Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning

Maribel Santos　　*Luís Amaral*

Information Systems Department
University of Minho
Campus de Azurém
4800-058 Guimarães
Portugal
{maribel, amaral}@dsi.uminho.pt
http://www.dsi.uminho.pt

Abstract. Human beings use qualitative identifiers extensively to simplify reality and to perform spatial reasoning more efficiently. Organisational databases usually store geographic identifiers, like addresses or postcodes, which spatial component is not incorporated in the knowledge discovery process. This paper addresses the process of Knowledge Discovery in Spatial Databases through a Qualitative Spatial Reasoning approach. The aim is the improvement of the referred process by the adoption of qualitative identifiers like *North*, *South*, *close*, *far*, etc. in the classification of spatial relations that exists between the geographic entities addressed. The proposed approach uses a spatial reasoning strategy that integrated direction and distance spatial relations in the reasoning process, allowing the inference of implicit spatial relations for the several levels of the considered geographic hierarchies. The integration of a geographic and a demographic database allowed the discovery of spatial patterns and general relationships that exist between the analysed spatial and non-spatial data.

1 Introduction

Spatial databases like geographic databases in geographic information systems usually store large amounts of spatial information. Queries in those systems involve the derivation of different spatial relations that are not explicitly stored. It is generally accepted that it's neither practical nor efficient to store the substantial amount of spatial relationships that can exist in space (Abdelmoty and El-Geresy, 1995). Qualitative spatial reasoning has been proposed as a complementary mechanism for the automatic derivation of spatial relations that are not explicitly stored. Qualitative reasoning is based on the manipulation of qualitative identifiers such as *close*, *North*, etc. as opposed to quantitative information like point co-ordinates or distance values.

The analysis of spatial information through the use of a qualitative approach is desirable when the precision of a quantitative representation (normally using co-ordinates) is not needed or when the input and output of the process is qualitative rather than quantitative (Papadias and Sellis, 1994).

This paper addresses the process of Knowledge Discovery in Spatial Databases (KDSD) through a qualitative spatial reasoning approach. The aim is the discovery of implicit relationships that can exist between spatial and non-spatial data.

Although spatial reasoning can handle relationships of several types, this work investigates the use of *direction* and *distance* spatial relations in the inference of new spatial knowledge and its subsequent use in the knowledge discovery process. The adoption of an integrated spatial reasoning strategy (Sharma, 1996) and geographic concept hierarchies allowed the inference of implicit spatial relations for the several levels of the considered hierarchies.

The integration of a geographic database, with the administrative subdivisions of Portugal at the municipality and district level (characterising the territory through geographic identifiers), and a demographic database, enabled the discovery of spatial patterns and general relationships that exist between the analysed spatial and non-spatial data.

The paper is organised as follows: Section 2 systematises different approaches in the process of KDSD; section 3 presents an overview of qualitative spatial relations and reasoning strategies. Section 4 describes the system architecture; section 5 reports the integration of the two databases and exhibits some of the discovered patterns. Section 6 concludes with some remarks and comments about future work.

2 Approaches in 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 general data characteristics not explicitly stored in spatial databases (Koperski and Han, 1995).

Spatial database systems are relational databases plus a concept of spatial location and spatial extension (Ester, et al., 1997). The explicit location and extension of objects define implicit relations of spatial neighbourhood. Thus neighbour attributes of some object may influence its behaviour and therefore must be consider in the process of knowledge discovery in databases (KDD). Traditionally, knowledge discovery in relational databases doesn't take into account this spatial reasoning, motivating the development of new algorithms adapted to the spatial component of spatial data.

The main approaches in KDSD are characterised by the development of new algorithms that treat the objects' position and extension mainly through the manipulation of its co-ordinates. These algorithms are then implemented, extending traditional KDD systems in order to accommodate them. In all, a quantitative approach is used in the reasoning process although the results are presented using qualitative identifiers.

Lu et al. (1993) proposed an attribute-oriented induction approach that is applied to spatial and non-spatial attributes using concept hierarchies. This allows the discovery of relationships that exists between spatial and non-spatial data. A spatial concept hierarchy represents a successive merge of neighbour regions into large regions. Two learning algorithms were introduced: i) non-spatial attribute oriented induction, which performs generalisation on non-spatial data first, and ii) spatial hierarchy induction, which performs generalisation on spatial data first. In both approaches, the classification of its corresponding spatial and non-spatial data is performed based on the classes obtained by the generalisation. Another peculiarity of this approach is that the user must provide to the system the set of relevant data, the concept hierarchies, the desired rule forms and the learning request (specified in a syntax similar to SQL).

Koperski and Han (1995) investigated the utilisation of interactive data mining for the extraction of spatial association rules. In their approach the spatial and non-spatial attributes are held in different databases, but once the user identifies the attributes or relations of interest, a selection process takes place and a unified database is created. An algorithm, implemented for the discovery of spatial association rules, analyses the stored data. The rules obtained represent object relationships described using spatial predicates like *adjacent_to* or *close_to*.

These two approaches are representative of the efforts made in the area of KDSD. One approach uses two different databases, storing spatial and non-spatial data separately. Once the user identifies the attributes of interest, an interface between the two databases ensures the selection and treatment of data without the creation of a new repository (that handles both spatial and non-spatial data). The other approach also requires two different databases, but the selection phase leads to the creation of a unified database where the analysis of data takes place. In both approaches new algorithms are implemented and the user is asked for the specification of the relevant attributes and the type of results pretended.

In this paper we present 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 location and extension is not always needed, we investigated how traditional KDD systems (and their generic Data Mining algorithms) for relational databases can be used in KDSD. Depending on the spatial reasoning 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. This approach is described in section 4.

3 Qualitative Spatial Reasoning

Spatial reasoning is the process by which information about objects in space and their relationships are gathered by various means, such as measurement or observation, and use to arrive to valid conclusions regarding the objects' relationships (Sharma, 1996).

Spatial relations have been classified in several types (Frank, 1996; Papadias and Sellis, 1994), including *topological relations* (that describe neighbourhood and incidence), *direction relations* (that describe order in space) and *distance relations* (that describe proximity in space). For each of these relations, composition tables facilitate reasoning with incomplete spatial information by allowing the inference of new information (see (Frank, 1996; Grigni, et al., 1995; Sharma, 1996) for more details). The composition tables referred represent a *homogeneous spatial reasoning* (Sharma, 1996) approach, because each one only allows the inference of spatial information of the same type (directional, topological or metric). Although it is useful, it has certain limitations.

Reasoning about qualitative distances necessarily involves *integrated spatial reasoning* about qualitative distances and directions. For example, the facts **A is very far from B** and **B is very far from C** do not permit the inference of the relationship that exists between **A** and **C**. **A** can be **very close** or **close from C** (if **B** and **C** are in opposite directions), or **A** may be **far** or **very far from C** (if **B** and **C** are in the same orientation).

Since an homogeneous spatial reasoning approach was already used for direction relations (Santos, et al., 1999) and the results obtained point out the validity of the adopted strategy for the inference of new spatial knowledge, the integration of distance and direction relations will now be addressed.

If the integration of qualitative distances and directions is required, the adoption of a set of identifiers, which allows the identification of the considered directions and distances and its respective intervals of validity, are needed. Hong et al. (1995) analysed some possible combinations for the number of identifiers and the geometric patterns that should characterise the distance intervals. The verification of the suggested rules, allow us to adopt a **localisation system** (Figure 1), based on **eight symbols for direction relations** (N, NE, E, SE, S, SW, W, NW) and **four symbols for the identification of the distance relations** (vc – very close, c - close, f- far, vf – very far).

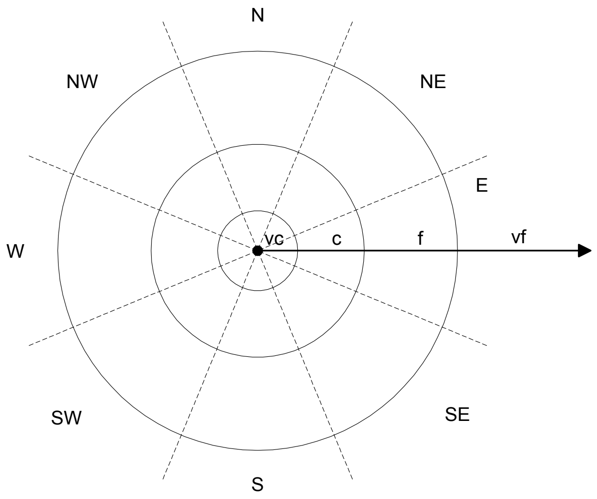


Figure 1 – The localisation system

The definition of the validity interval for each distance identifier must obey some rules. In these systems, as can be seen in Table 1, there should exist a constant ratio (**ratio = length (dist$_i$)/length (dist$_{i-1}$)**) relationship between the lengths of two neighbouring intervals.

The presented simulated intervals allow the composition of new distance interval sets by magnification of the original interval. For example, the set of values for a **ratio 4** can be increased by a factor of 10 supplying the values **dist$_0$ (0, 10]**, **dist$_1$ (10, 50]**, **dist$_2$ (50, 210]** and **dist$_3$ (210, 850]**. Since the same scale magnifies all intervals and quantitative distance relations, qualitative compositions will remain the same, regardless of the scale value.

Table 1 - Simulated intervals for four symbolic distance values (adapted from Hong, J.-H., M. J. Egenhofer, and A. U. Frank, *On the Robustness of Qualitative and Direction Reasoning*, Proceedings of Auto-Carto 12, Charlotte, North California, 1995).

Ratio	dist$_0$	dist$_1$	dist$_2$	dist$_3$
1	(0, 1]	(1, 2]	(2, 3]	(3, 4]
2	(0, 1]	(1, 3]	(3, 7]	(7, 15]
3	(0, 1]	(1, 4]	(4, 13]	(13, 40]
4	(0, 1]	(1, 5]	(5, 21]	(21, 85]
5	(0, 1]	(1, 6]	(6, 31]	(31, 156]
6	(0, 1]	(1, 7]	(7, 43]	(43, 259]
7	(0, 1]	(1, 8]	(8, 57]	(57, 400]
8	(0, 1]	(1, 9]	(9, 73]	(73, 585]
9	(0, 1]	(1, 10]	(10, 91]	(91, 820]
10	(0, 1]	(1, 11]	(11, 111]	(111, 1111]
20	(0, 1]	(1, 21]	(21, 421]	(421, 8421]
50	(0, 1]	(1, 51]	(51, 2551]	(2551, 127551]
100	(0, 1]	(1, 101]	(101, 10101]	(10101, 1010101]

It is important to know that the number of distance symbols used, and the ratio between the quantitative values addressed by each interval, plays an important role in the robustness of the final system, i.e., in the validity of the composition table for the inference of new spatial relations.

The final composition table, constructed following the suggestions made by Hong (1994) and presented through icons (Figure 2), is a 32x32 matrix. This matrix represents some of the knowledge needed, for the inference of new spatial information in the localisation system used in this study. Due to its great size, Table 2 exhibits an extract of the final matrix.

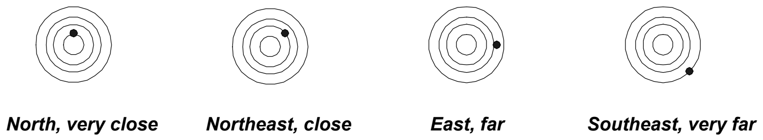
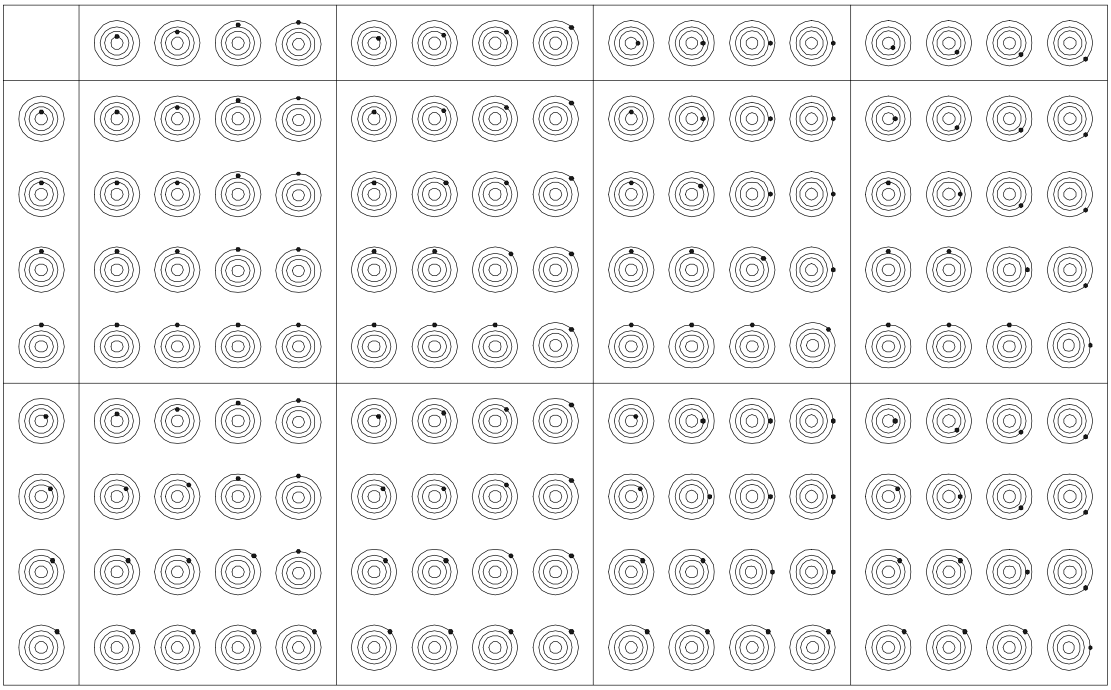


Figure 2 – Icons representing direction and distance spatial relations

Table 2 – An extract of the final composition table



This composition table is transformed in a set of facts, stored in a table, so that machine learning algorithms can assimilate them. The rules generated will later be used in the inference of new spatial relations needed in the knowledge discovery process. This assimilation and the inference mechanism are described in detail in Santos and Amaral (1999).

4 The System Architecture

The proposed system, a system for Knowledge Discovery in Spatial Databases based on Qualitative Spatial Reasoning, has three main components: data repositories, data analysis and results visualisation. Figure 3 presents a general overview of the system architecture.

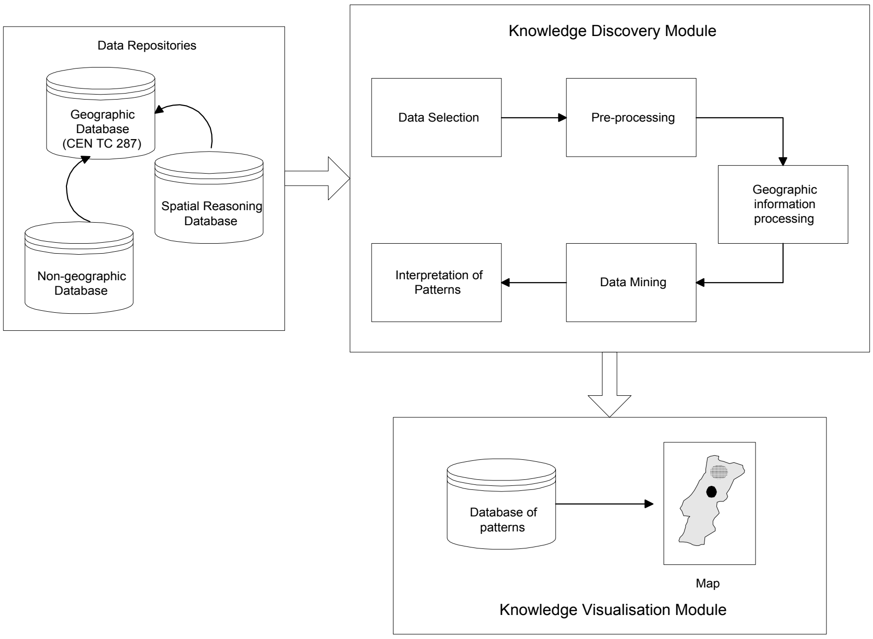

Data Repositories
Geographic Database (CEN TC 287)
Spatial Reasoning Database
Non-geographic Database
Knowledge Discovery Module
Data Selection
Pre-processing
Geographic information processing
Data Mining
Interpretation of Patterns
Database of patterns
Map
Knowledge Visualisation Module

Figure 3 – The System Architecture

1. The Data Repositories component aggregate three main databases:

- A ***geographic database*** constructed under the principles established by the European Committee for Normalisation in the CEN TC 287 standard for Geographic Information (CEN/TC-287, 1998a). Following its recommendations was possible to implement a geographic database in which the positional aspects of geographic data are obtained using a geographic identifiers system (CEN/TC-287, 1998b). In this system the administrative subdivisions of Portugal at the municipality and district level were characterised. Between other attributes, we point out the direction and distance spatial relations that exist between the considered municipalities and the geographic concept hierarchies that associate the several subdivisions adopted. More details about this database can be found in Santos and Amaral (1999).
- A ***spatial reasoning database*** that stores the inference mechanisms that allow the derivation of new spatial relations between regions. Between the knowledge available in this database, we point out the composition table, the set of identifiers and the validity of each distance interval for the localisation system described in section 3. This database is used in conjunction with the geographic database in the deduction of implicit spatial relations. This task is performed in the knowledge discovery module (data analysis component). More details about this database can be found in Santos and Amaral (1999).
- A ***non-geographic database*** that will be integrated with the geographic database and analysed in the knowledge discovery module. This procedure will allow the discovery of implicit relationships that exist between the geographic and non-geographic data. In this paper, we analyse a demographic database that collect the parish registers dated between

1690-1990 in one district of Portugal – *Aveiro* (this integration and the obtained results are presented in section 5).

2. The Data Analysis component is implemented through the knowledge discovery module. In this module, several steps characterise the knowledge discovery process:

- ***Data selection*** that picks out the relevant non-geographic and geographic data for analysis. This phase addresses the question of what relevant attributes are to be considered, the size of the sample needed and the period of time.
- ***Pre-processing*** of the selected data allowing the reduction of the sample set. Several tasks can here be effected. Data can be generalised attending to demographer's hierarchies (reducing the number distinct rows) or techniques of rough sets can be used to perform the reduction of the number of columns (Rodrigues, et al., 1999). To ensure proper data analysis, the data selected in the previous step must be checked to remove corrupt data and manage missing values.
- ***Geographic information processing*** verifies the availability of all spatial relations needed in the knowledge discovery process. As we said previously, spatial relations of interest might be implicit in the geographic database considered. In this case, and to ensure that all relevant geographic knowledge is available for the data mining algorithms, these implicit relations can be automatically transformed in explicit relations through the inference mechanisms described previously.
- ***Data Mining*** algorithms are now applied to the relevant geographic and non-geographic data in order the find relationships that exploit the intrinsic relations that exist between the analysed data.
- ***Interpretation of the discovered patterns*** is performed in order to evaluate their utility and importance, in this case study, to demographers. In this step it may be realised that relevant attributes were ignored in the analysis, suggesting that the process should be repeated using the missed values. Relevant relationships/patterns can be stored in the patterns' database allowing its subsequent use in further analyses or its visualisation in a more suitable way – a map.

3. The Results Visualisation component is responsible for the management of the database of patterns. This database is integrated in a geographic information system, for the visualisation of the discovered patterns in a map. At this stage, the knowledge visualisation module is not completely implemented, but allows the insertion of new patterns and its visualisation in an easy and friendly way.

In terms of technical characteristics, the data repository component and the database of patterns were implemented in a relational database system, being available to the knowledge discovery module through an ODBC (Open Database Connectivity) connection. This module was implemented in *Clementine* (ISL, 1998a; ISL, 1998b), a knowledge discovery tool that support all phases of the knowledge discovery process.

Clementine is a data mining toolkit based on visual programming, which includes machine learning technologies like rule induction and neural networks. Among its facilities are *data manipulation* (construct of new data items derived from existing ones), *browsing and visualisation* (display of data using interactive graphics), *statistics* and *hypothesis testing* (model constructs of how the data behaves and its verification). The knowledge discovery process is defined in *Clementine* through the construction of a *stream*, in which each operation on data is represented by a *node*.

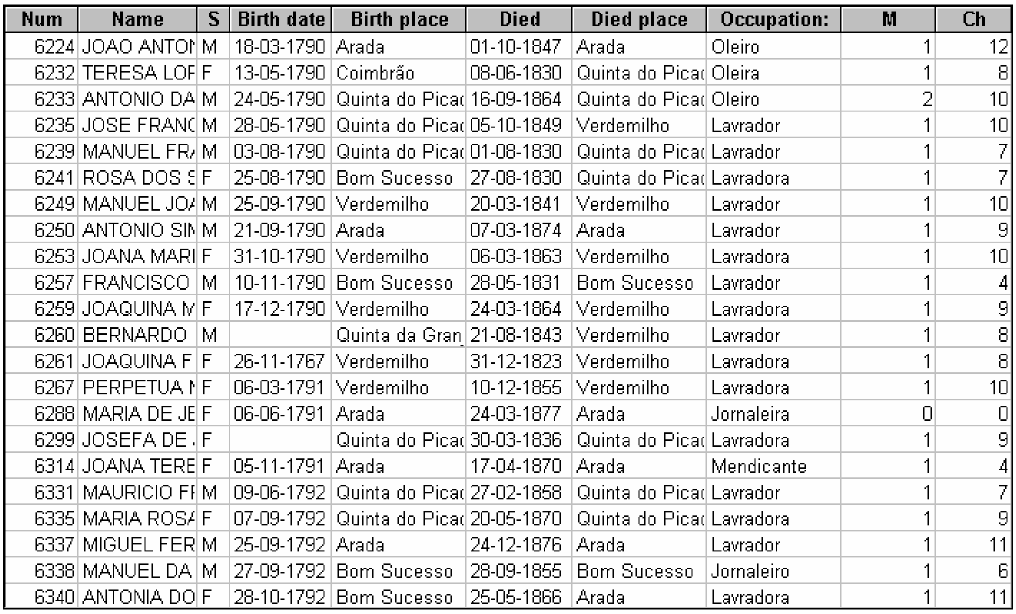
The geographic information system used is *Geomedia Professional v3* (Intergraph, 1999), allowing the visualisation of the knowledge discovery process results in a map. In this system is also possible the manipulation of the discovered patterns, in order to construct thematic maps that conjugate more than one result.

Next section present the integration of the demographic database in the knowledge discovery process and the results obtained by the data mining algorithms. The description is centred on tasks executed under the knowledge discovery module.

5 The knowledge discovery process with a demographic database

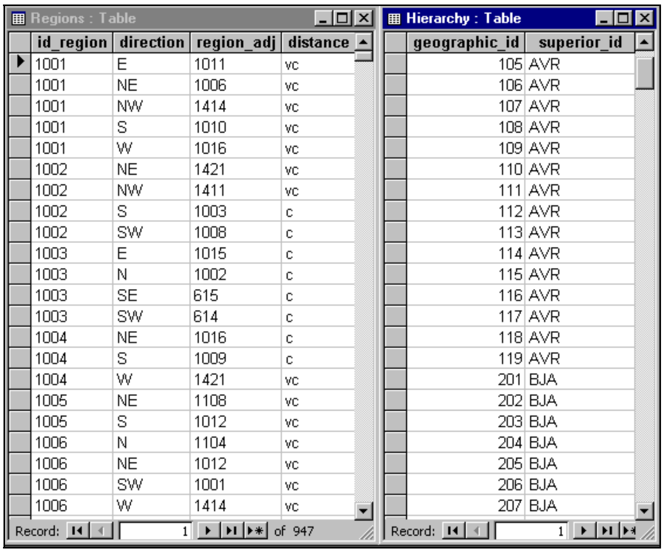
In systems of spatial referencing using geographic identifiers, a position is referenced to a real world location defined by a real world object. This object is termed a location, and its identifier is termed a geographic identifier (CEN/TC-287, 1998b). These geographic identifiers are present in the demographic database of the Aveiro district, in attributes like: place, parish, municipality and district. The integration of the geographic and non-geographic database is effected at the municipality level, since this is the level of aggregation considered relevant for the pretended results. Table 3 shows an extract of one table (*individual*) of the demographic database used, while Table 4 display some extracts of the spatial relations and hierarchies stored in the geographic database.

Table 3 – Demographic database: some attributes of *individual* table



Num	Name	S	Birth date	Birth place	Died	Died place	Occupation:	M	Ch
6224	JOAO ANTOI	M	18-03-1790	Arada	01-10-1847	Arada	Oleiro	1	12
6232	TERESA LOF	F	13-05-1790	Coimbrão	08-06-1830	Quinta do Picac	Oleira	1	8
6233	ANTONIO DA	M	24-05-1790	Quinta do Picac	16-09-1864	Quinta do Picac	Oleiro	2	10
6235	JOSE FRANC	M	28-05-1790	Quinta do Picac	05-10-1849	Verdemilho	Lavrador	1	10
6239	MANUEL FRA	M	03-08-1790	Quinta do Picac	01-08-1830	Quinta do Picac	Lavrador	1	7
6241	ROSA DOS S	F	25-08-1790	Bom Sucesso	27-08-1830	Quinta do Picac	Lavradora	1	7
6249	MANUEL JOA	M	25-09-1790	Verdemilho	20-03-1841	Verdemilho	Lavrador	1	10
6250	ANTONIO SIN	M	21-09-1790	Arada	07-03-1874	Arada	Lavrador	1	9
6253	JOANA MARI	F	31-10-1790	Verdemilho	06-03-1863	Verdemilho	Lavradora	1	10
6257	FRANCISCO	M	10-11-1790	Bom Sucesso	28-05-1831	Bom Sucesso	Lavrador	1	4
6259	JOAQUINA M	F	17-12-1790	Verdemilho	24-03-1864	Verdemilho	Lavradora	1	9
6260	BERNARDO	M		Quinta da Gran	21-08-1843	Verdemilho	Lavrador	1	8
6261	JOAQUINA F	F	26-11-1767	Verdemilho	31-12-1823	Verdemilho	Lavradora	1	8
6267	PERPETUA N	F	06-03-1791	Verdemilho	10-12-1855	Verdemilho	Lavradora	1	10
6288	MARIA DE JE	F	06-06-1791	Arada	24-03-1877	Arada	Jornaleira	0	0
6299	JOSEFA DE .	F		Quinta do Picac	30-03-1836	Quinta do Picac	Lavradora	1	9
6314	JOANA TERE	F	05-11-1791	Arada	17-04-1870	Arada	Mendicante	1	4
6331	MAURICIO FI	M	09-06-1792	Quinta do Picac	27-02-1858	Quinta do Picac	Lavrador	1	7
6335	MARIA ROSA	F	07-09-1792	Quinta do Picac	20-05-1870	Quinta do Picac	Lavradora	1	9
6337	MIGUEL FER	M	25-09-1792	Arada	24-12-1876	Arada	Lavrador	1	11
6338	MANUEL DA	M	27-09-1792	Bom Sucesso	28-09-1855	Bom Sucesso	Jornaleiro	1	6
6340	ANTONIA DO	F	28-10-1792	Bom Sucesso	25-05-1866	Arada	Lavradora	1	11

Table 4 – Geographic relations and hierarchies



Regions : Table

id_region	direction	region_adj	distance
1001	E	1011	vc
1001	NE	1006	vc
1001	NW	1414	vc
1001	S	1010	vc
1001	W	1016	vc
1002	NE	1421	vc
1002	NW	1411	vc
1002	S	1003	c
1002	SW	1008	c
1003	E	1015	c
1003	N	1002	c
1003	SE	615	c
1003	SW	614	c
1004	NE	1016	c
1004	S	1009	c
1004	W	1421	vc
1005	NE	1108	vc
1005	S	1012	vc
1006	N	1104	vc
1006	NE	1012	vc
1006	SW	1001	vc
1006	W	1414	vc

Record: 1 of 947

Hierarchy : Table

geographic_id	superior_id
105	AVR
106	AVR
107	AVR
108	AVR
109	AVR
110	AVR
111	AVR
112	AVR
113	AVR
114	AVR
115	AVR
116	AVR
117	AVR
118	AVR
119	AVR
201	BJA
202	BJA
203	BJA
204	BJA
205	BJA
206	BJA
207	BJA

Record: 1

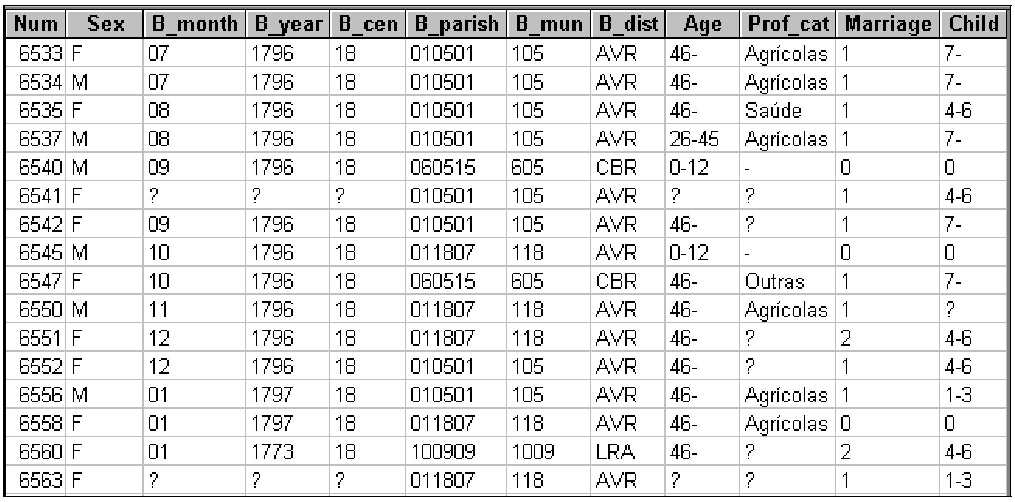
As can be seen in Table 3, the original table, which describes some relevant attributes about an individual, has some missing data fields. This is very usual since some data is from the XVII century. In order to avoid these faults, the unknown fields are marked with an '?', so they can be easily identified in the knowledge discovery module.

Another important task, and that must be performed in order reduce the search space and the variety of possible different values, is the generalisation and transformation of some attributes into discreet values. Following some suggestions (Rodrigues, et al., 1999) of possible concept hierarchies for the demographic data, Table 5 systematise the hierarchies/discreet values adopted. The resulting *individual* table is displayed in Table 6.

Table 5 – Hierarchies and discreet values adopted

Attributes		**Hierarchies / Discreet values**
Place		Place → **Parish** → **Municipality** → **District**
Date	Year	{1600..1699} → **17**, {1700..1799} → **18**, {1800..1899} → **19**, {1900..1999} → **20**
	Month	{January, February, March, April, May, June, July, August, September, October, November, December} → {**1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**}
Age		{0..12} → **0-12**, {13..25} → **13-25**, {26..45} → **26-45**, {46..110} → **46-**
Children		{0} → **0**, {1..3} → **1-3**, {4..6} → **4-6**, {7..16} → **7-**

Table 6 – The resulting *individual* table

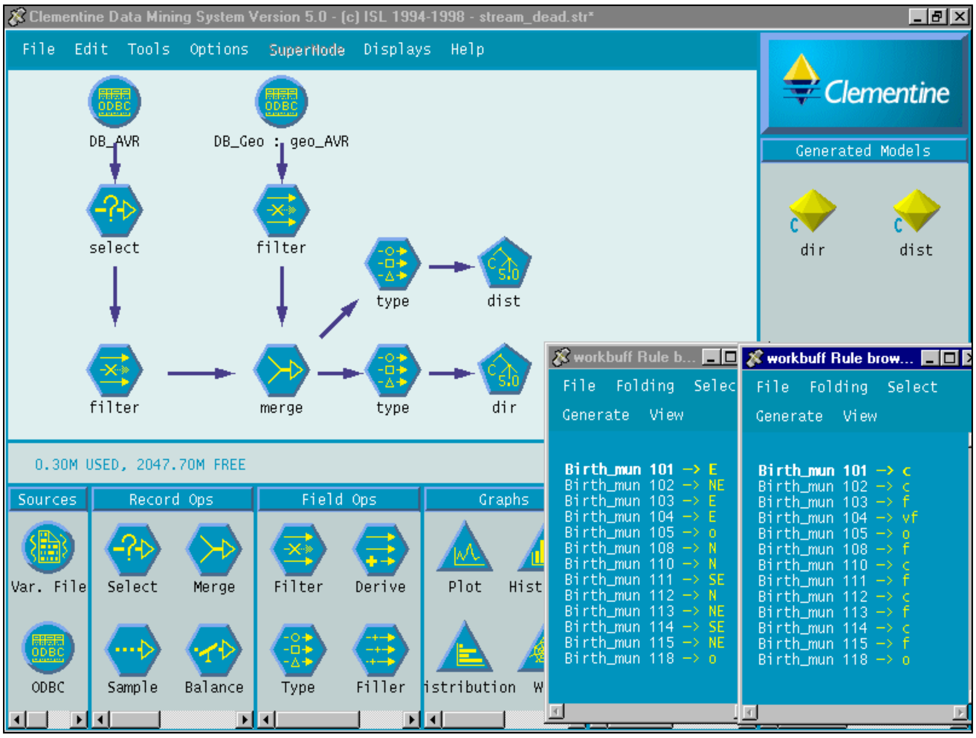


Num	Sex	B_month	B_year	B_cen	B_parish	B_mun	B_dist	Age	Prof_cat	Marriage	Child
6533	F	07	1796	18	010501	105	AVR	46-	Agrícolas	1	7-
6534	M	07	1796	18	010501	105	AVR	46-	Agrícolas	1	7-
6535	F	08	1796	18	010501	105	AVR	46-	Saúde	1	4-6
6537	M	08	1796	18	010501	105	AVR	26-45	Agrícolas	1	7-
6540	M	09	1796	18	060515	605	CBR	0-12	-	0	0
6541	F	?	?	?	010501	105	AVR	?	?	1	4-6
6542	F	09	1796	18	010501	105	AVR	46-	?	1	7-
6545	M	10	1796	18	011807	118	AVR	0-12	-	0	0
6547	F	10	1796	18	060515	605	CBR	46-	Outras	1	7-
6550	M	11	1796	18	011807	118	AVR	46-	Agrícolas	1	?
6551	F	12	1796	18	011807	118	AVR	46-	?	2	4-6
6552	F	12	1796	18	010501	105	AVR	46-	?	1	4-6
6556	M	01	1797	18	010501	105	AVR	46-	Agrícolas	1	1-3
6558	F	01	1797	18	011807	118	AVR	46-	Agrícolas	0	0
6560	F	01	1773	18	100909	1009	LRA	46-	?	2	4-6
6563	F	?	?	?	011807	118	AVR	?	?	1	1-3

After the demographic data treatment, and since the objective of this work is to show how qualitative spatial reasoning can be incorporated in the process of knowledge discovery, two exercises were effected. In both, the inference of new spatial relations was addressed. As previously described, the geographic database used was implemented under the principles established by the European standard for geographic information. Through it, only the spatial relations that exist between adjacent regions can be specified. All the other relations, existing between non-adjacent regions and needed in the knowledge discovery process, must be inferred. In *Clementine*, a rule induction algorithm learnt the inference rules explicit in the composition table that allows integrated qualitative reasoning about distance and direction relations. The decision tree created was subsequently applied in a process that cyclically infers unknown spatial relations. More details about the learning and inference processes can be found in Santos and Amaral (1999) and Santos, et al. (1999).

The first exercise carried out in *Clementine* had as its purpose characterises the **locality of birth** with respect to the **locality of death** of an individual, attending to the century in which the facts occurred. Geographically these attributes were analysed with respect to the **municipality** level. The stream constructed in *Clementine* that represents the knowledge discovery process is displayed in Figure 4. By the analysis of this figure it can be noted that two main steps constituted the process. The relevant demographic attributes for analysis were selected and after that, they were merged with the spatial relations coming from the geographic database[1]. The resulting table was submitted to the **C5.0** algorithm, in order to obtain expressive rules. In the same figure it is possible to see the tabulated results on the right side.

[1] Some of these relations were previously inferred, since not all the spatial relationships, which can exist between the municipalities of the Aveiro district, are explicit in the geographic database.



```
Birth_mun 101 -> E
Birth_mun 102 -> NE
Birth_mun 103 -> E
Birth_mun 104 -> E
Birth_mun 105 -> o
Birth_mun 108 -> N
Birth_mun 110 -> N
Birth_mun 111 -> SE
Birth_mun 112 -> N
Birth_mun 113 -> NE
Birth_mun 114 -> SE
Birth_mun 115 -> NE
Birth_mun 118 -> o
```

```
Birth_mun 101 -> c
Birth_mun 102 -> c
Birth_mun 103 -> f
Birth_mun 104 -> vf
Birth_mun 105 -> o
Birth_mun 108 -> f
Birth_mun 110 -> c
Birth_mun 111 -> f
Birth_mun 112 -> c
Birth_mun 113 -> f
Birth_mun 114 -> c
Birth_mun 115 -> f
Birth_mun 118 -> o
```

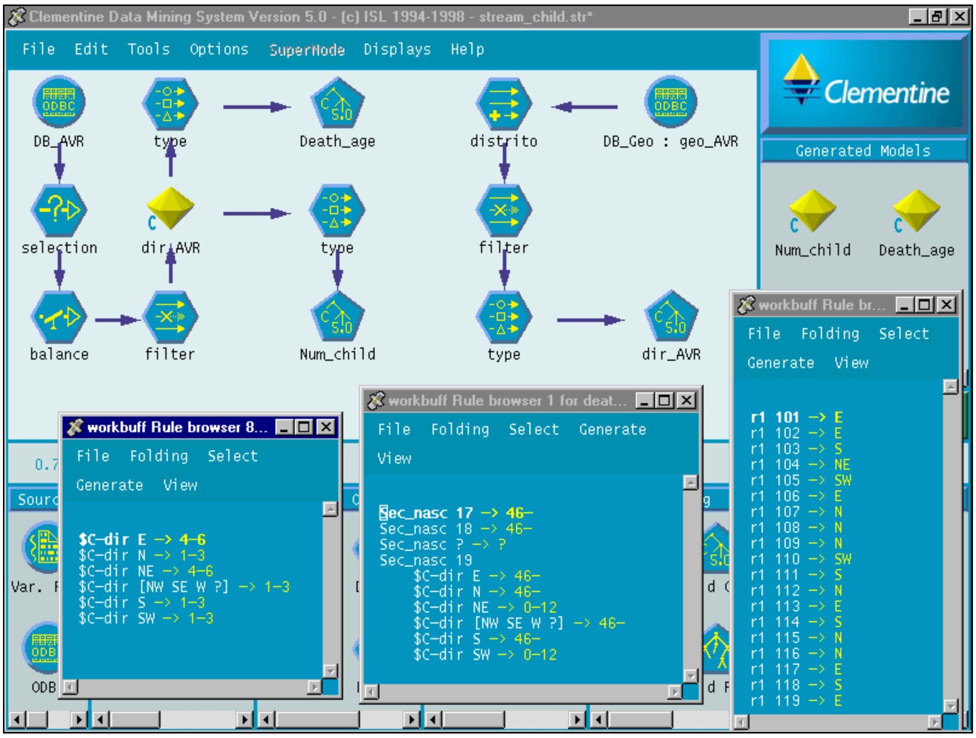
Figure 4 – Characterisation of the locality of birth with respect to the locality of death

The results achieved point out that there is a spatial movement to the Southwest in the analysed district. The two municipalities that represent the destiny of the population movement represent the exceptional cases. These two municipalities, **105** and **118**, have its West frontier with the ocean. This population movement in direction to regions located a Southwest and with the sea as neighbour must be explained by demographers.

The other exercise comprised the finding of a pattern that characterises the **age at death** and the **number of children** for the geographic region analysed. The streams constructed and the results obtained are presented in Figure 5. By the analysis of this figure it can be noted that two streams were needed to achieve the pretended results. The stream located at the right side had as purpose build a model that describe the geographical localisation of each municipality with respect to the whole district. The obtained model[2] (**dir_AVR**), browsed in the tabular table at the right of the figure, was subsequently used in the other steam, to characterise the two attributes (**Death_age** and **Num_child**) geographically.

Analysing the results obtained for the **Death_age** attribute (tabular table at middle of the figure), only in the XIX century exists a visible variation between regions. In the generated model, all municipalities located at Northeast to Southwest of Aveiro present a lower age at death, **0-12**, indicating that all regions at these locations had a great rate of infant mortality. The generated model for the **Num_child** attribute (tabular table at left of the figure) tell us that, regions with a high birth rate fall in the Northeast and East of Aveiro.

[2] Each generated model can be saved in order to be reused by others steams/work sessions.



```
$C-dir E -> 4-6
$C-dir N -> 1-3
$C-dir NE -> 4-6
$C-dir [NW SE W ?] -> 1-3
$C-dir S -> 1-3
$C-dir SW -> 1-3
```

```
Sec_nasc 17 -> 46-
Sec_nasc 18 -> 46-
Sec_nasc ? -> ?
Sec_nasc 19
    $C-dir E -> 46-
    $C-dir N -> 46-
    $C-dir NE -> 0-12
    $C-dir [NW SE W ?] -> 46-
    $C-dir S -> 46-
    $C-dir SW -> 0-12
```

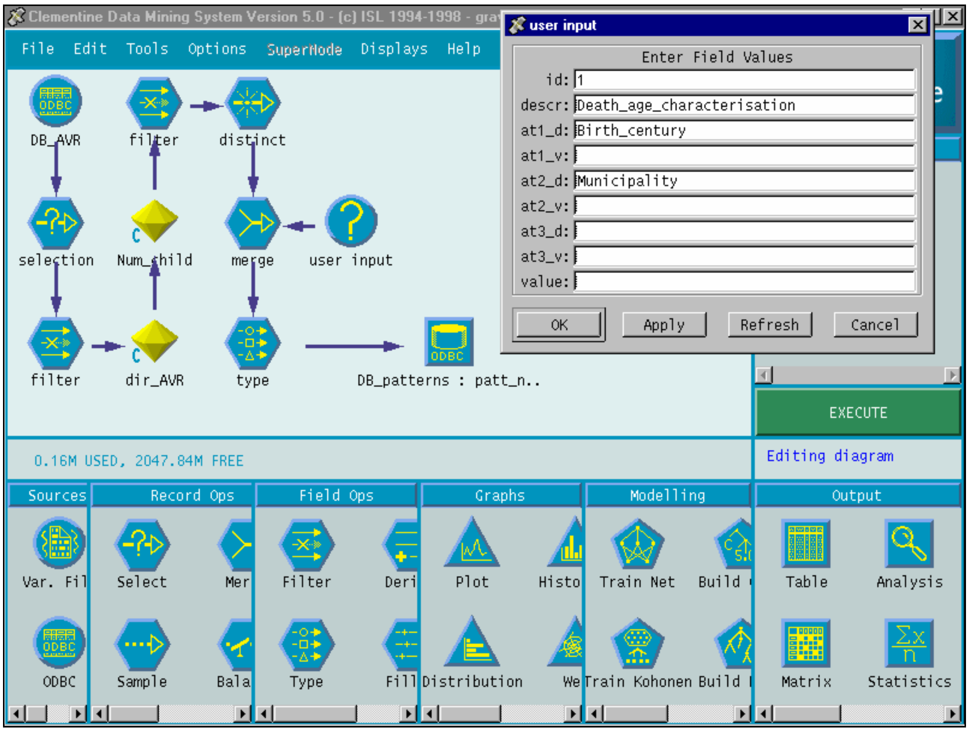
```
r1 101 -> E
r1 102 -> E
r1 103 -> S
r1 104 -> NE
r1 105 -> SW
r1 106 -> E
r1 107 -> N
r1 108 -> N
r1 109 -> N
r1 110 -> SW
r1 111 -> S
r1 112 -> N
r1 113 -> E
r1 114 -> S
r1 115 -> N
r1 116 -> N
r1 117 -> E
r1 118 -> S
r1 119 -> E
```

Figure 5 – Streams *Clementine* and the obtained results

After the analysis of the relevance of each discovered pattern, the selected patterns can be stored in the database of patterns. For this, the user only needs to load an ***input node***, with the structure of the database of patterns, in which the attributes to be stored are specified. An obligatory attribute is the geographic attribute that will allow the integration with the map. As we already said, in this case study it must be the municipality attribute.

Figure 6 shows the constructed stream, and the input window in which the user must specify the attributes that will be passed, through an OBDC connection, to the database of patterns. These patterns can then be visualised in a thematic map. The presented stream uses the generated model for characterisation of the **Num_child** attribute. The objective is the explicitation of the discovered rules. Figure 7 exhibits a map of the region that present the results obtained for the **Num_child** attribute.

It is important to say that although the presented exercises only used decision trees techniques (implemented through the C5.0 algorithm), the other data mining techniques available in *Clementine* (neural networks, association rules, clustering, etc.) could be used, if they were appropriate to the pretended results.



user input

Enter Field Values

id: 1

descr: Death_age_characterisation

at1_d: Birth_century

at1_v:

at2_d: Municipality

at2_v:

at3_d:

at3_v:

value:

0.16M USED, 2047.84M FREE

Editing diagram

Figure 6 – Definition of the attributes to be exported to the database of patterns

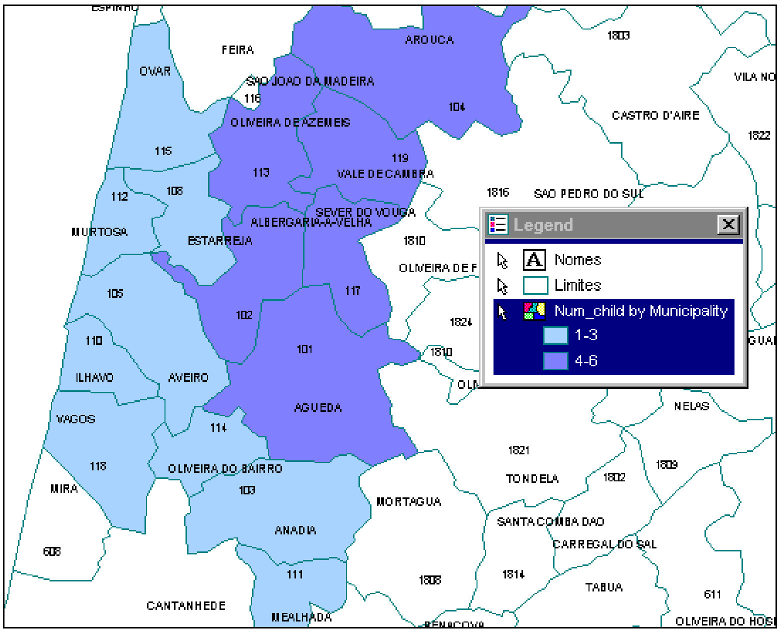


Figure 7 – Thematic map with the characterisation of the Num_child **attribute**

6 Conclusions and Future Work

Organisational databases usually store geographic identifiers, like addresses or postcodes, which geographic component is not incorporated in the process of knowledge discovery. This paper presented an approach for knowledge discovery in spatial databases, based on qualitative spatial reasoning, where the positions of geographic data were provided by qualitative identifiers.

For the municipalities of Portugal, some direction and distance spatial relations were defined. This knowledge and the composition table available for the localisation system used, allowed the inference of new spatial localisation relations between regions. The existence of domain knowledge, like geographic concept hierarchies, was also useful in the inference of localisation relations for the another levels of the considered hierarchies.

The integration of a demographic database (dated from 1690-1990 of the Aveiro district) with a geographic database (with the administrative subdivisions of Portugal), made possible the discovery of general descriptions that exploit the relationships that exist between the geographic and non-geographic data analysed.

The results obtained point out that traditional KDD systems, developed for relational databases and not having semantic knowledge linked to spatial data, can be used in the process of KDSD since some of this semantic knowledge and the principles of qualitative spatial reasoning were available as domain knowledge. *Clementine* was used in the assimilation of the geographic domain knowledge (like composition tables), in the inference of new spatial relations and in the spatial patterns discovery.

Many future tasks can be conceived. Two of them are the integration of topological spatial relations in the reasoning process, and the conclusion of the knowledge visualisation module implementation. This module plays an important role in the overall system, since users can use it to access the discovered patterns in a friendly way (via a map). Another priority is the study of the principles behind qualitative temporal reasoning, in order to integrate the temporal domain in the process of knowledge discovery (allowing the inference of so many unknown dates in the demographic database used).

7 Acknowledgements

This work has been partially supported by a grant from PRODEP II (*Acção 5.2 – Concurso nº3/98 Doutoramentos*).

8 References

Abdelmoty, A. I., and B. A. El-Geresy, *A General Method for Spatial Reasoning in Spatial Databases*, Proceedings of the fourth International Conference on Information and Knowledge Management, Baltimore, ACM, 1995.

CEN/TC-287, *Geographic Information: Data Description, Rules for application schemas*, European Committee for Normalisation, WI 006, 1998a.

CEN/TC-287, *Geographic Information: Referencing, Geographic Identifiers*, European Committee for Normalisation, prENV 12661, 1998b.

Ester, M., H.-P. Kriegel, and J. Sander, *Spatial Data Mining: A Database Approach*, Proceedings of the 5th International Symposium on Large Spatial Databases. Lectures Notes in Computer Science, Berlin, Germany, Springer-Verlag, 1997.

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

Grigni, M., D. Papadias, and C. Papadimitriou, *Topological Inference*, Proceedings of the International Joint Conference of Artificial Intelligence (IJCAI), Montreal, Canada, AAAI Press, 1995.

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

Hong, J.-H., M. J. Egenhofer, and A. U. Frank, *On the Robustness of Qualitative and Direction Reasoning*, Proceedings of Auto-Carto 12, Charlotte, North California, 1995.

Intergraph, *Geomedia Professional v3*, Reference Manual, Intergraph Corporation, 1999.

ISL, *Clementine, Reference Manual, Version 5.0*, 1998a.

ISL, *Clementine, User Guide, Version 5.0*, Integral Solutions Limited, 1998b.

Koperski, K., and J. Han, *Discovery of Spatial Association Rules in Geographic Information Systems*, Proc. 4th International Symposium on Large Spatial Databases (SSD95), Maine, 1995.

Lu, W., J. Han, and B. C. Ooi, *Discovery of General Knowledge in Large Spatial Databases*, Proc. of the 1993 Far East Workshop on Geographic Information Systems, Singapore, 1993.

Papadias, D., and T. Sellis, "On the Qualitative Representation of Spatial Knowledge in 2D Space", *Very Large Databases Journal, Special Issue on Spatial Databases*, 3, 4, 479-516, 1994.

Rodrigues, M. F., C. Ramos, and P. R. Henriques, *An Intelligent System to Study Demographic Evolution*, Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology, Orlando, Florida, 1999.

Santos, M., and L. Amaral, *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, Lisbon, Portugal, 24-26 November, 1999.

Santos, M., L. Amaral, and P. Pimenta, *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, Salvador, Brazil, 19-23 July, 1999.

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