n x n), where each element w_{ij} represents a measure of the proximity between A_i and A_j. This proximity measure can be calculated according to the following criteria.

- $w_{ij} = 1$ if the centroid of A_i is within some distance from A_j; on the contrary, $w_{ij} = 0$.
- $w_{ij} = 1$, if A_i shares a common side with A_j, on the contrary $w_{ij} = 0$.

- $w_{ij} = l_{ij} / l_i$, where l_{ij} is the length of the frontier between A_i and A_j and l_i is the perimeter of A_i.

Since the proximity matrix is utilized in the calculations of indicators during the exploratory analysis phase, it is very useful to normalize its lines, so that the sum of the weights of each line equals 1. This simplifies a lot many calculations of spatial correlation indexes, as we will see shortly. Figure 8 illustrates a simple example of spatial proximity matrix, where the values of the elements, that have been normalized, reflect the neighborhood criteria .



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Figure 8 – Spatial proximity matrix of first order, normalized by lines.

The idea of the spatial proximity matrix can be generalized to neighbors of higher order (the neighbors of the neighbors). With a criteria analogous to the one adopted for the first order spatial proximity matrix, one can construct the matrixes $\mathbf{W}^{(2)}$, …, $\mathbf{W}^{(n)}$. For example, in Figure 6, the areas A and C are neighbors in a second order spatial proximity matrix. In what follows, to simplify, the coefficients of the first order matrix are designated by w_{ij} and the ones from the order k matrix by $w_{ij}^{(k)}$ and the matrixes are normalized by lines.

Spatial Moving Averages

A simple form of exploring the variation in the data spatial trend is to calculate the mean of the neighbor's values. This reduces the spatial variability, for the operation tends to produce a surface with less fluctuation than the original data would. The moving average $\hat{\mu}_i$ associated to the attribute z_i, relative to the i-th area, can be calculated from the elements w_{ij} of the normalized spatial proximity matrix $\mathbf{W}^{(1)}$, simply taking the neighbors average:

$$\hat{\mu}_i = \sum_{j=1}^{n} w_{ij} z_i \qquad (1)$$

Figure 9 illustrates the use of the moving average estimator for the percentage of elders (more than 70 years old) in 96 districts of the city of São Paulo. These data are indicators of the great social inequalities in the city, with a great variation between downtown (where the proportion of elders reaches 8%) and the suburbs (where there are regions with less than 1%). The maximum value of elders' percentage is 8.2% and the minimum 0.8%, with a standard deviation of approximately 2%. With the local average there is a smoothing: the minimum value is 1% and the maximum is reduced to 6.8%. It can be noticed in the comparison of the two maps of Figure 9 that the local moving average provides a view of the great *trends* of the phenomena under study and, in the case of the percentage of elders, it shows a strong gradient downtown-suburbs.

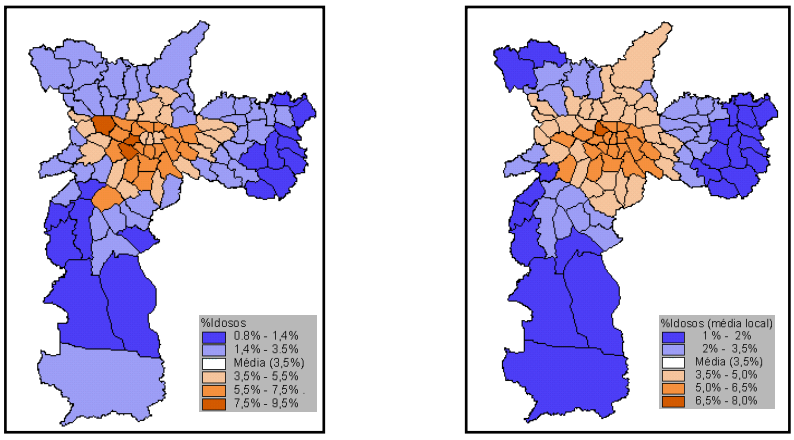


Figure 9 – Distribution of elders in the city of São Paulo (1991 Census). On the left, a presentation of the values by statistical distribution. On the right, the local moving average.

Global Indicators of Spatial Autocorrelation: Moran and Geary Indexes.

A fundamental aspect of the exploratory analysis is the characterization of the spatial dependency, showing how the values are correlated in space. Within this context, the functions used for estimating how much the observed value of an attribute is dependent on the values of this same variable in neighboring areas are the *spatial autocorrelation* and the *variogram*. Moran's global index *I* is the autocorrelation expression considering only the first neighbor.

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})} \qquad (2)$$

In the equation above, n is the number of areas, z_i the value of the attribute considered in area i, $\bar{z}$ is the mean value of the attribute in the region of study and w_{ij} the elements of the normalized spatial proximity matrix. In this case, the correlation will be computed only for the neighbors of first order in space, as established by the weights w_{ij}. The same calculation done for higher order proximity matrixes allows the estimation of the autocorrelation function for each order of neighborhood (or lag).

$$I^{(k)} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(k)}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{N}(z_i - \bar{z})} \qquad (3)$$

In general, Moran's index serves as a test where the null hypothesis is the spatial independence; in this case its value would be zero. Positive values (between 0 and +1) indicate a direct correlation, and negative values (between -1 and 0) an inverse correlation. Once calculated, it's important to establish its statistical validity. In other words, would the measured values represent a significant spatial correlation? To estimate the significance of the index it will be necessary to associate it to a statistical distribution, usually, the normal distribution. Another more usual approach, regardless the distribution, is a *pseudo-significance test*. In this case, different permutations of the attribute values associated to the regions are generated; each permutation produces a new spatial arrangement, where the values are redistributed among the areas. Since only one of the arrangements corresponds to the observed situation, one can build an empirical distribution of *I*, as shown in Figure 10. If the value of the index *I* originally measured corresponds to an "extreme" of the simulated distribution, then it will be a value of statistical significance.

In the case of the social inclusion/exclusion in São Paulo, shown in Figure 1, the global Moran's index measured is 0.642. A pseudo-distribution with 100 values is shown in Figure 10. In this case, the value of the significance associated is equal to 23, and that leads us to reject the null hypothesis (no correlation between the regions), with a significance of 99.5%. We can say that the social exclusion in São Paulo presents a strong spatial structure, partly a wide variation, or trend, partly spatial dependence among neighbors.

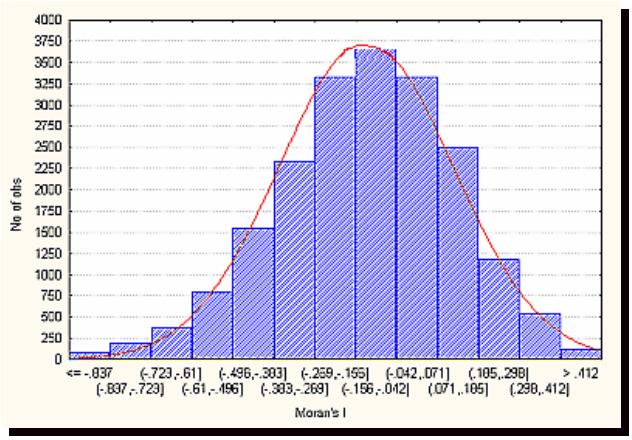


Figure 10 – Example of simulated distribution of the Moran index.

The implicit hypothesis of the calculation of the Moran index is the stationarity of first and second order, and the index looses its validity when calculated for non-stationary data. When there is a non-stationarity of first order (trend), the neighbors will tend to have closer values than the ones more distant because each value is compared to the global average, inflating the index. On the same way, if the variance is not constant, in places of higher variance the index will be lower, and vice-versa. When the data is non-stationary, the autocorrelation function continues to decay even after surpassing the distance where there are local influences. Some variations of this model are the Geary *C* test and the *Ipop* test. The first (Geary *C*) differs from the *I* test of Moran because it uses the difference between the pairs, while Moran uses the difference between each point and the global average. Thus, the *C* indicator of Geary resembles the variogram, while the *I* of Moran resembles the correlogram.

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n} z_i^{\,2}} \qquad (4)$$

The *Ipop* test is also used to detect deviations from a random spatial distribution, however it incorporates the variation of the population within the areas. It is thus sensitive to the occurrence of data clusters within the areas – that is, a high occurrence of cases in a small population of one municipality – besides the clusters among areas, where municipalities with many cases are adjacent. So the Ipop index can be decomposed in an intra-area and an inter-area component that can be presented as a percentage in the results. The null hypothesis (H_0) supposes that the geographical variation in the number of cases follows the geographical variation of the size of the

population, being particularly useful when the population of the areas is non-homogeneous.

$$Ipop = \frac{N^2 \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(e_i - d_i)(e_j - d_j) - N(1-2\bar{b})\sum_{i=1}^{m} w_{ij}e_i - N\bar{b}\sum_{i=1}^{m} w_{ij}d_i}{(X^2 \sum_{i=1}^{m}\sum_{j=1}^{m} d_i d_j w_{ij} - X\sum_{i=1}^{m} d_i w_{ij})\bar{b}(1-\bar{b})}$$

(5)

Where:

$m \rightarrow$ Number of areas

$N \rightarrow$ Total number of cases in all the areas.

$n_i \rightarrow$ Number of cases in area i

$e_i \rightarrow$ Proportion of cases in area i ($e_i = n_i/N$)

$X \rightarrow$ Total population in all the areas

$x_i \rightarrow$ Size of population in area i

$d_i \rightarrow$ Proportion of population in area *i* ($d_i = x_i/N$)

$Z_i \rightarrow$ Difference between the rate X_i and the average of X

$w_{ij} \rightarrow$ Assigned weights according to the connection between the areas i and j

$b \rightarrow$ Average Prevalence *(N/X)*

Table 2 presents the results of the tests for spatial clusters on the mortality due to homicides, in the state of Rio de Janeiro. Notice that the degree of significance of the *Ipop* test is greater than Moran's, and that approximately half of the aggregation is due to intra-municipality factors. That is, besides the nearby municipalities presenting similar patterns, an excess of cases exist within the violent municipalities that surpass the expected in comparison to the population.

TABLE 2

RESULTS OF TESTS FOR SPATIAL CLUSTERS

HOMICIDES IN RIO DE JANEIRO, 90-92

	Moran I	Ipop
Indicator	0.5861	0.00015
p-value	7.5091	88.9238
% between areas	-	54.3
% intra-areas	-	47

Variogram

We can use the variogram as indicator of the spatial dependence. To do that we associate the unique value of the attribute of each area to one point, usually the geometrical or populational center of the polygon. Based on these localizations, we calculate the variogram function. Notice that when the data is non-stationary, the variogram also does not stabilize, but keeps on increasing with the distance. As an example of the use of the variogram for areal data, Figure 11 illustrates the Human Development Index – HDI – for the state of São Paulo, calculated by IPEA, based on the 1991 census. Figure 12 presents the variogram of the HDI, computed from the centroid of each municipality.

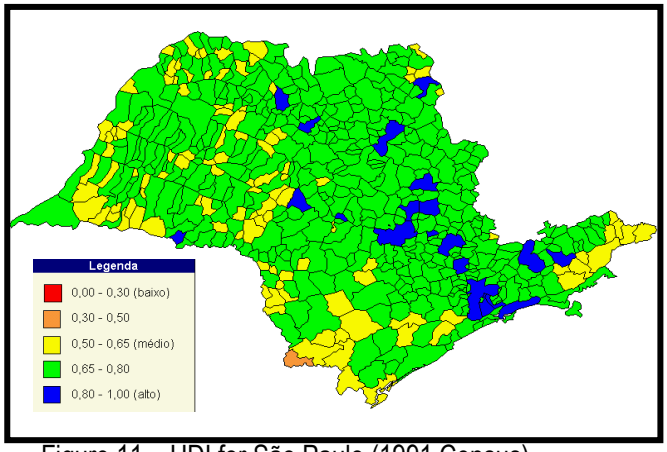


Figure 11 – HDI for São Paulo (1991 Census)

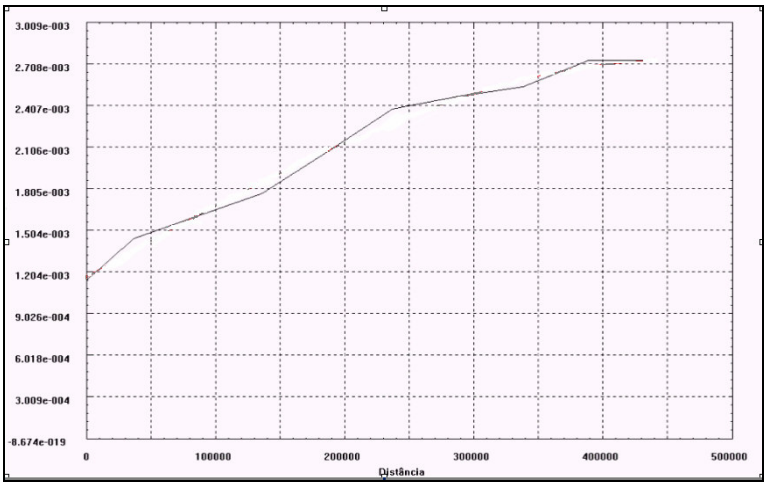


Figure 12 – Experimental variogram for the HDI of São Paulo (1991 census). Sample pass: 40 km (tolerance: 20 km).

What is shown in the variogram of Figure 11? On the X-axis, we present the distances between the municipalities, and on the Y-axis, the means of the square of the differences in the HDI, for municipalities separated by distance ranges, with 40-km intervals and 20 km tolerance. This way, the first point calculates the HDI difference between the municipalities whose distance between the centers falls between 20 and 60 km, and so on, up to a distance of 400 km. The graphic highlights a strong spatial dependency of the indicators of quality of life in the municipalities of São Paulo. This is a result of the occupation process of the state that followed regional perspectives. Starting with the logic of the expansion of the coffee plantations in the XIX century, we can observe today a region of strong farming production along the axis of the Anhanguera highway, the predominance of farming in the western region, and a strong industrial concentration in the São Paulo metropolitan region, in the ABC region and in the middle Paraiba Valley. Thus, all the historic processes point to a spatial dependence in the economic development of the state.

To consider a further example, take into account a study about the mortality due to homicide in the southeastern region, that are the cause of more than 20% of the deaths among men between 15 and 45 years old in Brazil. Figure 13 illustrates the spatial distribution of the mortality due to homicide, using as indicator the logarithm of the specific mortality coefficient per 100,000 residents within the same age group. Understanding the violence process as an "epidemic" of modern times that propagates in space, a simple visual observation allows us to identify a high incidence of violent deaths in the state of Rio de Janeiro (RJ), with a spatial trend capital-inland. In the case of the states of Espírito Santo (ES) and São Paulo (SP),

there is a concentration around the capital and big cities. However, in the state of Minas Gerais (MG), the most violent areas are located far from the capital and big cities, what indicates a distinct spatial pattern. Additionally, there is a well defined transition in the frontier between MG and RJ, indicating a change in the spreading conditions of the "violence epidemic". It's worth remembering that we used the logarithm of the indicators since its distribution is very concentrated around lower values, with a great tail to the right.

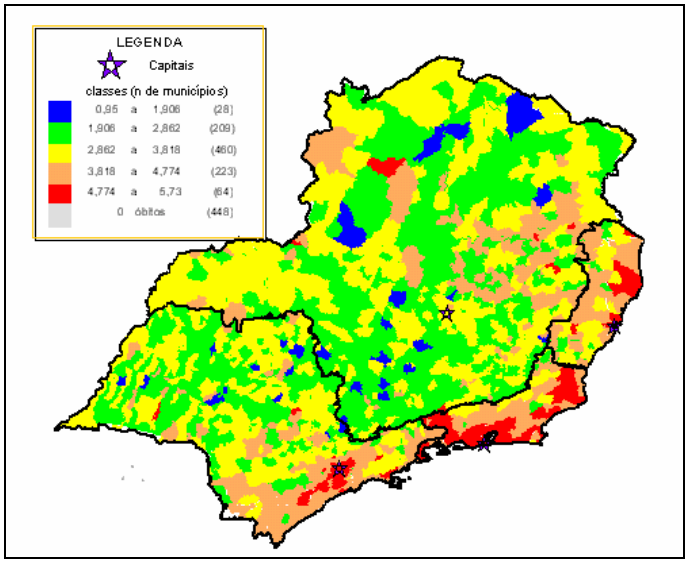


Figure 13 – Mortality due to homicide, Southeastern region of Brazil.

The correlogram of Figure 14 presents the spatial autocorrelation among the municipalities of each state, expressed in terms of the function defined by equation 3. The graphic indicates the existence of a strong spatial trend in RJ, for the autocorrelation function does not stabilizes with distance, but keeps on decreasing, on the contrary to MG, that does not present any clear spatial dependence. In other words, in RJ if the a municipality close by a city is violent, it's highly probable that this city is vilente too; the whole state presents a regionalized violence structure, and the violence decays in the inland. In MG this pattern is not observed: the violence seems to fluctuate randomly.

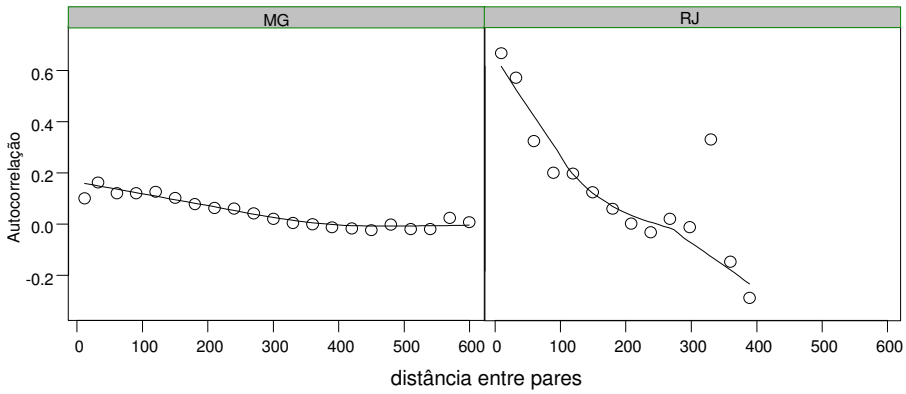


Figure 14 – Correlogram of the mortality due to homicide in the Southeast states .

The Moran Scater Plot

The Moran Scatter Plot is an additional tool for visualizing spatial dependence. If constructed upon normalized values (the attribute values subtracted of its average and divided by the standard deviation), it allows an analysis of the behavior of the spatial variability. The idea is to compare the normalized values of the attribute in an area with the average of their neighbors, constructing a bidimensional plot of *z* (normalized values) by *wz* (average of the neighbors), which is divided in four quadrants, as shown in Figure 15 for the social inclusion/exclusion index of São Paulo, 1991 Census. The quadrants can be interpreted as:

- Q1 (positive values, positive means) and Q2 (negative values, negative means): indicate points of positive spatial association, in the sense that a place has neighbors with similar values.
- Q3 (positive values, negative means) and Q4 (negative values, positive means): indicate points of negative spatial association, in the sense that a place has neighbors with distinct values.

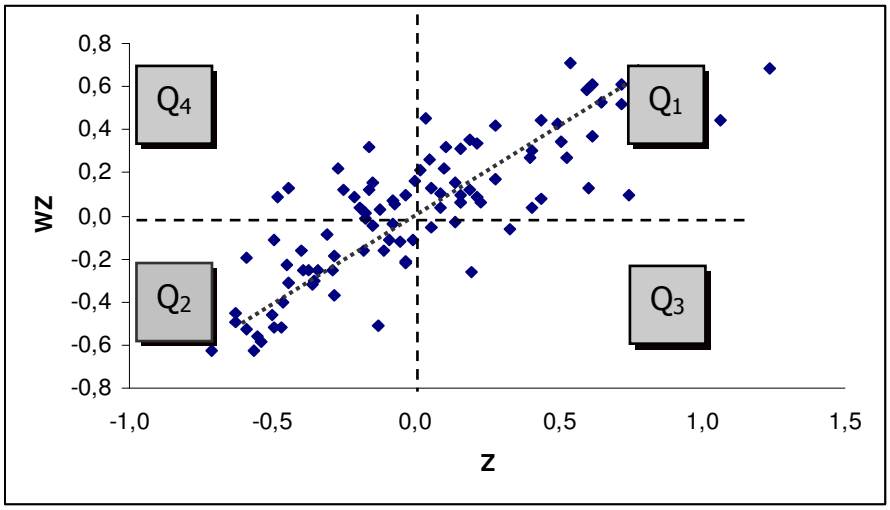


Figure 15 – Moran Scatter Plot for the social inclusion/exclusion index of São Paulo, 1991 Census.

The Moran Scatter Plot corroborates the results presented, where we indicate that Moran's global index for the social inclusion/exclusion indicator for the districts of São Paulo presented a statistically significant value. As shown in Figure 15 most of the districts of São Paulo are located in quadrants Q1 and Q2, that present a positive spatial association. The points located in quadrants Q3 and Q4 can be seen as regions that do not follow the same process of spatial dependence of the other observations. Evidently, the diagram reflects the spatial structure in the two scales of analysis: neighborhood and trend.

Moran's *I* index is equivalent to the linear regression coefficient that indicates the inclination of the regression line (α) of wz in z. In the case of the data presented in Figure 15, this coefficient is equal to 0.642, the same value calculated by applying the formula in equation 3. Moran's scatter plot can also be presented in the form of a bidimensional thematic map, where each polygon is presented by indicating its quadrant in the scatter plot, as illustrated in Figure 16, that shows the scatter plot of Moran's index for the social inclusion/exclusion index in the city of São Paulo in 1991. In this figure, "Alto-Alto", "Baixo-Baixo", "Alto-Baixo" and "Baixo-Alto" indicate quadrants Q1, Q2, Q3, and Q4, respectively, as shown in Figure 1 We observe a strong polarization downtown-suburb and observe that the districts in quadrants Q3 and Q4 (indicated in blue) can be understood as regions of transition between downtown (that tends to present positive values for the social inclusion/exclusion index) and the two great suburban areas of São Paulo (South zone and East zone).

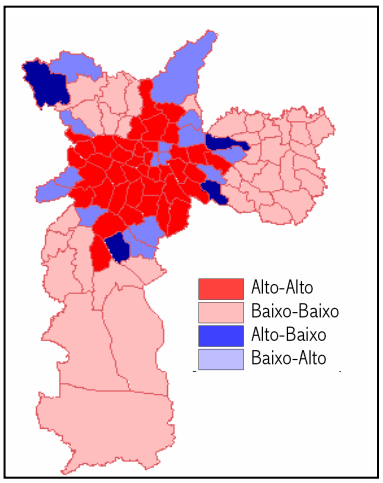


Figure 16 - Moran's scatter plot for the social inclusion/exclusion index of the city of São Paulo, 1991 census.

Local indicators of spatial association

The global indicators of spatial autocorrelation, such as Moran's index, provide a unique number as a measure of spatial association for the whole data set, which is useful for the characterization of the study area as a whole. When we deal with a great number of areas, it is very likely that different types of spatial association and that local maximums of spatial autocorrelation will appear where the spatial dependence is even stronger. Thus, many times it is desirable to examine these patterns in more carefully. To do that one needs to use indicators of spatial association that can be associated to the different localizations of a spatially distributed variable. The local indicators produce a specific value for each area, allowing the identification of groupings. Moran's local index can be expressed for each area i from the normalized z_i values of the attribute as:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2} \qquad (6)$$

The statistical significance of Moran's local index is computed similarly to the global index case. For each area, we calculate the local index and then randomly permute the values of the other areas until we obtain a pseudo-distribution that we can be used to compute the parameters of the significance. Once the statistical significance of Moran's local index has been determined it is useful to generate a map indicating the areas that present local correlation significantly different from the rest of the data. These regions can be viewed as non-stationarity "balloons" for they are areas with a specific spatial dynamics that deserve a detailed analysis. In the case of the

social inclusion/exclusion index of the city of São Paulo (1991 census), this map (Figure 17) shows clearly the aggregates of poverty and of wealth in the city. In the East and South zones of São Paulo there are critical regions, where the aggravation of the social conditions result in a significant degradation in life conditions.

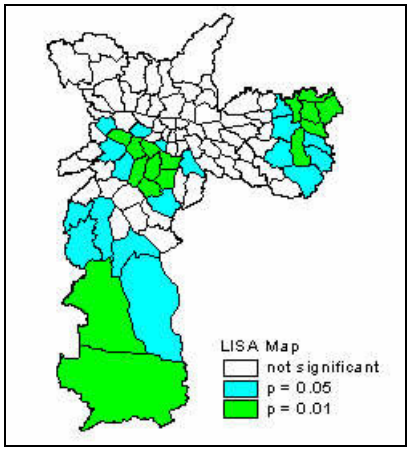


Figure 17 – Spatial autocorrelation indicator for the social inclusion/exclusion index for the city of São Paulo (1991 Census). Only the values with significance greater than 95% are shown.

ANALYSIS OF SMALL AREAS

We presented the problem of count aggregation in areas with the final recommendation of utilizing the best spatial resolution available. In practice, the use of such strategy requires an additional treatment of the data, especially in the case of small areas where we calculate the rates over a reduced population universe. To better understand the problem, consider Figure 18 that presents a thematic map of the infant mortality in the districts of Rio de Janeiro, in 1994. In this map, Rio de Janeiro has been divided into 148 districts, and the annual infant mortality rate for each district represents the number of deaths of children under one year, per 1,000 born alive.

It is shocking to read this map for the first time due to the high rates of mortality in various districts, with 15 of them presenting a rate higher than 40 deaths per 1,000 children born alive, and 2 cases with rates above 100 per thousand born alive. A reckless observer might conclude that all of these districts present a serious social problem. Actually, many of these extreme values occur in districts with reduced populations, for the city division used hides tremendous differences among the population at risk, varying from 75 up to 7,500 children per district. For example, consider a region with 15

children born and no deaths, what would apparently indicate an ideal situation. If only one child dies this year, the rate increase from 0 per 1,000 to 66 per 1,000!

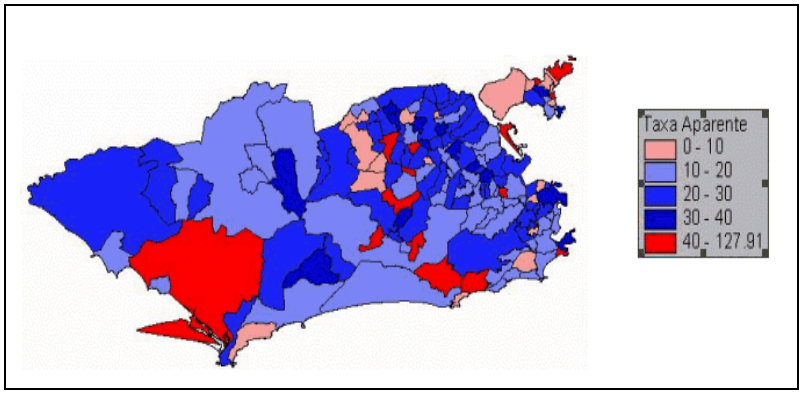


Figure 18 – Total infant mortality rate per each thousand born alive in Rio de Janeiro, in 1994.

Such problems are typical of the spatial partionings over political-administrative divisions, where areas with much different population at risk are analyzed. Various studies have shown that in political divisions like districts and municipalities present an inverse relation between area and population, that is, districts with greater population tend to have smaller areas, and vice-versa. For that reason, what frequently calls attention in a rate thematic map, the extreme values are many times the result of an extremely reduced number of observations, being thus less reliable, or just random fluctuation.

To smooth such random fluctuation, one considers that the estimated rate by simple division between the death count and the population – the observed rate – is just **one** realization of a non-observed process, and that it is less reliable the smaller the population is. Thus, we propose to re-estimate a rate that is closer to the real risk that a population is exposed to. The first step is to draw a graphic that expresses the rate as a function of the population at risk, as shown in Figure 19.

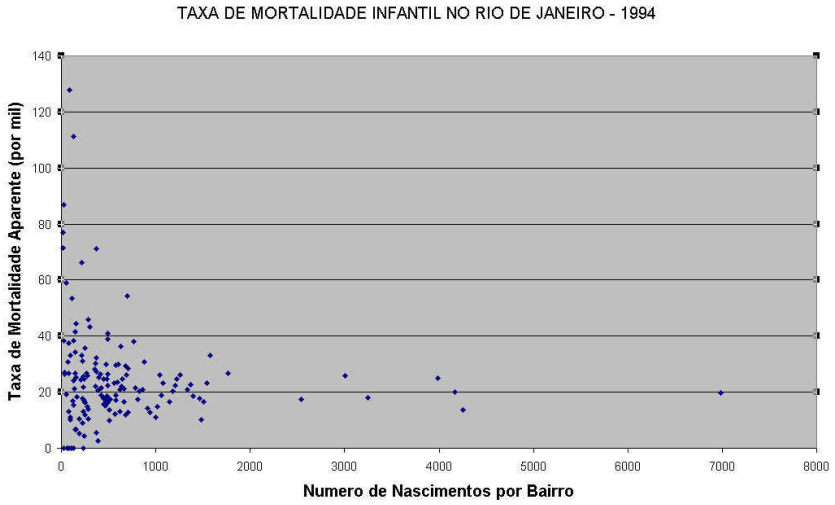


Figure 19 – Rate of infant mortality in Rio de Janeiro, in 1994, as a function of the number of births per district.

In the case of Rio the mean rate of infant mortality in the city, in 1994, was 21 deaths per thousand born alive. In the graphic we can observe that the districts with greater population present rates that are closer to the mean of the city. As the population at risk decreases, the fluctuation in the measured rate increases, forming what has been known as "funnel effect". In the districts with smaller population, such variation oscillated between 0 to almost 140 per thousand. It is reasonable to suppose that the rates in the different regions are autocorrelated, and to take into account the neighbor's behavior to estimate a more realistic rate for the regions of smaller population. Such formulation suggests the use of Bayes estimation techniques. In such a context, we consider that the "real" rate θ_i associated to each area is not known, and that we have an observed rate $t_i=z_i/n_i$, where n_i is the number of people observed while z_i is the number of events in the i-th area.

The idea behind the Bayes estimates is to suppose that the rate θ_i is a random variable, that has a mean μ_i and a variance σ_i^2. It can be demonstrated that the best Bayes estimates is given by a linear combination between the observed rate and the mean μ_i:

$$\hat{\theta}_i = w_i t_i + (1 - w_i)\mu_i \qquad (0.7)$$

the factor w_i is given by:

$$w_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu_i / n_i} \qquad (0.8)$$

The weight w_i is as small as smaller is the population under study in the i-th area and it reflects the degree of confidence with respect to each rate. In the case of reduced population, the confidence in the observed rate is

reduced and the rate estimation gets closer to our a priori model (that is, gets closer to μ). Regions with very small populations will have a greater correction, while large population areas will have little alteration in their rates. Thus, μ_i will be estimated, when n is small, with a greater weight for the neighboring mean.

At this point one must observe that the bayesian formulation requires the means and variances μ and σ_i^2 for each area. The simpler approach to treat the estimation of these parameters is the so called *empirical Bayes estimation*. This estimate is based on the hypothesis that the distribution of the random variable θ_i is the same for all the areas; this implies that all means and variances are equal. We can thus estimate μ and σ_i^2 directly from the data. In this case, we calculate μ_i from the observed rates:

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \qquad (9)$$

And we estimate the variance σ_i^2 from the variance of the observed rates in relation to the estimated mean:

$$\sigma^2 = \frac{\sum n_i (t_i - \hat{\mu})^2}{\sum n_i} - \frac{\hat{\mu}}{\overline{n}} \qquad (0.10)$$

The regions will have its rates re-estimated by applying a weighted mean between the measured value and the global mean rate, where the weight of the mean will be proportionally inverse to the population of the region. When we apply this correction to the infant mortality rate of Rio de Janeiro, we observe that there is a significant reduction in the extreme values. For example, the Cidade Universitária (Fundão Island), where 13 children were born in 1994, presented an apparent rate of 76 per thousand born alive and a corrected rate of 36 per thousand. Quarters with small population in the risk group presented similar reductions, while the most populated quarters kept the rates originally measured. The comparison between the primary rate and the estimated value is presented in Figure 18. In short, extreme care is needed when elaborating thematic maps, especially in cases where we present rates measured over small populations.

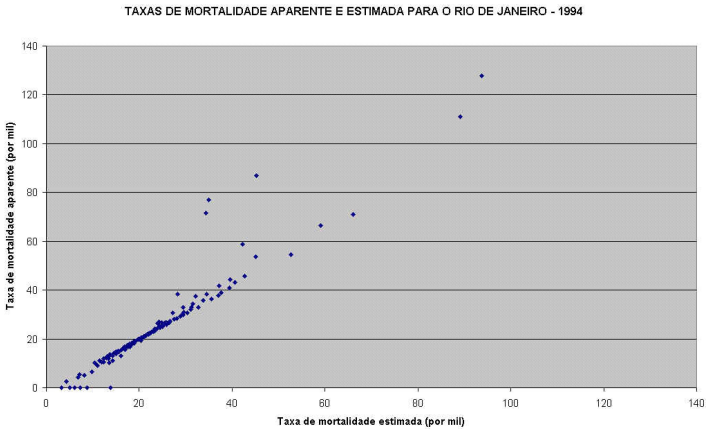


Figure 20 – Comparison between the observed infant mortality rate and the estimated rate using empirical Bayes method.

The empirical Bayes estimation can be generalized to include spatial effects. In such case, the idea is to make the Bayes estimation locally, converging to the direction of a local mean and not toa global mean. It is sufficient to apply the described method for each area considering its neighborhood as the "region". This is equivalent to assuming that the rates of the neighborhood of area *i* have common mean μ_i and variance σ_i^2. In this case, we are talking about a *local empirical Bayes estimate*. Next, we will present the detection of Hansen's disease in Recife (Figure 20) where we have used this local method to estimate the rate of the disease in the city quarters. Through the "corrected" map it was possible to indicate the priority quarters for the epidemiologic surveillance to act due to the high values they presented, even after the smoothing of the indicator.

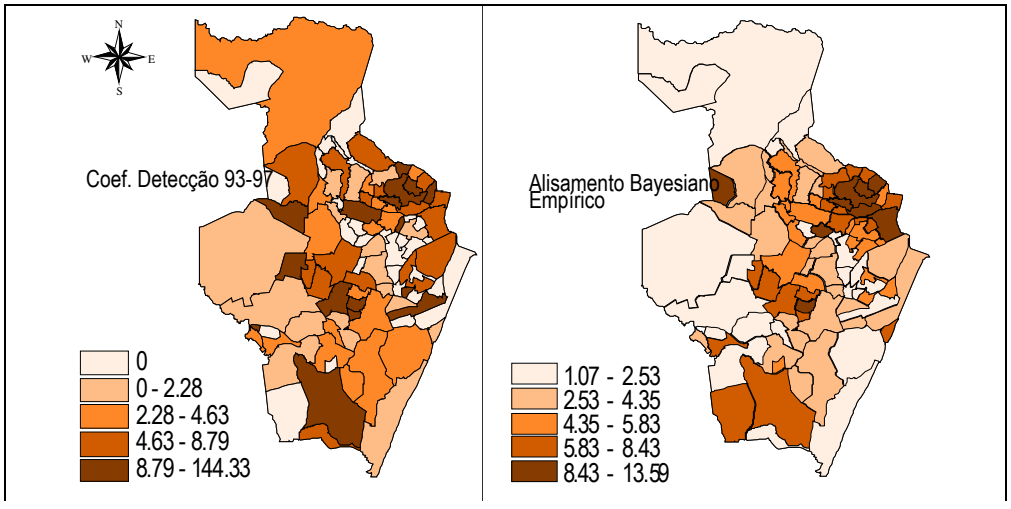


Figure 20 – Mean rates of detection of Hansen's disease among children under-15, 1993-1997 period, per quarter in Recife, with rates estimated through Bayesian smoothing.

As shown above, the *empirical Bayes estimate* is based on the hypothesis that the distribution of the random variable θ_i is the same for all the areas and that the mean θ_i and variance θ_i^2 for each area are equal. We must bear in mind that this hypothesis is not always realistic, for in social-economic statistics (as in the case of the discussed health data) the characteristics of the studied population are very heterogeneous. This way, in many cases it is desirable to make the hypothesis that each area has its own pattern (and the θ_i and θ_i^2 are distinct); and that implies in estimating the joint distribution $Z=\{Z_1, \ldots, Z_n\}$ of the random variables.

At first sight the estimate of the joint distribution may seem impossible, since we only have for analysis a sample of each random variable, that is, we only know the value collected in each unit of area. Nevertheless, the *full Bayes estimates* made it possible to solve the problem, through the utilization of simulation techniques based on MCMC – *Markov Chain Monte Carlo* – for the inference of the parameters of interest. Due to the complexity of the formulation, this book does not describe the Bayes estimates based on MCMC. The reader should refer to the bibliography in the end of the tutorial for more details.

REGRESSION MODELS

One of the most common studies with areal data lies in the use of regression models. A regression model is a statistical tool that utilizes the existing relationship between two or more variables in a way that one of them may be described or its value estimated from the others. In spatial data, when spatial autocorrelation is present, the estimates of the model must incorporate this spatial structure, since the dependence between each observation alters the explanatory power of the model. The significance of the parameters is usually overestimated, and the existence of large scale variations can even motivate the presence of spurious associations.

In this book, we won't present a detailed description of the traditional regression models, for they are available in various books, but we will only present a short description, necessary to the understanding of the spatial regression models. The general objective of a linear regression analysis is to quantify the linear relationship between a dependent variable and a set of independent variables, as expressed in the matrix equation:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \varepsilon, \ \varepsilon \sim \mathrm{N}(0, \sigma^2) \text{ ou} \qquad (11)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ .. \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & .. & X_{1k-1} \\ 1 & X_{21} & .. & X_{2k-1} \\ .. & .. & .. & .. \\ .. & .. & .. & .. \\ 1 & X_{n1} & .. & X_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ .. \\ .. \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .. \\ .. \\ \varepsilon_n \end{bmatrix} \qquad (12)$$

where Y is the dependent variable, composed by a (n x 1) vector of observations taken in each of the n areas, X is a (n x k) matrix with k-1 independent variables also taken in the n areas, $\boldsymbol{\beta}$ is a vector (k x 1) with the regression coefficients, and ε is a vector (n x 1) of random errors, or residuals.

Typically, when we make a regression analysis, we try to reach two objectives: (a) find a good adjustment between the values predicted by the model and the observed values of the dependent variable; (b) find which independent variables contribute in a significative way to this linear relationship. To achieve that the standard hypothesis is that the observations are not correlated, and consequently that the residuals ε_i of the model are also independent and uncorrelated with the dependent variable, have a constant variance, and present a normal distribution with zero mean.

However, in the case of spatial data, where spatial dependence is present, it is very unlikely that the standard hypothesis of uncorrelated observations is true. In the most common case, the residuals keep presenting the spatial correlation present in the data, and that could show up as systematic regional differences in the model relations, or even by a continuous spatial trend.

The investigation of the regression residuals in the search of signs of a spatial structure is the first step in a spatial regression. The usual graphical analysis tools and the residual mapping can provide the first indications that the observed values are more correlated than would be expected under a condition of independence. In this case, using the spatial correlation tests – Moran and Geary – on the regression residuals warn of its presence. If the autocorrelation exists we must specify a model that takes into consideration the interference caused by it.

In the rest of this section, we present various types of regression models that take into account the spatial effects, starting with the spatial structure in a global way (as the only parameter) until models that vary continuously in space.

Models with Global Spatial Effects

The explicit inclusion of spatial effects in regression models can be done in different ways. The simplest class of spatial regression models, called *global spatial effects models,* assume that it is possible to capture the structure of the spatial autocorrelation in one parameter only, that is added to the traditional regression model. In this case, we have two alternatives to treat the global autocorrelation in a regression model. First, the ignored spatial autocorrelation is attributed to the dependent variable *Y*. This approach is denominated *simultaneous autoregressive model (SAR)* or *spatial lag model*, since the spatial dependence is taken into account by the addition to the regression model of a new term in the form of a spatial relation for the dependent variable. Formally this is expressed as:

$$\boldsymbol{Y} = \rho \boldsymbol{WY} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{13}$$

Where *W* is the spatial proximity matrix, and the product *WY* expresses the spatial dependence in *Y* and ρ is the *autoregressive spatial coefficient*. The null hypothesis of the nonexistence of autocorrelation is that $\rho = 0$. The basic idea with this model is to incorporate the spatial autocorrelation as a component of the model. In terms of individual components, this model can be expressed as:

$$y_i = \rho \left(\sum_j w_{ij} y_j \right) + \sum_{i=1} x_i \beta_i + \varepsilon_i \tag{14}$$

The second type of spatial regression model with global parameters considers that the spatial effects are a noise, or perturbation, that is, a factor that needs to be removed. In this case, the effects of spatial autocorrelation are associated to the error term ε and the model can be expressed as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} = \lambda \boldsymbol{W} + \xi, \tag{15}$$

Where *W*ε is the error component with spatial effects, λ is the autoregressive coefficient and ξ is the uncorrelated error component with constant variance. The null hypothesis for the nonexistence of autocorrelation is that $\lambda = 0$, that is, the error term is spatially uncorrelated. Such model is also known as *spatial error model* or *conditional autoregressive model* – CAR.

From equation 15 we can show that the spatial error model can also be expressed as:

$$\boldsymbol{Y} - \lambda \boldsymbol{WY} = \boldsymbol{X\beta} - \lambda \boldsymbol{WX\beta} + \xi \tag{16}$$

Or even as:

$$(\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{Y} = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{X\beta} + \xi \tag{17}$$

what can be seen as a non-spatial regression in the "filtered" variables.

$$\boldsymbol{Y}^{*} = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{Y} \ , \quad \boldsymbol{X}^{*} = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{X} \tag{18}$$

In practice, the distinction between both types of spatial regression models with global parameters is difficult for, despite the difference in their motivation, they are very close in formal ways. These models are included in advanced spatial statistics environment, like the software SpaceSat™, S-Plus ™ and R, this one in the public domain. In the references in the end of this tutorial, the reader will find indications about how such models can be estimated and about the hypothesis about its behavior.

The spatial regression models with global effects are based on the principle that the underlying spatial process on the analyzed data is stationary. This means that the spatial autocorrelation patterns of the data can be captured in on parameter only. In practice, for census data sets of medium to large scale, the nature of the spatial processes is such that different patterns of spatial association can be present. This hypothesis, that can be verified, for example, by the local indicators of spatial autocorrelation, is in the origin of the models whose parameters vary in space, and are discussed as follows.

Regression Models with Local Spatial Effects

***(a)** Discrete Choice – Regression Models with Spatial Regimes*

When the spatial process is non-stationary the coefficients of regression need to reflect the spatial heterogeneity. To do that we have two alternatives: (a) to model the spatial trend in a continuous way, with parameters that vary in space; (b) to model the spatial variation in a discrete way, by dividing the space in stationary sub-regions, called *spatial regimes*.

The idea of spatial regimes is to divide the region of study in sub-regions, each one with its own spatial pattern, and to make separate regressions, one for each area. The observations are classified in two or more subsets, starting from an indicated variable, that is:

$$Y_1 = X_1\beta_1 + \varepsilon_1 \ , \ ind = 1 \tag{19}$$

$$Y_2 = X_2\beta_2 + \varepsilon_2 \ , \ ind = 2 \tag{20}$$

Although each regime has its own coefficient values, these values are estimated together, that is, the set of all the observations is used in the regression. For the determination of the spatial regimes, the techniques of exploratory analysis are very useful, especially the Moran scatter plot and the local indicators of spatial autocorrelation.

In practice, for the typical socioeconomic data of Brazilian cities, the model of spatial regimes tend to present better results than the models of

simple regression or spatial regression with global effects. This occurs due to the strong social inequalities in Brazil, that provoke sharp discontinuities in the studied phenomena, as in the case of the cut off between the slums and the rich area that frequently occur in Brazilian big cities.

Regression Models with Local Spatial Effects.

***(b)** Regression Models with Continuous Spatial Effects*

This class of models try to model non-stationary phenomena. Differently from the spatial regimes model, the spatial effects are modeled in a continuous way, with two hypothesis: (a) the existence of a smooth large-scale variation, without significant local effects or (b) the existence of continuous local variations, without a strong global trend. The first case corresponds to the trend surfaces. The *trends surface* model considers a spatial process where the value of the variable is a polynomial function of its position in space. The multiple regression model using vectorial notation is:

$$\boldsymbol{Y}(s) = \boldsymbol{X}(s)\beta + \varepsilon(s) \qquad (21)$$

Where, $Y(s) \rightarrow$ random variable representing the process at point s,

$X(s)\beta \rightarrow$ trend (that is, the mean value $\mu(s)$),

$\varepsilon(s) \rightarrow$ random error with zero mean and variance σ^2

Vector $x(s)$ consists of p functions of the spatial coordinates (s_1, s_2), of the sampled point s. For a surface of linear trend it is only $(1, s_1, s_2)$, for a quadratic one it is $(1, s_1, s_2, s_1^2, s_2^2, s_1.s_2)$, and so on. β is the vector $(p+1)$ of parameters to be adjusted. The basic presupposition of such model assumes that the errors have constant variance and are independent in each place, consequently, the covariance is zero: there are no second-order effects present in the process. In this context a model adjustment using ordinary least square is done. The trend surfaces model is useful mainly as a first approximation of the phenomena for, in practice, the cases where the spatial variation can be expressed in this way are limited. However, the residuals of these models are much informative about the nature of the local variations.

In the case of model of continuous local variations the idea is to adjust a regression model at each point observed, weighting all the other observations as a function of the distance from this point. This way, as many adjustments will be made as there are observations and the result will be a set of parameters, being that each point considered will have its own coefficients of adjustment. These parameters can be presented visually in order to identify how they behave spatially and the relationships between the variables. This technique is called *geographically weighted regression* (GWR). To apply the model GWR, the standard regression model is rewritten as:

$$Y(s) = \beta(s)X + \varepsilon, \tag{22}$$

Where ***Y(s)*** is the random variable representing the process at point ***s***, and ***β(s)*** indicates that the parameters estimated at point ***s***. To estimate the parameters of this model, the standard least squares solution for the non-spatial case that is given by:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{23}$$

is generalized using a method for the local adjustment:

$$\beta(s) = (\boldsymbol{X}^T W(s) \boldsymbol{X})^{-1} \boldsymbol{X}^T W(s) \boldsymbol{Y} \tag{24}$$

The local adjustment is done in way to guarantee a greater influence of the closer points, in a way similar to the kernel density estimators. An example is the use of a Gaussian function of the type:

$$w_{ij}(s,\tau) = \frac{1}{2\pi\tau} \exp\left(-\frac{{d_{ij}}^2}{2\tau^2} \right) \tag{25}$$

where τ represents the considered influence radius, and d_{ij} is the distance between the considered position and the *j-th* point. One can run hypothesis tests to verify whether the spatial variations have statistical significance or are random. For more details about the GWR model, the reader should refer to the bibliography at the end of the tutorial.

Diagnostics of Spatial Effects Models

The graphical analysis of the residuals is the first step to evaluate the quality of the regression adjustment. Mapping the residuals is an important stage in the diagnostic of the model, searching for signs of rupture in the presuppositions of independence. A high concentration of positive (or negative) residuals in a part of the map is a good indication for the presence of spatial autocorrelation. For a quantitative test, the most common is to utilize Moran's I index over the residuals.

Since the estimators and the traditional regression diagnosis do not take into account the spatial effects, the inferences, like for example the indications of quality of the adjustment based in R^2 (determination coefficient), will be incorrect. These consequences are similar to the ones that happen when a significant independent variable is omitted from the regression model. When one wishes to compare an adjustment obtained from a standard regression model with an adjustment obtained from one of the models whose specifications take into account the spatial autocorrelation, a measure like R2 is not reliable anymore.

The most usual method for the selection of regression models is based upon the values of *maximum likelihood* of the different models, weighting

the difference in the number of parameters estimated. In models with a dependence structure – spatial or temporal – one utilizes the *information criteria* where the evaluation of the adjustment is penalized by a function of the number of parameters. It's worth observing that it is necessary to take into consideration the number of independent parameters when including spatial functions in the models. For each new variable in a regression model a parameter is added. Usually the comparison of the models is done using the logarithm of the maximum likelihood, which has the best adjustment for the observed data. Akaike's information criteria (AIC) is expressed as:

$$AIC = -2 * LIK + 2k \qquad (26)$$

where *LIK* is the maximized likelihood log and *k* is the number of regression coefficients. According to this criteria, the best model is the one that has the smaller AIC value. Various other information criteria are available, most of them variations of AIC, with changes in the form of penalization of parameters or observations.

Illustrative example

As an illustrative example of the spatial regression techniques, we studied the relationship between income and longevity in the city of São Paulo, with data from the 1991 census. We're talking about two or three variables that compose the United Nation's HDI (Human Development Indicator). The dependent variable to be explained is named PERIDOSO (percentage of people older than 70 per district of São Paulo) and the independent variable is indicated by PERREN20 (percentage of family heads with income above 20 minimum wages a month). The spatial distribution of these variables is shown in Figure 21.

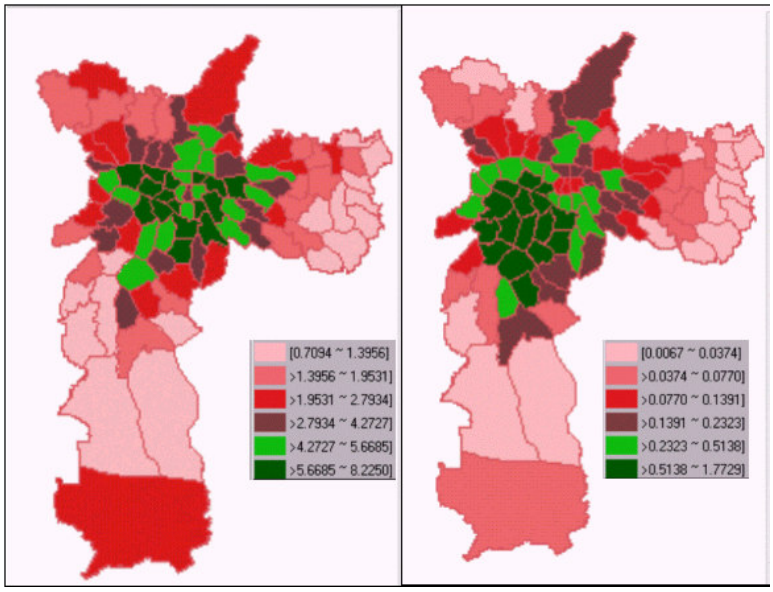


Figure 21 – Percentage of older (left) and family heads with income higher than 20 minimum wages a month (right) for the districts of São Paulo (1991).

Three regression models were compared: the non-spatial standard model, the spatial lag model, and the spatial regimes model. In the case of the spatial regimes three city regions were considered (downtown, suburbs, and the transition downtown-suburbs). The standard model is expressed as:

$$\text{PERIDOSO} = \beta_0 + \beta_1 \text{PERREN20} + \varepsilon \qquad (0.\ 27)$$

Using the neighborhood matrix W of the districts, the spatial lag model can be expressed as:

$$\text{PERIDOSO} = \beta_0 + \beta_1 \text{PERREN20} + \rho W(\text{PERIDOSO}) + \varepsilon \qquad (0.\ 28)$$

Using the neighborhood matrix W of the districts, the spatial lag model can be expressed as:

$$\text{PERIDOSO\_1} = \beta^1_0 + \beta^1_1 \text{PERREN20\_1},\ \text{reg}=1 \qquad (29)$$

$$\text{PERIDOSO\_2} = \beta^2_0 + \beta^2_1 \text{PERREN20\_2},\ \text{reg}=2 \qquad (30)$$

$$\text{PERIDOSO\_3} = \beta^3_0 + \beta^3_1 \text{PERREN20\_3},\ \text{reg}=3 \qquad (31)$$

The results of these regression models are presented in Table 3. In the traditional regression model, the relationship between income and longevity in São Paulo is very reduced, what gives support to the HDI idea that they are complementary dimensions of human development. However, when spatial effects are taken into account, we verify that there is a real dependence between both factors. In Figure 22, we present the spatial distribution of the regression residuals for the least squares and spatial lag models. A visual analysis of the residuals for the traditional regression models indicate a prevalence of positive residuals in downtown and negative

residuals in the suburbs, especially in the East and South zones. The numerical results confirm this analysis, for Moran's index for the residuals is highly significant. Concerning the global performance, the R^2 measures are limited indicators and should be dealt with care, and we should prefer the measures based on likelihood (LIK, AIC). In such case, the spatial lag model had a much better performance than the standard model. This effect is expected, because there is a significant Moran's index in the residuals, which is captured by the spatial effect coefficient (ρ).

The spatial regimes chosen for São Paulo are shown in Figure 23, together with the regression residuals considering these regimes. From the visual analysis of the residuals, we verify the non-existence of a strong spatial trend, which is evidenced by their low Moran's index, as indicated in Table 3. In general, the spatial regime model presented a better performance, by any criteria (R^2, LIK, AIC). The result reflects the strong polarization downtown-suburbs in the city of São Paulo, and that is compatible with studies that show the results of urban violence in the mortality rates, especially among men aged from 15 to 25.

Table 3 -Results of the Regression for Longevity and Income in São Paulo, 1991.

	Least Squares	Spatial Lag	Spatial Regimes
Adjusted R^2	0,280	0,586	0,80
Likelihood log	-187,92	-150,02	-124,04
AIC	379,84	306,51	260,09
Moran's index for residuals	0,620	-	0,020

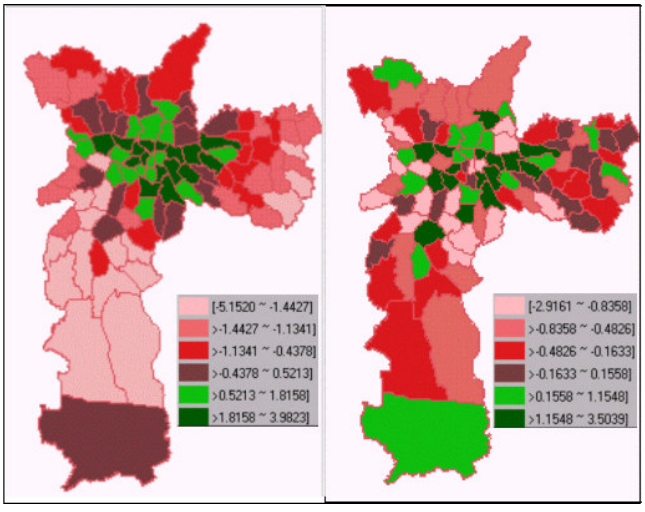


Figure 22 –Least squares regression residuals (left) and spatial lag regression residuals (right).

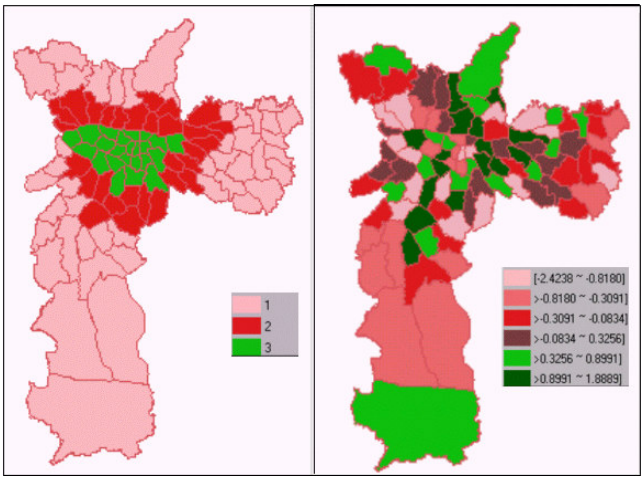


Figure 23 – Spatial regimes for the districts of São Paulo (left) and spatial regimes regression residuals (right).

CONTINUOUS MODEL ESTIMATION FROM AREAL DATA

The previous sections presented techniques for spatial analysis of areal data based on the model of *discrete spatial variation* , where each area is modeled by respecting its boundaries, surroundings and neighborhood. In this section we consider the *continuous space variation* model, that assumes a stochastic process $\{Z(x), x \in A, A \subset \Re^2\}$, whose values can be known in every point of the study area. The idea of continuous models for socioeconomic data stems from the fact that the censitary research many times impose area limits due to exclusively operational reasons, that doesn't have any relation to the modeled phenomena. This fact leads to the idea of dissolving the area limits in continuous surfaces, as a way of better modeling the real continuity of, for example, censitary sectors into densely populated urban areas.

In the case of surface estimators, the main options are the use of non-parametric techniques and the use of geostatistical interpolators.

Non-parametric intensity estimator

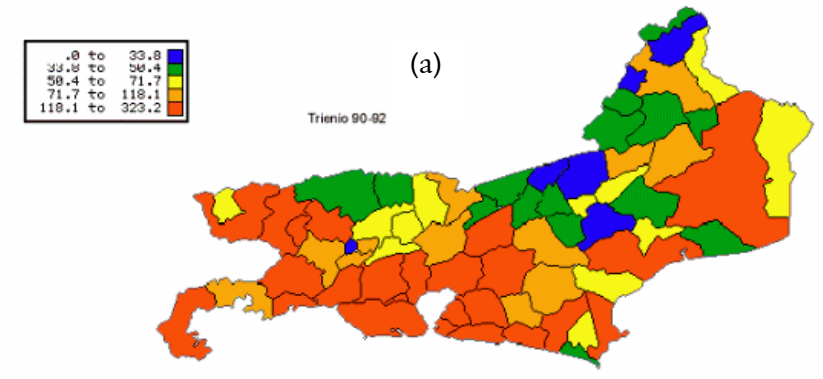
Similarly to the surfaces case, we can use the intensity estimator (kernel estimator) to provide us a first approximation of the spatial distribution of the phenomena or variable. In this case, when the observed values represent an "average" measurement like the mortality rate or percapita income, we can utilize an estimator that would allow us to calculate the attribute value per unit of area. For every position ***(x;y)*** which value we want to estimate, the intensity estimator will be computed from the values $\{z_1, z_2, .., z_n\}$ contained within a radius of length τ, according to the equation:

$$\hat{z}_i = \frac{\sum_{j=1}^{n} k(\frac{d_{ij}}{\tau}) z_j}{\sum_{j=1}^{n} k(\frac{d_{ij}}{\tau})}, \; d_{ij} \leq \tau \qquad (32)$$

In the equation above, function *k()* is a non-parametric interpolator, that could be, for example, a Gaussian kernel, where the reader may find a more detailed discussion about non-parametric intensity estimators. An example of the intensity estimator for rates can be seen in Figure 22, where data for mortality due to homicide are presented for the state of Rio de Janeiro, for the years 1990-1992 interpolated with the intensity estimator, which gives us an idea of the spatial distribution of the variable under study. In Figure 24(a) a map is presented with the values of the indicators of the mortality rate, grouped by municipality. In Figure 24(b), we present the results of the intensity estimator, that gives us a better idea of the spatial distribution of the studied variable.

When the observations in the areas represent counts, like the ones obtained in the census, the kernel estimator presented above is not appropriate. An "average" value of an attribute like "number of poor households" would make no sense, and one must think in terms of "number of poor households per unit of area". In this case, we can use the numerator of equation (32), divided by the area of the circle defined by search radius:

$$\hat{z}_i = \frac{1}{\pi\tau^2} \sum_{j=1}^{n} k(\frac{d_{ij}}{\tau}) z_j \;, \; d_{ij} \leq \tau \qquad (33)$$



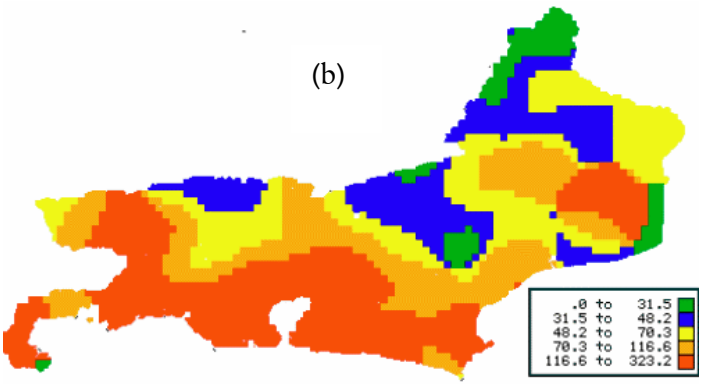


Figure 24 – (a) Mortality due to homicide in RJ (190-1992). Thematic map with values per municipality. (b) Surface obtained with non-parametric intensity estimator.

The use of geostatistical interpolators

The otraditional motivation f geostatistics is associated with physical data like mineral content or pollution rate. In the case of ordinary kriging, the underlying hypothesis is that the data present a Gaussian distribution, and in such case the optimal properties of the estimators (like the minimum variance of the result) are guaranteed. In the case of socioeconomic or public health data, the hypothesis of normality of the data is seldom realistic, being more common the assumption that the distribution is Poisson, for we are dealing with event counts. However, the optimal properties of the kriging estimator and its wide availability in different geographic information systems make it important to investigate its usefulness for socioeconomic data. In this case, the first step is to investigate how close to the normal distribution the data is; if necessary, appropriate transformations (like the logarithmic transformation) can be applied to "symmetrize" the empirical distribution and thus bring it closer to the normal one. To consider a concrete situation, Figure 25 presents the distribution of the homicide rate per 100 thousand inhabitants, for 96 districts of São Paulo in 1996, follwed by the normal probability graphic that indicates how much these data are close to a Gaussian distribution. From the analysis of both data, and considering that the mean (43,6) is sufficiently close to the median (39,3), and since the Shapiro-Wilk normality test indicates a value of 0,9653 (p-value of 0,012), the hypothesis of normality cannot be rejected and allow us to apply a kriging interpolator.

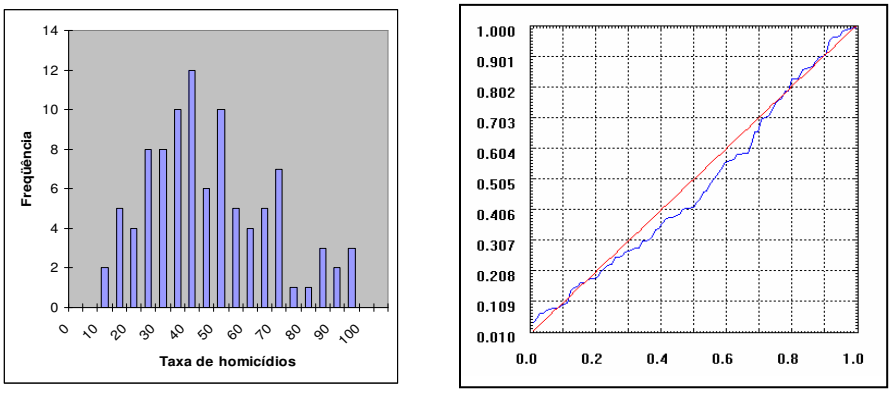


Figure 25 – Homicide rate per 100 thousand inhabitants for São Paulo in 1996. Right: relative frequency, left: normal probability graphic.

Based on these hypothesis, and with the objective of understanding the space-time patterns in São Paulo, we used ordinary kriging to produce surfaces of homicide rate for 96 districts of São Paulo for the years 1996 and 1999. To achieve it, a set of points obtained by the association of the parameter value for each area to its centroid was taken as a sample, used to compute a variogram that modeled the structure of spatial autocorrelation. The surface obtained is presented in Figure 26 and shows a significant drop in the areas with the lower homicide rates (less than 30 deaths by 100,000 people) in 1999 compared to 1996. Since the areas of lower homicide rate correspond to the wealthier areas of the city, the result shows a spatial spread of crime, with the violence progressively occupying the whole city.

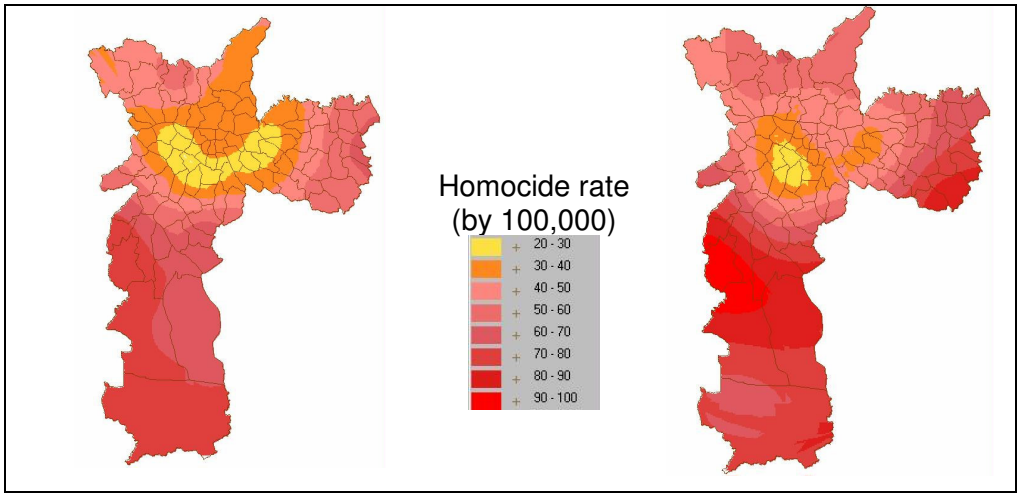


Figure 26 – Surface estimation for the homicide rate in São Paulo in 1996 (left) and 1999 (right).

FINAL COMMENTS

This tutorial showed that the spatial analysis techniques can considerably increase our capability of understanding the spatial patterns associated to areal data, especially when dealing with social indicators, that present global and local spatial autocorrelation. Exploratory techniques like Moran's indicators and Moran's scatter plots are very useful to show the spatial clusters and to indicate priority areas in terms of public policies. Bayes estimation methods for rates allow the correction of the effects associated to small populations. Regression models allow the establishment of the relationships between the variables, taking into account the spatial effects; in this case, the explicative power of the models can benefit from significant gains. The generation of surfaces is an efficient way of visually apprehending the spatial patterns. In short, researchers of social-economic data can substantially benefit from the techniques of this tutorial.

REFERENCES

The basic reference for most of the techniques presented in this tutorial is the book by Trevor Bailey, "*Spatial Data Analysis by Example*" (Bailey and Gatrell, 1995) and a general discussion about the distribution models for spatial data is presented in Diggle (2001). Peter Diggle's homepage (www.maths.lancs.ac.uk/~diggle) contains relevant material about spatial statistics.

In the case of spatial regression models, the SpaceStat by Luc Anselin, and the associated documentation (Anselin, 1992) present in detail the regression models with global effects (*spatial lag* and *spatial error*), and the model of spatial regimes. The SpaceStat was used to compute the models in the examples presented in the tutorial. The work of Luc Anselin in the field of local spatial autocorrelation indicators (Anselin, 1995; Anselin, 1996) are also important references. The SpaceStat site is at www.spacestat.com.

The GWR (*geographically weighted regression*) regression model was created by A.Stewart Fotheringham and is described in his book *Quantitative Geography* (Fotheringham et al., 2000) and other works (Fotheringham et al., 1996) (Brundson et al., 1996). More information can be found in the site at http://www.ncl.ac.uk/~ngeog/GWR/.

The discussion about the problem of the scale effects and the so-called "ecological fallacy" owes much to the work of Stan Openshaw; as an example, see Opensahw (1997). His works on the use of combinatorial optimization techniques to obtain more aggregate regions are also very important (Openshaw and Alvanides, 1999).

The issue of the surfaces generation from socioeconomic data owes much to the work of David Martin, in his book "*Geographic Information Systems: Socioeconomic Applications*" (Martin, 1995) and his works on censitary data in the United Kingdom (Martin, 1996; Martin, 1998). The empirical Bayes estimators were initially proposed in (Marshall, 1991). A general discussion about the subject, including a discussion on complete Bayes estimators, can be found in the excellent work of Renato Assunção (Assunção, 2001) or in the ample revision of Trevor Bailey, published in the "*Cadernos de Saúde Pública*" (Bailey, 2001).

The data about São Paulo from the 1991 census were extracted from the work "Mapa de Exclusão/Inclusão Social na Cidade de São Paulo" (Social Inclusion/Exclusion Map of the city of São Paulo) coordinated by professor Aldaíza Sposati, from PUC-SP (Pontifical Catholic University in São Paulo) (Sposati, 1996). The homicide rates for the districts of São Paulo in 1996 and 1999 were produced by Fundação SEADE and the generation of surfaces by kriging was done by José Luiz Rodrigues Yi.

Census data for Belo Horizonte for the year 1991 were provided by PRODABEL, and the study of the problem of the changes in the units of analysis was performed by Taciana Dias and Maria Piedade Oliveira.

The infant mortality data for the city of Rio de Janeiro were organized by FIOCRUZ and are presented in the work of Eleonora D'Orsi and Marilia Carvalho (D'Orsi & Carvalho, 1998). The data for the study about mortality due to homicide in the Southeastern Region were also published by the FIOCRUZ team, and can be reached in the personal pages of the authors: www.procc.fiocruz.br/~marilia and www.procc.fiocruz.br/~osvaldo.

The special issue of the *Cadernos de Sáude Pública* about the subject of spatial statistics in health (volume 17(5), October-November 2001), available in the internet (www.scielo.br) represent a good starting point on the subject, with various relevant studies.

1. ANSELIN, L. **SpaceStat tutorial: a workbook for using SpaceStat in the analysis of spatial data**. Santa Barbara, NCGIA (National Center for Geographic Information and Analysis), 1992.
2. ANSELIN, L. Local indicators of spatial association - LISA. **Geographical Analysis** v.27, p.91-115, 199
3. ANSELIN, L. The Moran scatterplot as ESDA tool to assess local instability in spatial association. In: M. Fisher, H. J. Scholten and D. Unwin (ed). **Spatial Analytical Perspectives on GIS**. London, Taylor & Francis, 1996. v., p.111-126.

4. ASSUNÇÃO, R. **Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia**. São Carlos, SP, UFScar, 2001. Disponível na homepage www.est.ufmg.br/~assuncao.

5. BAILEY, T. Spatial Statistics Methods in Health. **Cadernos de Saúde Pública** v.17, n.5,, 2001.

6. BAILEY, T. and A. GATTREL. **Spatial Data Analysis by Example**. London, Longman, 199

7. BRUNSDON, C. A.S. FOTHERINGHAM AND M.E. CHARLTON, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. **Geographical Analysis**, 28(4), 281-298, 1996.

8. CRUZ, O. C. **Homicídios no Estado do Rio de Janeiro: análise da distribuição espacial e sua evolução**. Dissertação de mestrado/Faculdade de saúde Pública-USP, 1996.
http://malaria.procc.fiocruz.br/~oswaldo/publi/ogc-diss.pdf

9. DIGGLE, P. **Spatial statistics in the biomedical science: future directions**. Lancaster, Lancaster University, 2001.

10. D'ÓRSI, E. and M. S. CARVALHO. Perfil de Nascimentos no Município do Rio de Janeiro - Uma Análise Espacial. **Cadernos de Saúde Pública** v.14, n.1, p.367-379, 1998.

11. FOTHERINGHAM, A.S., C. BRUNSDON AND M.E. CHARLTON, 2000, **Quantitative Geography**, London: Sage

12. FOTHERINGHAM, A.S., M.E. CHARLTON AND C. BRUNSDON, The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity. **International Journal of Geographic Information Systems**, 10: 60627, 1996.

13. GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (1995) **Bayesian Data Analysis** Chapman & Hall/CRC.

14. GILKS, W.R., RICHARDSON, S., SPIEGELHALTER, D.J. (orgs) (1998), *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

15. MARSHALL, R. Mapping disease and mortality rates using empirical Bayes estimators. **Applied Statistics** v.40, p.283-294, 1991.

16. MARTIN, D. **Geographic Information Systems: Socioeconomic Applications.** London, Routledge, 199

17. MARTIN, D. An assessment of surface and zonal models of population. **International Journal of Geographical Information Systems** v.10, p.973-989, 1996.

18. MARTIN, D. Optimizing census geography: the separation of collection and output geographies. **International Journal of Geographical Information Science** v.12, p.673-685, 1998.

19. OPENSHAW, S. Developing GIS-relevant zone-based spatial analysis methods. In: P. Longley and M. Batty (ed). **Spatial Analysis: Modelling in a GIS Environment**. New York, John Wiley, 1997. v., p.573.

20. OPENSHAW, S. and S. ALVANIDES. Applying Geocomputation to the analysis of spatial distributions. In: P. A. Longley, Goodchild, M. F., Maguire, D. J. and Rhind, D. W (ed). **Geographical Information Systems: Principles, Techniques, Management and Applications**. Chichester, Wiley, 1999. v., p.267-282.

21. SPOSATI, A. **Mapa de Exclusão/Inclusão Social de São Paulo**. São Paulo, EDUC, 1996.