SPATIAL DATA EXPLORATORY ANALYSIS AND USABILITY

D. Josselin

ESPACE, UMR 6012, CNRS, 74, rue Louis Pasteur, 84029 Avignon, France
Email: didier.josselin@univ-avignon.fr

ABSTRACT

In this article, we intend to show how useful Exploratory Spatial Data Analysis is in improving spatial data usability. We first outlined a general framework about usability using conceptual modelling, including Data, Users and Methodologies. We then defined keywords into classes and their relations. A central ternary relation is enhanced to describe usability. In the second section, we present ESDA with its fundamental basics: i.e. robustness and way(s) to handle data and related graphic tools. We also described the software package ARPEGE'. Through a concrete example, we demonstrate and discuss its relevance for exploratory spatial data analysis and usability.

Keywords: Spatial Data Usability, ESDA, Geovizualisation, Conceptual Modelling, ARPEGE'

1 INTRODUCTION

In November 2001, a very fruitful 'brainstorming workshop' about Spatial Data Usability, organised by the Center of Geo-Information in Wageningen (the Netherlands) and related to a new thematic working group of the European association for development of GI and GIS (AGILE) was held. This research field is very attractive to scientists for many reasons. Linked to the more general framework of 'decision making', this topic provides a wide coverage of several research subjects and includes geographers, statisticians, modellers, data users and end-users. Under this polysemous label, the word 'usability' is centered around the Human being whose objective is to manage his/her own life within his/her environment. Knowing the reliability and the quality of the underlying data, dealing with the methods in order to make a (as) 'good' (as possible) or 'adequate' decision. These are also common objectives for planners, experts or researchers.

In this paper, we first present our own understanding of what 'spatial data usability' is, related to the definitions found in the paper issued from the workshop (Wachowicz, Riedermann, Vullings, Suárez & Cromvoets, 2002). We defined a simple conceptual model that enhances the relationships between data, users and methods, using several keywords. We then viewed the different concepts associated with Exploratory Spatial Data Analysis and show how helpful they could be for improving spatial data usability. We finally describe a concrete application developed using a specific tool: ARPEGE', which illustrates spatial usable data management.

2 ABOUT SPATIAL DATA USABILITY

2.1 What are the Definitions and the Norms for Usability ?

Studying Spatial Data Usability leads us to ask two preliminary questions:
- What's (are) the definition(s) for Data Usability ?
- Is *Spatial* Data Usability specific, compared to more general Data Usability ?

For us, it is too early to clearly point out whether the spatial point of view will bring anything new or different to such a problem. We prefer to focus on the current definition(s) of data usability, from which there is still a lot to disentangle to mark the boundary of the research field tackled.

As mentioned in the paper written by Wachowicz *et al.* (2002), a few definitions exist for Usability. Two international official norms (ISO 9241-11, 1998; ISO 9126, 2001) have been established and a few authors gave complementary definitions. These definitions are generally centred on the user and his/her goals with ergonomics often linking these elements.

According to these definitions and to our own feelings, it seems that there are three main ways to consider whether certain components should be preferentially highlighted in usability studies:
- experts may emphasize the user in his/her specific activity, while setting aside the data and the methods as a more or less useful working environment, which has its own properties and utility; in this case, the usability is 'user-and-goals centred'.
- a second point of view has already been developed in the research about (spatial) data quality, centred on the data

and/or the methods. But this would partly ignore the user processes by considering that holistic information independent of the user exists, information which could be adapted to any purpose; this is the 'data-and-methods-centred' point of view.
- the components may also be associated in a single package, making the data and methods very important components, whose intrinsic quality will serve decision-making in a whole set of mutifarious elements of information; this point of view defines the usability as a relationship between its components; we shall call it the 'system-oriented' approach.

Beyond these definitions and official norms, the term 'usability' covers a very broad range of meanings, according to numerous points of view. The discussion during the workshop showed that this greatly depends on whether the individual is a data user, provider or researcher, but also on his/her own background and his/her relationship with the data handled and their environment (software, subject, professional). Many keywords make reference to usability. They are presented in the schematic we propose in the next section, which is 'system-oriented', according to the previous classification.

2.2 Three main Classes of Objects for Spatial Data Usability Modelling: Data, Users and Methods.

We propose a simple schematics (Figure 1) based on classical entity-relation conceptual modelling (Booch, Jacobson & Rumbaugh, 1999; Muller & Gaertner, 2001). This representation is 'system-oriented'. It includes classes of entities and their relations. For each of them, we attach a set of keywords which have already been used during the workshop or in related scientific or technical literature.

The first component is the class 'Data with their Characteristics'. The data are described by their *quality*, which may be *guaranteed* independently from users, their *quantity* and their *structure* (database). These data have a *cost* in the sense that they have been catalogued or measured. A 'good' spatial data quality implies knowledge of, for instance, the *accuracy*, the *precision* and the *completeness* of data. Data may be historical so that they would have a *shelf-life* and would be *used by date*. They can have a high level of confidentiality/protection and be *legally defensible*. The data are represented by their *core* and sometimes by a confusing *noise*. *Novelty*, *integrity* and *reliability* are very important characteristics of the data. All these characteristics concern the data themselves.

The second component is the class 'Users with their Objectives'. Here we try to describe what is directly associated with the user, such as his/her *profile* and *personal experiences*. (S)he states his/her own *point of view* and has his/her *habits*. (S)he may be specialist and have some specific competence in some *subjects*, in which his/her colleagues consider him/her as *credible* and *reliable*. (S)he may be a *data provider* or *consumer*, an *expert*, a *scientist*, a *planner*, a *technician*. (S)he defines specific *objectives* and gets some *satisfaction* (or not) during his/her activity. We think all these listed elements remain descriptors closely associated to the user.

The third component is the class 'Methods with their Design'. It encompasses some intrinsic capabilities of the methods to improve human knowledge. For example, *robustness* and *resistance* (to 'outliers', for instance) embody *statistical efficiency*. As the method must adapt to any configuration of data batches encountered, this property is assigned to the method, rather than to the data. That also seems to be the case about the following keywords. Methods may have a power for *decreasing noise*, *emphasizing trends* or *analysing individuals*. This class also lists the numerous methods and tools which are available for comprehending spatial problems and extracting knowledge e.g. *graphics* and *geovisualization*, *software interface*, *DBMS* within *GIS*, *neural networks*, *induction trees*, *hybrid systems* (Josselin, 1995).

To explain more thoroughly what is meant by spatial data usability, we now propose to examine the relations between the different classes of objects (data, users and methods). Notice that many of the terms relating to usability can be located in one or the other of the relations. First of all, let us have a look at the binary relations.

2.3 Binary Relations between Classes within the Spatial Data Usability Model

For instance, the relation linking Data and Users refers to several elements. Data can be *authoritative*, *exclusive*, *trusted* or *interesting* regarding a given user. *Compliance*, *pertinence*, *utility*, *appropriateness* or *validity* refer to the goals that the user has to achieve. The status of *metadata* is specific, in the sense that it is some information about data and that it mainly involves data providers or users. This is typically associated to the relation between data and user.

Another relation, this time between users and methodologies, encompasses several key-elements we can now list. *Fitness to use*, *usefulness* and, once again, *utility* and *trust*, tightly link the user to the methods and tools (s)he uses. Certainly, each method must be adapted to (a) defined purpose(s). Although *habits* correspond more to the user as a specific characteristic, *preference* is a concept connecting him(her) to the methodology chosen, through his(her) choice.

A third binary relation associates the data and the methodologies. A few of the elements resemble the previous ones. Indeed, the methods and tools used have to be *pertinent, appropriate* or *adequate* to the data (and *vice versa*). The data must sometimes follow the *constraints assumptions* required by the model. They must be structured to be *searchable* and *accessible*. The *statistical dependencies* within the data have to be studied and evidenced using various methods, considering the *trend* and *(marginal) points* as well, where it may be suitable to find innovation among outliers.

Usability is defined at two different levels in the schematics (Figure 1). It may involve all the described components by global behaviour. In this case, we can retrieve complementary and disseminated parts of usability, according to the different points of view about the data, user or methods and their relationships. This includes most of the definitions, norms and keywords which have been described previously.

2.4 A Central Ternary Relation within the Spatial Data Usability Model

Moreover, we propose to give a more accurate definition by focusing on the central ternary relation linking (i) data with their characteristics, (ii) the user with his/her objectives, (iii) methods and tools with their design.

This indeed appears to be a convenient place to explain where (spatial data) usability gets its consistency, by generating some *benefits* and a global *added-value* at a more abstract, decisional or conceptual level. This relation and its elements refer to a process that includes several components presented in the three classes. Combining user, data and methods provides properties and capabilities which surpass a simple juxtaposition of these three main blocks represented in the three classes. When drawing the boundaries of such a concept, we logically retrieve common and well-known elements. Their peculiarity is that, even if they remain more general compared to the keywords listed in the classes, they are closer to each others within the spatial processes and they have a higher level of abstraction. Thus, such a usability framework appears more homogeneous, but is in practice more extensive.

At the intersection of data, users and methods, we can also find the term *ergonomics*, for improving general quality of processes and materials, and *marketing* for spreading products and concepts. *Decision support* and *making*, *sustainability* generally refer to geographical planning. When a product, a process or a concept is worked out, it is crucial for it to be usable. This induces, depending on situations, a significant *capacity of integration*, *generalization*, *prediction*, *extrapolation*, or *interpolation*. This also requires the development of approach(es) for *data querying* and *mining*, specific process(es) for spatial *analysis* (*confirmatory* and *exploratory*) in order to extract the relevant *knowledge* and to *model* geographical space. All these terms appear to us as a complete and coherent framework around *usability*, or *spatial data usability* if applied to spatial information.

Exploratory Spatial Data Analysis (ESDA) belongs to such a data usability framework. Indeed, it does not completely occupy the scene of the framework. The core of ESDA is located in the central ternary relation and several of its components relate to the keywords expressed in the model we previously presented, within other classes and relations. Using an example, we now are going to show why and how ESDA may be a very helpful way for managing and improving spatial data usability.

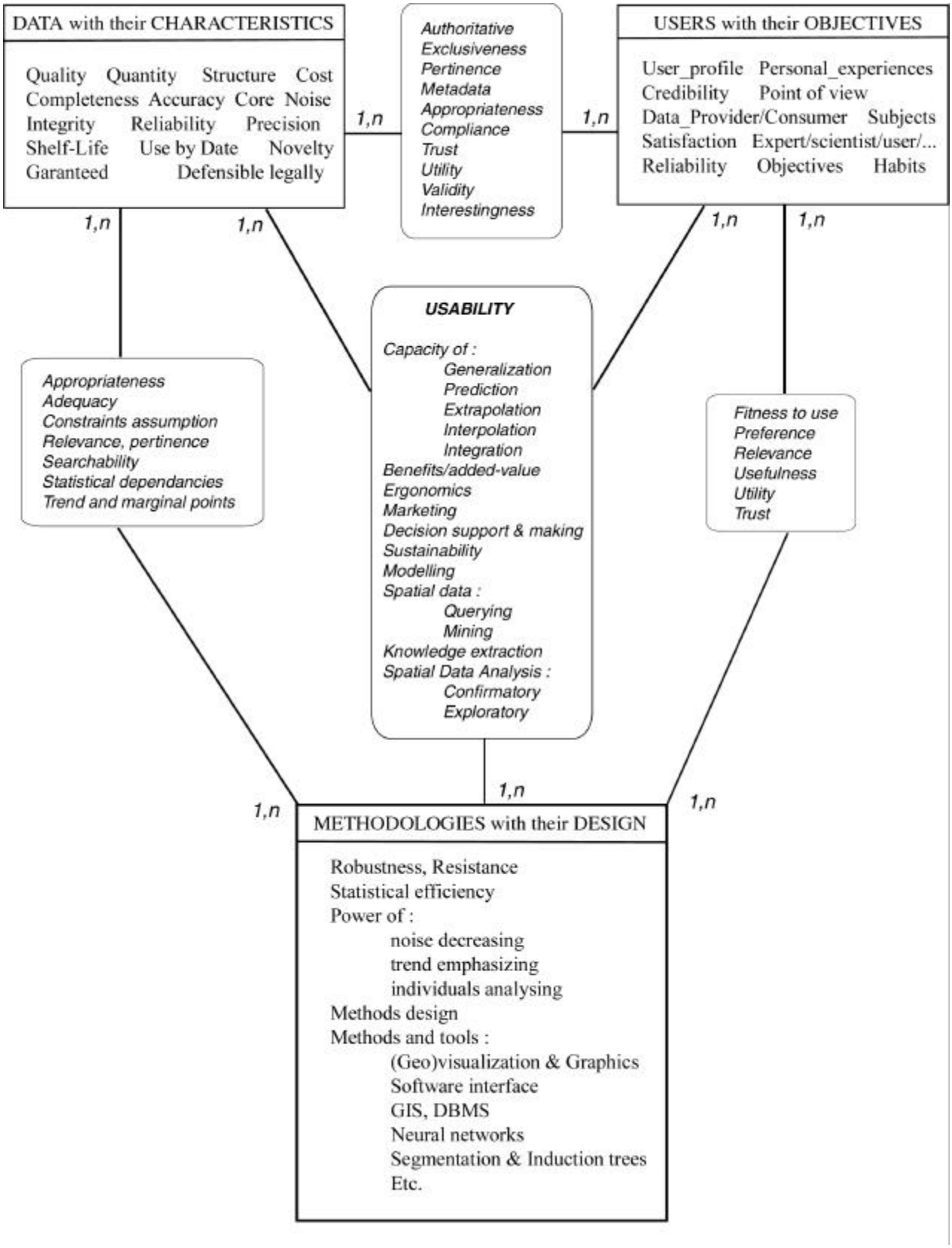

Figure 1. A general framework for the concept of usability: a bridge between data, methods and users

3 EXPLORATORY SPATIAL DATA ANALYSIS (ESDA)

3.1 Fundamental Principles of ESDA

Exploratory Data Analysis was initiated by John Tukey (Tukey, 1977). It is now used in many fields (Behren, 1997), and notably in geography, for tackling local geographical problems (Fotheringham, 1997; Fotheringham, Brunsdon & Charlton, 2000). However it remains quite marginal compared to confirmatory analysis in spatial analysis. ESDA lays down a few important concepts that may define a different approach or even 'philosophy' for data handling and analysing while exploring spatial data.

It's common nowadays to associate Exploratory Spatial Data Analysis and the close fields of Geographical Information Systems and Cartography (Kraak & Ormeling, 1996; Slocum, 1999; MacEachren & Kraak, 2001). However, if we look more carefully into the provided functionalities, it seems that users miss a few important possibilities. GIS often remains based on powerful but non graphical query languages. These functionalities are now being included in most of the GIS softwares, but they are still additional components. They usually provide a graphical exploration of only one class of geographical objects and are not very efficient in terms of statistical analysis capabilities, except if the user is able to develop his(her) own application within the associated programming environment. However, the good point is that they offer a high representation of the conceptual data model within Case Tools (Smallworld System, for instance).

On the other hand, Exploratory Spatial Data Analysis has provided a wide set of sophisticated statistical and graphic methods, which has been developed for many years. Written in different programming languages and environments (S+, R or XlispStat, Tierney, 1990), these methods are available in libraries which greatly improve the users capability to design his(her) own application. With this approach, researchers have several different ways of constructing useful geostatistical and geovisualisation environments for experts and users. The first way provides a dynamic link between two complementary softwares. This has been efficiently made in some cases. The second way is to include the available libraries in the GIS environment. The third way is to incorporate both mapping and statistical functionalities. It seems that there is still a lot of room left for improving the two last approaches.

A fundamental concept for exploratory spatial data analysis is robustness. Note that this word was previously mentioned in the 'methodologies with their design' class. From a statistical point of view, a good estimator is considered as robust (Hoaglin, Mosteller & Tukey, 1983; Huber, 1981; Hampel, 1986; Lecoutre & Tassi, 1987) (i) when it is only slightly affected either by a small number of gross errors or by a high number of small errors (resistance), (ii) when it is only slightly affected by small departures from the underlying statistical hypotheses (robustness). The efficiency of an estimator in terms of robustness is highly dependent on the local conditions related to data configuration. A large panel of robust estimators (Andrews, Bickel, Hampel, Huber, Rogers & Tukey, 1972), some of them based on the L1-norm (Dodge, 1987), have been developed in order to improve the poor efficiency of classical estimators, which are generally related to the mean and the standard deviation (referring to L2-norm). By extension, we can say a process, a method or an approach is robust when it deals well with the noise confusing the information, maintaining the trend(s) while either rejecting outliers or reducing those that could excessively affect the statistical estimate. In terms of decisions, this would provide a stable and efficient support for decision-making. It is at least very helpful to provide clear information for analysts. It is much better if this information generates a consensus of opinions and suitable decisions, because all local peculiarities have been taken into account.

However, robustness alone is not sufficient for defining a good framework. In some cases, robust methods can drastically eliminate small groups of potentially interesting marginal individuals. To counterbalance this, it is necessary to develop methods that provide a permanent connetion with raw or derived data. This is enabled by simple geovisualization functions such as dynamic links between lists of individuals in many graphics, different selection modes (selecting in a rectangle, brushing, for instance) and animations, or statistical methods (data transformation...) (Cleveland, 1993). These complementary views highlight differently tackled problems, enabling a systemic approach, providing an analysis of the objects in each class and their statistical, structural or functional relations. It is also possible to link complementary tools in the same operating system (Anselin & Bao, 1997; Josselin, 1999b; Banos, 2001).

Instead of grouping the whole batch of data in the same package and extracting a global trend whose relevance may be discussed, we may want to focus on subsets of typical data and explore the possible local models at different scales. The studied populations can easily be changed and the user can either develop a global analysis if considering the whole data, or a local exploration. The dialectic between these complementary levels of representation is very fruitful in terms of knowledge extraction. Even outliers and residuals may be handled as data and models. Thus, ESDA appears much more generic than confirmatory analysis: it doesn't require specific and strong assumptions because of its robustness, and because it is still possible to carefully study any model at any scale, including the one that involves the whole data set. Notice that individuals remain very important and cannot be exchanged or lose their identity. Even when we want to extract general laws and group individuals by their similarity, let us try to keep in mind the objects identity: objects are not interchangeable.

Geovisualisation, robustness, modelling were three important keywords that we mentioned previously. We believe that ESDA provides a good management of the decision process due to its own principles and functionalities. This approach avoids black-box problems and reduces input-output processes. They give the user and expert an opportunity to completely manage the spatial data, the associated methods and also the speed of the learning process. We believe this improves the data usability.

Now let us describe an example of a software dedicated to robust, exploratory spatial analysis: ARPEGE', which we think has the capacity to improve the analysis process due to an informed management of data usability.

3.2 ARPEGE'[1] Functionalities

3.2.1 Interactive Mining for Spatial Data

In the research field of Exploratory Spatial Data Analysis and within the Statistical Programming Environment XLISPSTAT, we developed a software called ARPEGE' for "Analysing Robustly in Practice and Exploring Geographical Environment" (Josselin, 2000). It is designed to explore geographical objects and their relations. It has been inspired by a deep need for a tight association between the supervised human analysis process and multiscalar data. It enables the combination of two learning 'temporalities' or phases: decision-making process and knowledge extraction. In other words, we think that the expert will be much more efficient at decision making if, at any time, (s)he knows (and can act on) the models parameters, the available (geo)statistical tools, the interesting places and geographical objects, in order to analyse and conduct his (her) spatial own investigation. This implies that dynamic and permanent links between any of the world related representations are mandatory, these representations being for instance, maps of geographical objects, statistical plots or geostatistical models. ARPEGE' has already been labelled a 'spatial data interactive miner' (Zeitouni, 1999), a part of the spatial data-mining field. The goals are indeed identical. Only the ways of obtaining the inference rules and extracting what we call 'composed/composite geographical objects' is different, because it includes an exploratory approach.

ARPEGE' provides dynamic links between maps and statistical plots, added to several robust spatial analysis methods to improve the users' decision-making and data management usability. It is original because of several methodological aspects.

---

[1] Analysing Robustly in Practice and Exploring Geographical Environment (ARPEGE') suggests the idea to play a 'nice music' with data as 'single notes' or as a 'coherent chord' ...

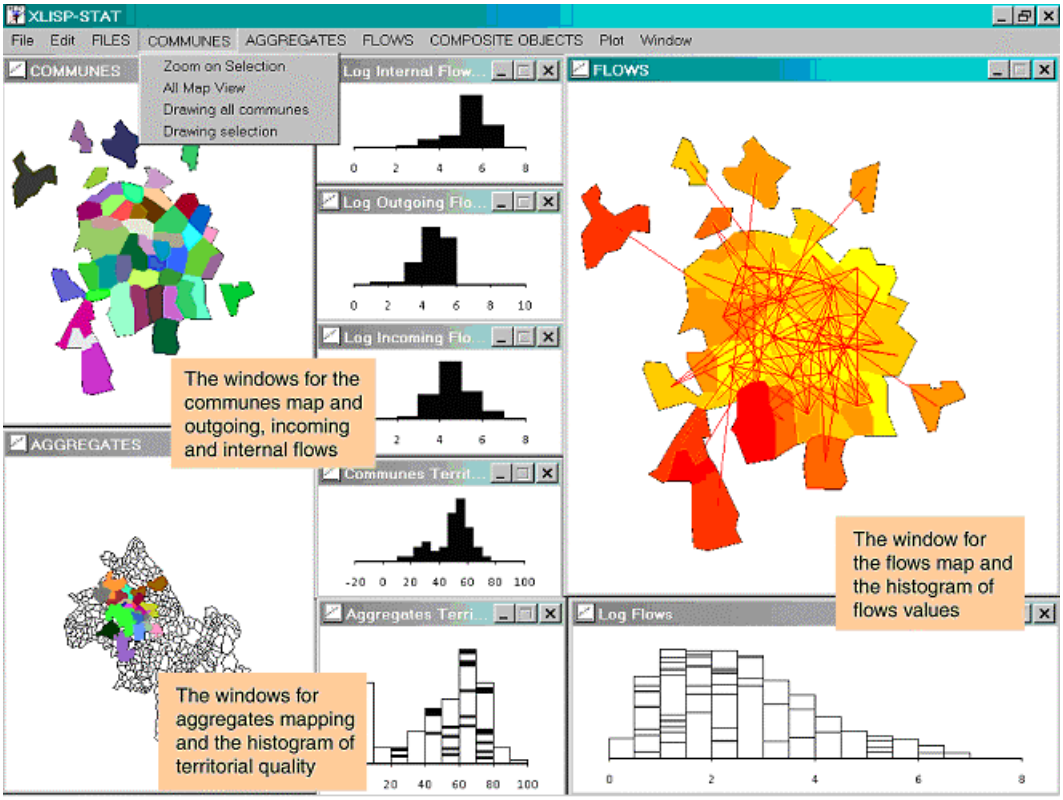

Figure 2. Exploratory analysis and spatial partitioning of French agricultural flows. An 'agricultural inter-communal flow' is a quantity exchanged between a 'source commune' and a 'target commune'. This quantity may be a commuting flow (from the residence to the work location) or an agricultural flow (the fact that a farmer uses parcels in different communes). For a given commune, flows can be 'internal' if they occur in the same commune, or 'external' ('outgoing' if the commune corresponds to the source commune, 'incoming' for the target commune). It is also possible to include communes in aggregates, such as French administrative 'cantons', 'departments' or 'regions'. In the example shown in this figure, we handle three types of objects: communes, aggregates (cantons, for instance) and flows. All of them are described by several attributes plotted in histograms.

3.2.2 Exploring and Modelling through Objects and their Relations

ARPEGE' implements relations between different objects classes (as in the Relational and Object Data Management Systems) whereas many visual tools do not allow a dynamical and systemic view of objects and relations. Mainly, the system implements (through indexes and pointers) the classical (i) association and (ii) aggregation relations (for example, some French communes belong to aggregates, as shown in Figure 2). This is made using indexing, where some individuals points to some others, with three possible types (one to one, one to many and many to many relations). Inheritance is also provided when some objects need to be classified following a hierarchical structure (that is a basic concept of Object Orientation) (iii): for instance, a waste land may be declared as a special form of agricultural land. Moreover, we added a 'behavioural' relation (iv), which triggers an object's behaviour (a calculation, movies, color changes, indeed any process) when another object is modified, selected or even activated. This provides dynamic interaction between objects. This relation is very useful but it is not very easy to deal with because of its high complexity. For example, moving a window on an image will dynamically modify the points involved and thus change the estimate of the local spatial autocorrelation or the shape of the variogram computed on a studied area (see Figure 3 for instance). These interactive graphic queries do not require a browser or sentences to write and are easy to apply for any user (Hasslet, Bradley, Craig, Unwin & Wills, 1991). The selection can be made on the screen. It is also possible to work interactively on sub-populations from the whole set of individuals in order to extract tightly related objects, which can become new classes of composite geographical objects. These selections can be performed within two complementary points of view: spatial (by maps) and statistical (by statistical representations and graphs).

The last functionality is a historical function which allows the user to store his(her) outcomes in a graphic or in a file. Indices are automatically created/updates at each step of the user's spatial investigation. Indeed, it provides a guide for keeping track of the relevant investigation by pointing to the status of the global structure of the relations and objects at any stored point. For instance, the user can look at a graphic on which every change of statistical calculation during his(her) spatial exploration is plotted. By selecting one of these points, (s)he can retrieve the location of the moving window and all its related statistical measurements and thus identify the configuration(s) which gave for example the lowest variance.

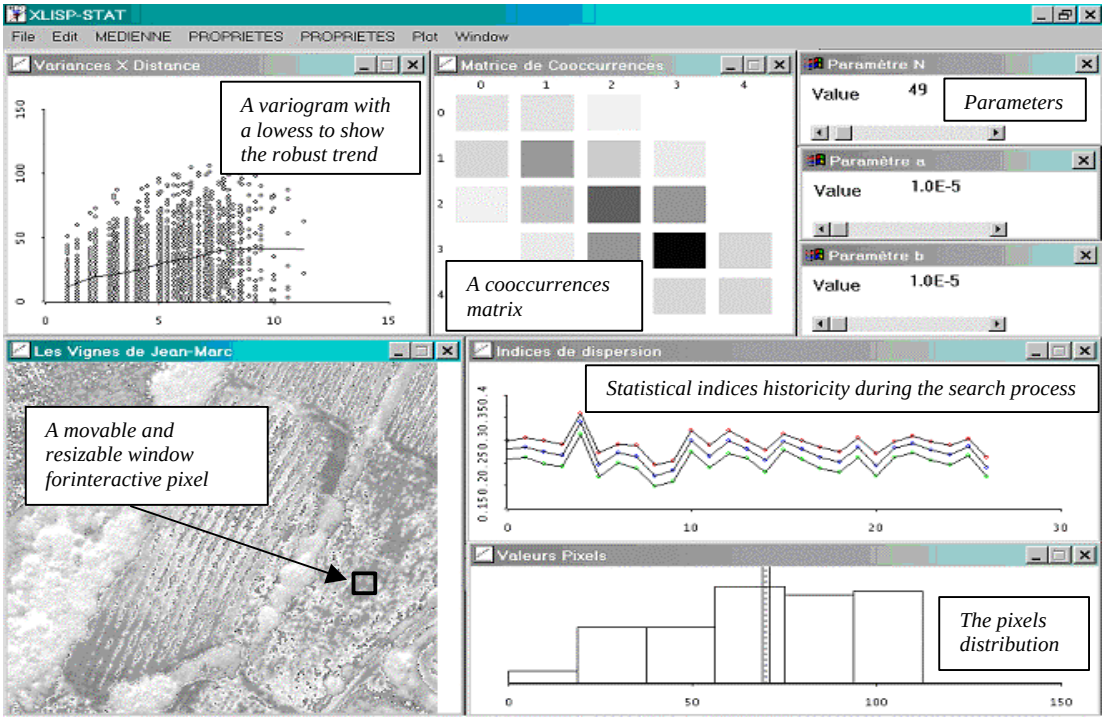

Figure 3. Interactive local variogram and cooccurrences matrix applied on interactive pattern recognition of wine ropes. This figure shows a similar dynamic environment. In this application (about wine ropes exploration and spatial patterns recognition), the user moves a window on the image, while the histogram of pixel values, a local variogram and a cooccurrences matrix are permanently reprocessed. Users can draw specific paths during exploration and extract statistical composed signatures s(he) might consider as relevant. S(h)e can also see a graphic which stores the key steps of his/her search track.

3.2.3 The High Potentiality of Models Coupling

ARPEGE' is also an original tool in the sense that it includes different kinds of spatial modelling in the same framework, as exemplified by Figure 4. The first group is composed of geometrical models: (i) raster, (ii) vectorial and (iii) topological. Indeed, these three spatial representations of the geographical world can be processed in the same environment and at the same time, which is quite novel among GIS software. We contend that it is a way to reconcile these two large fields of GIS applications, which generated a big discussion about spatial (dis)continuity for a few years (Worboys, 1995; Josselin, 2003). The user can explore the topological structure of his/her spatial database, assess the shapes of the pixels statistical distributions, calculate some index on areas or polylines. More than these basic functionalities, the software provides different ways to make the three models interact. For example, it allows a set of pixels in a polygon to be selected, to identify which arcs are connected to which nodes in a defined aggregate of pixels. Figure 4 shows its application to the Loire, a large French river. The available data include a topological map of the land use and the river at two separated dates (1973-1983), a remote-sensed SPOT image and aerial photographs. Many related graphs also interact. In this example, the relations are associations (the land use topological structure), inclusions (pixels of the images and the photographs included in polygons, for instance) and intersections (the spatial modifications of the Loire borders).

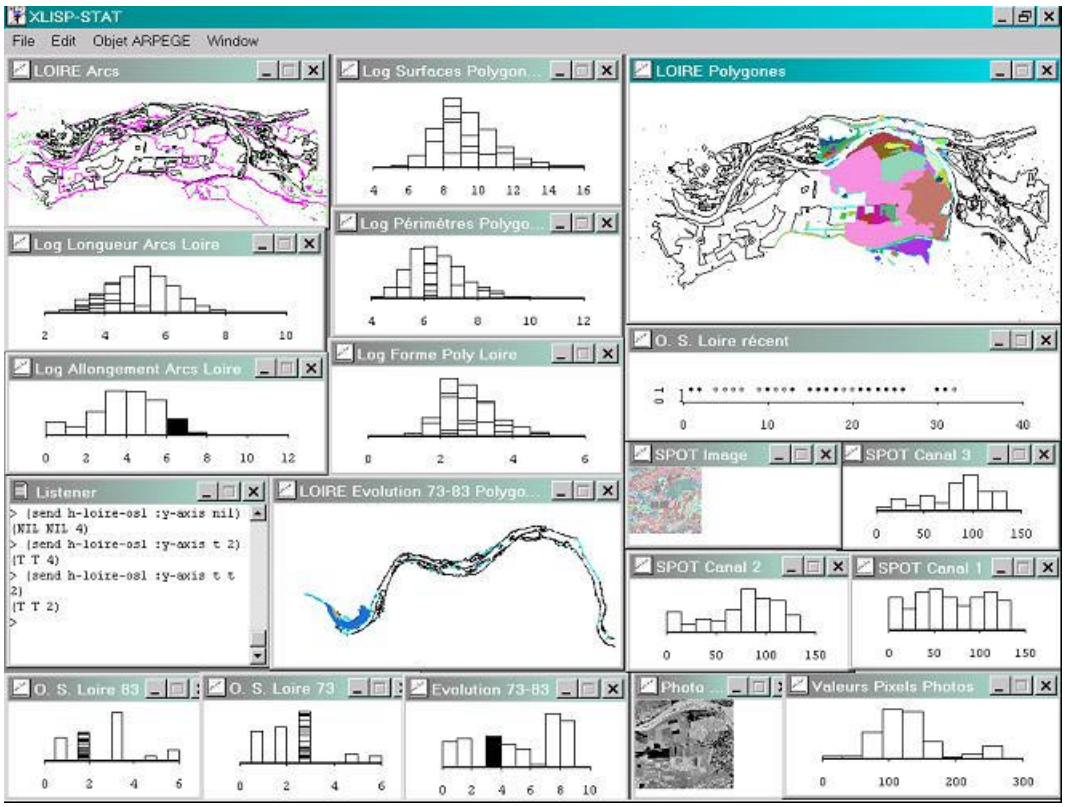

Figure 4. Exploring the Shapes of a French River through several complementary and linked Spatial Models: the case of the Loire.The combination and the interaction between raster, geometrical and topological models are presented here, in a context of pattern recognition for a French large river: the Loire. One can see different types of objects such as remote sensing data, air photographs, the linear topological structure of the river on two different dates. Associated with each of these a set of histograms which describe their relevant attributes for shapes description.

Because of its interactive and multifaceted viewing, the functions of exploratory datamining and knowledge extraction, ARPEGE' stays abreast of the research and progress in decision-making and usability. We can now detail an example that involves data quality, relationship enhancement and spatial partitioning.

4 MIXING DATA QUALITY ASSESSMENT AND EXPLORATORY TOOL, FOR IMPROVING SPATIAL DATA USABILITY: THE EXAMPLE OF FRENCH AGRICULTURAL FLOWS AND SPATIAL PARTITIONING.

4.1 What is the problem and what are we looking for ?

An agricultural inter-communal flow is a quantity of parcels (measured by their surfaces), which are exchanged between a source commune and a target one (Josselin, 1999a). It is in fact the areas used by a farmer in a commune differing from his/her housing one. Between two communes, the single flows involving each farmer are aggregated. For a given commune, flows can have different status: internal if they occur in the commune, external (outgoing if the commune corresponds to the source commune, incoming for the target commune). It is also possible to associate communes with different aggregates, such as French administrative ‘cantons’. In the example, we handle three types of basic objects: communes, aggregates (cantons, for instance) and flows (Figures 2 and 5). All these objects are described by several attributes plotted in histograms.

Figure 5 shows a theoretical model of these three features. Six aggregates are delimited (A, B, C, D, E and F). Each of them contains a few associated communes related by flows. The flows can involve different agricultural surfaces (expressed by different line widths). Communes are described by different variables (for example, the number of farmers, which is drawn by various greys). Aggregates, in this example, include several communes.

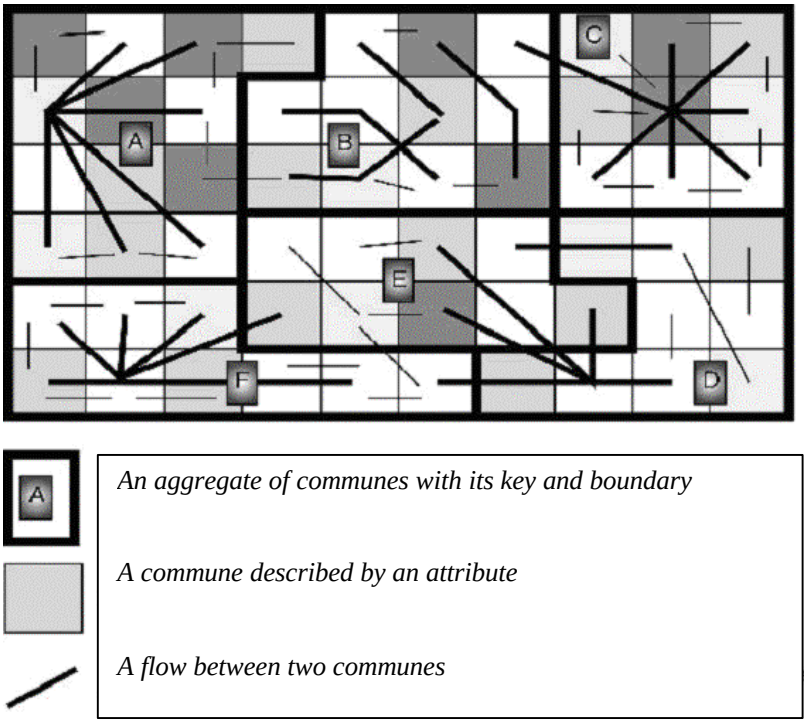

Figure 5. A model of a geographical space, composed by communes, aggregates, inter-communal flows and associated variables.

Let us imagine what questions could be asked about such an example? First, what are the aggregates which strongly look like each other, considering their communes and flows? A graphic analysis of the Figure 5 leads us to eliminate the aggregates which are too 'typical': E, poor in flows, B because of the spatial scope of its flows and the low values associated to its communes and D, whose outgoing flows are too numerous. Let us now find the couple of similar composite geographical objects between A, C and F. Regarding the commune attributes (squares in grey), the couple {A,C} may be chosen, due to its communes' low values noticed in F. However, if the expert wants to enhance patterns of polarized spatial flows, then (s)he would perhaps prefer to associate the couple {A,F}.

This finally identifies the basic features or associations of heterogeneous objects, characterised by particular attributes through spatial and reciprocal or non-symmetrical statistical relations. This is what we call 'composite geographical objects'. We believe this conceptual level of describing the spatial structure offers much meaning to experts during the decision-makingprocess, especially if we become able to build a user-friendly tool to provide such a capability. This is one way to improve spatial data usability.

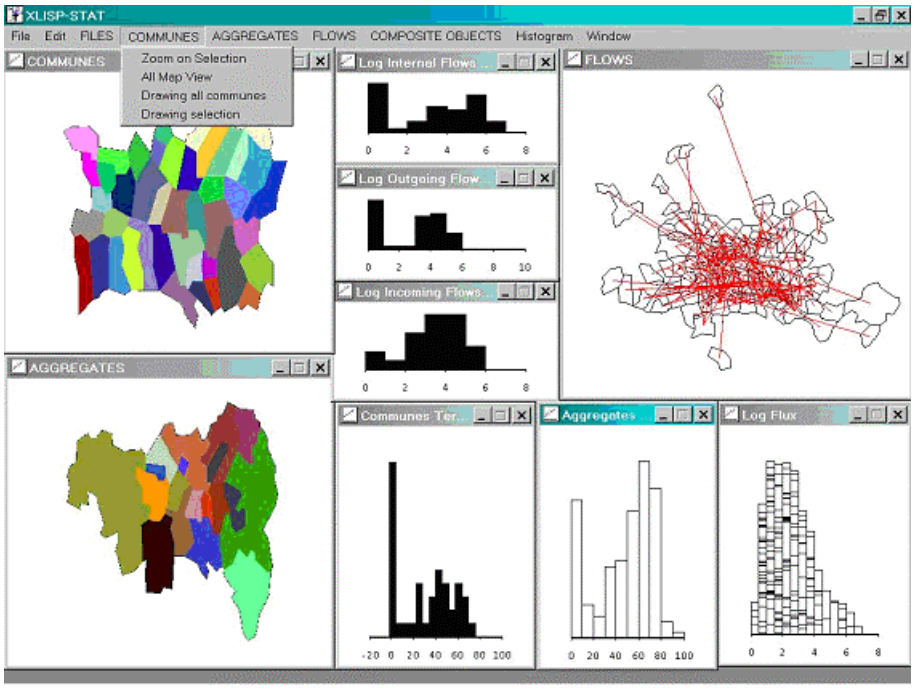

Figure 6. Focusing on different linked objects (maps, statistical graphs, notably) with ARPEGE'.
Each geographical object has a specific menu with a generic part (zoom, selection, all map view...) and a part that is dedicated to its own properties.

4.2 The Structure of the Database

In the application concerning the agricultural flows, ARPEGE' deals with 3 objects classes: communes, aggregates of communes and inter-communal flows (Bolot, Chatonnay & Josselin, 1999; Josselin & Bolot, 2000). The objective assigned by the applicants (the Regional Chamber of Agriculture) is triple (Josselin, 2000):
- exploring the flows, the relations between them, communes and aggregates, finding typical areas depicted by 'composite geographical objects';
- building an optimal partition whose aggregates include communes which have very high level of flow exchanges;
- aggregating the flows in order to overtake the administrative threshold below which the value of the flow must be confidential;
- while making the aggregates as small as possible.

A key variable measures the spatial partition quality via aggregates by computing an index of 'territorial pertinence'. The territorial pertinence is a proportion of internal flows among the all flows of the entity (commune or aggregate). It is processed by dividing the internal flows of aggregates (exchanged between its own communes) by the whole flows concerning it (internal, outgoing and incoming flows). This index provides an assessment of the aggregates' quality, a part of data usability that becomes manageable due to the interactive environment. This is a second link to the concept of spatial data usability. The user is closer to his/her spatial data than if (s)he would have apply a query using a statistical tool or a traditional GIS (input data -> output results -> statistical test -> decision). We believe this is a significant improvement of spatial data usability.

4.3 The Sequence of an Analysis Process

In such an analysis process, the analyst can handle multiple statistical distributions related to several attributes of the objects. (S)he can, as mentioned previously, focus on any group of individuals belonging to any class of objects. The selection may be processed in an histogram or on the map (Figure 6).

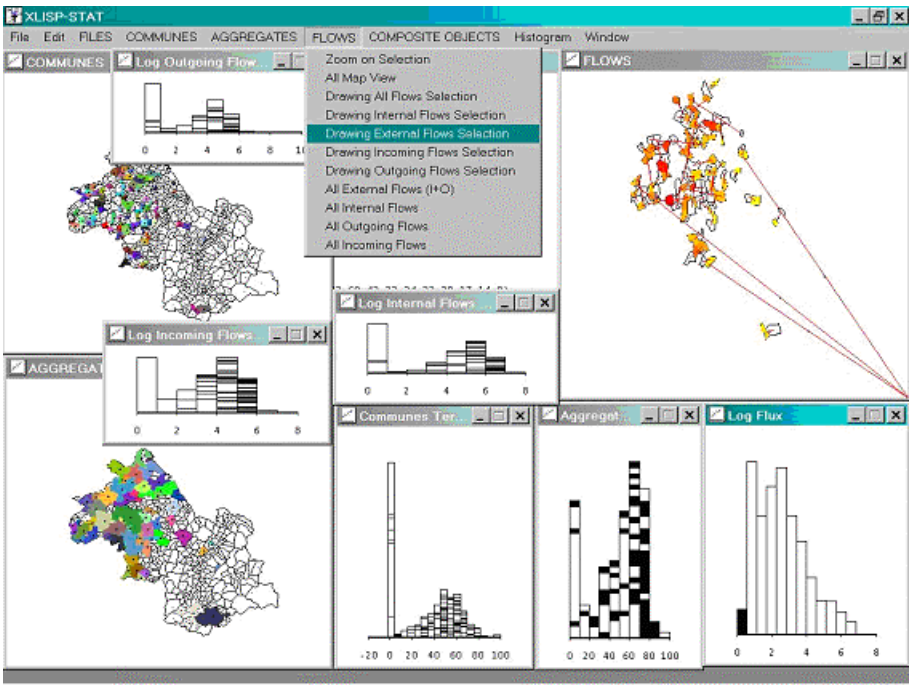

Figure 7. Statistical dependancies between flows values, and flows reported to communes (internal and external flows).

Using ARPEGE', it is possible to evaluate the statistical dependencies between different object variables. For example, in Figure 7, user selected flows with low values. (S)he can analyse their spatial spreading and their distribution in the other variables in the histograms associated with communes. (S)he then notices that most of these communes have low internal and outgoing/incoming flows. This also corresponds to aggregates with rather good territorial pertinence. All these relations are sufficiently neat to allow the user to extract a composite geographical object, which identifies the geographical areas characterised by high agricultural exchange and dynamism. This object refers to a subset of individuals which present statistical and/or spatial dependencies. Notice that the statistical dependency does not just occur between attributes of individuals of the same class, but also between attributes of individuals belonging to different classes, which is a novelty in similar software package. The whole set of objects, attributes and relations makes a composite geographical object. That is why the extracted object is quite similar to a production rule, which may be extracted using automated methods such as induction trees, neural networks or any data mining method. The only, but important, difference, lies in the capability user may develop to (in)validate, step by step, the geographical composite objects s(h)e finds. More than a simple tool for geovisualisation, we believe ARPEGE' is a powerful means of exploring data, through their multifaceted relations: notably through functional and structural relations, aggregations, statistical dependencies, affiliations and inheritance. These relationships may be the key for efficient spatial data usability: we tried to enhance its role and its systemic approach.

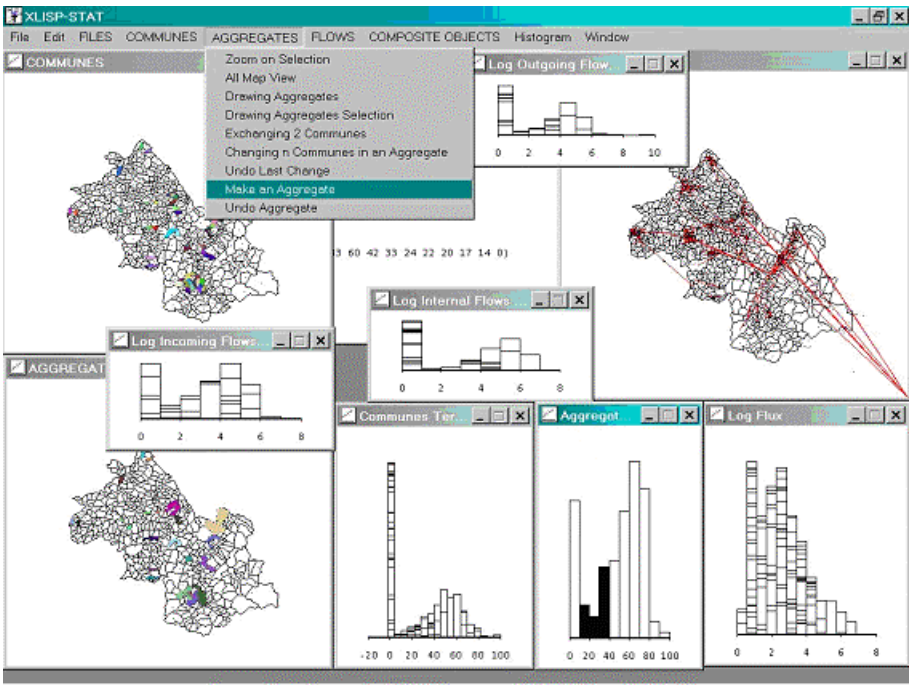

Figure 8. Interactive modifications of aggregates in ARPEGE'.

We previously showed how ARPEGE' facilitates the exploration of relations and objects. Another ARPEGE' functionality is provided that modifies a functional/structural relation in the database (Fig. 8). Due to this particular functionality, a relational modification may immediately be transmitted to the database, and can induce some predefined calculations (via triggers). For instance, if user is not satisfied by an aggregate's territorial pertinence, (s)he may then identify the communes and the associated flows to be processed. In order to improve the partition quality, (s)he can focus on a typical area, exchange 2 communes between different aggregates, move communes to other aggregates, or create a new one by merging selected communes. Due to dynamic links, the aggregate's territorial pertinence is reprocessed and histogram redesigned in 'real' time. Indeed, the status of some flows (internal *vs* external) may be changed following the modification of the spatial partition. This requires some local modifications in the database, involving aggregates and also relations. By doing this, user can modify by hand his/her partition, do tests, come back to some previous efficient trials, while taking permanently into account the partition and the aggregates' quality (Figure 8). In this case, the statistical distribution of the territorial pertinence is crucial information. Users can search to obtain a specific shape of distribution according to their own view of what is a 'reliable' or 'adequate' spatial partition. The fact that the user can always affect his/her data and their relational structure is very helpful for data usability and therefore decision making.

Once a user or expert finds a composite object, (s)he can save it in a named variable, and recall it at any time. This object has a specific structure (Figure 9). It is a set of different lists:
- a list of graph identifiers (objects prototypes);
- for each of these graphs, a list of selected individuals in the composite object, which allows the fast retrieval of all the individuals involved;
- for each of these graphs, a list of attribute values describing each individual;
- for each of these graphs, a list of statistical indices the user may define (mean, median, skewness, number of individuals by bin, etc.); this part of the object is useful for getting a pertinent description using statistics, but it is not necessary to recall the object onto screen.

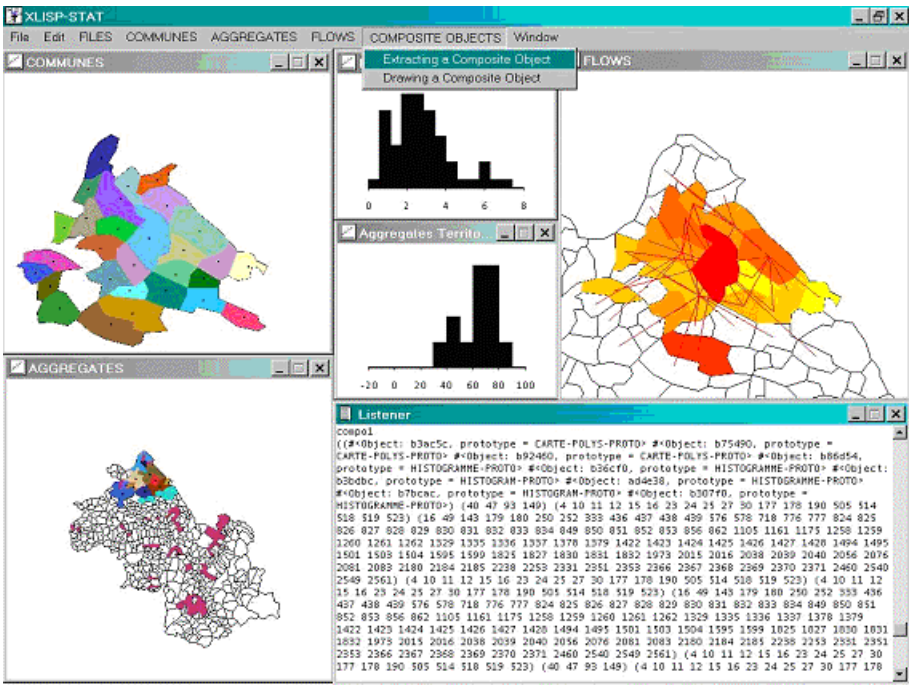

Figure 9. Compiling a geographical composite object within ARPEGE'.

As we saw earlier, a user can permanently and interactively tune lots of parameters and features of his/her handled data. Not only has ARPEGE' improved the data accessibility, but also the higher conceptual representation of them, due to its compilation of significant composite geographical objects. We believe spatial data becomes more usable in such a favourable environment.

4.4 Does ARPEGE' Improve the Spatial Data Analysis and Usability ?

What are the outcomes and the limitations of ARPEGE' ?

First there are some limitations. To be interactive, ARPEGE' needs to be fast. Even the increase in computer memory and power, there still remains to be developed methods for optimising the selection and application of the required (geo)statistics. The semiologic requirements are not yet very well supported. But the most important limitation is in the world's geographical complexity. In front of the user stands a complicated system made with objects and relations, offering so many trails to explore that a guide is necessary.

The important principle is to identify the relevant questions and the pertinent data required, and to draw up a plan to drive the exploration. Experts or users must have a prepared strategy for exploring, with questions to be asked during the process, hypotheses to be evaluated, and paths to be avoided, while keeping a large degree of freedom. Nevertheless, the user can set aside too rigid constraints due to assumptions in confirmatory statistics and allow a place for intuition within a step-by-step exploration. That's an important part of the research to be investigated: finding higher-level methods to help the user during his investigation. Another track to be investigated involves coupling exploratory and confirmatory analysis.

Is the tool used by the applicants? Normally, ARPEGE' would have been used in many French agricultural regions. But unfortunately, it has only been used in two regions. The reasons are quite paradoxical. Although the software provided appears quite attractive to the specialists, the head managers preferred to use the more tested classical methods, such as spatial clustering. Another constraint is the fact that the methods have to be equivalent and comparable within all the regions: ARPEGE' introduces more responsibility for the local specialist who would have the opportunity to shape the partitions according to his/her own subjective knowledge and use. For instance modifying the accuracy and global distribution of the aggregates, the shape of the partition and the statistical distribution of quality. A final reason is probably the loss of power from the national head managers if the local offices were able to make their own spatial partitions. As you can see, it is possible to improve the tools for data usability, without necessarily promoting their effective use!

However, this specific application was not representative of any spatial decision process. We still believe ARPEGE' remains an interesting way of managing and enhancing data usability. The advantages of ARPEGE' are numerous: robustness, association of several spatial models: integration of GIS and statistical functionalities; interactive and exploratory environment; permanent measurements; the capacity to focus on linked objects; an easy way to manage the

spatial partition quality and aspect; the possibility of extracting composite geographical objects, keeping in touch with the individuals. Some experiments have to be done in the future to compare the reliability of results in the decision making-process using either exploratory approach or confirmatory and automated methods (or both), and to assess the capacity of ARPEGE' to improve spatial data usability in different practical situations and applications.

5. CONCLUSION

On two occasions, we used conceptual modelling to explain a relevant way to improve spatial data usability. These occasions were not directly connected, even if as we showed, the second one (the ARPEGE' environment) refers to many keywords and ideas from the first one (because of general framework for the concept of usability). Therefore, it seems that such a conceptual environment is very useful when thinking about spatial data usability, because it involves spatial objects (which may be conceptual objects) and relationships. How can ESDA improve the usability of spatial data? According to the concrete examples we detailed using ARPEGE', we can lay the foundations of an exploratory approach that we think improves spatial data usability.

Here are our arguments:

1 – The robustness of the statistical methods allows all the data to be kept, including outliers: the user can keep in touch with his/her data that keep their identity;
2 – Objects and models are dynamically associated in the same environment: this highlights the problems tackled in all its forms/faces and provides critical points of view on the subject;
3 – One can extract and work on subsets of individuals: the global trend and some local groups may be enhanced, as well as their behaviour; local and global analysis can both be done;
4 – Dynamic links brings the expertise phase and the decision phase closer together: this makes the decision process more efficient;
5 – Managing objects and their relations through composite geographical objects: this opens up the world's complexity: data passes from the status of feature to the status of concept, which is more suitable for a 'fit to use' point of view;
6 – Due to composite objects and a 'historical' learning process (i.e. storage of 'memory check points' during the learning process) and the capacity to change the objects' relational structure, the user works in a dynamic virtual simulator: this may be useful for scenarios, planning and shaping the geographical space with usable spatial data.

6. ACKNOWLEDGEMENTS

I would like to thank a lot the reviewers and the Codata Journal and Maria for their help in improving this paper (especially the English).

7. REFERENCES

Andrews, F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972) *Robust Estimates of Location*, Princeton, NJ: Princeton University Press.

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.

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.

Behren, J. T. (1997) Principles and Procedures of Exploratory Data Analysis. *American Psychological Association*, *2*(2), 131-160.

Bolot, J., Chatonnay, P., & Josselin, D. (1999) Construction and evaluation of spatial partitions to describe geographical flows. *The International Symposium on Spatial Data Quality*, Hong-Kong.

Booch, G., Jacobson, I. & Rumbaugh, J. (1999) *The Unified Modeling Language User Guide*, Boston: Addison-Wesley.

Cleveland, W. S. (1993) *Visualizing data*, USA, New Jersey: ATT Bell Laboratories.

Dodge, Y. (Ed.) (1987) *Statistical Data Analysis based on the L1 Norm and Related Methods*, Y. Amsterdam: Elsevier Science Publishers B.V.

Fotheringham, S., (1997) Trends in quantitative methods: stressing the local. *Progress in Human Geography*, *21*(1), 88-96.

Fotheringham, S., Brunsdon, C. & Charlton, M. (2000) *Quantitative Geography, Perspectives on Spatial Data Analysis*, London Sage Publications.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986) *Robust Statistics: The approach based on influence functions*, New York: Wiley.

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.

Hoaglin, D., Mosteller, F. & Tukey, J.W. (1983) *Understanding robust and exploratory data analysis*, Series in probability and mathematical statistics, New-York: Wiley.

Huber, P. (1981) *Robust Statistics*, New York: Wiley.

ISO (1998) *ISO 9241-11:1998. Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*. Geneva: International Organization for Standardization (ISO)

ISO (2001) *SO/IEC 9126-1:2001 Software engineering -- Product quality -- Part 1: Quality model.* Geneva: International Organization for Standardization (ISO)

Josselin, D. (1995) *La déprise agricole en zone de montagne : vers un outil de modélisation spatiale couplant systèmes d'Induction et d'Information Géographiques*, Thèse de Géographie, Université Joseph Fourier, Grenoble, France.

Josselin, D. (Ed.) (1999a) *Les flux dans l'espace géographique*, N° Spécial de la *Revue de Géographie de l'Est*, XXXIX (4), Nancy, France: Association des Géographes de l'Est.

Josselin, D. (1999b) A la recherche d'objets géographiques composites avec le prototype ARPEGE'. *Revue Internationale de Géomatique, 9*(4), 489-505.

Josselin, D. (2000) Interoperating With GIS And Statistical Environment For Interactive Spatial Data Mining. *6th EC-GI&GIS Workshop, The spatial Information Society: Shaping the Future*, Lyon, France (CDROM). Retrieved February 13, 2003, from the European Commisssion GI-GIS website: http://wwwlmu.jrc.it/Workshops/6ec-gis/

Josselin, D. & Bolot, J. (2000) A semi-automatic method to build spatial partitions. *Proceedings of Geocomputation 2000*, Chattham, UK (CDROM), Retrieved February 13, 2003, from the Geocomputation 2000 website: http://www.geocomputation.org/2000/GC059/Gc059.htm

Josselin, D. (2003) (to be published) L'analyse des discontinuités spatiales avec le Distogramme. Contexte théorique, présentation, évaluation. In Josselin, D., Fabrikant, S. (Eds), N° spécial *"cartographie animée et dynamique"*, *Revue Internationale de Géomatique*, Paris: Hermès.

Kraak, M. J. & Ormeling, F.J. (1996) *Cartography. Visualization of Spatial Data*, Boston: Addison-Wesley.

Lecoutre, P. & Tassi, P. (1987) *Statistique non paramétrique et robustesse*, Paris: Economica.

MacEachren, A. M. (1995) *How Maps Work. Representation, Visualisation, Design*, New York: The Guilford Press.

MacEachren, A. M. & Kraak, M. J. (Eds) (2001) *Special Issue on Visualization, Cartography and Geographic Information Science, 28*(1).

Muller, P.A. & Gaertner, N. (2001) *Modélisation Objet avec UML*, Paris: Eyrolles.

Slocum, T.A. (1999) *Thematic Cartography and Visualization*, New Jersey: Prentice-Hall.

Tierney, L. (1990) *Lisp-Stat, an object oriented environment for statistical computing and dynamic graphics*, NewYork: Wiley, Interscience Publication.

Tukey, J.W. (1977) *Exploratory data Analysis*, Massasuchetts: Addison-Wesley.

Wackowicz, M., Riedermann, C., Vullings, W., Suárez, J. & Cromvoets, J. (2002) Workshop report on spatial data usability. *5th AGILE Conference on Geographical Information Science* (pp. 429-436), Palma, Spain.

Worboys, M. F. (1995) *GIS, A Computing Perspective*, London: Taylor&Francis.

Zeitouni, K. (Ed.), (1999) *Data mining spatial*. 9(4), *Revue internationale de Géomatique*, Paris: Hermès.