ON THE PARTICULAR CHARACTERISTICS OF SPATIAL DATA AND ITS SIMILARITIES TO SECONDARY DATA USED IN DATA MINING

Fernando BACAO, Victor LOBO and Marco PAINHO

It is frequently argued that spatial data has a number of characteristics which sets it apart from other types of data. In this paper we argue that this was the case when data analysis was essentially based on what is now called primary data. This is no longer true because in recent years a new type of data, called secondary data, has gained increasing relevance. This data has basically the same problems that have plagued spatial data. With the increase in the amount of available operational data due to the widespread use of computing devices and sensors, secondary data has emerged and gained importance. Secondary data is extensively used in data mining, and in fact it constitutes its "raison-de-être". In this paper we analyze the main characteristics of spatial data and draw a parallel with the major features of secondary data. Underlying our argument is the idea that the special nature of spatial data has lost relevance, as now it constitutes only one particular case of secondary data, which is widely used in many different areas. One consequence of this argument is that the GIScientist toolbox can be vastly increased by importing tools which have proved well in handling secondary data. This doesn't mean that these tools should not be adapted to the specific needs and perspective of GIScience analysis. Some variables, such as location, and principles, such as the First Law of Geography, play central roles within GIScience and researchers should focus on finding ways to introduce these concepts in the workings of data mining tools. This will potentially lead to a more powerful yet GIScience relevant analysis toolbox.

KEYWORDS

GIScience, spatial data, data mining, primary data, secondary data, spatial analysis.

INTRODUCTION

Much of the success of quantitative analysis in human and social sciences has been dependent on their ability to make use of tools provided by traditional statistical theory. The development of theories and models which can be translated into functional forms supporting inference objectives, much like in physics, has constituted an objective for many researchers in human and social sciences. These researchers tend not to accept the argument that human phenomena are eminently unpredictable. Basically, they reject the idea that human phenomena are much more complex than physical ones and that human behavior incorporates a level of complexity beyond the simplified models that science is capable of formulating.

In the particular context of GIScience the adoption of traditional statistical techniques has been troublesome. Problems arise from the fact that much of the statistical theory available was based on assumptions that were too strict for spatial data. In fact, spatial data has, from the traditional statistic theory perspective, particular characteristics which, among other consequences, have caused difficulties in developing a coherent and sturdy approach to spatial analysis based on traditional statistical techniques [1] [2].

Although in many other scientific fields and circumstances the strict statistical assumptions did not always hold, the deviations were rather small and the general attitude was to overlook these problems. In GIScience this was not the case. Many reasons have contributes to this fact, but probably the most relevant was the obvious contradiction between the assumptions underlying

traditional statistical models and the nature of spatial data.

It has been said that spatial data have special characteristics [3] [4] [5] which makes it a unique challenge to researchers, and for this warranting special tools. In the next section we will address in detail the special characteristics of spatial data. We will argue that the "uniqueness" usually associated with the word special, when used to describe spatial data, is no longer true. The adjective "special" has been used mainly in order to convey the idea that traditional statistical techniques are inadequate to deal with spatial data, because of the unique characteristics of spatial data. The consequence derived from the traditional idea that it is special is that GIScience needs specific tools, a completely new and different toolbox, in order to cope with the problems posed by spatial data. From our perspective this is no longer true.

Decades ago, when most data was collected for specific analysis purposes and data was scarce, the argument that spatial data was special, as explained in [3] [6], was appropriate. In those days, a lot of the statistical research efforts were focused on reaching conclusions based on minimal datasets. The fundamental idea was to define the conditions that would help us collecting the smallest amount of data and still be able to generalize the analysis results achieved based on that small sample. Today data availability and access is completely different.

When referring to the need to develop new analysis tools authors should be clear about the cause that underlines this need. It can be argued that new tools are needed because spatial data is special. On the other hand it can be argued that the need for new tools comes from the particular perspective of the GIScience analysis, and not from the constraints posed by data itself. Finally, one can also make use of both arguments, such as [6] *"It is very dangerous to overlook the features of geographical data that make it special. Geographical Data Mining (GDM) is to be regarded as a special type of data mining that seeks to perform similar generic functions as conventional data mining tools, but modified to take into account the special features of geoinformation, the rather different styles and needs of analysis and modelling relevant to the world of GIS, and the peculiar nature of geographical explanation. The focus here is on developing an explicitly Geographical Data Mining technology."*

In this paper we argue that new analysis tools are, in fact, needed but not because spatial data is special, this argument is no longer convincing. Spatial data has the same characteristics as most of the data used within data mining context, in many different research areas. If new tools are needed is because of the particular perspective of GIScience. Data mining techniques, more specifically machine learning (ML) tools, being essentially assumption free, can be safely used within the GIScience analysis context (for a review on the use of ML tools within GIScience see [2]). The problem of using standard ML techniques in GIScience analysis is not related with imposing unrealistic assumptions on data, but on accommodating the particular interests and analysis perspective of the GIScientist. To sum up we would say that there is nothing special in geographic data, assumption wise, what is truly special is GIScience analysis. What are needed are new tools to implement the particular perspective of GISicentists, not new tools to deal with geographic data.

In the following section we review the most relevant characteristics of spatial data. Next we will deal with the differences between primary and secondary data, trying to expose some of the similarities between the problems of secondary and spatial data. We finalize with some comments and conclusions on the consequences of the similarities found and point some possible research directions.

CHARACTERISTICS OF SPATIAL DATA

In the last decades GIScience have developed an in-depth understanding of some of the major features that characterize spatial data. The motivation for this is largely linked with the increasing use of computation in different sciences and also with the challenges prompted by the use of computers to represent geographic phenomena. Geographic Information Systems (GIS) forced researchers to look again at the nature and characteristics of spatial data.

Conceptual models allowing digital representation of spatial data emerged and provided an unparalleled environment for geographic research. GIS lead to a renewed enthusiasm about space and the role of location in many human, social and environmental problems. The fast paced developments also lead to an early understanding of the some major limitations and problems of GIS [7] [8]. There are two types of problems that are of particular interest for our purpose: problems related with the representation of spatial data, and problems related with the lack of tools to extract knowledge from GIS.

The most well-known characteristic of spatial data derives from the 1st law of geography [9], which states that "*everything is related to everything else but near things are more related than distant things*". The most consensual consequence of the 1st law of geography is that spatial data is characterized by the existence of spatial dependency. This means that values for a particular variable in a specific location are related with the values of that same variable in neighboring locations.

Let's assume that every phenomenon is defined by a process and expressed in a context. The process represents the factors underlying the phenomena and context represents the frame in which the phenomena are observed (e.g. space and time). Spatial dependency indicates that the context has an important impact in the process, in other words, the phenomenon in a particular location is a function of the underlying factors but also of the intensity of that same phenomenon in neighboring locations. This adds complexity to the analysis, for it would be much simpler to concentrate our attention on the underlying factors and assume a neutral context.

Traditional statistical theory usually assumes that observations are independent and that they follow identical distributions (i.i.d.). This is clearly an unacceptable assumption in the context of spatial data analysis, because of spatial dependency. Additionally, assumptions on the distribution of residuals are also affected by spatial dependency. Directly related with spatial dependency is the notion of spatial autocorrelation. In fact, spatial autocorrelation [10] can be seen as the computational expression of the concept of spatial dependency. The underlying idea was to develop an indicator which enabled the researcher to know the degree to which spatial autocorrelation is present in the data. There are different indicators of spatial autocorrelation, which can be grouped into two major sets: global measures [10] and local measures [11] [12]. While global methods estimate one parameter or index for the entire study region, local methods can provide as many parameters or indices as data points.

There is a second important characteristic of spatial data which is usually called heterogeneity. Heterogeneity results from the unique nature of each place, indicating that spatial data very rarely presents stationary characteristics [3]. Spatial heterogeneity is related with the lack of stability on the behavior of relationships over space. This characteristic is also known as nonstationarity. This means that the functional forms and parameters may vary (usually do) and are not homogeneous in different areas of the map. Closely related is the notion of isotropy which assumes that the pattern is similar in all directions. Clearly, a realistic perspective on most spatial data has to assume that in general most spatial processes are nonstationary and anisotropic.

Heterogeneity and nonstationarity create additional problems in analysis, emphasizing the local nature of space/process interaction. This notion implies that global models and global map statistics constitute poor analysis tools, which in most cases average out highly complex interactions between space and process. This has fueled interest in different forms of "*place-based*" analysis "*which allow results to vary spatially, rather than searching for a single universal result*" [13]. Another consequence is that errors and uncertainty in spatial data will tend to be spatially clustered. This means that certain areas of the map will present higher levels of error and uncertainty.

The term leverage is commonly used for an undesirable effect which is experienced in regression analysis and in other methods. It basically means that a single data point which is located well outside the bulk of the data (outlier) has an over-proportional effect on the resulting regression

curve. In the context of spatial data, leverage effects should be expected as particular areas are bound to present levels of error significantly higher than the overall dataset. This stresses the need to adopt what can be called error tolerant or noise robust approaches that graciously degrade in presence of outliers. This is surely not the case of least squares strategies or k-means approaches.

Geographic space is continuous and infinite (or at least compact). The discrepancies between the real world and the representations of the real world used as the basis of analysis can also affect the quality of the analysis. The need to achieve a discrete representation of geographic space creates (small) errors. Sometimes these errors endanger the quality of the analysis. The need to define crisp boundaries between spatial objects leads to data which apparently provides a rigorous and adequate representation of reality, but, in reality, may lead to serious limitations as far as accuracy and representativity are concerned. Spatial data is in its nature eminently fuzzy and this characteristic goes beyond representational models and issues. It is fuzzy also in terms of interactions between different spatial objects and in the way these interactions subsist in time.

A significant amount of geographical data is produced by sampling. The sampling procedures are not always the result of scientifically consistent processes but rather of operational limitations (*convenience* or *opportunity* samples) that tend to be superimposed on concerns about methodology. This results in what is usually referred to as selection bias and which leads to an uneven coverage and potentially to miss relevant space/process information. Additionally, the conjugated effects of spatial dependency and heterogeneity may well overthrow any attempts to produce a rigorous sketch of a particular geographic area. If we consider the dynamic nature of a lot of phenomena of interest in GIScience then the picture gets even darker.

Other times, as in the case of satellite images, data is collected in an extensive way. In these cases datasets are large in both size and dimensionality, although presenting a high degree of redundancy. Additionally, these datasets do not always have the appropriate resolution for the phenomena that they are meant to represent. Accordingly, its representativity may be compromised and thus taint later analyses.

Finally, spatially aggregated data, which is commonly used in socio-economic analysis, also has its fair share of problems. Largely resulting from statistical surveys such as census operations these data are distributed in aggregated form to comply with privacy issues and regulations. The aggregation encloses two major problems: the modifiable nature of the resulting areas [14] and the ecological fallacy [15].

PRIMARY AND SECONDARY DATASETS

We argue that spatial analysis is largely concerned with what is known as secondary data analysis as opposed to primary data analysis [16]. By primary data we refer to data that are collected with a particular analysis objective in mind. On the other hand the concept of secondary data is related with data that were collected with some other purpose but can also be used to perform analysis.

Typically, primary data results from a specific inference question. In this way the datasets are collected with this specific objective in mind and according to well established methodological directives, in accordance with the specific needs of the inference task. Secondary data results from different types of digital processing and operational systems, which, in the majority of the cases, are not concerned with inference.

While inferring from small and "clean" datasets is central in many research activities, new challenges emerged with the fast digitalization of our world. Every day the digital representation of the real world grows at an ever increasing rate. This representation is limited and certainly much less rich than the human experience. Nevertheless this "digital portrait" is becoming so large and detailed that it has the potential to enclose the answer to many questions and problems of our society [8]. It presents an opportunity to look back at the near past and "learn".

The down side to this immense potential is in the nature of the data. Contrary to the data usually used in statistics these datasets are very large and "dirty". The sheer size of these datasets constitutes the first problem. Not only are they large in terms of records (*n*) but also in terms of variables (*p*). These two aspects impose significant and complex computational problems which simply were not an issue a few decades ago. Even today, and regardless of foreseeable improvements in computing power, some problems remain intractable. On one hand, we have the "curse of dimensionality" [17] that states that as the dimensionality (*p*) of the problems grows, the search space grows so fast that no matter how much data we have, that search space will be very sparse. On the other hand most data processing algorithms have memory and time processing requirements that grow more than proportionally with the number of instances (*n*), and thus traditional data processing algorithms can not be scaled up to process these large datasets.

Another relevant issue concerning the increase in dimensionality (*p*) concerns the potential for "spurious correlations". In statistics the term "spurious correlation" is used when variables exhibit relation although missing causality. An example would be a correlation between mortality of hospital patients and data taken during the admission of the patients. One might conclude that taking fever measurements decreases the probability of death in the next 48 hours. Clearly, this correlation lacks causality, but it can be explained by the fact that most patients that are admitted through the emergency service, which due to their clinical situation have a higher risk of dying, are usually not monitored for fever. In secondary datasets this problem is very common and sometimes it occurs simply due to random numeric effects, which are bound to happen whenever the number of variables (*p*) is very large.

The problems faced when dealing with high dimensional datasets emphasize the need for appropriate pre-processing strategies and the dangers associated with purely inductive approaches. The idea of "letting data speak for themselves" is appealing but not completely safe. If we approach the modeling task based solely on "brute force" heuristic search, surrendering any domain knowledge, the results might well be disastrous. Further complications come from the meaningless of significance testing in datasets that contain millions of records [16]. In [2] the author addresses this problem, and rightly questions the relevance of significance testing in the context of approaches such as the Geographical Analysis Machine [18].

Particularly ironic is the fact that secondary datasets albeit deep, dimension wise, sometimes lack key variables for many modeling tasks. Often the challenge, when exploring or modeling these datasets, is to do the best with the available data, and confirm if the best is good enough for the task at hand. In many circumstances it is necessary to find proxies to replace unavailable but crucial variables. It is quite common to use ratios, indexes and pre-processed variables, with the objective of improving the information potential of the input patterns. Even so, the absence of key variables will usually introduce some fuzziness in the description of a given phenomenon.

There are numerous reasons to label secondary data as dirty. Most of these datasets comprise missing information. Data about certain variables is available for some records but not for others. This causes difficulties in the analysis but it can even get worse, such as the case when these missing values indicate systematic distortions. In [19] the author provides an example of such systematic distortion referring to road accidents recording, where more serious accidents, involving fatalities, are recorded with a greater accuracy than less serious ones.

Dependent observations are quite common in secondary data, where observations can exhibit multiple functional dependencies. In fact, association rules [20] constitutes a methodology which is essentially concerned with the detection of such "natural" dependencies. Supermarket transactions records constitute an example of datasets plagued with such dependency effects. Some of these dependency effects are "natural" (someone that buys paint has a greater probability of buying brushes) other are "induced" as they represent the effect of certain commercial initiatives (e.g. bundle of products, cross-selling strategies and promotions).

As [2] points out ML tools, usually, do not rely on assumptions of independence, as each variable is used to the extent to which it helps predicting the desired outcome. In this context, components of the input pattern which are highly correlated are of limited interest as the information contained in them is redundant. Typically, ML tools will use one of the components (the most informative one) and ignore all others with which it is highly correlated. This is achieved through the use of measures like the information gain [21] or, more commonly, through cross-validation techniques [20].

Another distinct feature between primary and secondary data is related with data accessibility. Typically, primary data is readily available on a convenient flat file. This is far from true in the case of secondary data. In secondary data, accessibility problems range from computational problems (e.g. incompatible data files), to different measurement units for the same variables, and all sorts of non-numeric data. All these issues need to be tackled, as they are bound to have negative impacts in the outcome of the analysis. In this context we can look at using primary data as deciding on the recipe and then go to the supermarket; using secondary data is more about "making do" with whatever ingredients are available.

Another important characteristic of secondary data is the evolving nature of the datasets. The data may come from operational processes that are continuously generating data, many times requiring real time analysis. Since the processes that are generating the data are usually subject to variations with time, the characteristics of the dataset are continuously changing. This is also known as population drift, and a typical example is the change in the population of bank applicants as a consequence of macro-economic cycles. Clearly this constitutes a form of nonstationarity.

Finally, secondary datasets often have unreliable, or erroneous data. In primary datasets, the value of data during its recording is clearly perceived, and thus serious effort if dedicated to filtering unreliable data. Since the exact analysis objectives are not clear when recording what will be used as secondary datasets, errors can easily go undetected.

Despite all these problems data mining tools, have emerged and grown tremendously in importance in recent years. In [16] the author describes the particular characteristics of data usually used in data mining tasks. Data mining is concerned mainly with analyzing these secondary datasets, so as to extract knowledge from these datasets. Most techniques used in data mining come from the machine learning community, and have proved to deal quite well with secondary data. Their orientation towards inductive strategies has eluded most of these problems and contributed to the establishment of data mining as an effective way to explore these datasets.

We argue that spatial data can be considered a specific set of secondary data from two perspectives. First, it is mainly obtained with no specific analysis objectives in mind, as is case of census data, administrative data or satellite images. Second, its characteristics are similar to those observed in typical secondary data, such as the dependency issues described above, selection bias, fuzziness, redundancy and even nonstationarity.

Before the craze with database analysis, when most analysis were done on (small and clean) datasets especially collected for that purpose, spatial data was in fact special but today its characteristics cannot be thought of as unique (or special), as most of these characteristics can also be found elsewhere.

CONCLUSIONS

We reviewed the basis for the claims that "*spatial data is special*" by analyzing the properties of spatial data. Next, we analyzed the most significant characteristics of secondary datasets usually used in the context of data mining. We concluded that most of the problems that affect spatial data can also be found within the realm of secondary data. Thus we argue that the unique nature of GIScience analysis is due not to the data itself, but to the particular perspective and interests of the GIScience analyst.

The premises that "*spatial data is special*" were true when only traditional statistical techniques concerned with primary data were used for analysis. A lot of the traditional statistic theory is concerned with primary data analysis. In fact entire sub-disciplines (e.g. sampling design, experimental design) were developed to facilitate the efficient collection of data so as to answer specific questions [16]. The objective is to be able to make reliable inference based on minimal datasets, "*make statements about a population when one has observed only a sample*" [19]. This is a valuable and powerful body of knowledge, which constitutes the corner-stone of much of today's research.

When trying to use traditional statistical tools on spatial data, much care should indeed be taken, for we have seen that spatial data does not have the properties that these methods normally assume. But to be fair, even traditional statistics has evolved to break free of overtly restrictive assumptions, and many techniques are available to deal with some of the characteristics found in spatial data .

One of the consequences of recognizing spatial data as an instance of the wider set of secondary data is the possibility of adopting tools that have proved to be adequate to deal with secondary data. Data mining encompasses a large set of tools which have been increasing as more organizations and researchers understand their value and potential. These tools have proved their value in numerous problems and even in landmark projects such as the human genome project.

Although essentially data-driven these tools are theoretically well founded and constitute the result of decades of research in many different fields. There is no fundamental reason for GIScience ignore them. It is much easier to use the basis provided by these tools, eventually improving or adapting then by introducing the GIScience reasoning, than to dismiss them as irrelevant and start from scratch. The idea that GIScience needs something completely new is not only unrealistic but also dangerous. It is unrealistic because a relatively small research community, such as GIScience, would, at best, take years of research to develop such tools. It is dangerous because GIScientist and Geographers are not the best equipped researchers to conceive and design such tools. Moreover, emphasizing the inductive nature of data mining tools as an excuse to ignore its workings, theoretical foundations and limitations constitutes a wrong approach. Addressing neural networks as essentially black boxes in which we include all the available data and wait for them to crunch it and return a relevant answer is truly misplaced.

Casting all data analysis problems as search problems which can be addressed through a "brute force" approach is uninteresting and will lead to infertile results. From our perspective the right approach consists on a close analysis of the different tools and finding ways to introduce the GIScience reasoning and unique perspective. This involves a thorough understanding of the potential and limitation of the tools and methodologies provided by data mining, but also the notion that there are some important paradigms within GIScience that should be respected.

It is not spatial data that is special, what is special is way in which GIScience "looks" at problems, with a particular perspective, with a unique interest in space and how space impacts phenomena with spatial representation. The challenge is to make good use of the new tools made available through data mining. Good use will surely mean different things for different GIScience researchers. For us it means taking advantage of the potential enclosed in these tools to deal with complex problems and large datasets, while remembering that space and location constitute the heart of GIScience. Solving practical problems is good, but is not good enough for a researcher. It is fundamental to "distill" knowledge, hopefully leading to a more axiomatic GIScience field. We feel that data mining tools have a role in this search for knowledge although the way in which they may contribute is not completely clear.

REFERENCES

1. Gould, P.R., Is statistix Inferens the geographical name for a wild goose ? Economic Geography, 1970. 46: p. 539-548.

2. Gahegan, M., Is inductive machine learning just another wild goose (or might it lay the golden egg)? International Journal of Geographical Information Science, 2003. 17(1): p. 69-92.

3. Anselin, L., What is special about spatial data? Alternative perspectives on spatial data analysis, in Spatial Statistics, Past, Present and Future, D.A. Griffith, Editor. 1990, Institute of Mathematical Geography: Ann Arbor, ML. p. 63-77.

4. Openshaw, S., What is gisable spatial analysis, in New tools for spatial data analysis, M. Painho, Editor. 1994, Eurostat: Luxembourg. p. 36-45.

5. Miller, H. and J. Han, Geographic Data Mining and Knowledge Discovery. 2001, London, UK: Taylor and Francis. 372.

6. Openshaw, S. Geographical data mining: key design issues. in GeoComputation '99. 1999.

7. Aangeenbrug, R.T., A critique of GIS, in Geographical information systems, D.J. Maguire, M.F. Goodchild, and D.W. Rhind, Editors. 1991, Harlow: Longman Scientific and Technical. p. 101-107.

8. Openshaw, S., Developing appropriate spatial analysis methods for GIS, in Geographical information systems, D.J. Maguire, M.F. Goodchild, and D.W. Rhind, Editors. 1991, Harlow: Longman Scientific and Technical. p. 389-402.

9. Tobler, W., A Computer Model Simulating Urban Growth in the Detroit Region. Economic Geography, 1970. 46: p. 234-240.

10. Goodchild, M.F., Spatial Autocorrelation. CATMOG 47,Geobooks. 1986, Norwich UK.

11. Anselin, L., Local indicators of spatial association - LISA. Geographical Analysis, 1995. 27: p. 93-115.

12. Ord, J.K. and A. Getis, Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis, 1995. 27: p. 286-306.

13. Goodchild, M.F., GIScience, Geography, Form, and Process. Annals of the Association of American Geographers, 2004. 94(4): p. 709-714.

14. Openshaw, S., ed. The Modifiable Areal Unit problem. CATMOG, Geo-abstracts. Vol. 38. 1984: Norwich.

15. Freedman, D.A., Ecological Inference and the Ecological Fallacy. 1999, International Encyclopedia of the Social & Behavioral Sciences.

16. Hand, D.J., Data mining: statistics and more? The American Statistician, 1998. 52: p. 112-118.

17. Bishop, C.M., Neural Networks for Pattern Recognition. 1995: Oxford University Press.

18. Openshaw, S., A. Cross, and M.E. Charlton, Building a prototype geographical correlates exploration machine. International Journal of Geographical Information Systems, 1990. 3: p. 297-312.

19. Hand, D.J., Statistics and data mining: intersecting disciplines. SIGKDD Explorations, 1999. 1: p. 16-19.

20. Mitchell, T.M., Machine Learning. 1997: McGraw-Hill.

21. Quinlan, J.R., C4.5: Programs for Machine Learning. 1993: Morgan Kaufmann.

AUTHORS INFORMATION

Fernando BACAO
bacao@isegi.unl.pt
ISEGI-UNL

Victor LOBO
vlobo@isegi.unl.pt
Naval Academy and ISEGI-UNL

Marco PAINHO
painho@isegi.unl.pt
ISEGI-UNL