CO-004

ARPEGE': A GRAPHIC TOOL FOR EXPLORATORY AND VISUAL DATA MINING AND SPATIAL ANALYSIS TEACHING

JOSSELIN D.
UMR ESPACE, AVIGNON, FRANCE

INTRODUCTION

Nowadays, the development of information society provides user-friendly and interactive tools for spatial analysis. After a period in GIS history where a huge quantity of maps have become available and popularized on the net (e.g. google map), the potential capacity of softwares in terms of functionalities has tenfold. It is said to the users that the GIS tools make great progress in Human-Computer interface, providing a kind of 'transparent' use of these softwares for spatial analysis. But what happens in reality? Paradoxically, this evolution tends to make the software interface 'opaque' as it becomes more and more difficult to manage the software complexity. In a sense, the distance that now exists between the user and the methods (s)he requires is increasing. It induces a certain difficulty to deeply understand the methods used, despite scientific references and a knowledge that should be shared by users. Indeed, we miss the confidence we can have in these methods because we don't really manage them, due to user interface and tool complexity.

In another hand, we must take into account and get benefits of such an evolution in the ease to make spatial analysis, especially when there exist tools that allow to permanently be in touch with the data and their multiple representations. That is why it is of crucial importance and interest to develop tools in this trend, but able to reinforce the link between the spatial expertise practice and the explored spatial data information. Exploratory Spatial Data Analysis can be useful for such a target objective and can be used in teaching, as it will be illustrated in this paper. After having presented the importance of robustness, complexity and interactivity, we'll briefly describe the concepts underlaying the software ARPEGE'. We'll end on discussed examples in teaching spatial analysis.

This paper has a content which is similar to the one published in the proceedings of GisPlanet (Josselin, 2005). The difference is that the talk is oriented and completed by a teaching point of view, after a few years of regular use in teaching spatial analysis.

1. GEOGRAPHICAL INFORMATION SYSTEM DEALING WITH ROBUSTNESS, COMPLEXITY AND INTERACTIVITY

1.1. Robustness in Spatial Analysis

There are different ways to improve the robustness (l. s.) of an exploration process. The user can first of all assess the robustness of the statistical indexes (s)he handles. This large topic is essentially tackled by mathematicians and statisticians (Hoaglin et al., 1983, 1985), developing fruitful research about robust metrics (Dodge, 1987) in geo-statistics, notably (Cressie, 1993). There also exist different means to assess the robustness of a statistical measure according to a given batch of empirical data (Huber, 1981, Hampel et al., 1986). Another way, more user-centered, consists in elaborating specific procedures in order to improve the user capacity to make a 'relevant analysis' and to provide a 'good decision'. According to this purpose, we developed an interactive GIS, called ARPEGE'. It includes some functionalities and graphic prototypes allowing to follow the exploration process.

1.2. A GIS supporting a Systemic Approach

The need for experts and decision makers to improve their knowledge and their acuteness during the planning process becomes more and more pressing. The questions of environment and urban sustainability require the decision makers to deal with very large (spatial) data bases. Nowadays, this is made easier by the software capacity for storing and querying the data, that enables to manage the complexity within a systemic approach. The systemic approach (Von Bertalanfy, 1980) is often considered as paradoxical to a Cartesian approach (Descartes, 1937). It is partially true. Both Cartesian and systemic approaches break the problem in many parts. But while the first method expects the global solution to emerge from the sum of the disaggregated pieces, the second one (the systemic) focuses on the relation and the importance of its part of explanation. We think the interactive GIS (Peterson, 1995, Andrienko, 2006, Cartwright et al., 2007, Cauvin et al., 2008) must take that position and consider the relations as key components of the system. So do we jointly manage two levels of complexity: the geographical application (i.e. the spatial data) and the exploration (i.e. the data base structure, the software and its functionalities) systems.

We need to interactively explore the data through a complex, accessible and understandable systemic model. Dynamic links provide such a crucial requirement (Hasslet et al., 1991). The user must be able to feel the data and the systemic model as (s)he explores them without worrying about any technical procedures required to allow the investigation. The idea is to provide a (as large as possible) set of efficient and fit-to-use methods of a high conceptual level to tune the parameters for visualizing and exploring the data, with different lights, keeping in mind that the maps and the graphs are not neutral vectors of information (Wood, 1992, 2010, Monmonier, 1993). Such a position leads to improve the efficiency and the speed of the methods, especially those dealing with large sets of data. We shall see further that we propose to implement different kinds of relations commonly composed by two elements, for selection and action. Both of them can operate using only a mouse or a key of the keyboard.

1.3. Toward an Interactive Geographical Information System

For users and experts in geography, the Data Base Management Systems and especially the Geographical ones provide very useful functionalities. The first one is the high capacity storage thanks to hardware progress. The second one relates to the Structured Query Languages which are extended to topological and geographical needs. They allow to write easily quite complex queries adapted to geographical features (Longley et al., 2001). Sometimes, advanced Case Tools are implemented for making easier the model designing (Oracle, for instance). The third interesting property corresponds to the graphical and cartographic functionalities in GIS software. The fourth functionality sets in the integration of both entity-relation and object oriented modelling, expressed in the famous Unified Modelling Language. ARPEGE' takes into account this progress. Let us notice that a few GIS software include such dynamic links and exploration methods in their core or in additive components (Anselin & bao, 1997, Josselin et al., 1999), that is a proof of their relevance.

We believe that the contribution also comes from the approaches which encompass the philosophy itself of exploration, because it holds the basic bricks to build dedicated applications, for geography or any other discipline. These research or software belong to either the statistical domain of the exploratory (spatial) data analysis (ESDA: Bailey & Gatrell, 1995, Fischer et al., 1993) or the (interactive) (geo)visualization (Hearnshaw, & Unwin, 1994, MacEachren, & Kraak, 2001, Banos, 2001, Josselin & Fabrikant, 2003, Cauvin et al., 2008, Peterson, 1995). For instance, Datadesk or LispStat or its affiliated software such as Arc (Cook & Weisberg, 1999) or Vista are developed in this spirit. In spatial analysis or data mining (Zeitouni, 1999), there also exist different research and applications, among which a few are developed in LispStat (Josselin, 1999, 2003, 2005). That is the case of ARPEGE'. As we previously emphasized the important role played by dynamic links during the geographical data exploration, we propose to access to the data and to their relation only by graphic selection and links, rather than being forced to write (sometimes in many steps) and execute a query. In practice, it highly accelerates the exploration process.

2. OBJECTS IN ARPEGE'

2.1. A M-to-M Generalized Relation Between Hierarchized Objects

We associate the object oriented and the entity-relation models in the following way: the object oriented model enables to build a hierarchical structure of objects more and more specialized inheriting from their parents. This approach allows to access to the numerous statistical functions, slots and methods of the parent prototypes from which are inherited the geographical prototypes which can nevertheless process their own methods (polymorphism). The entity-relation model managed notably by lists in LispStat, fulfill, until now, our needs in terms of relationship. Let us notice that in the text, the term object refers to an object class (described by a prototype) and as well to an instance of it.

The internal data model of ARPEGE' includes several complementary components that the exploration process requires (figure 1):

the graph prototypes in which the geographical data use to be depicted, that may be a map,

the spatial data (prototypes), those are geographical features with particular geometries, that may be polygons delineating geographical areas,

the relation prototypes between different graphs which make them dynamically interact, those able to trigger off different calculations and behaviour of different objects.

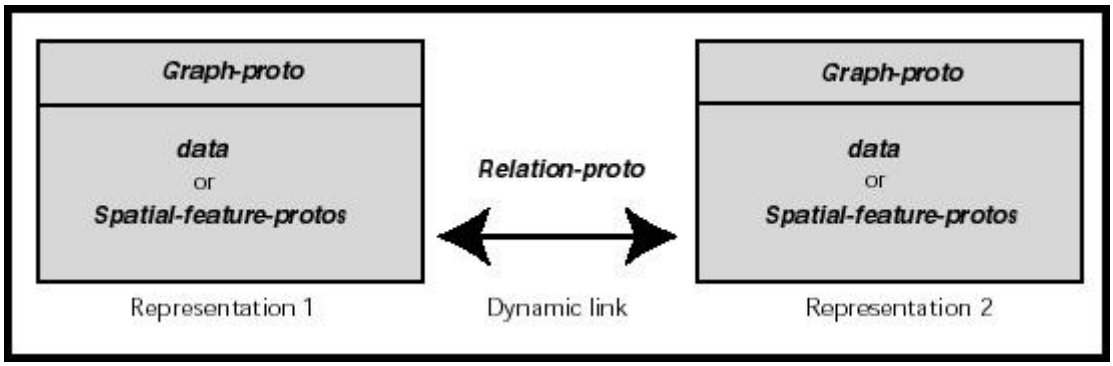

Figure 1: Graphs and dynamic relations

There exist a very large variety of information graphics (Harris, 1996). All the geographical graph prototypes developed in ARPEGE' inherit from the top parent, the graph-proto, depending on the number of their associated variables (figure 2). Most of them come from the "obgeo" prototype which is adapted from the scatterplot prototype. The prototype specialization depends successively on whether it deals with spatial or temporal information, on the aspects tackled (e.g. exploration vs application time) and the objects handled (e.g. maps vs topological structure) during the geographical phenomenon exploration process and on more concrete purpose (e.g. variogram) and data (e.g. polygons) to which the graph prototype is devoted. The hierarchical structure is presented in the figure 2.

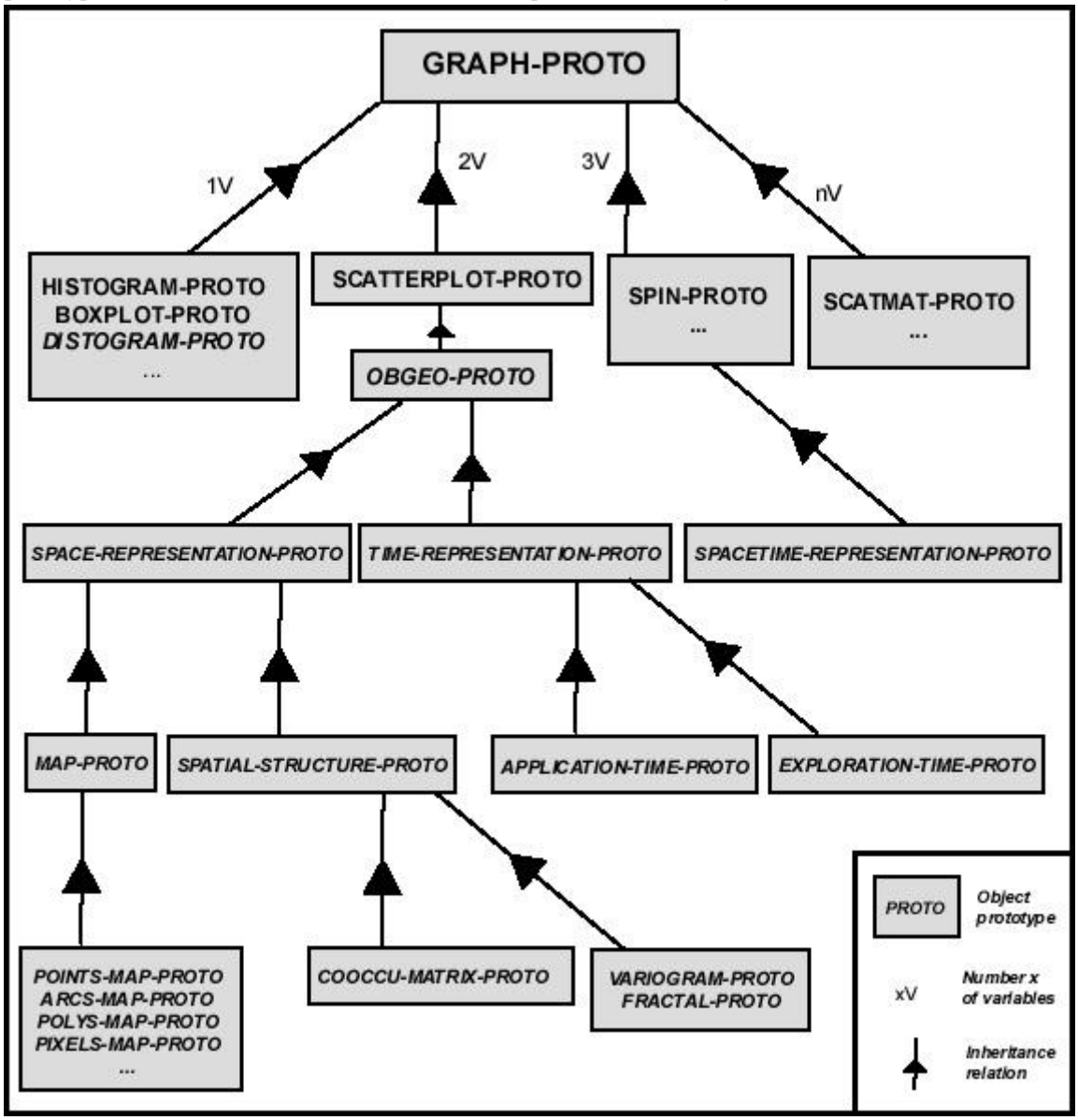

Figure 2: The graph prototypes

2.2. Some objects to represent Space

The geographical information is built on a set of primitives, such as points, polylines or polygons, depending basically on their points coordinates (figure 3) (Heywood, 2002) and also on the topological structure (vertex, nodes, arcs and polygons). We developed a generic model compatible with most of the spatial data models encountered in geography, however differing from them because of its high level of decomposition. While most of the common topological model separate the raster format and the vectorial format, we propose to associate them in the same framework, as proposed in other existing software (E-cognition).

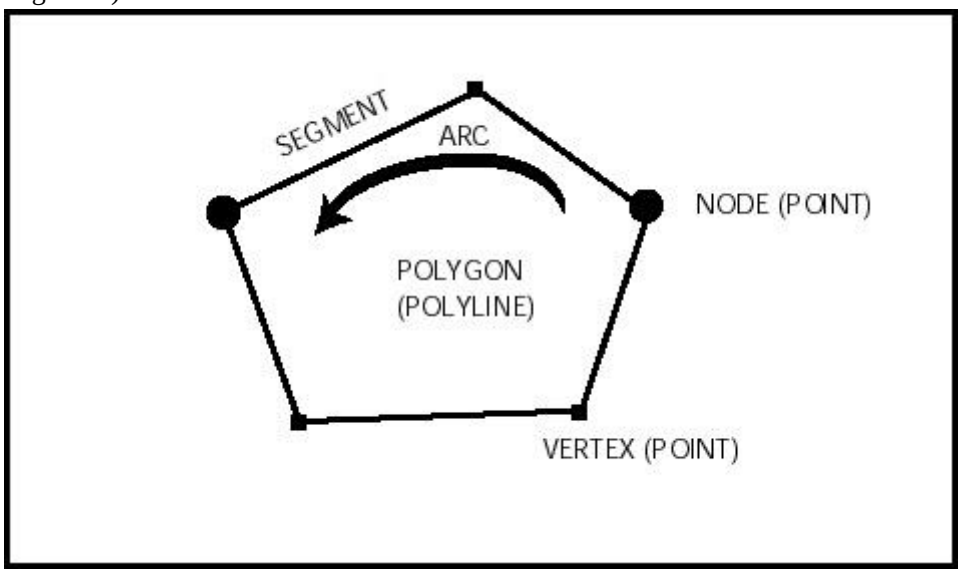

Figure 3: The topological model

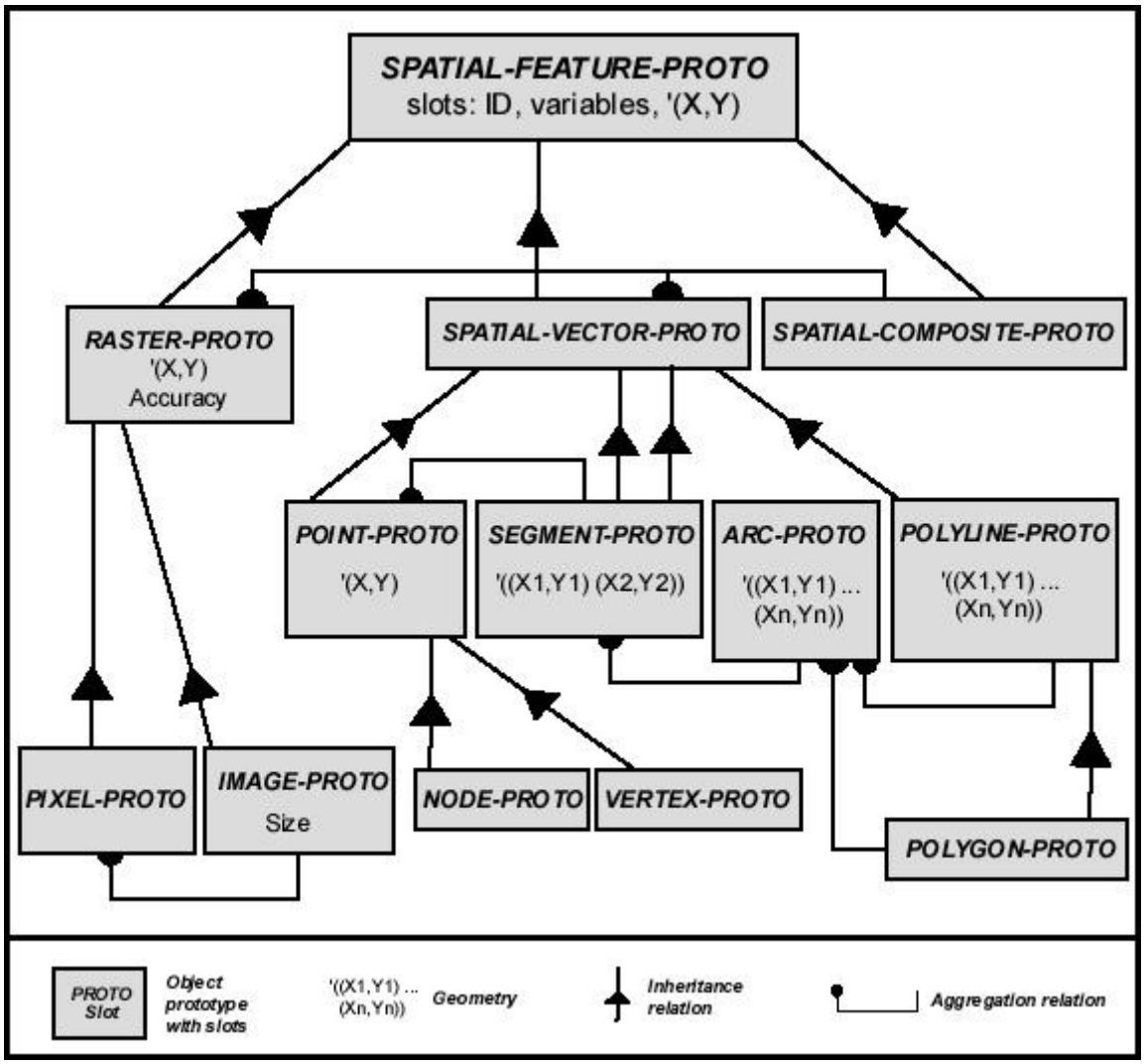

Figure 4: The prototypes of spatial features

We often think about maps to depict the spatial information. In practice, there exist a lot of other methods to help to understand how the geographical space is organized. Thus, the space representation prototype includes two main types :

the map-prototype, in which the spatial data are simply drawn,

the spatial-structure prototype, that provides complementary and synthetic views on spatial data, essentially based on their topological structure.

Most of the map prototypes we propose look like scatterplot prototypes because they are based on an Euclidian representation of space, crossing two particular variables in Y and X: latitude and longitude. As we shall see when presenting the spatial feature objects, these maps can support several kinds of spatial data, including points (points-map proto), arcs (arcs-map proto), polylines and polygons (polys-map proto), and pixels or images (pixels-map proto).

Sometimes, it is useful to have a description of the topological structure, that is to say, a clue, a statistical estimator or any method to emphasize the way the geographical objects are connected one to each others. For instance, a co-occurrence matrix gives the frequencies of all possible connexions between objects (every objects to all the others), a variogram crosses a variance with the distance between coupled objects, the fractal prototype does the same but transforms the two previous variables in log. The cooccu-matrix prototype behaves like an image whose pixels are disconnected. The variogram or fractal protos are plotted using a dynamic sampling when they deal with very numerous data and draw a non linear regression, such as a Lowess.

2.3. Composite Objects

The concept of (spatial) composite object is now widely approved by the researchers, the experts and the users in GIS (see for instance, Hornsby & Egenhofer, 1998, Josselin, 1999). A composite object is an

heterogeneous set of objects whose the association justifies the object existence itself. Our definition somehow differs from the current sense, because, beyond the set of objects, it includes a set of relations informed by a set of variables. The interest is that these variables can encode a list of objects (for selection in other objects), a statistical estimate (characterizing the statistical relation between two kinds of individuals belonging to two different objects), or even a behaviour (defined by a piece of code in a method assigned to the object).

For example, if an expert studies the occurrence of road accidents and the topological structure of routes, the frequency of accidents and their type can be crossed with the number, the size and the type of the road sections, even the network connectivity. This leads to analyse the statistical dependencies between the basic elements making the composite objects through their relations. So, the spatial composite objects include objects and relations, all of them carrying their own variables, data, slots and methods. Associated with other a-spatial information, these objects integrate many descriptors helpful to describe the studied territory from complementary aspects. This approach proceeds like an (interactive) spatial data mining. Every type of composite object can be grouped in a class and described with a lot of pointers and different geometries and semantics.

2.4. The Relation Object: a Way to Link Dynamically Graph Objects

One of the crucial functionalities of LispStat is to provide a very efficient dynamic link between objects, due to object orientation, triggers between objects and quick selection by lists. As in most of the software developed in this statistical programming environment, the objects interact from the individual to the individual in a simple sheet. We enlarge this capability to many ways of relationship, as database provide as well. We defined two principal objectives of the link:

the individuals selection: when a set of individuals is activated in a graph, a set of other individuals is selected and highlighted in (an)other graph(s);

the action: the individuals activation can also trigger off a more complex action having different possible impact on the object list.

Most of time, selection and action are related, but they can occur independently.

On the other side of the figure 5, we can see the action prototype, including three different prototypes. If the user wants to visualize individuals on a rather elaborated way (local zooms, specific colours, etc.) then s(he) can ask for the visualisation prototype containing several dedicated methods. Moreover, the need can also concern specific processes or calculation. In this case, the result won't be graphic but rather numeric (calculation-proto). Finally, the most complicated action on an object would be a change of its behaviour, such as starting an animation, changing the structure of its data, etc. The distinction between these three categories of actions is sometimes delicate to set, but this can help the user to decompose the relations in several parts easier to study.

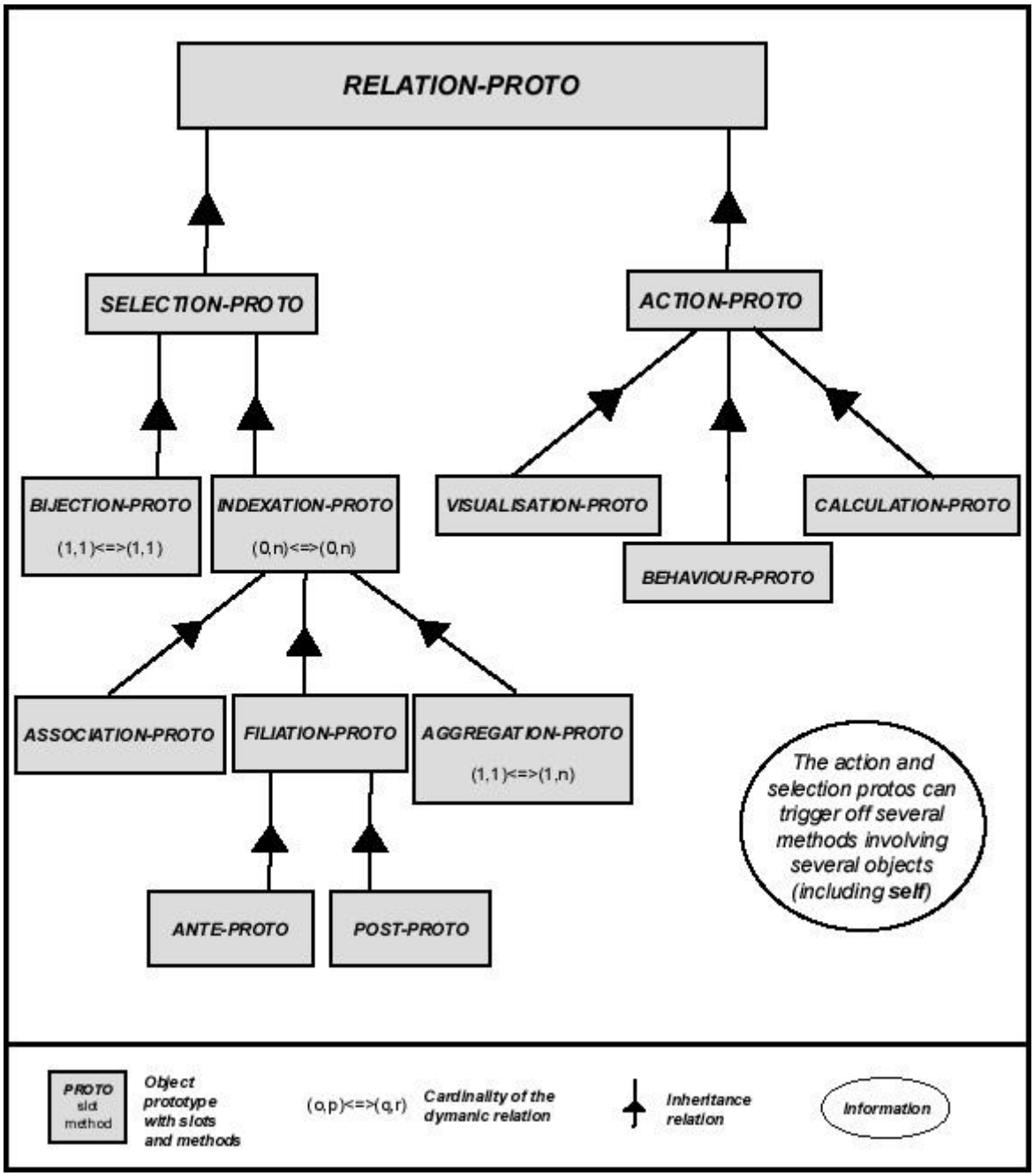

Figure 5: The Relation prototype

3. AN APPLICATION IN TEACHING SPATIAL ANALYSIS: EXPLORING HOLES AND DISCONTIUITIES IN WINE ROPES

To build an application using ARPEGE', we need first of all to identify the objects we want to study and the relevant associated statistical graphs related to discriminant variable. After having done that, the user has to populate and to update the index Many-to-Many for each couple of graph objects, especially for the indexation prototype.

The application illustrates a case of interactive image processing, whose aim is to find some relevant signatures of wine ropes organisation (figures 6 & 7). There is a double teaching objective: thematic and methodological. To do so, we activate a set of objects describing the studied territory spatial organization. The pixel values histogram shows a distribution of the radiometric levels. The co-occurrence matrix object calculates, for each cell of it, the frequency of the values corresponding to a given kind of pixels adjacencies (the pixel values are grouped in a few classes). The variogram object includes a new scaling dimension because it plots the absolute deviation between two pixel values with the distance that separates them. These three graph objects highlight several complementary aspects of the spatial structure.

When the user moves the window with the mouse, this activates the other graphs by a dynamic calculation which leads to generate a set of all the couples of contiguous selected pixels (N) and a sample made instantly from the whole set of pixels couples (N2, a selected pixel is computed with all the others) to improve the speed. Figure 6 shows which types of relations are involved.

This leads to an image on which the user can move a resizable window. In this window, the pixels are permanently reprocessed. Each step in the spatial analysis process allows to identify the window location in the image, the list of the involved pixels, the co-occurrence matrix values, the coordinates of the Lowess processed in the variogram graph, the summary of the pixels value distribution, etc. For instance, here is such a composite object, made with several basic objects and their signatures, including the prototype, the name, the individuals or parameters and a few statistical estimates:

> Explore26

((#<Object: 9195fa8, prototype = APPLICATION-TIME-PROTO> “Statistical indices” (6))

(#<Object: 91aaed8, prototype = PIXELS-MAP-PROTO> “Jean-Marc's wine ropes” ('Window size and position:') (50 42 10 10))

(#<Object: 9a5f54, prototype = HISTOGRAM-PROTO> 'Pixel values' ('Mean and median:' 'Min and Max:' 'Number of individuals per bin: ') ((84.09000000000002 95.0) (1.0 127.0) (15 2 4 4 23 26 26 0)))

(#<Object: 91c34c, prototype = COOCU-MATRIX-PROTO> 'Co-occurrence matrix' #2A((30 27 17 6 16) (23 55 47 10 17) (18 46 32 13 19) (7 12 14 1 14) (9 15 17 10 37))

(#<Object: 9297d8, prototype = VARIOGRAM-PROTO> 'Variances X Distance' (23.085053126206553 27.254046518948368 34.79982560872415 38.55202236703464 38.30347900588731 35.38433446053626 37.15191293213254 36.586033233644415 37.27753049647858 39.045634979229966 37.801220464964594 38.84617799163659 40.70703099217189 41.08640399393296 39.84716516488132 14.81013486598141)))

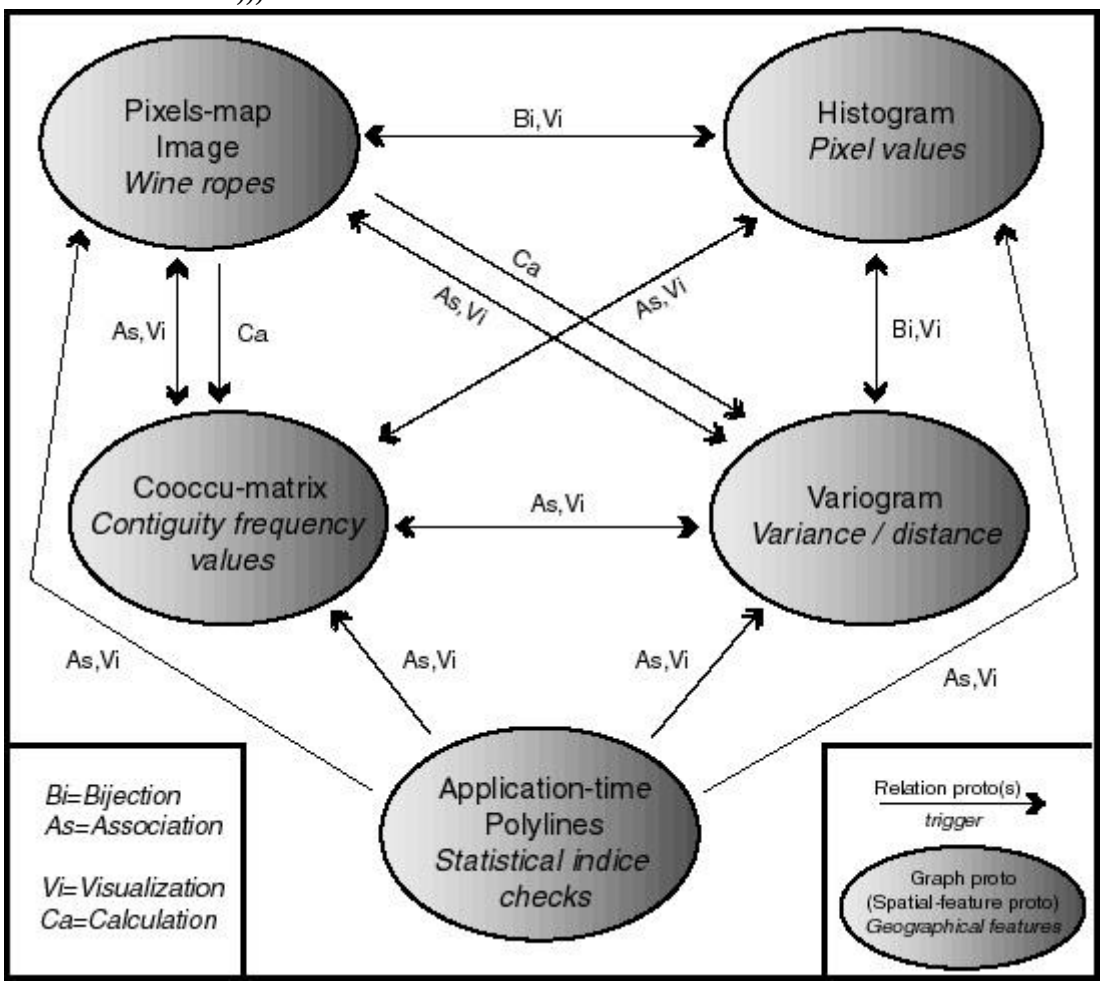

Figure 6: The spatial data model for exploring the wine ropes parcel

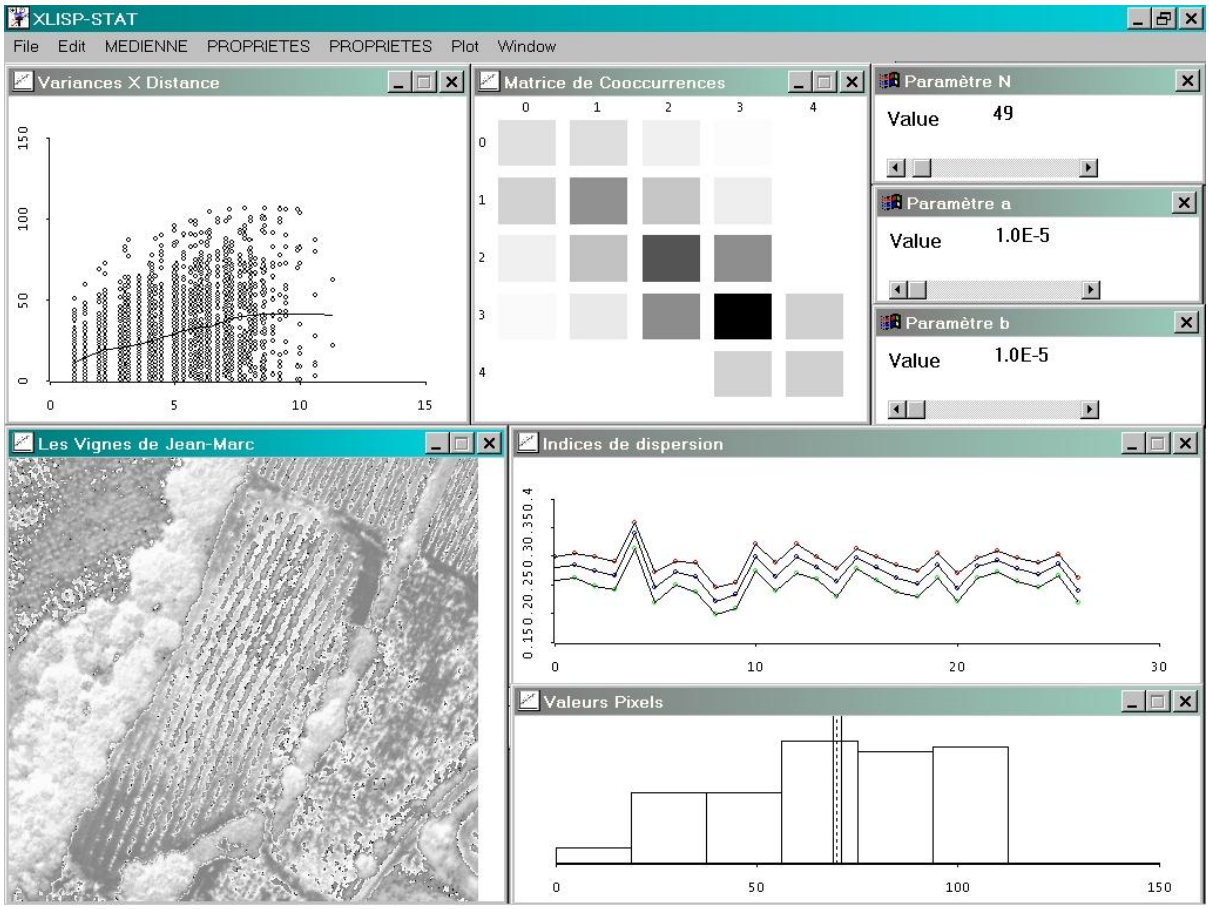

Figure 7: Exploring the Wine Ropes Oganization on a Image

DISCUSSION

ARPEGE' is often used for teaching spatial analysis. Different topics can be comprehended by students: exploratory spatial data analysis (deduction, induction, abduction, analogical processes of knowledge acquiring), Many to Many relationship and fundamentals of systemic approach, robustness in spatial analysis, spatial statistics (distributions, co-occurrence matrices, variogram). All these methods and concepts are not easy to understand, except for students in Master of Geomatics, who are a little more comfortable in such methodological approaches.

Although we didn't make any usability study about the use of ARPEGE' in teaching conditions (except for a specific graphic, called the Distogram, Josselin & Fabrikant, 2003), we use it for many years and can identify, through 300 students (about) who used it in methodological courses, different interests and limits of it. Indeed, it seems that there exist three main types of students in terms of profiles. A first (main) group includes students who are aware of spatial analysis and know about the flexibility of cartography and statistics to understand geographical space. For those, ARPEGE' is just another interactive tool to help in spatial analysis. They usually appreciate it and get easily the targeted concepts. A second group of students has been discouraged by maths and modelling during their schooling and preferred to develop their expertise in social sciences and literature or history. Usually, the exploratory point of view from ARPEGE' totally changes the way these students handle geostatistics. This 'science' then becomes accessible to them, because they can touch the data, test the robustness of the available methods, express their hypothesis in a graphical language, without SQL queries. At the opposite, the last (minor) group of students are good modeller who do not like at all the uncertainty included in statistics and subsequently applied on geographical data. For example, they cannot bear the fact there can exist kinds of 'fuzzy' signatures of composite spatial objects and that some objects can partially belong to a class or another. We usually loose these students who prefer go back to formal mathematics, based on deduction, assumptions, demonstration and exactitude. However, some of them sometimes come back because they realize that there can be some usefulness in exploring the geographical world and providing a way to take into account its variability and multiple aspects and scales.

REFERENCES

• Andrienko N. & Andrienko G., 2006, Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach, Springer.
• Anselin, L. & Bao, S., 1997, Exploratory spatial data analysis linking SpaceStat and Arcview. In Fisher M. & Getis A., (Eds), Recent Developments in Spatial Analysis, Berlin: Springer-Verlag.
• Bailey, T. C., Gatrell, A. C., 1995, Interactive Spatial Data Analysis, Longmann, Scientific &Technical, New-York.
• Banos, A., 2001, Le lieu, le moment, le mouvement : pour une exploration spatio-temporelle désagrégée de la demande de transport en commun en milieu urbain, Thèse de Géographie, Université de Franche-Comté, Besançon, France.
• Cartwright W., Peterson M.P., Gartner G. (Eds), 2007, Multimedia cartography, Springer.
• Cauvin C., Escobar F., Serradj A., 2008, Cartographie thématique 4. Des transformations renouvelées, Traité IGAT, Hermès lavoisier. 198 pages.
• Cook, RD. & Weisberg, S., 1999, Applied Regression Including Computing and Graphics, Wiley Series in Probability and Statistics, Wiley & sons, New-York.
• Cressie, N.A.C., Statistics for Spatial data, 1993, Wiley Series in Probability and mathematical statistics, New-York.
• Descartes, R., 1637, Le discours de la méthode, Booking International: Paris (Edition 1995).
• Dodge, Y. (Ed.), 1987, Statistical Data Analysis based on the L1 Norm and Related Methods, Y. Amsterdam: Elsevier Science Publishers B.V.
• Fisher, M., Scholten, H.J., Unwin, D., 1996, Spatial Analytical Perspectives on GIS, GISDATA 4, Taylor & Francis, European Science Fundation.
• Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986, Robust Statistics: The approach based on influence functions, New York: Wiley.
• Harris, R. L., 1996, Information graphics, a comprehensive illustrated reference, visual tools for analysing, managing and communicating, Management Graphics ed., USA.
• Hasslet ,J., Bradley, R., Craig, P., Unwin, A, Wills, G., 1991, Dynamics graphics for exploring spatial data with application to locating global and local anomalies. The American Statistician, 45(3), 235-242.
• Hearnshaw, H. M. & Unwin, D. J., 1994, Visualization in Geographical Information Systems, Wiley, New York.
• Heywood, I., Cornelius, S., Carver, S., 2002, An introduction to GIS, Prentice-Hall, UK.
• Hoaglin, D., Mosteller, F., Tukey, J.W., 1985, Exploring data Tables, Trends and Shapes, Wiley, New York.
• Hoaglin, D., Mosteller, F., Tukey, J.W., 1983, Understanding robust and exploratory data analysis, Series in probability and mathematical statistics, New-York: Wiley.
• Hornsby K. & Egenhofer M. J., 1998, Identity based change operations for composite objects, 8th International Symposium on Spatial data Handling, Vancouver, Canada, (Eds, Poiker & Chrisman), pp. 202-213.
• Huber, P., 1981, Robust Statistics, New York: Wiley.Josselin, D., 1999, A la recherche d'objets géographiques composites avec le prototype ARPEGE'. Revue Internationale de Géomatique, 9(4), 489-505.
• Josselin D., 2005, Interactive Geographical Information System using LISPSTAT : prototypes and applications. Journal of Statistical Software. February 2005, Volume 13, Issue 6, 20 pages.
• Josselin D., 2005, Interactive Geographical Information System, GisPlanet 2005, May 30-June 2, Estoril, Portugal, Proceedings, CDROM, ISBN : 972-97367-5-8, 12 pages.
• Josselin, D., Fabrikant, S., (Eds), 2003, Cartographie animée et interactive, Revue Internationale de Géomatique, Lavoisier, Paris.
• Josselin, D., 2005, Interactive Geographical Information System using Lisp-Stat. Prototypes and applications, Journal of Statistical Software, January 2005, Vol. 13, Issue 9., http://www.jstatsoft.org
• Josselin, D., 2003, Spatial Data Exploratory Analysis and Usability, Codata, Data Science Journal, http://journals.eecs.qub.ac.uk/codata/Journal/contents/2_03/2_03pdfs/DSS2.pdf.
• Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2001, Geographical Systems and Science, Wiley, New-York.

• MacEachren, A. M., Kraak ,M.-J., (Ed), 2001, Cartography and Geographic Information Science, vol. 8, n° 1, janvier 2001, Special issue on geovisualization, Journal of the American congress of on surveying and mapping.
• Monmonier, M. S., 1993, Comment faire mentir les cartes ou du mauvais usage de la géographie, Flammarion, Paris.
• Peterson, Michael P., 1995, Interactive and Animated Cartography, Published February, Engineering-Science-Mathematics, Prentice Hall, 464 pages.
• Tierney, L., 1990, Lisp-Stat, an object oriented environment for statistical computing and dynamic graphics, NewYork: Wiley, Interscience Publication.
• Von Bertalanfy, L., 1980, Théorie générale des systèmes, Paris : Dunod.
• Wood D., 2010, Rethinking the power of maps, The Guilford Press, London.
• Wood, D., 1992, The power of map, Guilford Press, London.
• Zeitouni, K. (Ed.), 1999, Data mining spatial. 9(4), Revue internationale de Géomatique, Paris, Hermès.