A DATA ORIENTED APPROACH TO GIS PROJECT

João MATOS and João BENTO

This paper presents an approach to GIS application projects guided by data management and specification issues. The approach is based on the assumption that a GIS has to incorporate data of different characteristics and different importance in what relates to the final objectives. These differences are associated with the nature, role and importance of each data type for both analysis procedures and maintenance of consistency.

The following taxonomy for data groups is proposed: structural, contextual, inventory, support and derived. This classification clarifies the relevance of each type of data for some of the important issues of GIS design, such as data accuracy, level of detail and updating constraints.

The quantification of appropriate levels of detail for anticipated specific uses is considered an issue of utmost importance. A contribution towards addressing this problem is put forward both from an empirical perspective and by resorting to more formal instruments such as information theory.

KEYWORDS: Project Specifications, Detail, Information theory, Data consistency.

1 INTRODUCTION

A major difficulty in the design of geographic information systems is the definition of the characteristics of their constituent data. Those characteristics must be defined taking in consideration the following aspects:

- adaptability of each theme in order to solve unavoidable inconsistencies, occurring both in updating procedures and in overlay with other themes.

- adequacy as input data, in order to enable the required operations;

This paper addresses both aspects and present an already tested approach to to the first as well as some research guidelines to the latter.

It is generally accepted and frequently stated that, in GIS, data and procedures should fit the objectives. Behind this self-evident truth there are nevertheless major difficulties to derive data characteristics and procedures based on objectives; even the establishment of clear objectives in a useful way shows to be a quite difficult task.

Considering the importance of project options over datasets, it is remarkable the small amount of effort put in its study. In certain situations, data management and acquisition may consume such a high level of resources that one can question the usefulness and the profitability of the GIS (Pornod, 1994). The aim of the present paper is a contribution to *methodological working tools* oriented to the activity of building and using a GIS.

2 THE ROLE OF INDIVIDUAL DATASETS IN THE PROJECT

From an operational point of view, typical objectives of a GIS can be divided in the following classes:

- data storage;
- production of printed outputs;
- management tool based on screen visualisation / interaction;
- query support tool;
- spatial analysis and modelling support.

This first approach is not explicit enough to allow for a definition of constraints on each data set. Not all the datasets have the same importance in the overall GIS project. Conditions to apply to each dataset depend on its intended role. Five major groups of roles are described below:

- Context data;
- Structural data;
- Inventory data;
- Support data;
- Derived data.

Context data is used only with descriptive purposes, where the detail and accuracy doesn't affect the results or the consistency of the overall datasets.

Structural data is the one that is not directly an object of calculations or spatial analysis but that assure the consistency of all the datasets and can also be used as an alphanumerical data aggregation base and as a base for the definition of new boundaries.

Inventory data is the class associated with objects that should be described exhaustively, without generalisation by omission.

Support data is used to perform calculations and spatial analysis and has direct influence in the results. The characteristics of support data are subject of option, balanced according to the intended characteristics of the resulting data.

Derived data results from analysis or statements and is not directly a subject of option. The characteristics of derived data are defined at support data and procedure levels, or they are just stated that way without any possibility modification.

Should a dataset be classified in more than one group, the most demanding classification had to be chosen. In Figure 1 is presented a decision tree to make the classification.

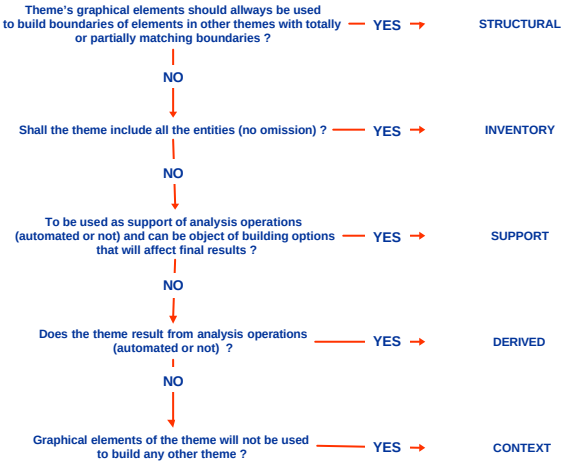

Figure 1- Decision tree to establish functional data classification.

Conditions over datasets, according to the classification above, are particularly important in the definition and usage of the structural data and in the support data. The other data roles does not present any special difficulty related with usage procedures or project options, so the focus will go to structural and support data.

3 USING STRUCTURAL DATA TO PRESERVE CONSISTENCY IN A MULTIPLE SOURCE DATA CONTEXT

Preservation of consistency in the geographical dataset, the task of structural data, is mainly related to the solution of the *4-set classification* problem. The *4-Set classification* problem is stated as follows:

Given two data sets, A and B, establish four sets using the following membership criteria:

Set 1 - elements of A not existing on B;

Set 2 - elements of B not existing on A;

Set 3 - elements of A existing on both A and B;

Set 4 - elements of B existing on both A and B.

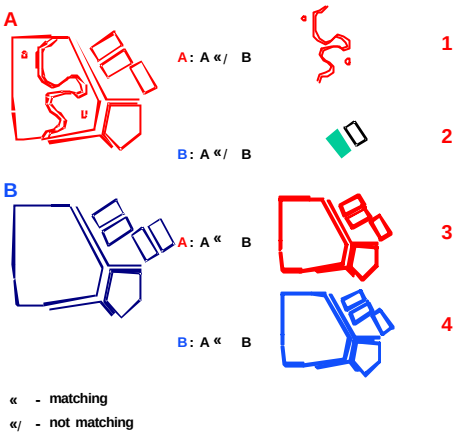

Figure 2 – 4-Set classification.

When overlaying A and B, if A is *structural* and B is not, all the elements from Set 4 should be replaced by the elements of Set 3. The topological relationships between elements of Set 2 and Set 4 should then be rebuilt, using elements of Set 3 instead of elements of Set 4.

Figure 3 – Regions adjusted to road axis used as structural data.

In Figure 3 a situation of application of structural data is illustrated, by opposition to a situation (Figure 4) where inconsistencies occur due to the inexistence of support data.

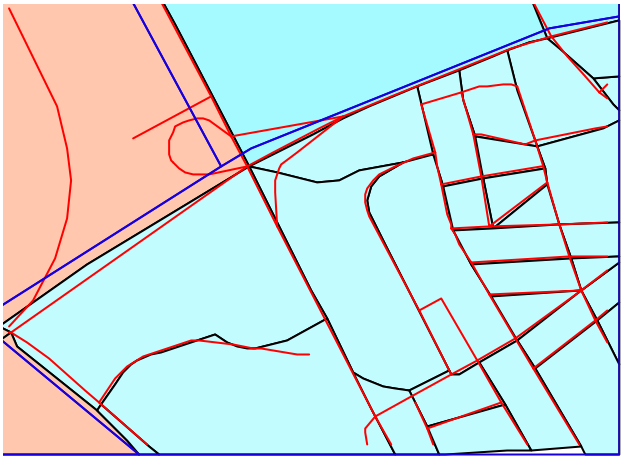

Figure 4 – Themes overlayed without any structural data constraint.

4 OPTIONS ON SUPPORT DATA

4.1 THE SET OF OPTIONS

The project options concerning support data are essentially related with the fitness of data accuracy and completeness to the desired results. Options made on these aspects are not independent of economical constraints, thus the need for optimization of the balance between cost and quality.

The choices can be made based upon empirical experience, from previous similar works, or by testing in pilot areas. Although testing is a common practice, the results might be valid only for that particular case and not for a general case. Even with testing and the use of empirical knowledge, it is sometimes difficult to support some of the options with objective reasoning, mainly when combining data of different natures (e.g., soil types, slopes, temperatures and water lines).

The practice of testing, including error propagation analysis, is useful but does not eliminate the need for information quantifiers. Those quantifiers should provide a vocabulary to describe the complexity of the data sets and its sensitivity to simplification operations and consequent results. They should allow for the comparison of characteristics between datasets of different nature and should take in consideration the dependency of modeled reality.

According to Couclelis (1996) the options available in the design of a geographic model may consider both the nature of the involved entities and the mode of observation. According to this author, one may consider the following two-valued characteristics with reference to the nature of the entities: atomic *vs. plenum*, homogeneous *vs.* heterogeneous, continuous *vs.* discontinuous, contiguous *vs.* sparse, solid *vs.* fluid, bi-dimensional *vs.* tri-dimensional, updated *vs.* outdated, permanent *vs.* variable, fixed *vs.* mobile and conventional *vs.* self-defined.

In what concerns the mode of observation, the same author proposes the following aspects: "scale", resolution, perspective, time, error and theory.

Considering that "perspective" and "theory" are controllable aspects and that changes in time are somehow quantifiable, three aspects remain: error, resolution and scale. Khun (1991) mention two dominating lines of thought on the problem of scale and resolution: "*the "pragmatists" that understand resolution in terms of map scale, acknowledging the limits of this concept; the "objectivists" look for geographic scale or dimensions in the real world*". The idea pursuited in this paper is to describe as objectively as possible the pragmatic knowledge.

The *pragmatic* way of thinking in terms of map scale present several problems at its own origin. In the first place, the concept just applies to themes that were traditionally represented in maps, and therefore they already have a generalisation frame of reference.

Another important remark is that, in geographic information systems, the quantity of information can be specified by theme and not for the whole that is being described. Such aspect is different from what happens in printed cartography, where the representation of objects from different themes could generate graphical conflicts. The solution for these conflicts requires the simultaneous generalisation of different themes.

The highly correlated concepts - *scale*, *error* and *resolution* - should now be analysed, as providing the main project constraints applicable to the support data. The term *detail* will be used herein after, aggregating these three aspects in a concept of data quantity and likelihood to the represented reality. The main problem in quantifying detail is its dependency of the reality to be represented (e.g. a DTM of a plain area has less complexity and its resolution can be decreased without significant increase of error). A detail index can only be built by comparison of different datasets of the same area (Fairbairn, 1998). In the following sections some of such topics are put forward, despite of requiring further consideration since they are not yet supported by experimental evidence.

4.2 TYPES OF OBJECTS

The definition of a unique type of detail measurement seems to be impossible; therefore, it must be defined according to objects nature. Considering only the bidimensional domain and a single epoch, detail measurements should be different for three different types of representation (Matos, 1998):

- objects that by their nature present well defined vertices (Figure 5);

- objects where vertices appear as a simplification of a smooth or fuzzy shape (Figure 5);
- object independent partitioning of the space (usually tesselations).

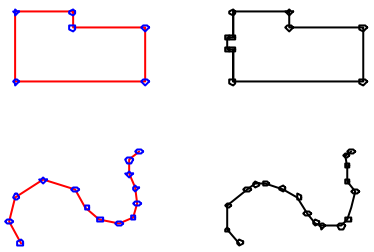

Figure 5 – Different object types and graphical descriptions.

In what concerns error, one has to distinguish between those errors associated to the definition of objects, i.e. those determined by the direct evaluation of coordinates (e.g. buildings) and those involving computations (e.g. isolines, slopes, etc.).

For all the three types of objects is necessary to define, even if arbitrarily, a reference for the highest detail level of cartographic interest. Without this assumption it is very difficult, if not impossible, to develop any further the subject. Usually the minimum granule of resolution (atomic resolution) is known empirically for the various cartographic representations. The problem consists, therefore, in comparing a dataset to its correspondent reference dataset.

4.3 Form of Reference

Objects of the first type are usually related with the type of representation used in the traditional large-scale maps. Importing this "pragmatic" concept to the "objectivist" field might be made taking as a reference the so-called 1/1000 cartography, whose associated error may not be smaller than 10 cm. Such hypothesis is reinforced by the fact that representations with a higher level of detail are applicable and useful only to engineering or architectural works of reduced spatial distribution.

The concept of *form of reference* can now be introduced. An object is said to be in its form of reference if, for a given value of maximum precision $\boldsymbol{e}$ ($\boldsymbol{e}$=10cm, as suggested above), any measurement of coordinates not belonging to the points used in the description of the form, is contained within a band of radius $\rho = \boldsymbol{ke}$ defined around the representation (Figure 6). The multiplying factor $\boldsymbol{k}$ corresponds to an amplification of the error band, to introduce the corresponding level of confidence.

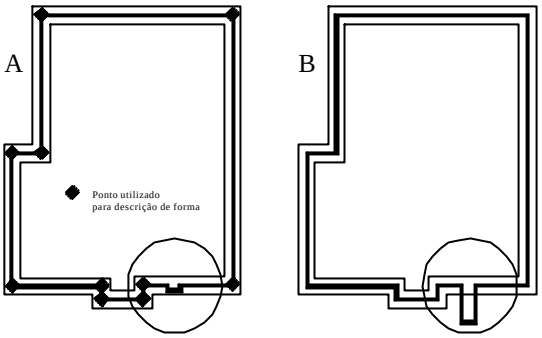

Figure 6– Object in its reference form (A) and the opposite situation (B), for a given e .

The form of reference concept reflects the existence of representations with no other simplification than that derived from the precision of the measurement instruments. The representation of objects in the 1/1000 cartography may be considered very close to this paradigm, hence its choice as a reference. A substitute for the traditional scale definition for detail might now be built using the percentage of line length *outside* the reference form band. Because only detail is supposed to be measured this way, any global positional error should be neglected.

The major inconvenient in this type of measurements consists of an equal result for *omission* or for *collapse* (e.g. changing from polygon to point). In this latter situation, not only the reference form concept should not be applied, but also the problem becomes more adequate to be solved using the concept below.

4.4 Quantity Variation

A simple coefficient (G), based on quantity variations under simplification, may be given for each dataset by:

$$G = \frac{total\ quantity - correctly\ represented\ quantity}{total\ quantity}$$

In the equation above, the referred quantity may be the area, the length or simply the number of entities. A *correctly represented quantity* is one not affected by omission or commission errors, relative to the *total quantity* in a reference map. This is, obviously an indicative parameter for its rigorous determination would require the existence of a reference cartography for all themes covering all the area. This indicator applies mainly to datasets of the second type.

Such an indicator brings in the following advantages:

- it may reflect the completeness of the dataset under analysis, i.e. the number of objects of a given theme, and its sensitivity to simplification operations;

- it can be calibrated, through known parameters, for each theme;
- it is applicable for all types of objects: points, lines or areas (with boundaries or continuous variation).

On the other hand, there are also associated disadvantages:

- it is not a transitive concept, i.e. one may not infer the relation between two different "scales" through the relation between them and a third one;
- its application requires that a reference dataset exists (at least in a sample area).

The coefficient is presented as a generic principle; its application strategy should fit individual theme characteristics, as exemplified in the examples of Table 1.

THEME	USABLE QUANTITIES
Road axis	Total line length Number of arcs in a graph structure
Buildings	Built area Number of polygons Built perimeter
Height (DTM)	Areas classified according to height intervals

Table 1 – Generalisation coefficient examples.

The example of DTM detail is especially important, as the DTM is a major example of data from which new themes are derived (e.g. erosion, water flow, visibility). This stands also for other themes represented as surfaces, where the complexity of variation and the minimal measurement unit will define the reference map.

These quantity variation parameters inform the project decisions about acceptable simplifications or necessary detail.

4.5 INFORMATION MEASURES

An interesting approach can be made resorting to an analogie with *Shannon information theory* (Shannon, 1949; Klir, 1988). Applications of *Shannon information theory* to cartography have already been studied (Bjorke, 1997), focusing on the communication characteristics in a "printed map" paradigm.

Shannon's entropy is defined by:

$$E(P) = -\sum_{i=1}^{n} p_i . \log_2 p_i \; ; P = (p_1, p_2, \cdots, p_n),$$

where P is a set of discrete probabilities, such that

$$\sum_{i=1}^{n} p_i = 1 \quad ; p_i \geq 0, i = 1, n$$

A weighted Shannon entropy can also be defined by:

$$WE(P) = -\sum_{i=1}^{n} w_i p_i . \log_2 p_i \quad .$$

This type of approach seems particularly interesting to objects of the third type (tesselations), providing a quantifier to measure the sensitivity to simplifications in resolution. The application analogy here presented is based on an image analysis application to soil characterisation(Ghalib, 1998).

Parameters such as *energy* (*E*), *local homogeneity* (*LH*) and *entropy* (*ENT*) can be defined by:

$$E = -\sum_i \sum_j [P(i, j \mid d)]^2 ;$$

$$LH = -\sum_i \sum_j \frac{1}{1+(i-j)^2} P(i, j \mid d) ;$$

$$ENT = -\sum_i \sum_j P(i, j \mid d) . \log_2 (P(i, j \mid d)) ;$$

where, for a given matrix I, :

$$P(i, j \mid d) = \frac{1}{R} \quad \{ \; f[(k,l),(m,n)] : k - m = 0, |l - n| = d + f[(k,l),(m,n)] : |k - m| = d, l - n = 0 + f[(k,l),(m,n)] : |k - m| = d, |l - n| = d \; \}$$

where $I(k,l) = i$, $I(m,n) = j$; $0 \leq k, m \leq colmax$, $0 \leq l, n \leq rowmax$. R is the total number of occurrences.

The probability $P(i,j \mid d)$ corresponds to the probability of existence of a cell valued j at a "distance" d from a cell valued i. In the application to image analysis the cells can be valued from 0 to 255, corresponding thus this process to fill a 256x256 matrix.

If applying to a DTM tesselation, instead of the 0-255 classification height intervals can be used and d can be a value related with the dimension of terrain features (e.g. one could consider d = 100m).

A smooth terrain will result in a matrix with higher values along the main diagonal. A flat surface will correspond to zero entropy. The value of the parameters provides a quantification of data complexity and differences between different DTM of

the same region can provide a detail level quantifier.

A vector data analogy can be easily built partitioning the space with a tesselation and associating to each cell the number of vertices it contains.

To measure line complexity one can use the absolute value of the angle between two consecutive line segments (subtracting 180°), rounded to the degree. Probabilities will be calculated for each value of the domain
{ 0, 1,..., 180°}. A sequence of segments in a straight line will correspond to zero entropy. Typically, isolines will correspond to low entropy values and lines in an urban area will produce high entropy values.

5 CLOSING REMARKS

The concepts presented in sections 2 and 3 have already been applied with success to several GIS projects within the DECART group, at IST and reference to some those application can be found elsewhere (Matos, 1997). The exploratory ideas about *support data* are presently being object of experimentation with real data and, at their present form, they should be considered as a discussion subject.

6 REFERENCES

Bjorke,J.T. *Map Generalisation. An Information Theoretic Approach to Feature Elimination. In Proceedings 18th ICA International Cartographic Conference, pp. 178-185. Ed. Swedish Cartographic Society, 1997.*

Couclelis,H. *Towards an Operational Typology of Geographic Entities with Ill-defined Boundaries.* In Geographic Objects with Indeterminate Boundaries, GISDATA 2, pp. 45-56. Ed. Taylor & Francis 1996.

Fairbairn, D *Determining and Using Graphic Complexity as a Cartographic Metric. In Proceedings 18th ICA International Cartographic Conference, pp. 480-486. Ed. Swedish Cartographic Society, 1997.*

Ghalib,A.;Hryciw,R.;Shin,S. *Image Texture Analysis and Neural Network for the Characterisation of Uniform Soils, 1997.*

Khun,W. *Are Displays Maps or Views ? In Proceedings AUTOCARTO 10, pp. 261-274.Ed. ACSM-ASPRS, 1991.*

Klir,G .;Folger,T. *Fuzzy Sets, Uncertainty and Information. Prentice Hall, New York, 1988.*

Matos,J.; **Costa**,**J.R.** *Especificações Técnicas para a Cartografia dos Planos de Bacia.* Relatório para Instituto da Água. Lisboa, 1997.

Matos,J.;Gonçalves,A. *Positional Error Measurement and Analysis. In Proceedings of the 8th International Symposium on Spatial Data Handling,pp. 151-160. Ed. IGU, 1998.*

Pornon,H. *Analyse Critique des Methodes de Conduite de Projet de Systemes d'Information Geographique. In Proceedings EGIS 94, pp. 1298-1304.Ed. EGIS Foundation, 1994.*

Shanon, C.E. and **Weaver, W.** *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

João MATOS
jmatos@civil.ist.utl.pt

João Matos is a lecturer at Instituto Superior Técnico, Lisbon. He is also secretary of the Technical Committee 134 on GI Standards.

His main research interests are in the fields of quality analysis and technical specifications on GI data.

Instituto Superior Técnico
DECivil
Av. Rovisco Pais
1096 LISBOA CODEX
PORTUGAL
Tel: +351-1-8418351
Fax: +351-1-8497650
URL: http://www.civil.ist.utl.pt/decart

João BENTO
joao@civil.ist.utl.pt

João Bento is an Associate Professor at Instituto Superior Técnico, Lisbon. Among other responsabilities he is the coordinator of the MSc Course on GIS, and has been deeply involved in the implementation of SNIG and EXPO'98's "Portugal Digital".

His main research interests are in the fields of application of Artificial Intelligence to Civil Engineering and data distribution in GIS.

Instituto Superior Técnico
DECivil
Av. Rovisco Pais
1096 LISBOA CODEX
PORTUGAL
Tel: +351-1-8418208
Fax: +351-1-8497650
URL: http://www.civil.ist.utl.pt/joao