Georeferenced data in epidemiologic research

Dados georreferenciados em epidemiologia

Guilherme Loureiro Werneck [1,2]

Abstract *This paper reviews some conceptual and practical issues regarding the application of georeferenced data in epidemiologic research. Starting with the disease mapping tradition of geographical medicine, topics such as types of georeferenced data, implications for data analysis, spatial autocorrelation and main analytical approaches are heuristically discussed, relying on examples from the epidemiologic literature, most of them concerning mapping disease distribution, detection of disease spatial clustering, evaluation of exposure in environmental health investigation and ecological correlation studies. As for concluding remarks, special topics that deserve further development, including the misuses of the concept of space in epidemiologic research, issues related to data quality and confidentiality, the role of epidemiologic designs for spatial research, sensitivity analysis and spatiotemporal modeling, are presented.*

Key words *Epidemiology, Medical geography, Epidemiologic methods, Small-area analysis, Ecological studies, Clustering*

Resumo *Neste artigo, faz-se uma revisão acerca de aspectos conceituais e práticos relacionados à aplicação de dados georreferenciados na pesquisa epidemiológica. Iniciando com a tradicional abordagem de mapeamento de doenças da geografia médica, discute-se heuristicamente com base em exemplos da literatura epidemiológica, tópicos como tipos de dados georreferenciados, implicações para a análise de dados, autocorrelação espacial e as principais estratégias analíticas, destacando-se os estudos de mapeamento da distribuição espacial de eventos de saúde, detecção de agregados espaciais de casos, avaliação de exposição em estudos de saúde ambiental e estudos ecológicos. Os comentários finais salientam tópicos especiais que merecem desenvolvimentos futuros, incluindo os dilemas relacionados à incorporação do conceito de espaço na pesquisa epidemiológica, aspectos relacionados à qualidade dos dados e confidencialidade, o papel dos estudos epidemiológicos na pesquisa com dados espaciais, análise de sensibilidade e modelos espaço-temporais.*

Palavras-chave *Epidemiologia, Geografia médica, Métodos epidemiológicos, Análise de pequenas áreas, Estudos ecológicos, Conglomerados*

[1] Departamento de Endemias Samuel Pessoa, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz.
Rua Leopoldo Bulhões 1480, Manguinhos. 21041-210
Rio de Janeiro RJ.
gwerneck@ensp.fiocruz.br
[2] Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.

Introduction

Georeferenced data, also known as spatial, geographical or geospatial data, are the basic pieces of information needed to identify the geographic location of phenomena across the Earth's surface. In general, georeferenced data consist of measurements or observations taken at specific locations (points referenced by latitude and longitude) or within specific regions (areal data). In epidemiologic research, this type of information is mainly used to investigate the relationship between georeferenced health events data and aspects related both to individual characteristics (e.g. genetic, behavioral and demographic) and contextual factors (e.g. socioeconomic neighborhood conditions, physical environment). Mapping disease distribution, detection of disease spatial clustering, evaluation of exposure in environmental health investigation and ecological correlation studies are some examples of possible applications of georeferenced data in epidemiologic studies.

The geographical distribution of disease has been considered a key element in epidemiologic research, as indicated by the importance given to the description of health events according to "person, place and time" in the classic epidemiology textbooks[1,2]. In fact, studies on the geographical distribution of diseases come back to the 18th and 19th centuries, when the term "medical geography" was devised[3].

Among the precursors of geographic studies of disease are the physicians James Lind[4], mainly recognized for his work on scurvy, and Leonhard Ludwig Finke[5], who published in 1792 what has been considered the most detailed conceptual contribution on medical geography written to that point and the first to systematize world-wide data[5,6]. A long tradition on disease mapping also started at that time. Aparently the first dot map applied to public health problems is due to Valentine Seaman, in 1798, describing the distribution of yellow fever cases in New York[7].

Early roots of epidemiology as a discipline can be found at this time, with research focusing mainly on the relationship between societal conditions and health[8]. Various researchers applied what modern epidemiology usually calls ecological study designs to address the problem of disease variation across different places. For instance, André-Michel Guerry, in 1833, explored the variation of suicide and homicide rates across regions of France[9] and Engels, in 1892, cited evidence of variation in mortality rates across different cities and streets surrounding Manchester[10].

For nearly every epidemiologist, however, John Snow´s investigation of the cholera epidemic in London is the most famous work demonstrating the importance of studying the geography of disease[11]. Considered a classical example of epidemiologic reasoning, which led to development of a water-borne theory of cholera transmission, his work also became famous in medical geography because he used a dot-map to plot the location of cholera deaths around the Broad Street pump in London Soho´s district. Nevertheless, the belief that Snow used the dot-map to determine the source of the cholera outbreak and to make a causal connection between the removal of the pump and the end of the outbreak seems not to be supported by evidence[12]. As a matter of fact, Snow already had his theory on the transmission of cholera before collecting data to test it[13].

Despite the longstanding tradition of disease mapping and geographically oriented research, during the first half of the 20th century epidemiologists were more inclined to focus their research on the time dimensions of disease distribution, and progressively more and more on individual characteristics, a result of the rising emphasis on the biological causes of disease[8,14].

Advances in geographic information systems (GIS) permitted a remarkable increase in the efficiency of processing and analysis of complex georeferenced data involving different variables at a variety of geographical scales, providing new tools for epidemiologists to incorporate place and space in their investigations[15]. Since the 1970s, GIS and related technologies, such as remote sensing, have spread rapidly to many scientific and technical fields, including public health and epidemiology[14-17]. Today GIS and remote sensing are considered important tools in environmental health research and disease surveillance[15,18], and have been used to investigate patterns of disease spread and inform control strategies for infectious diseases, in particular vector-borne and zoonotic diseases[19,20], to help define the boundaries of the communities or neighborhoods where study participants reside in multilevel studies[21,22], and to estimate socio-demographic and environmental variables[21,23]. This process of increasing incorporation of GIS in epidemiologic research should not be considered simply as a result of technological forces, but must be put in the context of the renewed efforts detected after the 1960s to integrate social sciences and epidemiology[8].

In this review, I intend to introduce the basic approaches used by epidemiologists to deal with georeferenced data. I will focus on the strategies for analysis of this type of data, building upon examples from different areas of epidemiologic research. I wish to keep statistics apart as much as possible, emphasizing the conceptual issues behind the techniques.

For the sake of completeness, it is necessary to mention that many authors use the terms "spatial epidemiology" or "geographical epidemiology" to define this area of investigation[24-28]. Nonetheless, I prefer not to support an unnecessary autonomy to what I see as a field of epidemiologic practice that just carry within its roots a criticsm concernig the excessive focus of modern epidemiology on the individual causes of disease.

Types of georeferenced data

Broadly speaking, geospatial data can be point referenced or area referenced. Point referenced data are observations registered at specific locations that might be identified by latitude and longitude, for instance the location of cases of disease and the location of air pollution monitoring stations. Area referenced data are observations specific to a region (e.g. census tracts, neighborhoods). For both you may also have measured attributes, such as demographic characteristics of the disease case, specific measurements of pollutants in each monitoring station, rates of disease by census tracts, socioeconomic variables for neighborhoods and so on. Hence, georeferenced health data combine the usual information available in epidemiologic studies, that is, values for attributes of some object, with information about their locations.

When deciding to choose the appropriate analysis strategy for georeferenced data, one needs to consider the statistical model for the spatial process underlying the available data. A possible approach is to distinguish data types according to the nature of the spatial domain in which the data is observed[29,30]. By spatial domain I informally mean ***where*** things are being observed. The spatial domain might be continuous or discrete (do not confuse this with the attribute being measured, that is, whether the variable measured is continuous or not) and fixed or random[30].

A continuous spatial domain means that what you are observing can be, theoretically, measured everywhere within that domain. For instance, imagine that you are interested in obtaining data on temperature. You will probably gather this data from monitoring stations that are located at specific points in the space. However, temperature could have been measured at any place, hypothetically there are an infinite number of places that you can place monitoring stations, but unfortunately you have, in general, only a sample of possible locations to get data on temperature. On the other hand, a discrete spatial domain means that you can count the number of locations in which observations are taken. An example is the number of dengue cases recorded in neighborhoods of Rio de Janeiro in January 2008. Here the spatial domain is discrete: you can count the number of neighborhoods that configure the region of analysis.

A fixed spatial domain is the one that does not change from one realization of the spatial process to the next[30]. Heuristically, a realization of a spatial process is simply a set of georeferenced observations. Imagine you asked your research assistants to do a fieldwork using some kind of instrument that measures soil contamination by Arsenic, a known carcinogenic substance. The results of your study are a set of measures taken at different spatial locations, and might be considered one realization of a spatial process. However, if you send other assistant researchers to do the same work, then they will probably take soil samples from different locations, and this will be another realization of the spatial process. Anyway, although there are different samples (realizations), the spatial domain has not changed between the first and second study.

A random spatial domain can be conceived in situations where the focus has switched from studying the attribute itself to studying the locations. Imagine you label your data as 0 or 1, being 1 whenever the Arsenic level exceeds a given threshold (say the level above which soil contamination is considered unacceptable) and 0 otherwise. Throw away all points returning a value of 0 and keep only those with values of 1. Now there is no interest in studying the attribute, because all points represent the same condition (Arsenic concentration above the threshold). The interest is now on the spatial arrangement of observations, which is, together with the number of points observed, the outcome of the spatial random process[30]. Here, the spatial domain is random since it will change in every new realization. Even when attributes are available at each location, the important statistical feature of these data is the random domain[30].

Based on the nature of the spatial domain, georeferenced data can be divided in three subtypes: geostatistical data, areal or regional data and spatial point patterns[29,30].

Geostatistical data

Consist of measurements that can potentially be taken in any location in space, although the actual data are sampled at specific locations in a spatial continuum (fixed and continuous spatial domain). Statistical approaches to analyze this type of data are commonly known as ***geostatistics***. Geostatistics is a branch of applied mathematics developed in the early fifties to help obtain better predictions in mineral prospection[31]. Usually the geostatistical approach uses the observed data to estimate values at unsampled locations and produce a continuous surface showing the variation of the attribute across space. The spatial regression technique known as ***kriging*** is the usual interpolation method used in this setting[31]. Typical geostatistical approaches have been used in epidemiologic research for predicting the spatial distribution of insect vectors of infectious disease[32-34] and of air and soil pollutants[35-37].

However, one should be aware that, in general, data on health events of individuals are only usually registered where population exists. Therefore, strictly speaking, geostatistics is probably not the canonical approach for disease data, since cases can occur only in selected inhabited parts of larger geographical regions. Because of the patchiness of population distribution, a continuous surface of disease rates would not make a sense in many epidemiological applications. Of course it is possible to partially circumvent the problem by delimiting study areas that are completely populated and by relaxing the assumption of a continuous spatial domain. Actually, it seems that the geostatistical approach is flexible enough to support different types of geospatial data, and recent developments have added more and more flexibility to this type of analysis[38,39]. In fact, there are many examples of applications of kriging and other smoothing techniques to produce continuous surfaces for data that are not typically continuously distributed across space, such as the prevalence and parasitic load of ascariasis obtained by household coproparasitologic surveys in Duque de Caxias (Brazil)[40,41], positivity for rotavirus infection in stool specimens collected at laboratories across the United States[42], prevalence of malaria infection in different survey sites in Mali[43], incidence of visceral leishmaniasis in census tracts of Teresina (Brazil)[44], epidemiologic surveillance data of influenza-like illness in France[45] and Japan[46], and breast and cervical cancer mortality in New England counties (United States)[38].

Areal, regional or lattice data

Involve observations associated with spatial regions. Data are not measurable at any location in space and are artificially gathered at sites or areas usually defined for statistical or administrative purposes. The whole spatial domain in exhaustively divided in areas (fixed and discrete spatial domain). Regional data can be regularly or irregularly spaced. An example of regular regional data is that obtained by remote sensing. Sensing devices on board of satellites measure reflected or emitted electromagnetic energy from the earth surface, and data are displayed in a series of small rectangles (pixels) of the same size[47]. However, most data used in epidemiologic research are irregularly spaced, for example, rates of disease or socioeconomic indicators measured at county, neighborhood or census tracts level.

Regional data form the basis of the so-called multiple-group ecological study[48] in which the units of analysis are, in general, geographically defined areas. The aims of ecological studies are to describe geographical patterns of disease frequencies and risk factors, and estimate putative ecological correlations between variables. In the purely descriptive ecological study no data on exposure is available, and the major interest is to map the spatial distribution of disease rates, aiming at detecting areas with a significant higher (or lower) incidence as compared to some expected rate. The most common approach for this objective is to use a ***choropleth map***, which is a class of quantitative thematic maps[49]. A choropleth map, also called area or shaded map, is a cartographic representation employing color or shading schemes, graded in intensity from light to dark, to depict variability of data distribution across regions[49,50]. The name derives from the Greek words ***choros*** (place), and ***pleth*** (value, quantity)[49,50]. Choropleth maps appear frequently in epidemiologic studies, and many national atlases of mortality and disease distribution have used this approach[51-53].

Ecological or geographic correlation studies are those that examine the spatial variation of environmental, socioeconomic, demographic and lifestyle factors in relation to health events, all measured on a geographic (ecologic) scale[24]. Exposure variables defined for regions are classified in three categories: aggregate measures based on individual data (e.g. proportion of smokers, mean income), environmental measures (e.g. temperature, air pollution), and global measures or contextual attributes (e.g. population density, social cohesion)[48]. Outcomes are generally expressed as incidence or mortality rates for that region. This type of approach has long tradition in sociology[54] and comes from that discipline one of the most famous example: Durkheim´s study on suicide[55]. Durkheim compared suicide rates between Prussian provinces classified according to the proportion of population that was Protestant. He found that suicide rates were higher at provinces with higher proportion of Protestants. Although Durkheim had not actually concluded from that evidence that suicide was more frequently among Protestants, this individual-level inference coming from an ecological study became "the" example of a typical bias called "ecological fallacy". The ecological fallacy refers to the fact that the degree of association between an exposure and disease may differ in ecological data, as compared to the same association measured using data obtained on individuals[56]. Since none of the regions were entirely Protestant or non-Protestant, it is not possible to exclude the hypothesis that are just the minority (Catholics or Jews) that are committing suicide in the provinces with higher proportion of Protestants.

There are several examples of ecological correlation studies in epidemiologic research. For instance, various ecological studies investigated the association between per capita consumption of specific alcoholic drinks and mortality from heart disease across countries[57]. Most of them suggested that wine was more effective in reducing risk of mortality than beer or spirits, but evidences from individual studies indicate that the benefits are attributable primarily to the alcohol content rather than to other components of each drink[57]. In Northeast Brazil, an ecologic study in 165 municipalities of the State of Ceará found that, among others, the level of inequality, population growth, and the presence of a railroad in the municipality were predictors of the incidence rate of leprosy[58].

Spatial point patterns

Comprise a set of locations of events, usually indexed by geographic coordinates (latitude and longitude), in a defined study region[59]. A spatial point pattern is an example of the above-mentioned situation of a random spatial domain. In this case the locations of the events themselves are the phenomena of interest, and the investigation focuses on whether the pattern is exhibiting complete spatial randomness, clustering, or regularity[59].

In the simplest situation, spatial point patterns include only information on the location of events (e.g. the location of disease cases), and are referred as ***unmarked*** patterns[30,59]. However, if you have additional variables attached to the locations (e.g. socio-demographics, time since diagnostic for cases) then it is called a ***marked*** process. A case-control study in which the geographical location of both cases and controls is known and additional variables are collected is a typical marked point pattern frequently used in epidemiologic research.

Spatial point patterns arise in many fields of investigation, for example, in plant ecology to study the spatial distribution of shrubland[60] and very large trees in a forest[61], in the analysis of urban land use to study the distribution of fast-food restaurants around schools[62], in criminology to study patterns of urban crimes[63], and in geology to investigate spatial patterns of volcano eruptions[64]. In these areas of investigation, having the background hypothesis of a complete spatial random distribution might be interesting, at least for a starting point, because there is no ***a priori*** obvious reason for spatial heterogeneity in the distribution of these events[59]. In epidemiology, however, population variation across space is a major factor leading to geographical clustering of disease cases and should always be considered in the analysis, since we are interested in spatial arrangements that are not explained by this specific feature. Urban land occupation and the heterogeneous distribution of risk factors across space are other important factors that might explain spatial patterns.

Although methods for spatial analysis were developed specifically for each of these three types of data, it is common to see methods originally developed for one type being applied to analyze data of a different category. For instance, regional data might be artificially allocated to some point inside the area (e.g. the seat of the state capital or the centroid, defined as the center of gravity of the region). Doing so, data will be displayed as points and might be analyzed as geostatistical data or point patterns. Also, the location of cases might be aggregated in regions and analyzed as lattice data. These approaches, when applied, need to have their underlying assumptions clarified and justified.

Spatial autocorrelation: a key concept

The key issue in the analysis of georeferenced data is that they often exhibit some spatial structure in the sense that there is a tendency of observations closer together to be more alike than observations farther apart[29]. The property that geographically nearby values of a variable tend to be similar on a map, that is, high values tend to be located near high values, and low values near low values, is called spatial autocorrelation[65]. In this case we are saying that exist some ***positive*** spatial autocorrelation, because a ***negative*** spatial autocorrelation means that nearby regions or points tend to be different, that is, one showing high values of the attribute and the neighbor low values, and vice-versa.

In ordinary statistical analysis, researchers often use the correlation coefficient to measure the direction and strength or degree of the relationship between a pair of quantitative variables[66]. A variant of conventional correlation is the serial correlation, which refers to the correlation between measures of a single variable over successive time intervals[65]. The geographic version of serial correlation is called spatial autocorrelation (the prefix "auto-" means self), the relationship between observations on the same variable taken at different locations in space[65].

To assess the nature and degree of spatial autocorrelation, it is necessary to represent the spatial arrangement of observations in order to get a sense of how close or distant there are apart from each other[67]. Then we use a set of rules, called ***weighting function***, to express the degree of proximity between observations[67]. For instance, for each possible pair of observations in space we may attribute a value of one if the observations are nearby (sometimes we say that they are ***neighbors***) and zero otherwise. There are many other options for defining these weights, and they may be based on distances between points (geostatistical and point pattern data) or centroids (areal data). In this case, pairs of observations might be defined as neighbors using a dichotomous classification (yes/no) or entering the actual distance as a measure of the degree of proximity between observations. For areal data it is common to use indicators of proximity based on whether the regions share a boundary or not. In any case, it is important to consider the fact that any measure of spatial autocorrelation will be influenced by the choice of the neighboring weights.

There are basically two major mechanisms responsible for spatial autocorrelation to occur with disease data: reaction and interaction[68,69]. The reaction mechanism implies that neighborhoods or nearby observations behave similarly because they share a common background risk. Spatial autocorrelation arises as a result of the heterogeneous distribution of risk factor across space. On the other hand, spatial autocorrelation due to interaction mechanisms means that areas or observations close together are more alike because proximity facilitates transmission. The reaction mechanism is the major mechanism underlying spatial autocorrelation in non-transmissible diseases, and interaction is implicated in spatial clustering of transmissible diseases. These two mechanisms may operate at the same time, in particular for infectious diseases. The term ***spatial dependence*** seems to be more appropriate in the case of the interaction mechanism, since we are dealing with the so-called dependent happenings, that is, disease incidence in individuals or regions depends on the prevalence of the infection in the population[70]. However, some authors do not make this distinction and use the terms spatial autocorrelation and spatial dependence interchangeably[25,71].

There are two major reasons for taking spatial autocorrelation in consideration in epidemiologic analysis of georeferenced data. First, most conventional statistical approaches assume that observations are independent, which is clearly violated when spatial autocorrelation exists (actually, the statistical assumption refers to the error structure, but lets leave formalities behind)[65]. In this situation, it is necessary to take spatial autocorrelation among observations into account in order to obtain valid estimates of regression coefficients, confidence intervals, and significance levels[29]. In the statistical jargon, positive spatial autocorrelation increases the likelihood of the null hypothesis rejection when it is true[65]. For instance, take a study on the association between social deprivation and the incidence of breast cancer, both variables measured at the municipality level (regional data). Consider that the distribution of both variables show spatial autocorrelation. Even if the truth is that there exist no association between deprivation and breast cancer incidence (null hypothesis), the results of a study ignoring the lack of independence between observations might well find as statistically significant such association, thus rejecting the null hypothesis when this is true. Because of spatial autocorrelation, there is some redundancy in the information provided by georeferenced data, meaning that more spatially autocorrelated than independent observations are needed to attain similar information[65]. Werneck and Maguire[72] show an example on how ignoring spatial autocorrelation may bias regression coefficients and standard error estimates.

The second reason to consider spatial autocorrelation is conceptual. The detailed description of how things are distributed in space might be used to support prediction. For instance, based on available georeferenced tuberculosis data at the municipality level, Braga[73] used geostatistical approaches to obtain better estimates in areas where surveillance reporting of cases was considered inappropriate. Lagrotta ***et al***.[74] used the spatial patterns of entomological parameters related to the distribution of ***Aedes aegypti*** to identify areas for targeting control actions. Explicitly accounting for spatial autocorrelation in a statistical model might also be used as a proxy for unknown or unmeasured variables and improve model specification[65]. In other situations, spatial autocorrelation should be considered as alternative explanations for interpreting study results. Suppose that you implement a community trial for controlling vector-borne disease in which the intervention proposed is spraying houses and the peridomestic environment with insecticides. The design you choose is some kind of community intervention trial, in which the area under study in divided into blocks, some of them randomly allocated to receive intervention and the other not. Interpretation of eventual changes in the incidence of the infection in the study blocks should consider the possibility of a spillover effect, that is, an area with no intervention might benefit from being a neighbor of an area in which spraying was performed (or, on the contrary, spraying would increase even more mosquito population in surrounding areas due to a repellent effect of insecticide).

As a matter of fact, spatial autocorrelation has a dual nature[67]. Sometimes it is considered a statistical nuisance asking for new analytical methods. At times, it is regarded as an intrinsic characteristic of spatial processes carrying essential information to be considered when interpreting results of studies using georeferenced data[65,67]. In a certain way, this duality represents the focus of statisticians and geographers, respectively. Epidemiologists have only recently incorporated spatial analysis in their framework, and still need to develop meanings for spatial autocorrelation that are appropriate for their field of investigation.

Analytical approaches

Statistical methods for spatial data analysis generally are used for either characterization of spatial structure or model adjustment. Characterization of spatial structure is well suited for descriptive purposes, hypothesis generation, prediction and forecasting. One main interest in epidemiology is prediction of future occurrences at specific geographic locations, which can be implemented through space-time models[75,76]. Model adjustment involves the use of the autocorrelation structure to obtain more accurate measures of effect and standard errors estimates.

Gatrell and Bailey[77] proposed that the analyses of georeferenced data may be divided into three broad classes: mapping, exploratory techniques and modeling methods.

Mapping

Mapping is the most elementary technique employed to describe the basic spatial features of disease data. For point patterns data, a dot map might be used to make simple displays of the location (usually the residency) of disease cases[78]. One should be aware that visual clustering of dots might reflect only the heterogeneous distribution of population across space. Interpretation of dot maps is also problematic since the mapped location might not reflect the place where the causal mechanism leading to disease actually operated. This is especially important for non-infectious diseases with long latency and induction periods. Even for some infectious diseases, such as tegumentary leishmaniasis, the risk of infection is mainly associated with economic and leisure activities in forested areas, including fishing, hunting, and ecotourism, and the household does not indicate where transmission occurred. In any case, a dot map is an interesting option to show spatial density of phenomena, but two critical choices should be made: dot size and dot value or quantity to be represented by each dot[49]. Using a small dot size leads to the sense of a sparse distribution, but choosing a large dot size may be unattractive because it increases too much the density of the distribution. If a dot will represent multiple events (then dot location does not necessarily relate to the specific location of a phenomena), choose a dot value so that the dots just coalesce in the most dense area on the map but a few dots are still represented in the sparse areas of the map[49]. Consider also using different symbol or colors to represent variations in the attribute being mapped, for instance to discriminate between cases and controls in a case-control study[78].

Choropleth maps is the most commonly strategy for the visual display of regional data. Disease maps of this type usually show directly age- and sex- standardized rates or standardized mortality (or morbidity) ratios (SMR), achieved by indirect standardization[24]. Many problems emerge when mapping areal data, in particular concerning the choices regarding scale or resolution, number and boundaries of classes, and the color or shading scheme[24,78]. The same data drawn in different levels of resolutions (e.g. census tracts and municipalities) may lead to different patterns and interpretations. Aggregation of small areas usually lead to loss of information, and the result is a spatial pattern that is pushed towards the values of the more populated regions[24]. A related question is the differences in geometry and size of areal units, larger ones dominating the depicted pattern. The use of gradations of color and shading helps simplifying the message to be transmitted, but arbitrary color intensities and types may induce the reader to focus on specific areas, drawing attention to some features and leaving other aspects less evident[78].

Most mapping softwares supply different ways to generate data classes. If data has a rectangular distribution, classes showing equal intervals might be a good choice. The natural breaks technique scrutinizes the actual data distribution to find cut points for creating classes. The equal area procedure forces the classes to represent approximately the same area in the map, but the number of units may vary across classes. Cutoff points based on percentiles oblige the classes to have approximately the same number of units. If the data has normal distribution you can create classes representing deviations from the mean value. All these techniques may be useful in some situations, but if there exists a clear epidemiological meaning for defining classes (e.g. an accepted threshold point for surveillance of infectious diseases) this should be considered. Concerning the number of classes, there is no universal accepted rule, but it is rare to see maps showing less than four and more than eleven classes. In general more classes are used if you have more geographical units. Anyway, the method used to generate data classes and the number of classes to be displayed are critical points to be considered because they may have strong influence on interpretation of such maps.

Small-area mapping studies have the additional problem of instability of rate estimates for areas with small populations. In this case, the addition or deletion of one or two events can cause dramatic changes in the observed value, and the most extreme estimates in a map tend to be provided by the least reliable data[79,80].

One way to address this problem is to produce a probability map[81]. Assuming that the number of cases of disease in each region follows a Poisson distribution with a constant mean, a map is created which shows those areas with unusually high or low rates. Although useful as an exploratory tool, the Poisson model assumes independent observations and ignores the possibility of spatial autocorrelation between neighboring areas[29].

A proposed solution to reveal spatial patterns that may be hidden by noisy data is to smooth the incidence rates using Bayesian methods[79,80,82]. The smoothed estimates represent a compromise between the actual observed value and the mean value of the whole region or some local value taking into consideration the possible dependency between neighboring areas[79,80,82]. Although calculations might be complex the idea is simple, that is, the technique estimates the incidence rate in a given region as a weighted average of the rates in this region and its neighbors. However, one should be concerned about the degree of smoothing to employ, because it will determine a tradeoff between smoothing away areas with truly high incidence (low sensitivity) or not identifying correctly areas with low risk (low specificity)[24].

Exploratory methods

Models for describing georeferenced data often decompose spatial variation into two main components[83,84]: data = large-scale variation + small-scale variation.

Large-scale variation, first order effect or lack of stationarity, is a regular variation, analogous to secular trend in time-series, but taking place in space[83]. It is also called spatial gradient and refers to the variation in the mean value of the process (e.g. incidence rate) in space and can be represented as a function of geographical coordinates or of variables that have their spatial distribution akin to the outcome[81,83]. For point patterns data, first order effect is often describes as the "intensity" of the process, defined as the mean number of events per unit of area at a spatial location[59]. The so-called ***kernel*** estimation is the usual technique employed for examining large-scale variation in spatial point patterns[59].

Explanations for large-scale variation usually rely on variables that vary slowly across large geographical regions, such as altitude, temperature, vegetation, and some socioeconomic characteristics. For instance, the decreasing north-south gradient in prostate cancer mortality in the United States is inversely associated to the ultraviolet radiation gradient[85]. Montenegro ***et al***.[86] found a spatial gradient in the incidence rates of leprosy in the State of Ceará, Brazil, with a tendency of high rates to be concentrated on the north-south axis in the middle region of the state, which was, at least in part, attributed to the process of urbanization and the heterogeneous distribution of underlying factors such as crowding, social inequality, access to health services or environmental characteristics that determine the transmission of ***Mycobacterium leprae***.

Small-scale variation around the gradient, or second order effect, results from clustering of high or low values across the region and from spatial autocorrelation[81,83]. To study small-scale variation accurately, it is often necessary to remove the effects of large-scale variation[83]. Unfortunately, this decomposition is not unique, and researchers have too much flexibility to "explain" spatial variation as almost totally due to small-scale variation or to assume a certain degree of large-scale variation and leave less to be explained as local clustering[83].

Once the spatial gradient has been identified and removed by using trend surface analysis, median polishing techniques, or other smoothing techniques[87,], the residual spatial variation or small-scale variation might be evaluated by using, for instance, the variogram (geostatistical data), the Moran or Geary autocorrelation coefficients (regional data), and the K-function (point patterns)[81]. For example, Jones ***et al***.[88] used the K-function in a case-control study to determine the degree of clustering in road traffic accidents outcomes. In Teresina, Brazil, after removing the spatial gradient by using smoothing techniques, Werneck ***et al.***[44] employed the Moran autocorrelation index to identify small-scale variation in the incidence of visceral leishmaniasis across census tracts.

Modeling

The objective here is to describe associations between exposure variables and some health outcome, for instance, between incidence rates and socioeconomic variables, using regression models that explicitly take into consideration the spatial structure of data.

The ecological study is the typical epidemiologic design used to study such associations when variables are measured at the group level, most frequently defined as geographical areas. As previously stated, georeferenced data used in these studies are usually spatially autocorrelated and explicitly taking care of the spatial autocorrelation is both a statistical necessity and an additional source of information on how the process evolve across space. Most ecological regression studies, however, do not include spatial autocorrelation parameters in their analytical models. In this sense, these are not spatial studies, since the geographical structure of data is not being allowed for in the analysis. Non-spatial ecological analysis work as if the geographical areas are independent from each other (no spatial autocorrelation structure), and results from this type of analysis would be the same if you randomly rearrange geographical units across space. In epidemiologic research, one of the first approaches to include spatial autocorrelation in the analysis of ecological data is due to Cook and Pocock[89]. They re-analyzed data from an ecological study in Great Britain during 1969-73 based on 253 towns aiming at uncover the possible contributions of drinking water quality, climate, air pollution, blood groups, and socioeconomic factors on the regional variations of cardiovascular mortality[90]. They showed that previous results based on ordinary regression methods overstate the significance of the regression coefficients[89].

There are basically two major ways in which spatial autocorrelation can be managed in regression modeling: as spatial-lag and spatial error models[84]. Spatial-lagged regression models take care of spatial autocorrelation by considering that the dependent (outcome) variable in one area is affected by variables in nearby areas[84]. For instance, when studying mortality from respiratory diseases in areas as a function of air pollution, it is possible to include as exposure variables not only the measure of pollution in that area but also measures taken at neighborhoods. An alternative is to include the outcome as lagged variables. For example, the incidence of tuberculosis in one area might be expressed in a regression model as a function of socioeconomic variables and the incidence of tuberculosis in nearby regions. The first model is called a regression model with spatial lagged explanatory variables and the last as a regression model with spatial lagged response variable[84].

Alternatively, when there exist some unmeasured spatially correlated variables that have an effect on the outcome or when the major source of autocorrelation in the dependent variable is an interaction mechanism, a spatial error model is recommended[84,91]. In this case, the error for the model in one area is correlated with the error terms in its neighboring locations[84,91]. Okwi ***et al***.[92] using spatial-lag response and spatial error models found that a variety of geographic factors, such as soil type, distance/travel time to public resources, elevation, type of land use, and demographic variables significantly explain spatial patterns of poverty in rural Kenya.

Regression approaches for spatial point patterns are less common in the epidemiologic literature, but some relatively recent developments propose methods for analyzing georeferenced case-control studies using nonparametric binary regression models using kernel and generalized additive model methods[93]. Webster ***et al***.[94] used similar approaches to investigate the association between residence and breast cancer on Cape Cod, Massachusetts (USA) using data from population-based case-control studies.

Facing up to the future

The vast accumulation of methods and techniques for analyzing georeferenced data in the last years is also posing many challenges to epidemiologists willing to use such approaches in their studies. Beyond those problems already mentioned above in the text, there are some other that deserve further consideration.

The concept of space

It is well recognized that there is an acritical incorporation of "space" as a category of analysis in epidemiologic research. Most of the time, space is considered just as geometric space indicating where things occur[95,96], or just used for the differentiation of social conditions[96], or as a circumstance of spatial factors inducing risk[96]. Therefore, space in many epidemiological applications is a concept with no social or historical dimensions[97]. Space, however, is both the medium and the outcome of social relations[98], a social construct resulting from the human action, organized in a society, over landscape[99]. Barcellos and Sabroza[100] used the expression "the place behind the case" when analyzing environmental risk conditions for leptospirosis, which I regard as an inspired terminology to stress that the links between risk conditions and occurrence of disease are directly determined by socioeconomic, cultural and social factors operating in space[96]. To be more useful, epidemiologic studies focusing on georeferenced data should embrace the concept of social space or, at least, explicitly indicate the concepts of space they are using.

Data availability and quality

In Brazil and some other countries, national registries of sociodemographic and health data are important sources of information for georeferenced studies. However, the availability and quality of essential information is an issue of major concern[24]. For instance, missing data on socioeconomic variables is a common problem, as well as, the lack of coverage of the whole population and diagnostic errors. Reliability and validity studies of large databases are essential for judging the possibility of bias in spatial investigations based on secondary data.

Confidentiality

There are two major issues regarding confidentiality[24]. The first is related to the possibility of using the address recorded in secondary registry data (e.g. mortality or hospital admittance data) to assess the location of disease cases. Since these are personal data, the issue of whether individual consent is necessary arises. Without no doubt the use of this type of information should only be allowed on the basis of specific rules and conditions, but the need for individual consent would greatly impair the development of research in this field. The second, is the use of very small-area data which might permit the identification of neighborhoods and even city blocks. If data will be used to map environmental hazards or clusters of disease one should be concerned about the impact not only on community perceptions but also on the value of properties[24].

Epidemiologic design

Most analyses of georeferenced data are based on aggregated or secondary data. It means that there is no research planning for defining what and how variables will be collected. In general, these studies lack essential information for understanding disease processes and are based on available data collected for other purposes than the specific research. The development of spatial sampling techniques is an important feature to be implemented in surveys aiming at describing spatial variation of health events[101,102]. Case-control studies have been used more frequently, but still there are some issues of concern. For instance, unmatched designs are more clearly interpreted since the assumption of cases and controls coming from the same spatial distribution leads to the expectation that their spatial patterns will not differ, unless some social or environmental risk factor exist[103]. However, the same reasoning will not directly apply in matched case-control studies[103].

Sensitivity analysis

Since spatial data analysis involves a series of choices (e.g. neighboring weights, issues in mapping, variety of data types and models), sensitivity analysis is an area that deserves more attention. Using sensitivity analysis to check whether the results of a study are too much rooted or dependent on specific choices is a way to incorporate a quantitative approach to improve the qualitative judgments that are expected in the discussion of epidemiologic studies[104].

Space-time models

In infectious disease epidemiology, heterogeneities in space and time has been considered responsible for the increase in transmission. In general, such heterogeneities have been more frequently incorporated in the dynamical modeling framework, but diffusion models based on extensions of the autorregressive models to the spatiotemporal context have also some tradition. Although there exists a tradition on the study of space-time clustering of non-infectious disease[105], investigations using space-time modeling in this area are still infrequently seen in epidemiology. The difficulties in specifying and interpreting models with a complex autocorrelation structure that goes not only in space and time alone, but also considers the interaction between the two components (e.g. incidence today in one specific area depends not only on what happened in this area in the past and what happens today in the neighboring areas, but also on what happened in the past in nearby areas) is probably the main reason for not using such models regularly in epidemiologic research.

Spatial analytical techniques were developed first for use in geology, ecology, and agriculture, and only recently have been largely integrated to the epidemiologic agenda of investigation. At this point in time, when there is a wide-range of available software for applying such models in an almost mechanical way, a critical incorporation of methods for the analysis of georeferenced data is a major challenge for epidemiology.

Acknowledgements

This work was partially supported by CNPq - 483675/2006-9 - 410590/2006-1 - 308889/2007-0.

References

1. MacMahon B, Pugh TFH. ***Epidemiology: principles and methods***. Boston: Little Brown & Co.; 1970.
2. Lilienfeld AM, Lilienfeld DE. ***Foundations of epidemiology***. 2nd ed. New York: Oxford University Press; 1980.
3. Meade M, Florin J, Gesler W. ***Medical geography***. New York: Guilford Press; 1988.
4. Barret FA. 'SCURVY' Lind's medical geography. ***Soc Sci Med*** 1991; 33:347-353.
5. Barrett FA. A medical geographical anniversary. ***Soc Sci Med*** 1993; 37:701-710.
6. Light RU. The progress of medical geography. ***Geogr Rev*** 1944; 34:636-641.
7. Barrett FA. Finke's 1792 map of human diseases: the first world disease map? ***Soc Sci Med*** 2000; 50:915-921.
8. Krieger N. Epidemiology and social sciences: towards a critical reengagement in the 21st century. ***Epidemiol Rev*** 2000; 22:155-163.
9. Guerry A-M. ***Essai sur la statistique moral de la France***. A Translation of Andre-Michel Guerry's Essay on the Moral Statistics of France (1883): a sociological report to the French Academy of Science; edited and translated by Hugh P. Whitt and Victor W. Reinking. Lewinston: Edwin Mellen Press; 2002.
10. Engels F. ***A situação da classe trabalhadora em Inglaterra.*** Porto: Afrontamento; 1975.
11. Snow J. ***Sobre a maneira de transmissão da cólera***. Rio de Janeiro: USAID; 1967.
12. McLeod KS. Our sense of Snow: the myth of John Snow in medical geography. ***Soc Sci Med*** 2000; 50: 923-935.
13. Vandenbroucke JP, Eelkman Rooda HM, Beukers H. Who made John Snow a hero? ***Am J Epidemiol*** 1991;133:967-973.
14. Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology. ***Epidemiol Rev*** 1999; 21:143-161.

15. Cromley EK. GIS and disease. ***Annu Rev Public Health*** 2003; 24:7-24.
16. Krieger N. Place, space, and health: GIS and epidemiology. ***Epidemiology*** 2003;14:384-385.
17. Clarke KC, McLafferty SL, Tempalski BJ. On epidemiology and geographic information systems: a review and discussion of future directions. ***Emerging Infectious Diseases*** 1996; 2:85-92.
18. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. ***Environ Health Perspect*** 2004; 112:1007-1015.
19. Rogers DJ, Randolph SE. Studying the global distribution of infectious diseases using GIS and RS. ***Nat Rev Microbiol*** 2003; 1:231-237.
20. Beck LR, Lobitz BM, Wood BL. Remote sensing and human health: new sensors and new opportunities. ***Emerg Infect Dis*** 2000; 6:217-227.
21. Tatalovich Z, Wilson JP, Milam JE, Jerrett ML, McConnell R. Competing definitions of contextual environments. ***Int J Health Geogr*** 2006; 5:55.
22. Santos SM. ***A importância do contexto social de moradia na auto-avaliação de saúde*** [tese]. Rio de Janeiro (RJ): Escola Nacional de Saúde Pública; 2008.
23. Correia VR, Monteiro AM, Carvalho MS, Werneck GL. Uma aplicação do sensoriamento remoto para a investigação de endemias urbanas. ***Cad Saude Publica*** 2007; 23:1015-1028.
24. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. ***Environ Health Perspect*** 2004; 112:998-1006.
25. Rezaeian M, Dunn G, St Leger S, Appleby L. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. ***J Epidemiol Community Health*** 2007; 61:98-102.
26. Lawson AB. ***Statistical methods in spatial epidemiology***. New York: Wiley; 2001.
27. Elliott P, Cuzik J, English D, Stern R, editors. ***Geographical and environmental epidemiology***. Oxford: Oxford University Press; 1996.
28. Bailey TC. Spatial statistical methods in health. ***Cad Saude Publica*** 2001; 17:1083-1098.
29. Cressie N. ***Statistics for Spatial Data***. New York: John Wiley; 1993.
30. Schabenberger O, Gotway CA. ***Statistical methods for spatial data analysis***. Boca Raton: Chapman & Hall/CRC; 2005.
31. Wackernagel H. ***Multivariate geostatistics***. New York: Springer; 1995.
32. Nicholson MC, Mather TN. Methods for evaluating Lyme disease risks using geographic information systems and geospatial analysis. ***J Med Entomol*** 1996; 33:711-720.
33. Allen TR, Lu GY, Wong D. Integrating remote sensing, terrain analysis, and geostatistics for mosquito surveillance and control. In: ***Proceedings of the American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Conference***; 2003; Anchorage, Alaska, USA. p. 1-10.
34. Carbajo AE, Curto SI, Schweigmann NJ. Spatial distribution pattern of oviposition in the mosquito Aedes aegypti in relation to urbanization in Buenos Aires: southern fringe bionomics of an introduced vector. ***Med Vet Entomol*** 2006; 20:209-218.
35. Leem JH, Kaplan BM, Shim YK, Pohl HR, Gotway CA, Bullard SM, Rogers JF, Smith MM, Tylenda CA. Exposures to air pollutants during pregnancy and preterm delivery. ***Environ Health Perspect*** 2006; 114:905-910.
36. Jerrett M, Buzzelli M, Burnett RT, DeLuca PF. Particulate air pollution, social confounders, and mortality in small areas of an industrial city. ***Soc Sci Med*** 2005; 60:2845-2863.
37. Ersoy A, Yunsel TY, Cetin M. Characterization of land contaminated by past heavy metal mining using geostatistical methods. ***Arch Environ Contam Toxicol*** 2004; 46:162-175.
38. Goovaerts P. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. ***Int J Health Geogr*** 2005; 4:31.
39. Goovaerts P. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. ***Int J Health Geogr*** 2006; 5:52.
40. Campos MR, Valencia LI, Fortes BP, Braga RC, Medronho RA. Distribuição espacial da infecção por Ascaris lumbricoides. ***Rev. Saude Publica*** 2002; 36:69-74.
41. Fortes BP, Ortiz Valencia LI, Ribeiro SV, Medronho RA. Modelagem geoestatística da infecção por Ascaris lumbricoides. ***Cad Saúde Pública*** 2004; 20:727-734.
42. Török TJ, Kilgore PE, Clarke MJ, Holman RC, Bresee JS, Glass RI. Visualizing geographic and temporal trends in rotavirus activity in the United States, 1991 to 1996. ***Pediatr Infect Dis J*** 1997; 16:941-946.
43. Kleinschmidt I, Bagayoko M, Clarke GP, Craig M, Le Sueur D. A spatial statistical approach to malaria mapping. ***Int J Epidemiol*** 2000; 29:355-361.
44. Werneck GL, Costa CH, Walker AM, David JR, Wand M, Maguire JH. The urban spread of visceral leishmaniasis: clues from spatial analysis. ***Epidemiology*** 2002; 13:364-367.
45. Carrat F, Valleron AJ. Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France. ***Am J Epidemiol*** 1992; 135:1293-1300.
46. Sakai T, Suzuki H, Sasaki A, Saito R, Tanabe N, Taniguchi K. Geographic and temporal trends in influenzalike illness, Japan, 1992-1999. ***Emerg Infect Dis*** 2004; 10:1822-1826.
47. DeMers MN. ***Fundamentals of Geographic Information Systems.*** New York: John Wiley & Sons; 1997.
48. Morgenstern H. Ecologic studies. In: Rothman KJ, Greenland S, editors. ***Modern epidemiology***. 2nd edition. Philadelphia: Lippincott-Raven; 1998.
49. Dent BD. ***Cartography: thematic map design.*** 3rd edition. Dubuque, Iowa: Wm. C. Brown; 1993.
50. Cliff AD, Haggett P. ***Atlas of disease distributions: analytic approaches to epidemiologic data.*** Oxford: Blackwell; 1988.
51. Walter SD, Birnie SE. Mapping mortality and morbidity patterns: an international comparison. ***Int J Epidemiol*** 1991; 20:678-689.
52. Pickle LW, Mungiole M, Jones GK, White AA. ***Atlas of United States mortality***. Hyattsville, Maryland: National Center for Health Statistics; 1996.

53. Brasil. Ministério da Saúde. ***Atlas de Saúde do Brasil***. [acessado 2008 Mai 02]. Disponível em: http://www.saude.gov.br/svs/atlas
54. Robinson WS. Ecological correlations and the behavior of individuals. ***Am Sociol Rev*** 1950; 15:351–357.
55. Durkheim E. ***Suicide: a study in sociology***. New York: Free Press; 1951.
56. Walter SD. The ecologic method in the study of environmental health. II. Methodologic issues and feasibility. ***Environ Health Perspect*** 1991; 94:67-73.
57. Rimm EB, Klatsky A, Grobbee D, Stampfer MJ. Review of moderate alcohol consumption and reduced risk of coronary heart disease: is the effect due to beer, wine, or spirits. ***BMJ*** 1996; 312:731-736.
58. Kerr-Pontes LR, Montenegro AC, Barreto ML, Werneck GL, Feldmeier H. Inequality and leprosy in Northeast Brazil: an ecological study. ***Int J Epidemiol*** 2004; 33:262-269.
59. Gatrell AC, Bailey TC, Diggle PJ, Rowlingson BS. Spatial point patterns and its application in geographical epidemiology. ***Trans Inst Br Geogr*** 1996; 21:256-274.
60. Haase P. Spatial Pattern Analysis in Ecology Based on Ripley's K-Function: Introduction and Methods of Edge Correction. ***J Veg Sci*** 1995; 6:575-582.
61. Neeff T, Biging GS, Dutra LV, Freitas CC, Santos JR. Modeling spatial tree patterns in the tapajós forest using interferometric height. ***Revista Brasileira de Cartografia*** 2005; 57:1-6.
62. Austin SB, Melly SJ, Sanchez BN, Patel A, Buka S, Gortmaker SL. Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. ***Am J Public Health*** 2005; 95:1575-1581.
63. Craglia M, Haining R, Wiles P. A Comparative Evaluation of Approaches to Urban Crime Pattern Analysis. ***Urban Studies*** 2000; 37:711-729.
64. Bishop MA. Point pattern analysis of eruption points for the Mount Gambier volcanic sub-province: a quantitative geographical approach to the understanding of volcano distribution. ***Area*** 2007; 39:230-241.
65. Griffith DA. ***Spatial autocorrelation***. [acessado 2008 Mai 02]. Disponível em: http://www.utdallas.edu/~dagriffith/Taiwan_lectures/background%20readings/ESM-2005.pdf
66. Griffith DA. ***Spatial autocorrelation: a primer***. Washington, D.C.: Association of American Geographers; 1987.
67. Odland J. ***Spatial autocorrelation***. Beverly Hill: Sage; 1988.
68. Cliff AD, Ord JK. ***Spatial Processes***. London: Pion; 1981.
69. Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. ***J R Statist Soc A*** 1991; 154:421-441.
70. Halloran ME, Struchiner CJ. Study design for dependent happenings. ***Epidemiology*** 1991; 2:331-338.
71. Lloyd CD. ***Local models for spatial analysis***. Boca Raton: CRC Press; 2007.
72. Werneck GL, Maguire JH. Spatial modeling using mixed models: an ecologic study of visceral leishmaniasis in Teresina, Piauí State, Brazil. ***Cad Saúde Pública*** 2002; 18:633-637.
73. Braga JU. ***O uso da modelagem espacial na estimativa dos dados da tuberculose no Brasil*** [tese]. Rio de Janeiro (RJ): Instituto de Medicina Social, UERJ; 1997.
74. Lagrotta MT, Silva WC, Souza-Santos R. Identification of key areas for Aedes aegypti control through geoprocessing in Nova Iguaçu, Rio de Janeiro State, Brazil. ***Cad Saúde Pública*** 2008; 24:70-80.
75. Linde A van der, Witzko KH, Jockel KH. Spatial-temporal analysis of mortality using splines. ***Biometrics*** 1995; 51:1352-1360.
76. Cliff AD, Haggett P, Smallman-Raynor M. An exploratory method for estimating the changing speed of epidemic waves from historical data. ***Int J Epidemiol*** 2008; 37:106-112.
77. Gatrell AC, Bailey TC. Interactive spatial data analysis in medical geography. ***Soc Sci Med*** 1996; 42:843-855.
78. Lawson AB, Williams FLR. ***An introductory guide to disease mapping***. Chichester: Wiley; 2001.
79. Devine OJ, Louis TA, Halloran ME. Empirical Bayes methods for stabilizing incidence rates before mapping. ***Epidemiology*** 1994; 5:622-630.
80. Assunção RM, Barreto SM, Guerra HL, Sakurai E. Mapas de taxas epidemiológicas: uma abordagem Bayesiana. ***Cad Saúde Pública*** 1998; 14:713-723.
81. Bailey TC, Gatrell AC. ***Interactive spatial data analysis***. New York: Longman; 1995.
82. Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. ***Stat Med*** 1992; 11:983-1007.
83. Richardson S. Statistical methods for geographical correlation studies. In: Elliot P, Cuzick J, English D, Stern R, editors. ***Geographical and enviromental epidemiology***, Oxford: Oxford University Press; 1996. p.181-204.
84. Haining R. ***Spatial data analysis in the social and environmental sciences***. Cambridge: Cambridge University Press; 1990.
85. Schwartz GG, Hanchette CL. UV, latitude, and spatial trends in prostate cancer mortality: all sunlight is not the same (United States). ***Cancer Causes Control*** 2006; 17:1091-1101.
86. Montenegro AC, Werneck GL, Kerr-Pontes LR, Barreto ML, Feldmeier H. Spatial analysis of the distribution of leprosy in the State of Ceará, Northeast Brazil. ***Mem Inst Oswaldo Cruz*** 2004; 99:683-686.
87. Cressie N, Read TRC. Spatial data analysis of regional counts. ***Biom J*** 1989; 31:699–719.
88. Jones AP, Langford IH, Bentham G. The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. ***Soc Sci Med*** 1996; 42:879-885.
89. Cook DG, Pocock SJ. Multiple Regression in Geographical Mortality Studies, with Allowance for Spatially Correlated Errors. ***Biometrics*** 1983; 39:361-371.
90. Pocock SJ, Shaper AG, Cook DG, Packham RF, Lacey RF, Powell P, Russell PF.British Regional Heart Study: geographic variations in cardiovascular mortality, and the role of water quality. ***Br Med J*** 1980;280:1243-1249.
91. Anselin L. Spatial econometrics. In: Baltagi B, editor. ***Companion to Theoretical Econometrics***. Oxford: Blackwell; 2001. p.310–330.
92. Okwi PO, Ndeng'e G, Kristjanson P, Arunga M, Notenbaert A, Omolo A, Henninger N, Benson T, Kariuki P, Owuor J. Spatial determinants of poverty in rural Kenya. ***Proc Natl Acad Sci USA*** 2007; 104:16769-16774.

93. Kelsall JE, Diggle PJ. Spatial variation in risk of disease: a nonparametric binary regression approach. ***Appl Statist*** 1998; 47:559-573.
94. Webster T, Vieira V, Weinberg J, Aschengrau A. Method for mapping population-based case-control studies: an application using generalized additive models. ***Int J Health Geogr*** 2006; 5:26.
95. Curtis S, Jones IR. Is there a place for geography in the analysis of health inequality? ***Soc Health Illness*** 1998; 20:645-672.
96. Barcellos C. Elos entre geografia e epidemiologia. ***Cad Saúde Pública*** 2000; 16:607-609.
97. Costa MCN, Teixeira MGLC. A concepção de "espaço" na investigação epidemiológica. ***Cad Saúde Pública*** 1999; 15:271-279.
98. Barcellos C, Bastos FI. Geoprocessamento, ambiente e saúde, uma união possível? ***Cad Saúde Pública*** 1996; 12:389-397.
99. Kearns RA, Joseph AE. Space in its place: developing the link in medical geography. ***Soc Sci Med*** 1993; 37:711-717.
100. Barcellos C, Sabroza PC. The place behind the case: leptospirosis risks and associated environmental conditions in a flood-related outbreak in Rio de Janeiro. ***Cad Saúde Pública*** 2001; 17(Suppl):59-67.
101. Gyapong JO, Remme JH. The use of grid sampling methodology for rapid assessment of the distribution of bancroftian filariasis. ***Trans R Soc Trop Med Hyg*** 2001; 95:681-686.
102. Arbia G. The use of GIS in spatial surveys. ***International Statistical Review***1993; 61:339-359.
103. Glass GE. Update: spatial aspects of epidemiology: the interface with medical geography. ***Epidemiol Rev*** 2000; 22:136-139.
104. Greenland S. Basic methods for sensitivity analysis of biases. ***Int J Epidemiol*** 1996; 25:1107-1116.
105. Werneck GL, Struchiner CJ. Estudos de agregados de doença no espaço-tempo: conceitos, técnicas e desafios. ***Cad Saúde Pública*** 1997; 13:611-624.

Artigo apresentado em 08/05//2008
Aprovado em 30/05/2008
Versão final apresentada em 20/06/2008