Geocomputation techniques for spatial analysis: are they relevant to health data?

Técnicas de geocomputação para análise espacial: é o caso para dados de saúde?

Gilberto Câmara ¹
Antônio Miguel Vieira Monteiro ¹

Abstract  Geocomputation is an emerging field of research that advocates the use of computationally intensive techniques such as neural networks, heuristic search, and cellular automata for spatial data analysis. Since increasing amounts of health-related data are collected within a geographical frame of reference, geocomputational methods show increasing potential for health data analysis. This paper presents a brief survey of the geocomputational field, including some typical applications and references for further reading.

Key words  Spatial Analysis; Geographical Methods; Computational Technics

Resumo  A geocomputação é um campo de pesquisa emergente que propõe o uso de técnicas intensivas em computação, tais como redes neurais, busca heurística e autômatos celulares para análise de dados espaciais. Com o aumento do volume de dados de saúde coletados dentro de um referencial geográfico, os métodos geocomputacionais demonstram um potencial crescente para a análise desses mesmos dados. Os autores apresentam uma revisão breve do campo da geocomputação, apresentando algumas aplicações típicas e sugestões bibliográficas.

Palavras-chave  Análise Espacial; Métodos Geográficos; Técnicas Computacional
Introduction

In recent years, the use of computer-based techniques for spatial data analysis has grown into an important scientific field, combining techniques from geographic information systems and emerging areas such as neurocomputing, heuristic search, and cellular automata. In order to distinguish this new interdisciplinary area from the simple extension of statistical techniques to spatial data, Oppenshaw & Abrahart (1996) coined the term “geocomputation” to describe the use of computer-intensive methods for knowledge discovery in physical and human geography, especially those involving non-conventional data clustering and analysis techniques. More recently the term has been applied in a broader sense to include spatial data analysis, dynamic modeling, visualization, and space-time dynamics (Longley, 1998).

This paper is a brief survey of geocomputational techniques. The review should not be considered exhaustive, rather attempting to provide an overview of the concepts and motivation behind the term “geocomputation”. Our prime motivation is to draw the attention of the public health community to the new analytical possibilities offered by geocomputational techniques. We hope this discussion will serve to broaden their perceptions of new possibilities in the spatial analysis of health data.

Motivations for research in geocomputation

Simply defined, geocomputation “is the process of applying computing technology to geographical problems”. According to Oppenshaw & Abrahart (1996:665), “Many end-users merely want answers to fairly abstract questions such as ‘Are there any patterns, where are they, and what do they look like?’”. Although this definition is generic, it points to a number of motivating factors, like the emergence of computerized data-rich environments, affordable computational power, and spatial data analysis and mining techniques.

The first motivation (data-rich environments) has come about through the massive collection of socioeconomic, environmental, and health-related data, increasingly organized in computerized databases with geographical references such as census tracts or postal codes. Even in Brazil, a country with a limited tradition of public availability of geographical data, the 2000 Census is being described as the first such initiative where all data collection will be automated and georeferenced.

The second motivation (computational power) has materialized in two forms: the emergence of the Geographic Information Systems (GIS) technology and of a set of algorithmically-driven techniques such as neurocomputing, fuzzy logic, and cellular automata.

The third motivation (data analysis and mining techniques) has been heavily driven by the application of data analysis techniques to spatial statistics, a research topic of considerable importance in recent decades.

The broad nature of challenges and approaches to geocomputational research is perhaps best illustrated by four different yet complementary approaches: computer-intensive pattern search, exploratory spatial data analysis, artificial intelligence techniques, and dynamic modeling, as described in the following sections.

Focus 1 – Computer-intensive pattern search

GAM – The Geographical Analysis Machine

One of the most typical examples of the computer-intensive approach to geocomputation is the Geographical Analysis Machine (GAM) developed by Stan Oppenshaw and co-workers at the Centre for Computational Geographics at the University of Leeds. For a recent survey of the GAM, see Oppenshaw (1998). The following description is largely based on Turton (1998).

GAM is basically a cluster finder for point or small-area data. Its purpose is to indicate evidence of localized geographical clustering in cases where statistical distribution of the phenomenon is not known in advance. For the GAM algorithm, a cluster is like a localized excess incidence rate that is unusual in that there is more of some variable than might be expected. Examples would include: a local excess disease rate, a crime hot spot, an unemployment black spot, unusually high positive residuals from a model, the distribution of a plant, surging glaciers, earthquake epicenters, patterns of fraud, etc (Figure 1).

The basic idea of the GAM is very simple. Within the study region containing a spatial point pattern, GAM works by examining a large number of circles of varying sizes that completely cover the region of interest. The circles overlap to a large degree to allow for edge effects and to provide a degree of sensitivity...
Within each random circle, one counts the number of points and compares this observed value with an expected value based on an assumption about the process generating the point pattern (usually that it is random). Ideally, the population at risk should be used as the basis for generating the expected value, such as using a Poisson probability model with the observed mean and the population at risk within each circle. Once the statistical significance of the observed count within a circle has been examined, the circle is drawn on a map of the region if it contains a statistically significant cluster of points. The process is repeated many times until a map is produced containing a set of circles centered on parts of the region where interesting clusters of points appear to be located.

A GAM application to infant mortality in Rio de Janeiro

Oppenshaw (1998) makes a strong case for performance of the GAM algorithm to locate clusters of diseases, including a comparison with other cluster-finding techniques. To better assess and understand the potentials and limitations of GAM, Teruiya et al. (1999) conducted an investigation using data from the study Spatial Analysis of Live-Born Profile and Socioeconomic Conditions in Rio de Janeiro, by D’Orsi & Carvalho (1998). This study assessed the spatial birth and socioeconomic patterns in subdivisions of the city of Rio de Janeiro, aiming to identify the main groups of infant morbidity and mortality risks and the selection of prime areas for preventive programs.

In order to apply the GAM algorithm, the values had to be converted from areal-related patterns to point variables. The authors selected some of the basic attributes used by D’Orsi & Carvalho and converted each area unit (corresponding to a city district) to a point location which received the value of the areal unit it represented, as illustrated by Figure 2.

The GAM algorithm was applied to the values for the live-born quality index for all neighborhoods of Rio. GAM found three clusters of high values for this index, located approximately in the Botafogo, Barra da Tijuca, and Ilha do Governador regions (Figure 3). The results were concentrated in what is perceived by the algorithm as “extreme” events of high values for the index, disregarding cases which are not “significant” enough. As a basis for comparison, the traditional choropleth-map is shown in Figure 4, where the areal-based values are grouped by quintiles.

It should be noted that we have used the Rio de Janeiro birth patterns merely as an ex-
Figure 2

Location of Rio de Janeiro urban sub-divisions.

Source: d’Orsi & Carvalho, 1997.

Figure 3

Clusters of high APGAR index values found by GAM.

GAM = Geographical Analysis Machine; APGAR = APGAR scoring for newborns.
ample to illustrate the computational behavior of the GAM technique. It is important to note that the algorithm was only searching for clusters of high values for the live-born quality index. Clusters of low values are disregarded by GAM, since the algorithm was originally conceived to find clusters of high disease incidence. We hope to motivate health researchers to apply the GAM techniques to problems closer to its original intended use, such as sets of epidemiological events.

Focus 2 – Exploratory spatial data analysis

Local spatial statistics

Statistical data analysis currently is the most consistent and established set of tools to analyze spatial data sets. Nevertheless, the application of statistical techniques to spatial data faces an important challenge, as expressed in Tobler’s (1979) *First Law of Geography*: everything is related to everything else, but near things are more related than distant things. The quantitative expression of this principle is the effect of *spatial dependence*: the observed values will be spatially clustered, and the samples will not be independent. This phenomenon, also termed *spatial autocorrelation*, has long been recognized as an intrinsic feature of spatial data, and measures such as the Moran coefficient and the semi-variogram plot have been used to assess the global association of the data set (Bailey & Gatrell, 1995).

Most spatial data sets, especially those obtained from geo-demographic and health surveys, not only possess global spatial autocorrelation, but also exhibit significant patterns of spatial instability, which is related to regional differentiations within the observational space. As stated by Anselin (1995), "the degree of non-stationarity in large spatial data sets is likely to be such that several regimes of spatial association would be present".

In order to assess the degree of spatial instability, various local spatial statistics indicators have been proposed, such as the Moran Local $I_i$ (Anselin, 1995), the Moran scatterplot (Anselin, 1996) and the $G_i$ and $G_i^*$ statistics (Ord & Getis, 1995). For a recent review, see Getis & Ord (1996). Although local spatial statistics can be seen as a branch of spatial statistics, they have been highly praised by geocomputational proponent Stan Oppenshaw: "It is absolutely fundamental that we can develop tools able to detect, by any feasible means, patterns and localized association that exist within the map" (Oppenshaw & Abrahart, 1996:665).
Spatial statistics as a basis for zoning: social exclusion/inclusion in São Paulo

In order to assess the validity of local spatial statistics, the authors have conducted a project to study the potential of such indicators as a basis for the design of administrative zoning systems for the city of São Paulo. As is well known, zone design is a major challenge for urban and regional planners, since it involves major decisions on how to distribute public resources.

São Paulo is one of the world’s largest cities and presents a major challenge to urban planners and public administrators. Given its present size (over 13 million inhabitants) and enormous socioeconomic inequalities, rational planning of the city requires a careful division of the urban space into administrative regions that are homogenous by some objective criteria. Unfortunately, the current regional division of São Paulo has been driven by historical and political forces and fails to reflect a rational attempt to challenge the city’s disparities.

As a basis for a zoning design for São Paulo, we have taken the “Social Exclusion/Inclusion Map of São Paulo”, a comprehensive diagnosis of the city coordinated by Prof. Aldaiza Sposati of the Social Research Group at the Catholic University of São Paulo. This map used 49 variables obtained from Census data and local organizations to quantify the social apartheid in 96 districts of São Paulo (Sposati, 1996).

The main results of the “social exclusion/inclusion map” were indicators of social exclusion and disparities in quality of life in São Paulo. Figure 5 shows the map of the social exclusion index $I_{ex}$, where the values vary from -1 (maximal social exclusion) to +1 (maximal social inclusion) with a value of 0 indicating the attainment of a basic standard of inclusion (Sposati, 1996). Note from the map that two-thirds of the districts in São Paulo have acceptable living standards.

Taking the social exclusion index as a basis, the proposed task was to group the 96 districts into a set of administrative zones, each containing a significant number of districts and homogeneous with respect to social exclusion status. We used two exploratory spatial analysis tools: the Moran Scatterplot Map (Figure 6, left) and the local Moran index significance map (Figure 6, right). The basis for these local spatial statistics indicators is the use of a neighborhood or contiguity matrix $W$ whose elements are $w_{ij} = 0$ if $i$ and $j$ are not neighbors and non-zero otherwise.

The Moran scatterplot map is a tool for visualizing the relationship between the observed values $Z$ and the local mean values $WZ$, where $Z$ indicates the array of attribute values (expressed as deviations from the mean) and $WZ$, is the array of local mean values, computed using matrix $W$. The association between $Z$ and $WZ$ can be explored to indicate the different spatial regimes associated with the data and to display graphically as indicated by Figure 6 (left). The Moran Scatterplot Map divides spatial variability into four quadrants:

- Q1 (positive values, positive local means) and Q2 (negative values, negative local means): indicate areas of negative spatial association.
- Q3 (positive values, negative local means) and Q4 (negative values, positive local means): indicate areas of positive spatial association.

Since the $I_{ex}$ variable exhibits global positive spatial autocorrelation (Moran $I = 0.65$, significance = 99%), areas in quadrants Q3 and Q4 are interpreted as regions that do not follow the same global process of spatial dependence, and these points indicate transitional regions between two different spatial regimes.

The local Moran index $I_i$ is computed by multiplying the local normalized value $z_i$ by the local mean (Anselin, 1995):

$$I_i = \sum_{j \neq i} w_{ij}(z_j - \bar{Z}) (z_i - \bar{Z})$$

$$I_{ex}$$ variable exhibits global positive spatial autocorrelation (Moran $I = 0.65$, significance = 99%), areas in quadrants Q3 and Q4 are interpreted as regions that do not follow the same global process of spatial dependence, and these points indicate transitional regions between two different spatial regimes.

The local Moran index significance map indicated three “hot spots”, two of which related to low values of inclusion (located to the South and East of the city) and one related to high values of inclusion (located in the Center of the city). These patterns correspond to the extreme regions of poverty and wealth in the city and were chosen as “seeds” in the zoning procedure.

The remaining regions were defined interactively, taking into account the Moran scatterplot map, which clearly indicates a number of transition regions between the regions of Q1 and Q2 locations (to so-called “high-high” and “low-low” areas), some of which are indicated by the ellipses. These regions were grouped into separate zones. The work proceeded interactively until a final zoning proposal was produced, which can be confronted with the current administrative regions (Figure 7).

In order to assess the resulting map, a regression analysis was performed. This regression analyzes the correlation between the per-
percentage of houses with proper sewage facilities (as independent variable) and the percentage of people over 70 years of age (as dependent variable). The rationale behind this choice was that social deprivation is a serious impediment to healthy living, as measured by the percentage of elderly in the population. Three OLS (ordinary least squares) regression analyses were performed: the first, taking all districts of the city overall; the second, using the current administrative division as separate spatial regimes; and the third, using the proposed new zoning as spatial regimes. The results as summarized in Table 1.

These results are a positive indication of the possible use of local spatial statistics as a basis for zoning procedures and show how indicators such as the social exclusion index of Sposati (1996) can be used as a support for urban planning.

Focus 3 – Neural networks and geographic analysis

Introduction

An Artificial Neural Network (ANN) is a computer paradigm inspired by the way the brain processes information. The key element in this paradigm is a processing system composed of a large number of highly interconnected elements (neurons) working in unison to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process (Gopal, 1998).

In principle, ANNs can represent any computable function, i.e., they can do everything a normal digital computer can do. In practice, ANNs are especially useful for classification and function approximation and mapping problems which are tolerant of some imprecision and have plenty of training data available. Almost any mapping between vector spaces can be approximated to arbitrary precision by feedforward ANNs (which are the type most often used in practical applications) if there are enough data and enough computing resources.

Given the capabilities of ANNs as exploratory tools in data-rich environments, there has been considerable interest in their use for spatial data analysis, especially in remote sensing image classification (Kannelopoulos, 1997; Leondes, 1997). Other geographical applications include: spatial interaction modeling (Gopal & Fischer, 1996; Oppenshaw, 1993) and classification of census data (Winter & Hewitson, 1994).

Neural networks for spatial data integration: an economical-ecological zoning application

To illustrate the potential of ANN for spatial data analysis, we have selected one example: the use of neural networks for the integration of multiple spatial information for an environmental zoning application (Medeiros, 1999). Although the chosen application does not involve health data, the integration procedure shown is relevant to health-assessment applications, which involve multiple data sets as possible sources of epidemiological risk.

One of the more important problems in geographical data analysis is the integration of separate data sets to produce new spatial infor-
mation. For example, in health analysis, a researcher may be interested in assessing the risks associated with a disease (such as malaria) based on a combination of different conditions (land use and land cover, climatology, hydrological information, and distance to main roads and cities). These conditions can be expressed as maps, integrated into a common geographical database by means of GIS technology.

Once the data has been organized in a common geographical reference, the researcher needs to determine a procedure to combine these data sets. Taking a hypothetical example, a health researcher may want to calculate a risk map for malaria based on known disease incidence, climate, distance to cities, and land cover, where the conditions are such that a region is deemed “high risk for malaria” if it rains more than 1000 m/year and the land cover is “inundated forest” and is located less than 50 km from a city.

The main problem with these map inference procedures is their ad hoc, arbitrary nature: the researcher formulates hypotheses from previous knowledge and applies them to the data set. The process relies on inductive knowledge of the reality. Additionally, when the input maps have many different conditions, the definition of combinatory rules for deriving the output may be difficult. For example, if an input map has eight different conditions (e.g., land cover classes) and five maps are to
be combined, then $8^5 = 32,768$ different situations have to be taken into account.

There are two main alternative approaches to this problem. One is to use fuzzy logic to combine the maps (Câmara et al., 2000). In this case, all input data are transformed into fuzzy sets (in a $[0,1]$ scale) and a fuzzy inference procedure may be used. Alternatively, the use of neural network techniques aims at capturing the researcher’s experience, without the need for the explicit definition of the inference procedure. The application of neural networks to map integration can be done using the following steps:

- Create a georeferenced database with the input (conditional maps)
- Select well-known regions as training areas. For these areas, indicate the desired output response (such as health risk).
- Use these training areas as inputs to a neural network learning procedure.
- Using the trained network, apply the inference procedure for the entire study region.
- Evaluate the result and redo the training procedure, if necessary.

This idea was applied by Medeiros (1999) in his study of the integration of natural resources data as a basis for economical-ecological zoning in the Amazon region. Medeiros used five data sets as input: vegetation, geology, geomorphology, soils, and remote sensing images. Medeiros (1999) compared the result obtained

**Figure 7**

Focus 4 – Cellular automata

The computer representation of geographical space in current GIS technology is essentially static. Therefore, one important research focus in geocomputation aims to produce models that combine the structural elements of space (geographical objects) to the processes that modify such space (human actions as they operate in time). Such models would free us from static views of space (as centuries of map-making have conditioned us) and to emphasize the dynamic components as an essential part of geographical space.

This motivation has led to the use of cellular automata as a technique for simulation of urban and regional growth. Cellular automata (CA) are very simple dynamic spatial systems in which the state of each cell in an array depends on the previous state of the cells within a neighborhood of the cell, according to a set of transition rules. CA are very efficient computationally because they are discrete, iterative systems that involve interactions only within local regions rather than between all pairs of cells. The good spatial resolution that can thus be attained is an important advantage when modeling land use dynamics, especially for planning and policy applications (White & Engelen, 1997).

A conventional cellular automaton consists of: (a) a Euclidean space divided into an array of identical cells; (b) a cell neighborhood; (c) a set of discrete cell states; (d) a set of transition rules which determine the state of a cell as a function of the states of cells in the neighborhood; and (e) discrete time steps, with all cell states updated simultaneously.

The application of CA to geographical systems was first proposed by Tobler (1979). More recently, a number of researchers have proposed modifications of the original CA idea to accommodate geographical constraints. The most important characteristic to be discarded is the homogeneous cell space, replaced by a space in which each cell has its own inherent set of attributes (as distinct from its single state) which represent its relevant physical, environmental, social, economic, or institutional characteristics. These advances have been accompanied by an increase in the models’ complexity (Couclelis, 1997; White & Engelen, 1997).

This modification has allowed CA models to be linked both conceptually and practically with GIS. Since the CA is running on an inhomogeneous cell space (essentially identical to what would be found in a raster GIS), the CA may be thought of as a sort of dynamic GIS (Batty & Xie, 1994). At present, however, CA models developed in GIS remain simple, because GIS do not yet provide operators with sufficient flexibility to define complex CA transition rules, and in addition they lack the simulation engines needed to run complex models at practical speeds. The more practical approach is to couple GIS to special-purpose CA software modules, and possibly other models as well. White et al. (1997) have developed several CA and CA-based integrated models designed as prototypes of Spatial Decision Support Systems for urban and regional planning and impact analysis (demos of several of these models can be downloaded from <http://www.riks.nl/RiksGeo/freestuff.htm>).

In conclusion: geocomputation as a set of effective procedures

This survey has examined some of the main branches of research in geocomputation, and we conclude the paper with an attempt to provide a unified perspective of this new research field.

We propose that a unifying perspective for geocomputation is the emphasis on algorithmic techniques. The rationale for this approach is that the emergence of data-rich spatial databases motivated a new set of techniques for spatial data analysis, most of them originally proposed under the general term “artificial intelligence”, such as neural networks, cellular automata, and heuristic search.

Since there are fundamental differences in the perspectives of the set of techniques used by geocomputation, the only unification per-
Neural network based inference.

### Table: ANN Modeling and Modeling via LEGAL Map Algebra

<table>
<thead>
<tr>
<th></th>
<th>Background</th>
<th>Stable</th>
<th>Moderate-stable</th>
<th>Mean-stable/vulnerable</th>
<th>Vulnerable</th>
<th>Sum. Xj</th>
<th>Xj/Sum. Xij (%)</th>
<th>Producer exactness (%)</th>
<th>Omission error (%)</th>
<th>K %</th>
<th>Overall Exactness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background</strong></td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>8</td>
<td>16</td>
<td>142</td>
<td>210</td>
<td>0.47</td>
<td>0.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td><strong>Stable</strong></td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>21</td>
<td>0.5</td>
<td>11.76</td>
<td>91.89</td>
<td></td>
</tr>
<tr>
<td><strong>Moderate-stable</strong></td>
<td>1</td>
<td>0</td>
<td>19861</td>
<td>434</td>
<td>782</td>
<td>159</td>
<td>21237</td>
<td>47.26</td>
<td>93.52</td>
<td>6.448</td>
<td></td>
</tr>
<tr>
<td><strong>Mean-stable/vulnerable</strong></td>
<td>0</td>
<td>0</td>
<td>473</td>
<td>2598</td>
<td>17</td>
<td>6</td>
<td>3094</td>
<td>6.88</td>
<td>83.97</td>
<td>16.03</td>
<td></td>
</tr>
<tr>
<td><strong>Moderate-vulnerable</strong></td>
<td>0</td>
<td>0</td>
<td>671</td>
<td>26</td>
<td>10166</td>
<td>106</td>
<td>10969</td>
<td>24.41</td>
<td>92.68</td>
<td>7.32</td>
<td></td>
</tr>
<tr>
<td><strong>Vulnerable</strong></td>
<td>0</td>
<td>2</td>
<td>219</td>
<td>13</td>
<td>122</td>
<td>9053</td>
<td>9409</td>
<td>20.94</td>
<td>96.22</td>
<td>3.78</td>
<td></td>
</tr>
<tr>
<td><strong>Sum. Xj</strong></td>
<td>1</td>
<td>17</td>
<td>21268</td>
<td>3079</td>
<td>11103</td>
<td>9472</td>
<td>44940</td>
<td>90.00</td>
<td>100.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Xj/Sum. Xij</strong></td>
<td>0.00</td>
<td>0.04</td>
<td>47.33</td>
<td>6.85</td>
<td>24.71</td>
<td>21.08</td>
<td>100.00</td>
<td>41693.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Producer exactness (%)</strong></td>
<td>0.00</td>
<td>88.24</td>
<td>93.36</td>
<td>84.36</td>
<td>91.56</td>
<td>95.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Omission error (%)</strong></td>
<td>100.00</td>
<td>11.76</td>
<td>6.62</td>
<td>15.62</td>
<td>8.44</td>
<td>4.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ANN** = Artificial Neural Network; **Sum.** = summation; **LEGAL** = Map Algebra Language.
spective is the computational side: such techniques can be thought of as a set of effective procedures that, when applied to geographical problems, are bound to produce results. Whatever results are obtained need to be interpreted in light of the basic assumptions of these techniques, and it may be extremely difficult to assign any traditional "statistical significance" criteria to them.

Therefore, the authors propose a tentative definition: "Geocomputation is the use of a set of effective computing procedures to perform spatial data analysis, whose results are dependent on the basic assumptions of each technique and therefore are not strictly comparable".

According to this view, geocomputation emphasizes the fact that the structure and data dependency inherent in spatial data can be used as part of the knowledge-discovery approaches, and the choices involve theory as well as data. This view does not den the importance of the model-based approaches, such as the Bayesian techniques based on Monte Carlo simulation for the derivation of distribution parameters on spatial data. In fact, in this broader perspective, the use of Bayesian techniques that rely on computationally intense simulations can be considered a legitimate part of the geocomputational field of research.

In conclusion, what can public health researchers expect from geocomputation? When used with discretion, and always bearing in mind the conceptual basis of each approach, techniques such as GAM, local spatial statistics, neural nets, and cellular automata can be powerful aids to a spatial data analysis researcher, attempting to discover patterns in space and relations between its components.

We hope this article serves as inspiration to health researchers and that it will have broadened their notions about what is possible in spatial data analysis.

Acknowledgments

The authors wish to thank Marilia Carvalho and Oswaldo Cruz at Fundação Oswaldo Cruz for the invitation and challenge to apply GIS techniques to health data. We also acknowledge the important contribution of our fellow colleagues and students at Instituto Nacional de Pesquisas Espaciais (INPE): Eduardo Camargo, José Simeão de Medeiros, Marcos Corrêa Neves, Roseli Teruiya, Virginia Correa, and to the entire SPRING software team at INPE.

Further reading

For readers interested in more information on geocomputation, we provide a set of references, organized by topics. We suggest that prospective readers begin with Longley (1998) and then proceed to their specific area of interest.

References


The authors are to be commended on this interesting and thought-provoking review of the new analytical possibilities offered by "geocomputation" techniques, and I concur with them that there is considerable potential for useful applications of such techniques in the spatial analysis of health data. I agree wholeheartedly that the wider dissemination of such methods along with associated software tools may ultimately benefit many areas of geographical health and environmental research. We do indeed face a "data-rich" future in those fields of study and one where data will be not only voluminous but also complex. I mean both complex in content (e.g. in the topographic and geographical detail provided by GIS and remote sensing) and also complex in structure (e.g. data from disparate sources relating to different geographical scales and reference frameworks that need to be integrated in the study of many issues of interest in health research). Indeed one suspects that the future may already be with us! Traditional spatial analysis methods are not designed to handle such data complexity (e.g. many make little use of anything more sophisticated than simple Euclidean distance or the contiguity of areal units in order to reflect proximity, many assume some form of stationarity in the processes modeled, and few can handle data sources at different levels of spatial aggregation). We undoubtedly do need new analysis methods that are capable of exploiting more complex concepts. The authors convince me that geocomputational research offers some promising avenues for achieving that, and this paper and the work referenced in it therefore deserves serious and careful consideration by those involved in geographical health research.

However, while generally enthusiastic about the possibilities offered by the techniques discussed and agreeing with much that is said in the paper, there are some issues which I would like to take up from the perspective of an applied statistician with an interest in spatial analysis, and I restrict my remaining remarks to those.

First, I do not consider that we need to think of geocomputational techniques as an alternative to more traditional statistical methods and models, but rather as a complement to them. Modern statistical analysis is itself a broad church and no stranger to computer-intensive methods. To establish a dividing line between many existing forms of descriptive or exploratory statistical analysis and geocomputation may be useful in order to focus attention and promote the use of novel forms of algorithmic approach. However, from the point of view of a practicing statistician, such a distinction is somewhat artificial. Many existing forms of visualization and projection techniques used in statistics, particularly those employed in the analysis of high-dimensional data, have little to do with traditional notions of statistical inference. Statisticians are quite comfortable and familiar with using essentially algorithmic methods where appropriate and have been doing so for many years. What matters in exploring data is that the analyses conducted are careful and thorough, not what type of algorithms are employed to achieve that. So I do not see geocomputation as competing with my current statistical exploratory tool box, but rather as adding to it (in fact I consider two of the methods discussed in this paper, the GAM and local indicators of spatial association, to already be a part of it, although I am happy to see them re-branded as geocomputation if it encourages their use!).

However, I do stress that I see geocomputational techniques as essentially exploratory, and that brings me to my second point. Answers to the questions: Are there any patterns, where are they, and what do they look like? Are undoubtedly of value, but ultimately they are a preliminary to the more important ones of Why are they there, will they happen again, and how will they change if we intervene in a particular way? The answer to this second set of questions requires a scientific explanation of the phenomenon under study, and given the intrinsically stochastic nature of most social, environmental, and health-related phenomena, the best tool for this will, I suspect, remain the statistical model. I am not suggesting that such models will be true; the very word model implies simplification and idealization and I fully appreciate that that complex geographical health and environmental systems cannot be exactly described by a few formulae. However, the construction of idealized representations of important aspects of such systems consistent with the existing substantive epidemiological or public health knowledge should remain the ultimate goal. I therefore see the primary value of geocomputation as assisting in the sta-
tistical model-building process and not circum-
venting it.

This view of the role of geocomputation (which I freely admit may be narrower than that held by the authors) leads me to my third point, which relates to various concerns over the practical use of some kinds of geocomputational algorithms. The process of model-building is ideally both interactive and iterative. The analyst needs to try out ideas on the data, and this requires exploratory tools that can be guided or steered towards particular chosen ends or hypotheses. At present, many geocomputational algorithms appear too much of a “black box” to make this possible. The very nature of the algorithms makes it difficult to provide simple, readily understood control parameters which enable them to be “steered” towards answering particular questions which one might wish to ask of the data. In a sense they provide an answer in the absence of a question. This detracts from their value as exploratory tools for the model builder. In that sense what is often termed “artificial intelligence” might be better referred to as “artificial un-intelligence”. There is also the problem of whether such techniques produce robust results as opposed to ones which are pure artifacts of the data. I appreciate that traditional notions of statistical significance and standard error cannot and perhaps should not be looked for in relation to these algorithms and that different algorithmic approaches will naturally reveal different aspects of the data. However, the sensitivity of the results from any one of them (e.g. to starting conditions or in repeated application to various subsets of the data) needs to be investigated and is often not. If the data are to be mined then we need to establish whether a vein of gold has been found or a vein of fool’s gold, and currently the algorithms are weak on the diagnostics that would enable us to measure that.

In summary I do not wish to appear as a dogged defender of existing spatial statistical models and methods. I am well aware how deficient many of those are. For example, traditional spatial models largely involve space in terms of glib abstractions – “distances”, “boundaries”, and “edge effects”. Of course in reality the areas over which analyses are being conducted are vastly complex, criss-crossed with natural boundaries such as forests, rivers, or ranges of hills, or else human constructions such as roads, industrial estates, recreational parks, and so on. Many commonly used spatial statistical methods and models should be viewed in the cold light of their spatial simplic-

Epidemiologists, after several decades of favoring non-spatial statistical models, are increasingly realizing the importance of understanding socioeconomic and ecological contexts in the interpretation of disease patterns in populations (McMichael, 1999). As the questions we are asking change in both scope and nature, input from scholars in non-health fields with expertise in studying spatial patterns, such as this paper, are a welcome addition to the health literature.

The authors state that their intent is to “draw the attention of the public health community to the new analytical possibilities offered by geocomputational techniques”. While the introduction of these techniques to health researchers is laudable in and of itself, I would like to throw out some cautionary notes, based on some experience working with interdisciplinary teams where these techniques have been proposed.
None of the motivations listed by the authors, apart from what they refer to as the “abstract” searching for patterns, are grounded in asking scientific or scholarly questions. Certainly the identification of disease patterns is an important first step, but without carefully thinking through the nature of the disease and how it is spread, combining maps of various outcomes and characteristics can be both misleading and also dangerous, to the extent that it leads to misdirection of funds to attack diseases in particular ways.

As they themselves acknowledge, the major motivating force behind the use of many of these techniques is simply “the emergence of computerized data-rich environments” and the availability of “affordable computational power.” My experience has been similar, and as a scientist I am very skeptical of such motivation. It leads to researchers confusing their units of analysis, slipping between individuals, communities, and regions, or combining them in the same maps, and making false inferences across scales. The determinants of cases and the determinants of incidence rates are often quite different (Rose, 1985). I shudder to think that we are training young scholars who are driven by a mere fascination with technology and who have forgotten how to frame clear, important questions and design studies to answer them.

For instance, the example they give of regressing percentage of people over 70 years on percentage of houses with proper sewage facilities is based on the problematic assumption that the populations and sewage disposal of urban neighborhoods have been stable over time. Older people may have grown up in the countryside and only moved to those urban areas as adults (poverty often being associated with old age): thus migration patterns may be the major determinants of percentage of people over 70. Or increasing population densities may have interacted with sewage disposal methods to create problems over time; in this case it is most important to understand demographic and sewage production and disposal dynamics of those urban neighborhoods over the past seven decades. It seems to me that before jumping into the computational techniques, the researchers need to propose a clear theoretical framework and a biological and socially substantive logic which leads to specific questions to be answered in the research.

A further concern I have with the focus on these newer techniques of analyses is that researchers sometimes ignore the sources and quality of data, how they were collected, and their real spatio-temporal distribution. Data collected from referral hospital and health center records, based on diagnostic tests and questionnaires with a wide range of sensitivities, specificities, and precision cannot simply be lumped together with satellite data to produce meaningful information. Sometimes simple hand-drawn maps combined with intensive community survey or focus group work may be what is needed most.

Health researchers are facing important and often unprecedented questions in the 21st century. How can we create sustainably healthy societies? What are the relationships between economic policy, environmental change, and human health? How might global warming affect changes in regional disease patterns? I have no doubt that geocomputational techniques can make important contributions to answering these questions. The authors recognize that “the results are dependent on the basic assumptions of the technique”, and that researchers should use these techniques “with discretion, and always bearing in mind the conceptual basis of each approach”. I only wish they had spent a little more time and space exploring those assumptions and concepts, to enable those of us who are novices to more carefully select those techniques most suitable to the questions we seek to answer.

neural networks, fuzzy logic, genetic algorithms, and cellular automata for spatial data analysis.

The study of spatial and spatial-temporal epidemiological data is a timely issue which is driven by both decreasing technology costs and increasing availability of information. For example, it is becoming increasingly possible to access georeferenced public health data in a speedy manner through the Internet for analyzing and merging with other information. Several models and methods to work with spatial health-related data have cropped up in the literature in the last twenty years. Most of these were developed in other areas, like geostatistics, which originated in the mining industry and was later borrowed to help understand and explain the spatial distribution of health events. As is common in many applied sciences, the method is first introduced in an intuitive way, and once the heuristic results prove encouraging, there is major involvement by mathematical and statistical theorists to get the technique soundly established. The wave of progress following this pattern continues with Câmara and Monteiro’s paper, presenting a basic review of existing possibilities for the use of different computing procedures to perform spatial health data analysis.

On the application side, I would partially support the motivating statement of the paper citing Oppenshaw (1996), that “many end users merely want answers to fairly abstract questions ...”. However, some care should be exercised here. Some twenty years ago I heard in a Brazilian workshop on statistical methods for epidemiologists, particularly on multiple regression, that the basic concepts are cumbersome and difficult to be understood by public health workers, and that they should be more involved in collecting good data to be analyzed by the “foreigners”, i.e., specialists in statistics. Obviously, the authors of the paper would not wish us to merely engage in using these “black box” tools (which are well understood in the artificial intelligence community) but rather, that we begin close collaboration to both further the knowledge of these new methods and convince ourselves that they could be included in the analytical tool box of epidemiologists and public health professionals.

The authors provide examples of real analyses in the hope of giving a genuine applied flavor to the methods reviewed. I wish to make some comments on these applications. The first concerns the use of the GAM (Geographical Analysis Machine) to find clusters in data that are originally areal data. Although the authors emphasized that it is only an example, there is no mention of the large differences in area sizes and population distribution in Rio de Janeiro’s districts, which I believe could substantially influence the results. If one uses some sort of altered or transformed data set, one must interpret it with caution and be certain that the alteration is stated clearly to avoid misuse by newcomers to this field of spatial analysis research.

My other point concerns Section 3, on neural networks and geographical analysis, where the authors present a classification problem to produce a map of environmental vulnerability. One of the most fundamental aspects of neural network modeling is the requirement of “plenty of training data”, which is properly identified in the paper. Neural networks are “adaptive computing” in that they learn from data to build a model. Therefore, the training data set should contain all examples of possible sets of explanatory and outcome variables if one uses the workhorse of neural network modeling: a feed forward network with a back propagation algorithm. Users interested in applying this new technology should be aware of this important aspect. In addition, analysts must be willing to both tolerate the large amount of time for training and have a “black box” model which unfortunately does not provide the ability to explain the reasoning used to arrive at a result. This still limits the usefulness of this technique in some areas, particularly when one is interested in measuring the effects of input variables rather than prediction.

Recent developments in computing performance have provided a wealth of opportunities for advancement of new analytical approaches to spatial data analysis. These include the increasing use of Bayesian thinking, particularly with the introduction of the Monte Carlo Markov Chain (MCMC) approach to tackle intractable integrals. For the unfamiliar reader, the paper provides a brief introduction to various techniques. Some of these techniques were derived from the so-called intelligent systems, and it is the hope of the authors, and also mine, that they may assist our capability to convert data into information.
Câmara and Monteiro have the merit of drawing the attention of epidemiologists to new spatial analysis techniques. They have done a good job of summing up the main methodologies recently developed and presenting examples of their use, along with recent bibliographical references. As a statistician, I wish to focus my comments on the relationship between the area referred to as geocomputation by the authors and the usual statistical methods.

Before emphasizing differences, it is necessary to identify commonality. A discipline is merely a label for a set of knowledge and practices exercised by people who use that label to refer to themselves. Such practices and knowledge change dynamically, and statistics is no exception. Currently, many of the methods described by the authors are found in the best and most traditional statistics journals. In particular, the first two topics, GAM (Geographical Analysis Machine) and local spatial statistics are topics of articles and books by statisticians interested in spatial analysis. The other two, neural networks and cellular automata, are less common but not totally absent. As the authors point out, the presence of these topics is due to the current combined availability of data and computational power. I still do not feel comfortable in identifying geocomputation as a defined field of work, since most of the techniques presented emerged in traditional contexts (as in the case of the first two topics) or non-geographical ones (the last two). But this is not relevant for using and learning the techniques presented by the authors, which are useful regardless of labels.

Although the latter two topics, neural networks and cellular automata, are not absent from the statistical literature, they are less present than might be expected nowadays. Thus, one might ask, what can statisticians learn from researchers in neural networks and cellular automata? I believe that we should be less concerned with asymptotic results and optimality and seek methods that function well with large databases. We should deal with large and difficult problems involving a large number of parameters and depending little on hypotheses that cannot be verified. We should be alert to algorithms like the steepest descent with learning rates that can be highly useful in order to avoid over-adjustment in models with many parameters. This could be useful mainly for Bayesian models, which have become increasingly important (Assunção et al., in press).

On the other hand, what can researchers of neural networks and cellular automata learn from statisticians? I believe they should be a little more concerned with the statistical properties of their methods and perhaps slightly more with their optimality. They should make greater effort to compare their methods with others, including simpler traditional statistical methods. Often a linear regression can have a performance similar to that of a multi-layer perceptron. Contrary to what the authors state at the end of their article, the results of techniques like those presented are comparable. A clear example of this is the book by Alexander & Boyle (1997), which presents various techniques, including GAM, used by their respective creators in a set of simulated maps which might or might not present disease foci. The process of generating maps was described to the researchers, but not what each particular map contained. This simulation exercise, although displaying limitations, served to clearly demonstrate that some methods should be abandoned once and for all, and the GAM was not among them.

The techniques and examples presented in the article are a good sample of what new computer-intensive methods can offer. Research on these methods is increasing continuously and will no doubt continue over this decade. The authors are to be congratulated for having raised the topic and for having motivated health researchers to take interest in these new methods.


What do public health researchers expect of geocomputation?

The article by Gilberto Câmara and Antônio Miguel Vieira Monteiro is highly interesting and objective. While it introduces the concept of geocomputation in a clear and didactic way, demonstrating its potential as a tool for analyzing spatial data, it also invites the reader to answer the question at the end with the same clarity as the authors: what do we expect from geocomputation?

Epidemiology seeks to improve the methods and techniques that allow it to describe, explain, and predict health and disease phenomena in populations, with a view towards prevention. Therefore it plays a fundamental role in public health. From this perspective, the analysis of spatial distribution of diseases has contributed to the production of knowledge in the field and should not be seen as a “second-class” replacement for studies focusing on the individual as the unit of analysis (Susser, 1994a).

Depending on the problem one wishes to solve, the ecological approach has its indications and specificities. Thus, studies can focus on mapping the geographical distribution of diseases with the identification of spatial clusters of cases and the analysis of associations between the incidence of diseases/events and environmental or contextual exposures related to the collective sphere.

How can geocomputation help improve such studies? We must first ask if we really understand what is being offered to us.

Reading the article was certainly enlightening, providing us with the scope of development of techniques and the analytical possibilities offered by the various methods. The authors facilitated an understanding of the concepts by giving a detailed development of the theme through examples of health-related and environmental situations. For example, we are left with the idea that the four types of approaches presented by the authors have different premises and objectives but can be viewed as complementary.

Thus the use of GAM (the Geographic Analysis Machine) is capable of revealing clusters of events/diseases and constructing maps when the excess rates found are statistically significant. For example, this would be a useful technique for detecting priority areas for public health interventions, and would not aim at helping explain the occurrence of phenomena.

Meanwhile, techniques for the detection of “spatial autocorrelation”, measured by the Moran coefficient or through semi-variograms, would detect dependence between geographically proximate events, “explicitly considering the possible importance of their spatial arrangement in the analysis or interpretation of the results” (Bailey & Gatrell, 1995). There are thus specific indications for this type of research, for example: when one’s point of departure is the hypothesis that the event at issue is generated by environmental factors that are difficult to detect at the individual level.

The other two approaches described by the authors involve more sophisticated techniques, incorporating functions aimed at contemplating the complexity of the phenomena. The authors explain that an Artificial Neural Network (ANN) can be used as an exploratory tool in data-rich environments and that it is capable of integrating different types of nature in a single geographic data base using Geographic Information Systems (GIS) technology. The information to be introduced into the model should be chosen by the researcher, which obviously presupposes the existence of an underlying theoretical basis.

Meanwhile, cellular automata move even further in the sense of incorporating dynamic elements into the models. These models “would free us from static views of space” and would be capable of representing the change in space over time as the product of human actions.

We may be closer to achieving the ambitious objective identified by Susser (1994b) (speaking of the logic of the ecological approach): to understand how the context affects the health of individuals and groups. In other words, it appears increasingly possible to develop studies that reveal the effects not only of the structural elements of space but also those of the processes, not perceptible within the sphere of studies whose unit of analysis is the individual. Hence epidemiology turned to critical geography for the concept of “socially organized space”.

Finally, the authors point out that computational technology for solving health problems should always be applied keeping in mind the conceptual underpinnings of each approach. This concern has its counterpart in the health field. The conceptual basis to be considered in studies should be related to the theoretical and methodological issues of public health. This underscores the need for an interdisciplinary dialogue, where the respective challenge for the public health researcher is to guarantee the epidemiological content of the
studies, allowing for better knowledge of the target phenomenon, prediction of new occurrences, and the organization of interventions aimed at prevention.


Claudio J. Struchiner

First I would like to express my appreciation to the authors for this impressively wide-ranging paper. It is a review paper that provides an introduction to geocomputation techniques, i.e., computer-intensive techniques for knowledge discovery in physical and human geography. The authors seem to favor the view that this new interdisciplinary area is to be distinguished from the simple extension of statistical techniques to spatial data. My comments are motivated by questions I have posed to myself after reading their review: How do such methods compare to established techniques? What are their advantages and disadvantages? What are their ranges of applications? Do the new techniques challenge or extend any of the existing paradigms in data analysis?

The computational dimension appears to be the common denominator of the techniques described in this review and goes into the definition of the key concept at stake, geocomputation. Faster and more powerful computers and advances in software engineering have had a profound impact on all areas of statistics. Bootstrap and Monte Carlo Markov Chain (MCMC) methods, for example, allow the estimation of parameters in richer and more realistic model-based representations of natural phenomena, thereby freeing the imagination of the scientific community. In this context, the boundaries of statistical models and statistical theory have been extended, while preserving the current paradigms, i.e., good statistical thinking is based on solid philosophical principles.

Algorithmic thinking also plays an important role in other areas of science. Complex systems can be generated through the use of very simple building rules, which resemble the functioning of DNA chains. In this context, computer-intensive algorithmic techniques are intimately related to the mechanisms of pattern formation that supposedly occur in nature. In opposition, the procedures under the heading of geocomputation also seek to uncover pattern formation, but their search mechanisms are general in nature and do not bear any relationship to the various possible mechanisms that generate those spatial patterns.

In my view, the geocomputational methods reviewed in this paper do not share the same principles as these extensions. These algorithmic techniques appear to be a computerized version similar, in spirit, to a once very fashionable set of techniques developed by J. Tukey and known as Exploratory Data Analysis. Other statistical techniques put together under different headings such as Data-Driven Procedures and Data Mining attempt to answer similar questions raised here, i.e., “Are there any patterns, what are they, and what do they look like?”

The literature on quantitative methods has acknowledged, at least since the beginning of this century, the existence of two dimensions in research practice, i.e., exploratory versus analytical. For example, R. Ross opposed the concepts of a priori versus a posteriori pathometry in his Theory of Happenings. Most textbooks make a distinction between descriptive and analytical epidemiology. The debate seems endless and can be naively put by such questions as: “Are there purely descriptive studies? Without knowing what one is looking for, how can one tell when one has found it? If there is some previous knowledge or intuition of a subject, why not make it explicit in a model and see how the available empirical evidence modifies this knowledge or intuition? Do pattern-discovery algorithms carry some sort of built-in intelligence?”

Therefore, by analogy with other computer-intensive techniques mentioned above, one could wonder whether geocomputation, and other modern exploratory data analysis techniques, could benefit from incorporating a causal structure or more specific pattern-formation mechanisms.
The specificities of spatial health data analysis

The article by Gilberto Câmara and Antônio Miguel Monteiro describes various recently developed spatial analysis techniques which have been applied mainly to environmental, geological, and land cover/land use problems, etc. Their use in the collective health field is still not very frequent and can present some analytical limitations. I wish to touch on some of these problems based on the question contained in the title, i.e., questioning the specificities of health data and problems as compared to other areas where these techniques have been applied.

In the first place, all health events – birth, infection, illness, death – manifest themselves in persons. These individuals are not randomly distributed in space. Thus, when one works with health records to evaluate risks, one should estimate the probability of an event occurring, weighted by the population distribution. The most common way to consider population distribution in risk evaluation is to group demographic and health data in wertight spatial units and to subsequently calculate epidemiological indicators. This strategy poses serious limitations, such as ignoring interactions between spatial units and the instability of indicators created in small areas (King, 1979). However, this is not the only way to consider population distribution. For example, one can calculate case density (the number of cases per area), producing a surface of probabilities where areas with more proximate cases present greater risk. Analogously, one can calculate the density of persons (inhabitants per area, or simply population density) as a continuous surface to be used as the denominator for calculating rates. A third strategy to evaluate the spatial distribution of these events is to test the randomness of the “cases” in relation to a set of “controls” obtained by survey or drawing from a population with a similar profile. Population density is always an implicit variable in all spatial analyses of health. However, this variable is not neutral. At least in Brazil, it is associated with concentration of wealth and a particular way of life. This variable is the result of human capacity, through the territorial division of labor, to produce surpluses and technology and to organize power structures. In addition, population clustering can have important repercussions on the spread of diseases, especially transmissible ones. For example, the initial years of the AIDS pandemic were characterized by the rapid dissemination of the virus in large cities and by its spread through a downwardly hierarchical network of cities. These cities, considered “central”, concentrate people, income, and cases, as well as fostering an intense exchange among individuals, a condition for HIV transmission. Thus, the population of a given place is both the denominator for evaluating risks and one of the conditioning factors for the spread of diseases, which could be expressed mathematically as a differential equation.

In addition, the macro-determinants of diseases, whether environmental, social, or economic, occur “outside” of persons. It is interesting to note that the World Health Organization defines the environment as “the totality of external elements that influence the health conditions and quality of life of individuals or communities”. Therefore, if we intend to relate health problems to their determinants, we should combine health data, referenced in the population, to environmental data, referenced to something “external” to the population, with each coming from different information systems. Geographic Information Systems (GIS) can allow for this type of data relationship by superimposing layers of health event incidence rates on other layers relevant to this association (Vine et al., 1997).

Third, in Brazil, epidemiological data are collected according to the territorial logic of the Unified Health System (SUS), with increasing hierarchical levels and primarily administrative objectives. Thus, data location is conducted based on the spatial reference of these units, which display a wide variation in dimensions and resident populations. These dimensions, as well as the form of the reference spatial units, can have a major impact on the visual and statistical results. The Geographical Analysis Machine (GAM), for example, searches for excess points in relation to an expected number within circles generated by the program. However, in various situations one should consider non-circular risk locations, non-Euclidean distances between cases (and between the latter and sources of risk), like the bands around power transmission lines, where exposure to low-frequency radiation can cause damage to human health. By selecting indicators, one should search for a territorial division that maximizes the variances of both exposure and the measured effects on the population. One explains – or makes explicit – the environmental and social determinants on the scale in which the greatest variability in indicators is found (Cleek, 1979). Form can be an important factor for constructing a risk model due to its
Influence on the “exposure geometry”, studied through landscape ecology (Frohn, 1998). A more elongated unit can have more neighbors, while compact units have a smaller perimeter and can have less neighborhood relations with other units.

In general, studies in medical geography have been characterized by the search for explanatory factors for a given spatial distribution of diseases, viewing space as an *a posteriori* factor. This approach can produce theoretical simplifications through the association of climatic, cultural, and social characteristics with epidemiological ones, which led a major portion of studies by pioneers in medical geography to conclusions that ideologically reinforced colonialism (Bennett, 1991). The use of neural networks, as suggested by the authors, can reverse the direction of analyses of socio-spatial disease determinants, seeking combinations among factors, constituted *a priori*, to explain this distribution. This approach requires that researchers formally present their hypotheses and construct a series of “layers” representing human spaces and which, when combined, best characterize the places where these diseases occur.

With the improvement of information systems, the inclusion of addresses on health records, and the growing use of satellite positioning equipment or Global Positioning Systems (GPS) in health surveillance activities, one can access these health events as points on a map with a local scale. The main advantage of the data georeferencing strategy is the possibility of producing different forms of data aggregation, constructing indicators in different spatial units according to the study’s purpose. The same point (health event) can be contained in different types of spatial units: a neighborhood, a river basin, a health district, etc., defined by polygons on the maps. This characteristic incorporates the principles of simultaneity and interaction between scales for spatial analysis. This property also involves adopting a geometrical rigor that must be present in the planning phase and construction of the mapping base. In order for there to be a univocal relationship between the point and the polygon, the spatial units must cover the entire working area, and one area cannot be covered by more than one polygon, i.e., there cannot be empty places between units or overlapping of them. Each spatial unit represents a slice of space, containing populations at risk of diseases and displaying disease rates. Geographic Information Systems allow one to construct rates for different exposure conditions by superimposing layers of disease data (points) and population data (polygons). These technical requirements in the handling of both tabulated and mapped data hinder the adoption of less rigid criteria for spatial studies, restricting the concept of space to watertight units. By using network analysis techniques, interpolation, and smoothing of spatial data, one can dissolve previously established boundaries between spatial units. The adoption of fuzzy boundaries for spatial units, ideal for studying place, is jeopardized by the operational norms of information systems (Oppenshaw, 1996).

Spatial analysis is defined as the capacity to generate new information based on existing spatial data (Bailey, 1994). To this end, software applications have been developed that facilitate the search for patterns and exceptions in space. Such techniques do not replace the researcher. Spatial analyses applied to health allow one to study health problems where they manifest themselves. Although this statement may sound obvious, it is important to recall that these analyses are only made possible through the increasingly deep knowledge of both the health problem and the health place.

---


The authors reply
Os autores respondem

The authors wish to thank the reviewers for their extensive and careful comments on our paper. We found that most of the remarks complement our work, either further developing certain issues or elucidating differences in our approach to spatial statistical theory. We fully concur with Bailey’s remarks that geocomputational techniques should be used to complement more traditional statistical approaches, not as alternatives to them. Indeed, some of the techniques we presented (such as the GAM algorithm) can be used as *a priori* data-mining techniques to investigate data-rich environments. After a significant pattern is found, model-fitting approaches can be applied more effectively. We take note of Bailey’s thoughtful comments on the dangers of applying techniques that are not easily associated with statistical measures of sensitivity and robustness, such as neural network or cellular automata.

Although we understand the cautious approach of many of our reviewers, we wish to indicate, as pointed out by both Nobre and Struchiner, that there are circumstances in which exploratory and non-robust geocomputational techniques are useful. Indeed, there are cases where the statistical alternatives are either extremely complex to apply or have yet to be fully developed. Let us consider two types of problems: *multi-dimensional spatial data analysis* and *dynamic spatio-temporal modeling*. As pointed out by Albuquerque, these types of problems arise when we are interested in studying not only the structural elements of space but also the effects of processes.

As for multidimensional spatial data analysis, our paper presents a typical situation in which a health researcher searches for areas prone to the incidence of a disease, given a number of possible environmental factors. This problem can be described in general terms as one of prediction, when it is assumed that a causal structure is in place. We proposed to use neural networks as one of several possible solutions to this problem. In his comments, Nobre made the important point that neural networks are “black boxes” in the sense that they do not provide the ability to explain the reasoning used to arrive at a result. Nevertheless, they provide a viable practical alternative to an otherwise difficult problem with traditional techniques, since establishing a spatial correlation structure for such a problem may prove almost intractable from a statistical viewpoint.

In the case of dynamic spatio-temporal modeling, a researcher may be interested in representing geographical space in a detailed way, e.g., a matrix of cells where each cell has unique characteristics. If one is interested in establishing the conditions of disease propagation in such an environment, the sheer size of the problem and the number of variables required to provide a realistic prediction may make the statistical approach unfeasible. In this case, approaches such as cellular automata guided by econometric equations provide a first approximation to an answer.

Assunção makes the important point that *ad hoc* techniques such as neural networks and cellular automata would benefit substantially from the use of statistical techniques for establishing optimality properties and for a better characterization of their variability and the relative impact of each factor. We heartily agree and certainly consider this an important research topic in geocomputation.

A final comment is in order. Geocomputation will not provide any “silver bullets”. Indeed, some of its early proponents have in a sense jeopardized its acceptance by the scientific community at large by making rather preposterous claims. For problems where clean and robust statistical techniques are available, these should be used instead of *ad hoc* approaches like neural networks or cellular automata. However, we hope to have pointed out that there are many situations in which current state-of-the-art statistical methods are not applicable. In such cases, researchers should be encouraged to use the modeling capabilities of computers and allow for different representations of geographical space. It is inevitable and indeed desirable that such a pragmatic approach will bring a new understanding to the analysis of health data sets.