16th ESRI-EMEA User Conference

Knowledge discovery method for propagation phenomena modeling

Nuno Neves
Universidade de Évora
Departamento de Planeamento Biofísico e Paisagístico
E-mail:nneves@uevora.pt

Abstract:

Propagation phenomena is one of the most complex application areas in GIS as the variety of models is frequently affected by the lack of knowledge in many input parameters, specially those regarding the behavior of geographical elements or geographical characteristics. As model development involves a systemic approach to the universe of relations in geographical and environmental scenarios, it is crucial to develop tools that can be used to learn and retrieve knowledge that can be further applied in application design and model implementation.

In this paper, spatial data mining methods are applied to extract knowledge about propagation phenomena, using geographical minimal elements (GME) as the basis for built up of experimental scenarios for knowledge discovery. The use of GME provides a minimal spatial unit of homogeneous characteristics integrating composite information, fully maintaining its analytical potential and providing an operational basis with ideal characteristics for this type of application.
The application presented in this paper is integrated in GEOMETA project, approved in the framework of POCT program, supported by the Portuguese Foundation for Science and Technology.

Introduction

Considering spatial data as the data related to objects or elements in a given dimensional space, spatial databases provide the infrastructure environment for management and analysis, storing either the objects or the elements and the relationships among them.

Spatial elements data frequently include different levels of abstraction and informational procedures to provide the spatial relationships, such as topological information or

distance information, in order to access the characteristics and the different scenarios of spatial interaction (Böhm C., *et all*, 2000).

Spatial information databases are usually organized according to the establishment of spatial indexing structures that can be accessed by analytical functions. These analytical functions can provide a large number of accepts on the relationships between elements or phenomena. However, frequently, the tools used to explore are based in classical statistical procedures, trying to retrieve the relations (e.g., correlation's), for relational databases, and then represent these relations in a graphical way in a GIS.

As the up mentioned methods were not appropriate for a large number of spatial analysis operations, it led to the emergence of a specific research field known as knowledge discovery of spatial data mining. This field involves the integration of several different approaches, such as machine learning, data visualization, statistics and information theory, (Koperski, K., *et all*, 1998). From the GIS point of view knowledge discovery is also related with the development of applications specially those oriented for simulation or decision support purposes.

The extraction of implicit relations represented by functions, algorithms or graphical pattern's aims to provide interesting and regular knowledge presented in an explicit way. Some of these explicit ways are related with the definition and construction of new metaphors, especially those that can subsequently be used for new analysis and knowledge representation.

Classical geography is frequently criticized for the lack of structuration of the analytical problems. As many of the geographical problems and phenomena are ill structured, it is difficult to establish algorithms or defined criteria for spatial analysis. The goals are not well defined and large areas of knowledge remain peritial and not explicit.

The question is: Is it possible to explain or explicitate such complex problems involving a large number of variables, integrating so many different concepts and aiming for such a general representation? In fact, this is one of the main goals of knowledge discovery techniques: trying to relate, represent, express the relation. Then represent again, continuously trying to extract new information in naturally increasingly abstract scenarios.

The general idea that everything is related in a geographical scenario immediately generates the goal of discovering these relations, establishing connections between phenomena and spatial characteristics. Relations and their identification are essential for

model formulation and to build theoretical approaches that can be accepted to describe the general behavior of natural and socio-economical scenarios.

Simulating propagation phenomena - case study

In general, environmental models seek to simulate or re-create a given natural phenomenon. Computer software like GIS can be seen as a virtual computational environment that provides the adequate metaphor for model implementation and look-alike simulation. In fact, modern simulation models are also testing tools for general knowledge prospective evaluation. It is much easier to develop and implement computer models than to acquire direct field knowledge about natural phenomena.

Propagation models are fundamentally related with the spread across a given surface, bidimensional or tri-dimensional, of a specific phenomenon or substance. They are often characterized by a complex set of rules and formulas resulting from large efforts, mainly empirical and statistical, on defining the parameters of these formulas. However, the models defined and implemented are too dependent on the definition of isotropic conditions, and this purified scenario is not adequate for the inherent complexity of the real world conditions; so, these models frequently lack precise information for the desired modeling process.

In the case study presented in this paper we will provide an approach to knowledge discovery techniques applied to the modeling of propagation phenomena. In particular, we will focus the need for knowledge discovery in order to create virtual laboratories dedicated to searching for spatial relations and properties in geographical elements.

Problem

The general problem can be better understood considering the similarities in the different types of propagation phenomena:

- Propagation phenomena are associated with a spreading movement across a given scenario or surface;
- The propagation scenario or surface has a variable “impedance” to movement that generates different types or values on propagation patterns;

- The “impedance” is usually associated with a variable number of factors frequently difficult to distinguish (e.g. mechanical resistance, gravity, absorption, cumulative phenomena, etc)

In order to build analytical models to simulate propagation patterns, it is necessary to conceive a structure that makes it possible to isolate the different factors in a reproducible way. The phenomena must be reproducible to allow further simulation in different scenarios or locations.

In the case study presented we considered two major factors conditioning propagation:

- A gravity factor resulting from the different elevation of terrain and the associated slope;
- A specific “impedance” associated to the composite characteristics of each point in a given geographical scenario. Composite in the sense that fundamental characteristics are grouped in a codified way, creating homogeneous areas of soils, geology, land use, etc.;

The fundamental goal is to create a system that is able to retrieve the characteristics of a given situation of propagation, isolate the factors considered and learn about the “rules” of the phenomena. Than, these rules must be transformed in knowledge (e.g. models) that can be further associated to the geographical elements.
The hypothesis can be resumed like this:

- There are major factors in propagation phenomena that can be related;
- The relation between factors can be measured and described through a model;
- There are analytical functions that are similar to propagation and can be used as variables or co-factors;
- If we can retrieve this relations and isolate the factors it will be possible to reproduce (e.g. simulate) the acquired knowledge.

The analytical function used in this case study as similar function was CostDistance from ArcView Spatial Analyst, vers.1.1a.

Information

The information base used in the modeling process was previously integrated in a composite basis of geographical minimal elements and subsequent derived information was generated from that base.

As the GME as essentially points with composite data, they are an ideal basis for retrieving information with a very high detail level. The GME are codified according to their different characteristics (e.g. composite information).

The composition code describes a unique combination of characteristics and specific behavior that will be measured and retrieved accordingly.

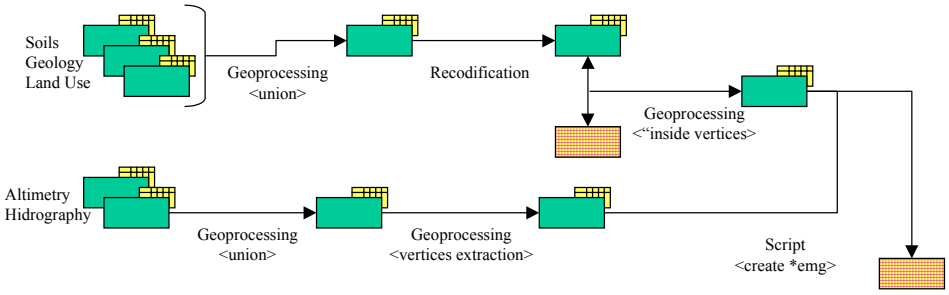

Fig 1 - Creation of geographical minimal elements (GME)

The information used can be grouped in three categories:

- Geographical minimal elements containing elevation data, hydrographic data, soils, land occupation and geology. All these types of information are integrated in a unique data set and were generated from official cartography at scale 1:25000;
- Propagation surface generated from fictitious data simulating propagation phenomena through a real surface. Fictitious data were used because there is no adequate information available, but the results of the modeling process are not affected;
- Point source or initial location of propagation process. This location can be given by the user in order to simulate a specific process.

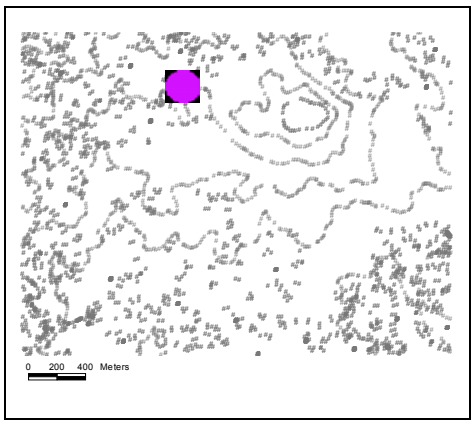

Fig 2 - GME for the study area

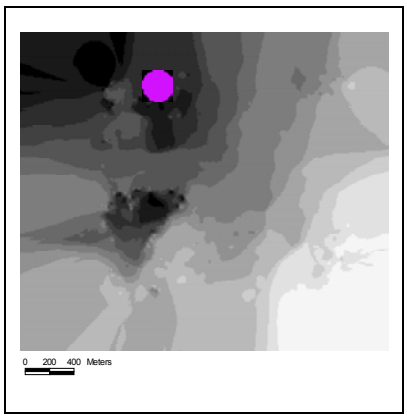

Fig 3 - Source and simulated propagation surface

Spatial Analysis and geographic modeling

The spatial analysis operations implemented are initial steps in the generation of data with which we aim to identify and retrieve relations. As the GME integrate different types of geographical information, they also allow the generation of a large variety of derived data that will become part of their attributes for each specific purpose.

So, in fact, GME base information integrates a primary set of information, but they were created also with the purpose of retrieving all the information generated with them, information that is needed for each specific analytical process.

Fig. 4 describes a geographic modeling flux diagram illustrating the creation of “useful slopes”. This information can be considered as an attempt to identify the importance of the slope in the propagation process. The slope can not be considered “as it is” because the same slope can be favorable or non-favorable for propagation, depending of the its direction.

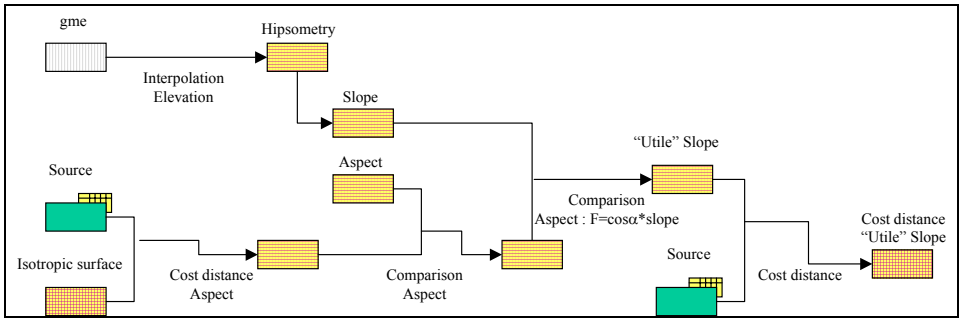

Fig 4 - Geographic modeling diagram for the creation of "useful slope"

Considering propagation as a radial function, the relative directions to the source location can be compared with the orientation of the relief. This comparison can measure the "useful effect" of the slope considering a specific propagation scenario. The "useful slope" is a comparison between the orientation of the relief and the orientation of the propagation scenario.

Basic trigonometry and Newton mechanics immediately suggest the formula:

Useful slope = (cos. (aspect propagation – aspect) * slope)

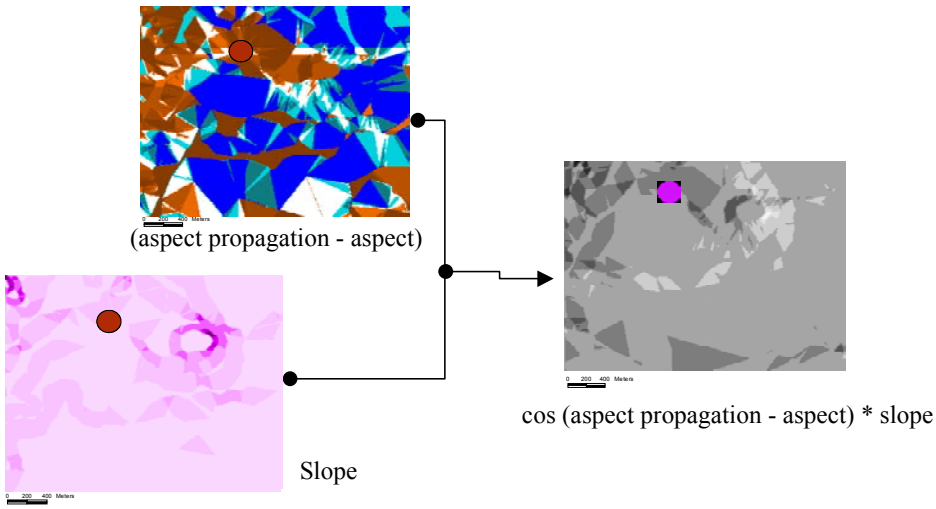

Fig. 5 – Generation of "useful" slopes

Exploratory spatial analysis

Exploratory spatial analysis can be considered as a preliminary attempt to identify similarity between generated data and specific phenomena. It involves a component of

intuition and experience on imagining which spatial analytical functions and its results can have and provide a relation, which can be measured and used in modeling and knowledge discovery.

The characterization process of the geographical minimal elements in this type of modeling process is related with the extraction or mining of a complex set of multiple-level abstraction descriptors to be connected to the interpretation process. As the basis or the fundamental structure of the elements only provide a limited set of primary data, all the other information required for each specific type of phenomena or analysis **must** be generated through a set of analytical procedures.

As we can always generate propagation surface based on the relief characteristics, more precisely based on slope and relative orientation, this variable or model component can be considered "stable" or normalized in the modeling process.

The real question is: What do we miss? What is the difference between the resistance or impedance associated to the slope (e.g. gravity) and the other types of impedance associated with specific characteristics of GME? In this simulation we considered a type of aggregated impedance that we called "specific impedance".

Fig. 6 describes the process to obtain "specific impedance" for the case study propagation scenario.

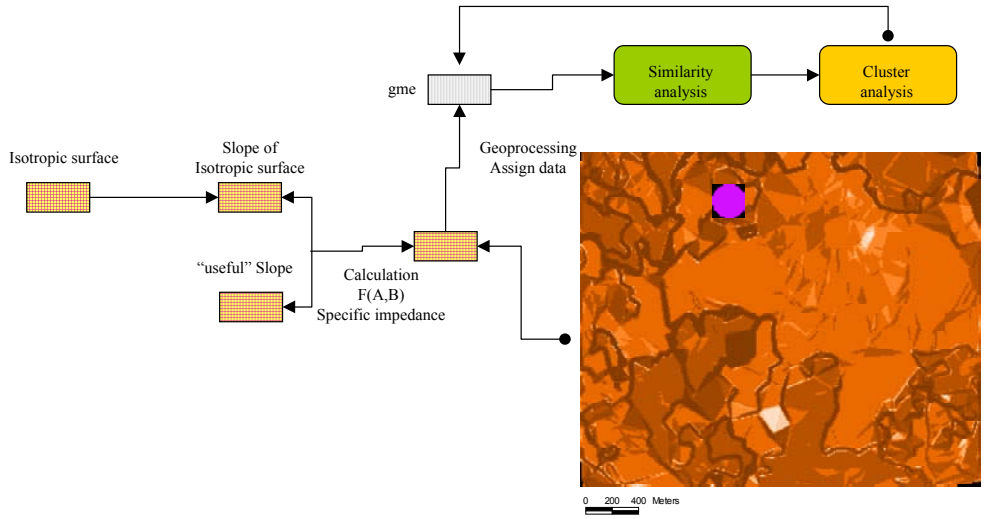

Fig 6 - Obtaining and retrieving a specific impedance.

Specific impedance is the impedance associated to each type of GME. It is defined as a relation or coefficient between the reference situation and the useful slope in each location. This coefficient is retrieved and associated to each GME code for each studied phenomenon. The coefficient can be a stable or linear relation or can be a non-linear relation.

If the specific impedance is stable, the value or linear regression function can be immediately retrieved and become part of the model associated to that type of GME. Otherwise, the relation can be described through a non-linear estimation with the most adequate model.

Comparison and identification of elements

The comparison and identification process is based on the characteristic descriptors of each element according to a comparison method. The methods used try to reach similarity according to the precise variable (e.g. soils) or evaluation values assigned from spatial analysis generated information.

As the GME are stored in a unique data set, it is simpler to compare their characteristics and to implement data mining methods. This procedure also avoids the need of considering a primary set of data or layer as a target class and another layer as a contrasting class.

A crucial issue on the comparison methods is related with the set of information used and the types of spatial analysis implemented.

As the knowledge retrieved from an experiment is frequently insufficient for acquiring the characteristics and behavior of all types of elements, it is important to implement similarity measures with indirect information in order to achieve some “proximity” information elements. The cluster analysis is the classical mining analysis to be implemented in this condition.

Cluster analysis as a classical component of exploratory data analysis, integrates a set of analytical functions or statistical equations to measure the “distances” between the characteristics or descriptors of the elements. In the case study presented, cluster analysis was used to evaluate similarities between GME in order to allow information retrieval for those elements (e.g. types, codes of elements) that were not present in the experiment location. This kind of indirect association has naturally different levels of confidence that must be precisely evaluated through correlation analysis.

Comparison of functional behavior

Knowledge discovery methods are in general associated to the mining relation's process among large volumes of data existing in a database. What is proposed is to establish knowledge discovery processes among data and information that does not exist. This information needs to be created according to a preliminar and exploratory spatial analysis.

The first issue is related with the evaluation of patterns that are similar to the phenomena being studied. This process is inherently connected with the experience in the field and imagination. It is important in this phase to have the contribution of experts in a specific field, in order to evaluate the reasonability of the proposed analysis.

General knowledge discovery tools are designed to compare and relate data and not to propose similar information and patterns to be generated. It is however possible to retrieve experiences in this domain to further use in similar processes.

Once the similar patterns, functions or generated information are evaluated as similar or related through simple correlation analysis, it is possible to start the effective process of knowledge discovery.

The classical functions to be used in this phase are linear and non-linear regression analysis, trying to identify and to make explicit the relations needed to create a polynomial function, in the case of linear regression, or other type of model, expressing in an adequate way the relations among variables.

An important issue in this phase is to identify stable and independent relations from the nature of elements and to retrieve other as characteristics of the elements. To simulate a process after creating a model, it's important that relations are independent of the specific situation (e.g. position, proximity, etc), or at least that we know how that they are related.

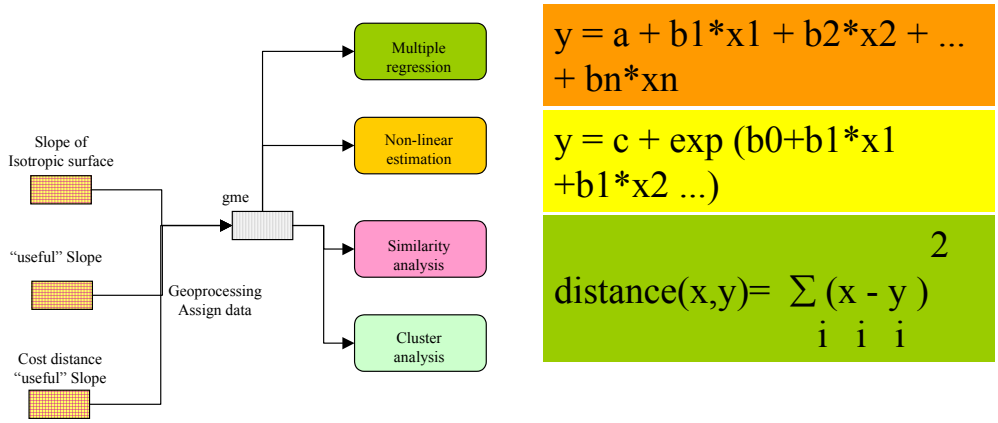

Fig 7 - Knowledge discovery and associated basic equations.

Building the model

In the process of building this model, we were interested in whether and how the dependent variable (e.g. propagation surface) was related to a list of independent variables. Our independent variables were the coefficient of impedance and similar function cost distance through "useful slope". Our goal was to establish a function of the type y = F (x1, x2, ..., xn) where the term F (x) means that y is the dependent or response variable, and y is a function of the x1, x2, ...xn, that are the independent variables.

However some of the x variables can also be a function of other independent variables. In fact our in similar models, coefficient of impedance can be described through a linear regression equation of the type y = a + bx, a multiple regression equation such as y = a + b1*x1 + b2*x2 + ... + bn*xn, or for instance a nonlinear estimation based in an exponential model like y = c + exp. (b0+b1*x1 +b1*x2 ...).

In our study, coefficient of impedance assumed a linear behavior and the best model to describe the relation between independent variables and the dependent variable was multiple regression described by the equation:

Y=F (Specific impedance (Ki), Cost distance “useful slope” (S))

Y=a+(b1*Ki)+(b2*S)

Y = 4216,39597039021+(0,22714*Ki)+(1,627796*S)

The comparison between the observed values and the values predicted by the model can be observed in Fig. 8 and Fig. 9. A curious result of this simulation is that the values predicted by the models seem more reasonable than some of the reference values. This can be explained with the inadequacy of simulated values of the reference situation. However some locations have intentional incorrect values, and that doesn’t affect the model equation generation. This can also be observed in the graphical representation of the relation between observed values and predicted values in Fig. 8 and Fig. 9.

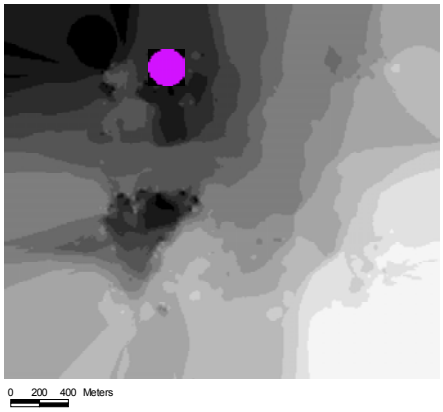

Fig. 8 – “Observed” values

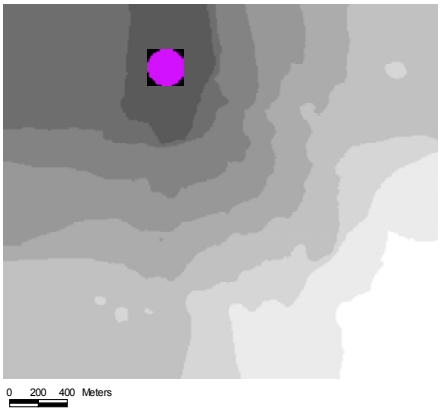

Fig. 9 – Predicted values

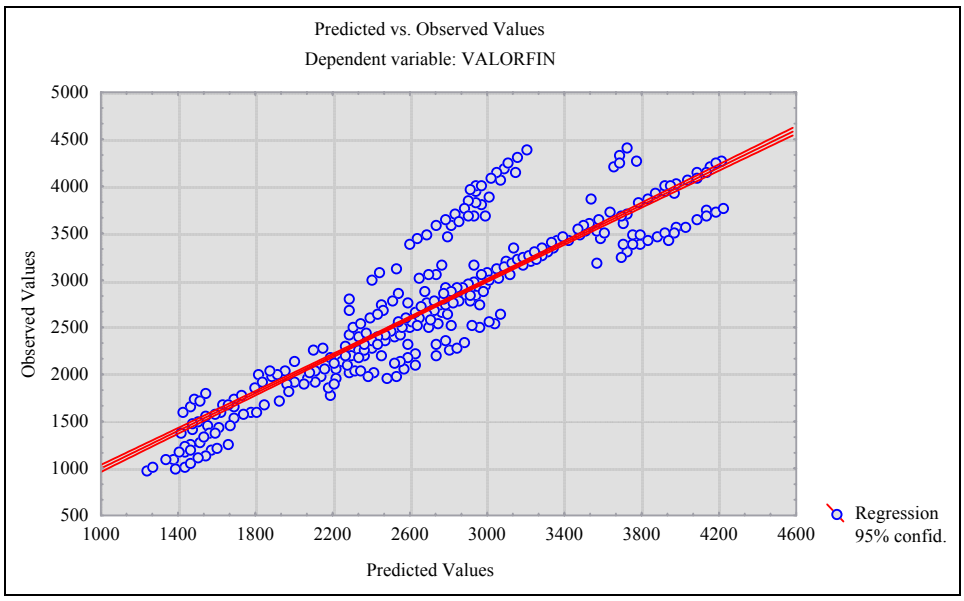

Fig 10 – Observed and predicted values.

Some graphical based methods are being tested in order to achieve new identification parameters. The generation of graphical metaphors provides an appropriate way to implement new forms of similarity and cluster analysis. It is particularly important to establish an organized process of knowledge acquisition for different types of phenomena, comparing the results of the modeling process and providing continuous validation of the methods applied.

The methods described in this paper offer a simple and effective approach to knowledge discovery in propagation phenomena modeling in a simplified way. They can be used in environmental impact assessment situations as comparative methods simulating the different alternatives and scenarios.

Acknowledgements:

The Portuguese Ministry of Science and Technology supports the research project GEOMETA (POCT/GEO/35129/99-00) within the framework program of Science and Technology.

References and bibliography:

Andrienko, G. and Andrienko, N.: "Intelligent Visualization and Dynamic Manipulation: Two Complementary Instruments to Support Data Exploration with GIS", In Proceedings of AVI'98: Advanced Visual Interfaces Int. Working Conference, L'Aquila – Italy, May 24-27, 1998, , ACM Press, pp.66-75.

Andrienko, N., "Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment", In Proceedings of Workshop on Status and Trends in Spatial Analysis, Santa Barbara, California: December 10-12, 1998, NCGIA

Böhm C., Braunmüller B., Kriegel H.-P.:"The Pruning Power: Theory and Heuristics for Mining Databases with Multiple k-Nearest-Neighbor Queries", In Proceedings of International. Conference. on Data Warehousing and Knowledge Discovery (DaWaK 2000), Greenwich, U.K., 2000, pp. 372-381.

Brodley, C.: "Addressing the Selective Superiority Problem: Automatic Algorithm / Model Class Selection. In Machine Learning", In Proceedings of the 10th International Conference, University of Massachusetts, Amherst - San Mateo, USA, June 27-29, 1993, Morgan Kaufmann, pp.17-24.

Gama, J. and Brazdil, P.: "Characterization of Classification Algorithms", In Progress in Artificial Intelligence, LNAI 990, Berlin, 1995, Springer-Verlag, pp.189-200.

Gebhardt, F.: "Finding Spatial Clusters", In Principles of Data Mining and Knowledge Discovery PKDD'97, LNCS 1263, Berlin, 1997, Springer-Verlag, pp.277-287.

Koperski, K., Han, J., and Stefanovic, N.: "An Efficient Two-Step Method for Classification of Spatial Data", In Proceedings SDH'98, Vancouver, Canada, 1998, International Geographical Union, pp.45-54.

Neves. N., "GEOMETA - Elementos Mínimos Geográficos para Análise Territorial e Ambiental" in Coordenação dos SIG e OIT para o desenvolvimento de espaços rurais de baixa densidade, ÁMDE, Évora, 16 November 2000.

Wrobel, S., Wettschereck, D., Sommer, E., and Emde, W. (1996) "Extensibility in Data Mining Systems", In Proceedings of KDD'96 2nd International Conference on Knowledge Discovery and Data Mining, 1996, AAAI Press, pp.214-219.

Deleted: .