Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

# A qualitative spatial reasoning approach in knowledge discovery in spatial databases

Maribel Santos & Luís Amaral
*Information Systems Department,*
*University of Minho, Portugal.*

## Abstract

Spatial data mining refers to the extraction of implicit knowledge, spatial relationships, or others interesting patterns not explicitly stored in spatial databases. Spatial databases like geographic databases in geographic information systems usually store large amounts of spatial information. Queries in these systems involve the derivation of different spatial relations that are not explicitly stored. Qualitative spatial reasoning has been proposed as a complementary mechanism for the automatic derivation of these relations. Qualitative spatial relations are specified using of a small set of symbols, like *North, close, etc.*, and are manipulated through a set of inference rules. In this paper is presented a new approach to the process of knowledge discovery in spatial databases, where geographic identifiers give the positional aspects of geographic data. These identifiers are manipulated using qualitative reasoning principles, allowing the inference of new spatial relations required in the knowledge discovery process. The exploration, with this approach, of a demographic database allowed the discovery of implicit spatial relations that characterise the analysed geographic and demographic data. The overall system, including the spatial reasoning mechanisms and the discovery process, was implemented in *Clementine.* Between others facilities, the system allows the integration of the relevant discoveries in a database of patterns, manipulated by a geographic information system, enabling its visualisation in a map.

## 1 Introduction

Knowledge Discovery in Spatial Databases (KDSD) refers to the extraction of interesting spatial patterns and features, general relationships that exist between spatial and non-spatial data, and other data characteristics not explicitly stored in spatial databases.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

Spatial database systems are relational databases plus a concept of spatial location and spatial extension [1]. The explicit location and extension of objects define implicit relations of spatial neighbourhood. The neighbour attributes of a given object may influence its behaviour and therefore must be considered in the process of knowledge discovery. Knowledge discovery in relational databases doesn't take into consideration this spatial reasoning, motivating the development of new algorithms adapted to the spatial characteristics of spatial data.

The main approaches in KDSD are characterised by the development of new algorithms that treat the objects' position and extension through the manipulation of its co-ordinates [2][3][4][5][6]. These algorithms are subsequently implemented, extending traditional knowledge discovery systems. In all, a quantitative spatial reasoning approach is used, although the results are presented using qualitative identifiers.

This paper presents a new approach to the process of KDSD based on qualitative spatial reasoning. Since the use of co-ordinates, for the identification of a spatial object's location and extension is not always needed, we investigated how traditional knowledge discovery systems (and their generic data mining algorithms), for relational databases, can be used in KDSD. Depending on the discovery task, geographic concept hierarchies and spatial knowledge are needed in order to ensure the inference of new spatial relations, not explicitly stored in the geographic database used.

The integration of a *geographic database*, with the administrative subdivisions of *Portugal* at the municipality and district level, and a *demographic database*, storing the parish registers of the *Aveiro* district, allowed the discovery of implicit relationships that exist between the analysed geographic and demographic data.

This paper is organised as follows: Section 2 explains how qualitative identifiers are used in the spatial referencing of geographic data. Section 3 describes the use of qualitative spatial reasoning in the inference of new spatial relations. Section 4 presents the geographic and demographic database integration, and the results of the knowledge discovery process. Section 5 concludes with some comments about the presented work.

## 2 The use of a geographic identifiers system

The positional aspects of geographic data are provided by a spatial reference, which relate the data to a given position on the Earth's surface. Spatial references fall into two categories: based on *co-ordinates* or on *geographic identifiers*. In systems of spatial referencing using geographic identifiers (Figure 1), a position is referenced to a real world location defined by a real world object. This object is termed a *location*, and its identifier is termed a *geographic identifier* [7]. These systems shall comprise a structured collection of location classes, location instances and their corresponding geographic identifiers.

The geographic database used in this study was constructed under the principles established by the European Committee for Normalisation in the CEN TC 287 standard for Geographic Information [8]. In this database, the positional aspects of geographic data are obtained using a *geographic identifiers system*.

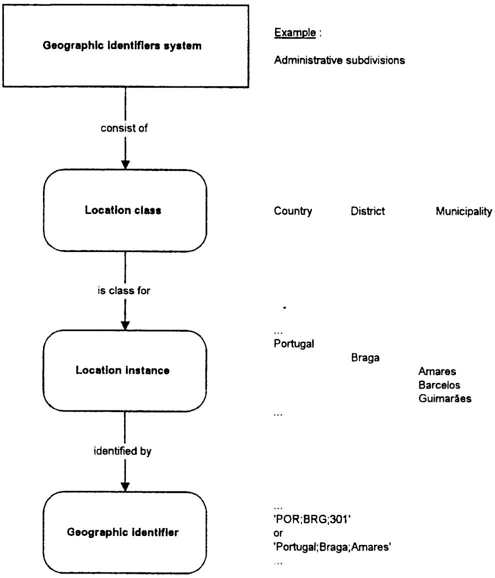


Figure 1: A geographic identifiers system (adapted from [7] p. 6)

The *geographic identifiers system*, characterising the administrative subdivisions of *Portugal* at the municipality and district level and proceeding from the *geographic identifiers schema*, is integrated with the *spatial schema* [9], allowing the specification of the spatial relations that exist between the addressed geographic entities. Between others attributes, this integration permits the identification of the direction, distance and topological spatial relations, existing between the considered municipalities.

The *spatial schema* only allows the specification of spatial relations for adjacent regions. All the others relations, existing between non-adjacent regions and needed in the knowledge discovery process, must be inferred. The rules required and the inference processes are explained in the next section. More details about this geographic database can be found in Santos et al. [10].

## 3 The inference of new spatial relations

Spatial reasoning is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference, and used to arrive to valid conclusions regarding the objects' relationships [11].

Spatial relations have been classified in several types [12][13], including *direction relations* [14] (that describe order in space), *distance relations* [15] (that describe proximity in space) and *topological relations* [16] (that describe neighbourhood and incidence). The *representation* of these relations, using qualitative identifiers, requires the adoption of a small set of symbols, like *North, South, close, far, disjoint* or *meet.*

The inference of new spatial relations, existing between the addressed geographic entities and implicit in the used geographic database, can be achieved through the construction of qualitative rules (compiled in a composition table).

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

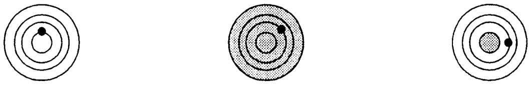
These rules allow the manipulation of the adopted qualitative identifiers. For example, knowing the facts, **A North, very far from B** and **B Northeast, very close from C**, it is possible, consulting the composition table for integrated direction and distance spatial reasoning [17], the inference of the relationship that exist between **A** and **C**: **A North, very far from C**.

Qualitative rules can be constructed using quantitative methods [17] or manipulating qualitatively the set of identifiers adopted [12] (through the definition of axioms and properties for the spatial domain).

The adoption of a mixed approach [18] allowed the integration of direction, distance and topological relations, under the principles of qualitative spatial reasoning. Quantitative and qualitative methods enabled the construction of the composition table used in this work, in the inference of new spatial relations.

The adopted qualitative identifiers were: **N**, **NE**, **E**, **SE**, **S**, **SW**, **W** and **NW** for direction relations; **very close**, **close**, **far** and **very far** for distance relations; and **disjoint** and **meet** for topological relations. With respect to topology, these are the only two topological relations that can exist in the geographic domain characterised (as they are administrative regions they can not be overlap).

The composition table constructed is represented using graphic symbols like the presented in Figure 2. In order to exemplify the explicit knowledge stored in the final composition table, Table 1 exhibits an excerpt of the rules contained in it.



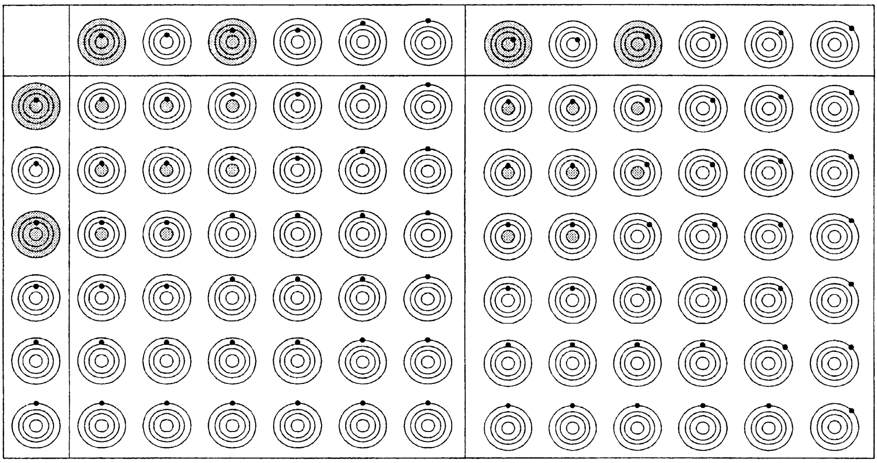
*North, very close, disjoint*

*Northeast, close, meet*

*East, close, disjoint* or, *East, close,meet*

Figure 2: Icons representing direction, distance and topological relations

Table 1: Excerpt of the composition table



Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

The set of facts stored in the final composition table is subsequently assimilated by machine learning algorithms. In *Clementine* [19], the knowledge discovery tool used, a rule induction technique was applied, allowing the construction of a decision tree that assimilated the inference rules explicit in the composition table. This decision tree is used in a process that cyclically infers unknown spatial relations needed in the knowledge discovery process. More details about the learning and inference processes can be found in Santos and Amaral [20].

# 4 The knowledge discovery process using qualitative spatial reasoning

The constructed system, for KDSD using a qualitative spatial reasoning approach, has three main components: data repositories, data analysis and results visualisation [21]. The *data repositories component* aggregate three main databases: the geographic database, the spatial reasoning database and a non-geographic database. In this paper, the non-geographic database is represented by a demographic database that collects the parish registers, dated between 1690 and 1990, in the *Aveiro* district. The three databases were implemented in a relational database system, being available to the others components through an OBDC (Open Database Connectivity) connection.

The *data analysis component* is implemented through the knowledge discovery module and is constituted by the five steps: data selection, data treatment (pre-processing), geographic information processing, data mining and interpretation of the discovery patterns. This module was completely implemented in *Clementine*. In this application, the knowledge discovery process is defined through the construction of *streams*, in which a node represents each operation on data.

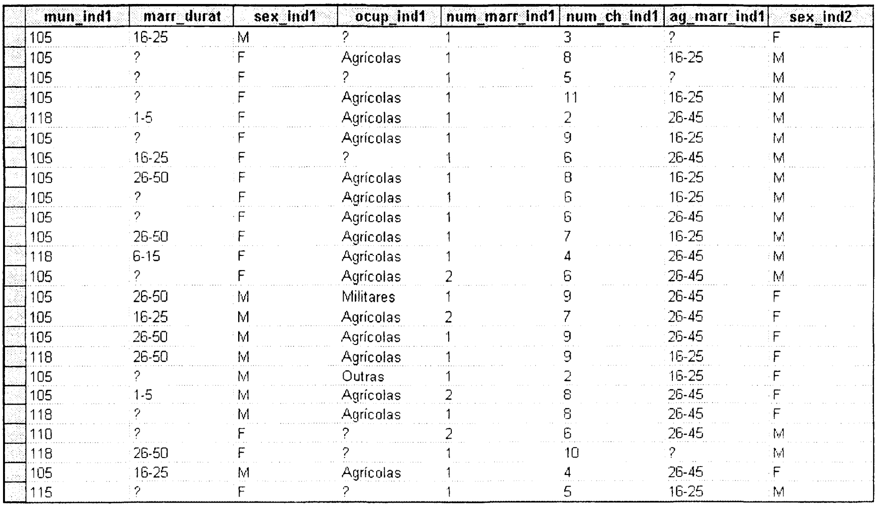
The *results visualisation component* is the responsible for the management of the database that stores the relevant patterns/discoveries. The database of patterns was integrated in a geographic information system (*Geomedia Professional* [22]), allowing the visualisation of the discovered patterns in a map.

In order to exemplify the results that can be achieved with the proposed approach, the demographic database was explored and the obtained results discussed. As already mentioned, the demographic database stores the parish registers collected in the *Aveiro* district. This database stores attributes like *birth date*, *birthplace*, *death date*, *death place*, *occupation*, *number of descendants*, *number of marriages*, etc. The integration of the geographic and the demographic databases occurred at the municipality level.

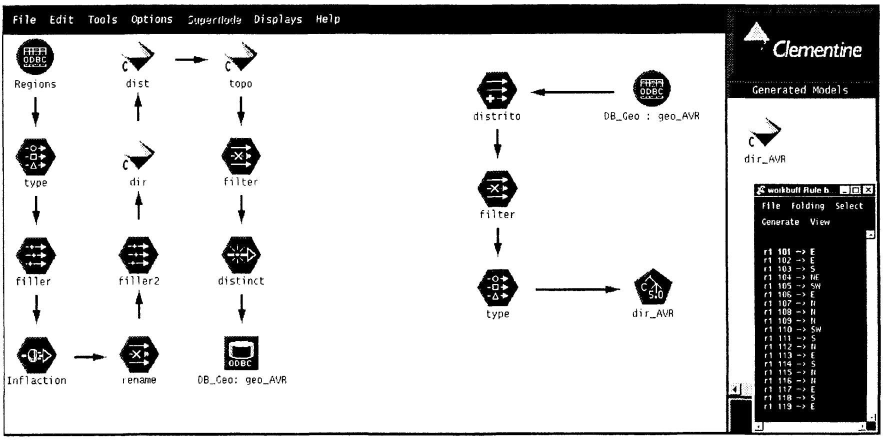
The exercise carried out in *Clementine* had as its purpose characterises the ***age at marriage*** (first marriage) of the individuals, attending to their ***sex***, the ***century*** and ***municipality*** in which they lived, and the ***number of marriages*** verified by each one. Table 2 presents the necessary attributes, some of them already transformed into discreet values or generalised attending to their hierarchies.

After the data selection and data treatment phases, the geographic information-processing step allowed the creation of a model in which the municipalities of the analysed district, *Aveiro*, were classified in terms of their directions. For this classification, two tasks where implemented. Firstly, the explicit relations, existing for this district in the geographic database, where selected. A cyclically process used these relations in the inference of all unknown spatial relations. The inferred relations were saved in a table (of the geographic database), in order to be used, second task, in the construction of the model (**dir_AVR**) that distributes geographically all the municipalities. The two streams constructed in *Clementine*, for the implementation of these two tasks, are presented in Figure 3. The stream at the left side represents the first step, while the stream located at the right side allowed the construction of the **dir_AVR** model (browsed in the tabular table also at the right side).

Table 2: Some attributes of the *marriage* table



| mun_ind1 | marr_durat | sex_ind1 | ocup_ind1 | num_marr_ind1 | num_ch_ind1 | ag_marr_ind1 | sex_ind2 |
|---|---|---|---|---|---|---|---|
| 105 | 16-25 | M | ? | 1 | 3 | ? | F |
| 105 | ? | F | Agrícolas | 1 | 8 | 16-25 | M |
| 105 | ? | F | ? | 1 | 5 | ? | M |
| 105 | ? | F | Agrícolas | 1 | 11 | 16-25 | M |
| 118 | 1-5 | F | Agrícolas | 1 | 2 | 26-45 | M |
| 105 | ? | F | Agrícolas | 1 | 9 | 16-25 | M |
| 105 | 16-25 | F | ? | 1 | 6 | 26-45 | M |
| 105 | 26-50 | F | Agrícolas | 1 | 8 | 16-25 | M |
| 105 | ? | F | Agrícolas | 1 | 6 | 16-25 | M |
| 105 | ? | F | Agrícolas | 1 | 6 | 26-45 | M |
| 105 | 26-50 | F | Agrícolas | 1 | 7 | 16-25 | M |
| 118 | 6-15 | F | Agrícolas | 1 | 4 | 26-45 | M |
| 105 | ? | F | Agrícolas | 2 | 6 | 26-45 | M |
| 105 | 26-50 | M | Militares | 1 | 9 | 26-45 | F |
| 105 | 16-25 | M | Agrícolas | 2 | 7 | 26-45 | F |
| 105 | 26-50 | M | Agrícolas | 1 | 9 | 26-45 | F |
| 118 | 26-50 | M | Agrícolas | 1 | 9 | 16-25 | F |
| 105 | ? | M | Outras | 1 | 2 | 16-25 | F |
| 105 | 1-5 | M | Agrícolas | 2 | 8 | 26-45 | F |
| 118 | ? | M | Agrícolas | 1 | 8 | 26-45 | F |
| 110 | ? | F | ? | 2 | 6 | 26-45 | M |
| 118 | 26-50 | F | ? | 1 | 10 | ? | M |
| 105 | 16-25 | M | Agrícolas | 1 | 4 | 26-45 | F |
| 115 | ? | F | ? | 1 | 5 | 16-25 | M |



```
r1 101 -> E
r1 102 -> E
r1 103 -> S
r1 104 -> NE
r1 105 -> SW
r1 106 -> E
r1 107 -> N
r1 108 -> N
r1 109 -> N
r1 110 -> SW
r1 111 -> S
r1 112 -> N
r1 113 -> E
r1 114 -> S
r1 115 -> N
r1 116 -> N
r1 117 -> E
r1 118 -> S
r1 119 -> E
```

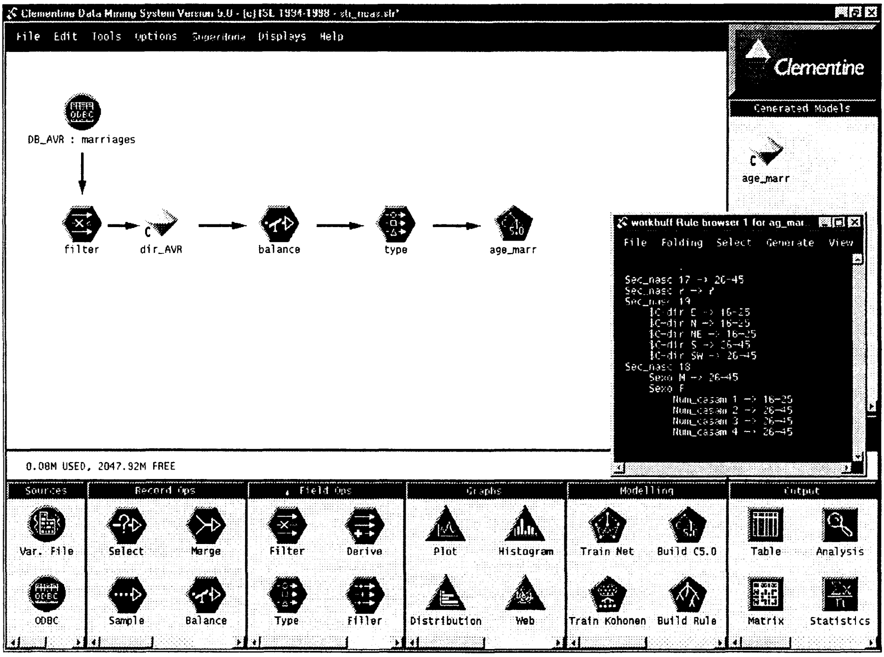
Figure 3: Geographic information processing

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

By the analysis of the Figure 3 can be noted that the stream on the left side used three models, **dir**, **dist** and **topo**. These models, previously generated, contain the qualitative rules for the inference of new spatial relations. *Clementine* allows the storage of the generated models and its subsequent re-use in other streams.

After the geographic information processing, the demographic and geographic data can be analysed as a whole. The constructed stream, for the discovery of the pattern that characterises the ***age at marriage*** attribute, is showed in Figure 4. This figure also presents the generated rules (tabulated results on the right side).

The results achieved point out that in the seventeen-century, the individuals presented an ***age at marriage*** classified by the **26-45** class (the first marriage occurred between they 26 and 45 years old). In the eighteen-century, the ***age of marriage*** for men were again classified at the **26-45** class, while that for women, the ***number of marriage*** influenced the results (curiously, the age at the first marriage increased for more than one marriage). Only in the nineteen-century appears a pattern with a geographical distribution well defined. All the municipalities located at North, Northeast and East presented an ***age at marriage*** classified in the **16-25** class. Regions at South or Southwest of the *Aveiro* district were classified in the **26-45** class. This geographic pattern may be justified by the fact that, most of the regions at South and Southwest have frontiers with the sea, indicating that due to the profession of the majority of the men, fishermen, they married later.



```
Sec_nasc 17 -> 26-45
Sec_nasc ? -> ?
Sec_nasc 19
    1C-dir E -> 16-25
    1C-dir N -> 16-25
    1C-dir NE -> 16-25
    1C-dir S -> 26-45
    1C-dir SW -> 26-45
Sec_nasc 18
    Sexo M -> 26-45
    Sexo F
        Num_casam 1 -> 16-25
        Num_casam 2 -> 26-45
        Num_casam 3 -> 26-45
        Num_casam 4 -> 26-45
```

0.08M USED, 2047.92M FREE

Figure 4: Stream *Clementine* for the classification of the ***age at marriage*** attribute

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

The discovered pattern can be stored in the database of patterns, *results visualisation component*, in order to be visualised in a map. The results are passed to this database through an OBDC connection. For that, the user only needs to load a *source node*, with the database structure. In this node, only is required the definition (name) of the attributes to be stored. This process is explained in detail in Santos and Amaral [21] (where also can be found a thematic map created for the visualisation of a given discovery).

Although the presented exercise only used directions relations, the others spatial relations, distance and topology, could be used if they were required for the pretended results. Since they inference rules were already assimilated, they inclusion in the knowledge discovery process can be effected like presented here for direction relations.

## 5 Conclusions

This paper presented an approach for knowledge discovery in spatial databases based on qualitative spatial reasoning. In this approach, the positional aspects of geographic data were provided by a spatial reference gave by a geographic identifier. The geographic database used was constructed under the principles established by the European standard for geographic information.

For the municipalities of Portugal, the geographic database stores the *direction*, *distance* and *topological* spatial relations existing between some of the referred regions. This knowledge and the composition table, with the qualitative rules for the inference of new spatial relations, enabled the deduction of implicit spatial relations needed in the knowledge discovery process.

The integration of this geographic database with a demographic one allowed the discovery of implicit relationships that exist between the geographic and demographic data analysed.

The obtained results validate the proposed approach. Traditional knowledge discovery systems, developed for relational databases and not having semantic knowledge linked to spatial data, can be used in the process of KDSD, if this semantic knowledge and the principles of qualitative spatial reasoning are available as domain knowledge.

## Acknowledgements

This work has been partially supported by a Portuguese grant from PRODEP II (*Acção 5.2 – Concurso n°3/98 Doutoramentos*).

## References

[1] Ester, M., Kriegel, H.-P. & Sander, J. Spatial Data Mining: A Database Approach. *Proceedings of the 5th International Symposium on Large Spatial Databases*, Springer-Verlag: Berlin, 1997.

[2] Ester, M., Kriegel, H.-P. & Xu, X. A Database Interface for Clustering in Large Spatial Databases. *Proceedings of the first International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp. 94-99, 1995.

[3] Ester, M., Frommelt, A., Kriegel, H.-P. & Sander, J. Algorithms for Characterization and Trend Detection in Spatial Databases. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1998.

[4] Lu, W., Han, J. & Ooi, B.C. Discovery of General Knowledge in Large Spatial Databases. *Proc. of the 1993 Far East Workshop on Geographic Information Systems*, pp. 275-289, 1993.

[5] Koperski, K. & Han, J. Discovery of Spatial Association Rules in Geographic Information Systems. *Proc. 4th International Symposium on Large Spatial Databases (SSD95)*, pp. 47-66, 1995.

[6] Koperski, K., Han, J. & Stefanovic, N. An Efficient Two-Step Method for Classification of Spatial Data. *Proceedings of the International Symposium on Spatial Data Handling (SDH'98)*, 1998.

[7] CEN/TC-287. *Geographic Information: Referencing, Geographic Identifiers*, European Committee for Normalisation, European pre-standard prENV 12661, 1998.

[8] CEN/TC-287. *Geographic Information: Data Description, Rules for application schemas*, European Committee for Normalisation, European pre-standard WI 006, 1998.

[9] CEN/TC-287. *Geographic Information: Data Description, Spatial Schema*, European Committee for Normalisation, European pre-standard prENV 12160, 1996.

[10] Santos, M., Amaral, L. & Pimenta, P. A Descoberta de Conhecimento em Bases de Dados Geográficas através da Explicitação Semântica *(in Portuguese). GISBrasil'99 - V Congress and Exhibition of Latin America Geo-processing Users*, 1999.

[11] Sharma, J. *Integrated Spatial Reasoning in Geographic Information Systems: Combining Topology and Direction*, PhD thesis, University of Maine, 1996.

[12] Frank, A.U. Qualitative Spatial Reasoning: cardinal directions as an example. *International Journal of Geographical Information Systems*, **10(3)**, pp. 269-290, 1996.

[13] Papadias, D. & Sellis, T. On the Qualitative Representation of Spatial Knowledge in 2D Space. *Very Large Databases Journal, Special Issue on Spatial Databases* , **3(4)**, pp. 479-516, 1994.

[14] Freksa, C. Using Orientation Information for Qualitative Spatial Reasoning. *Theories and Methods of Spatio-Temporal Reasoning in Geographic space, Lectures Notes in Computer Science 639* , eds. Frank, A.U., Campari, I. & Formentini, U., Springer-Verlag: Berlin, 1992.

[15] Hernández, D., Clementini, E. & Felice, P.D. Qualitative Distances. *Spatial Information Theory - A Theoretical Basis for GIS, Proceedings of the International Conference COSIT'95, Lectures Notes in Computer Science 988* , eds. Frank, A.U. & Kuhn, W., Springer-Verlag, pp. 45-57, 1995.

[16] Egenhofer, M.J. Deriving the Composition of Binary Topological Relations. *Journal of Visual Languages and Computing* , **5(2)**, pp. 133-149, 1994.

[17] Hong, J.-H. *Qualitative Distance and Direction Reasoning in Geographic Space*, PhD thesis, University of Maine, 1994.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

[18] Santos, M. *O Raciocínio Espacial Qualitativo na Descoberta de Conhecimento em Bases de Dados Geográficas (in Portuguese)*, PhD thesis (in finalisation), University of Minho, 2000.

[19] ISL. *Clementine, User Guide, Version 5.0* , Integral Solutions Limited, 1998.

[20] Santos, M. & Amaral, L. As Normas de Informação Geográfica e o Raciocínio Espacial Qualitativo na Inferência de Informação Geográfica Qualitativa *(in Portuguese). Proceedings of the V Geographic Information Systems Meeting* , 1999.

[21] Santos, M. & Amaral, L. Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning. *PADD'00 Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining* , 2000.

[22] Intergraph. *Geomedia Professional v3*, Reference Manual. Intergraph Corporation, 1999.