

KNOWLEDGE DISCOVERY IN SPATIAL DATABASES

– The PADRÃO's qualitative approach

Maribel Santos and Luís Amaral

Abstract

Knowledge discovery in databases is a complex process concerned with the discovery of relationships and other descriptions from data. Knowledge discovery in spatial databases represents a particular case of discovery, allowing the discovery of relationships that exist between spatial and non-spatial data, and other data characteristics that aren't explicitly stored in spatial databases.

This paper describes the conception and implementation of PADRÃO, a system for knowledge discovery in spatial databases. PADRÃO presents a new approach to this process, which is based on qualitative spatial reasoning. The *spatial semantic knowledge* and the *principles of qualitative spatial reasoning* needed for the spatial reasoning process are available in the PADRÃO's *geographic database* and PADRÃO's *spatial knowledge base*, allowing the integration of the geo-spatial component, associated with the analysed non-geographic data, in the process of knowledge discovery.

Résumé (in French)

...

Maribel Santos, Luís Amaral

Information Systems Department and Algoritmi Research Centre

University of Minho, Campus de Azurém

4800-019 Guimarães, PORTUGAL

Tel.: +351 253 510259

Fax: +351 253 510250

e-mail: {maribel, amaral}@dsi.uminho.pt

Knowledge Discovery in Spatial Databases: the PADRÃO's qualitative approach

Knowledge discovery in databases is a complex process concerned with the discovery of relationships and other descriptions from data. Knowledge discovery in spatial databases represents a particular case of discovery, allowing the discovery of relationships that exist between spatial and non-spatial data, and other data characteristics that aren't explicitly stored in spatial databases.

This paper describes the conception and implementation of PADRÃO, a system for knowledge discovery in spatial databases. PADRÃO presents a new approach to this process, which is based on qualitative spatial reasoning. The *spatial semantic knowledge* and the *principles of qualitative spatial reasoning* needed for the spatial reasoning process are available in the PADRÃO's *geographic database* and PADRÃO's *spatial knowledge base*, allowing the integration of the geo-spatial component, associated with the analysed non-geographic data, in the process of knowledge discovery.

INTRODUCTION

Large amounts of operational data concerning several years' operation are now becoming available, mainly in middle-large sized organisations. Knowledge Discovery in Databases (KDD) is the key to access the strategic valued of the organisational knowledge buried in databases, usable both for daily operation, general management and strategic planning.

The process of KDD automates the discovery of relationships and other descriptions from data. Data mining is one of the steps of this process, concerned with the application of specific algorithms for extracting patterns from data (Fayyad et al., 1996b). The main recognised advances in the area of KDD (Fayyad et al., 1996a) are related with the exploration of relational databases. However, in most organisational databases exists one dimension of data, the *geographic* (associated with addresses or postcodes), which semantics is not used by traditional KDD systems.

Knowledge Discovery in Spatial Databases (KDSD) is related with "*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*" (Koperski and Han, 1995).

Spatial database systems are normally 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. 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 takes into consideration this spatial reasoning, motivating the development of new algorithms adapted to the 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 (Ester et al., 1998, Lu et al., 1993, Koperski and Han, 1995, Koperski et al., 1998). 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 (like *far, close, North, ...*).

This paper describes the conception and implementation of PADRÃO, a system for KDSD. PADRÃO presents a new approach to the process of KDSD based on qualitative spatial reasoning and was implemented recurring to a traditional knowledge discovery system. The *spatial semantic knowledge* and the *principles of qualitative spatial reasoning* needed for the spatial reasoning process are available in the PADRÃO's *geographic database* and PADRÃO's *spatial knowledge base*, allowing the integration of the data geo-spatial component in the process of knowledge discovery.

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 one district of Portugal, allowed to PADRÃO the discovery of implicit relationships existing between the analysed geographic and demographic data.

This paper is organised in several sections. In them, qualitative spatial reasoning is defined and described how its concepts are used in the knowledge discovery process. The architecture of PADRÃO' is presented, describing its main components, the several steps associated with it, and its implementation. The application of PADRÃO to the demographic domain is also illustrated, referring the type of discoveries that can be achieved with it.

QUALITATIVE SPATIAL REASONING

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 (*indirect* referencing systems), 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, 1998). These geographic identifiers are very common in organisational databases, allowing the integration of the spatial component associated with it in the process of knowledge discovery.

The adoption of an indirect geographic reference system imposes the use of *qualitative spatial reasoning* strategies, able to deal with the spatial semantic not explicitly associated with the adopted geographic identifiers. *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 (Sharma, 1996). *Qualitative spatial reasoning* (Abdelmoty and El-Geresy, 1995) is based on the manipulation of

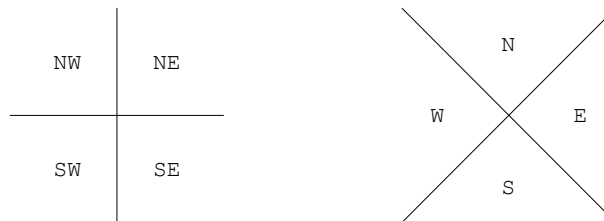
qualitative spatial relations, for which composition tables facilitate reasoning, allowing the inference of new spatial knowledge.

Spatial relations have been classified in several types (Frank, 1996, Papadias and Sellis, 1994), including *direction* relations (Frank, 1996, Freksa, 1992) (that describe order in space), *distance* relations (Hernández et al., 1995) (that describe proximity in space) and *topological* relations (Egenhofer, 1994) (that describe neighbourhood and incidence). These spatial relations are briefly described in the next subsections.

Direction relations

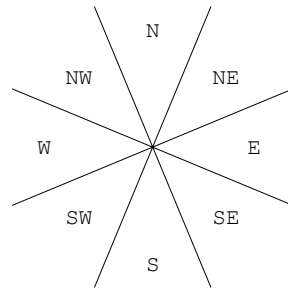
Direction relations describe where objects are placed relative to each other. Three elements are needed to establish an orientation: two objects and a fixed point of reference (usually the North Pole) (Frank, 1996, Freksa, 1992). Cardinal directions can be expressed using numerical values specifying degrees (0° , 45° ...) or using qualitative values or symbols, such as North or South, those have an associated acceptance region. The regions of acceptance for qualitative directions can be obtained by projections (also known as half-planes) or cone-shaped regions (Figure 1).

Figure 1: Directions definition by projection and cone-shaped systems



A characteristic of the cone-shaped system is that the region of acceptance increases with distance, which makes it suitable for the definition of directional relations between extended objects (Sharma, 1996). Another benefit is that this system allows the definition of finer resolutions, permitting the use of eight (Figure 2) or sixteen different qualitative directions. This model uses triangular acceptance areas that are drawn from the *centroid* of the reference object towards the primary object (In the spatial relation A North B, B represents the reference object, while A constitutes the primary object). Since the geographic domain analysed in this work characterises administrative subdivisions, the cone-shaped system will be used in the identification of the direction relations existing between the several geographic subdivisions.

Figure 2: Cone-shaped system with eight regions of acceptance

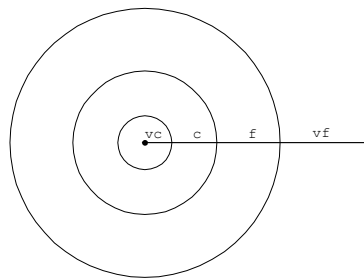


Distance relations

Distances are quantitative values determined through measurements or calculated from known co-ordinates of two objects in some reference system. The most familiar definition of distance is the length of the shortest possible path between two objects, also known as Euclidean distance. Usually a metric quantity is mapped onto some qualitative indicator such as *very close* or *far* for human common-sense reasoning (Hong, 1994, Hernández et al., 1995, Zimmermann, 1995).

Qualitative distances must correspond to a range of quantitative distances specified by an interval, and should be ordered so that comparisons are possible. Another requirement is that the length of each successive qualitative distance should be greater or equal to the length of the previous one (Figure 3).

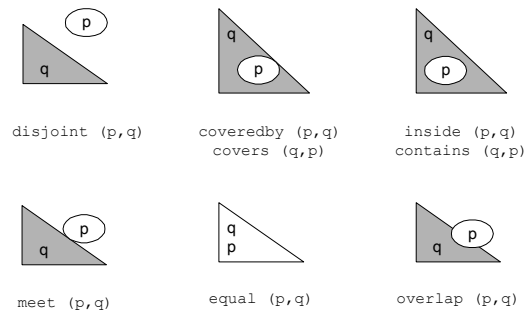
Figure 3: Qualitative distances



Topological relations

Topological relations are those relations that are invariant under continuous transformations of space such as rotation or scaling. There are eight fundamental relations that can exist between two planar regions: *disjoint*, *contains*, *inside*, *equal*, *meet*, *covers*, *covered by* and *overlap* (Figure 4). These relations can be defined considering intersections between the two regions, their boundaries and their complements (Egenhofer, 1994).

Figure 4: Topological relations



In some exceptional cases, the geographic space can't be characterised, in topological terms, recurring to the eight primitives presented above. One of these cases is related with application domains in which the addressed geographic regions are administrative subdivisions. Administrative subdivisions can only be related through the topological primitives *disjoint*, *meet* and *contains* (and its inverse *inside*), since they can't have any kind of overlapping. The topological primitives used in this work are *disjoint* and *meet*, once the implemented qualitative inference process only considers regions of the same geographic level.

Integration of direction, distance and topological spatial relations

Reasoning about qualitative directions necessarily involves integrated spatial reasoning about qualitative distances and directions. Particularly in objects with extension, the size and shape of objects, and the distance between them, influence the directions. One of the ways to determine the direction and distance¹ between regions is calculating them for its respective *centroids*. The extension of the geographic entities is somehow implicit in the topological primitive used to characterise its relations.

Qualitative spatial reasoning requires the adoption of a set of qualitative identifiers. In the implemented approach, integrated spatial reasoning with direction, distance and topological spatial relations, the adopted set of qualitative identifiers were:

Direction = {N, NE, E, SE, S, SW, W, NW}

Distance = {very close, close, far, very far}

Topological = {disjoint, meet}

For each of these qualitative identifiers (direction and distance, since topological relations are quantitatively by nature), is required the definition of the validity interval that limits quantitatively the region of acceptance of each one. In the case of

¹ Defining distances between regions is a complex task, since the size of each object plays an important role in determining the possible distances. Sharma (Sharma 1996) enumerates as possible ways to the definition of distances between regions: i) taking the distance between the *centroids* of the two regions; ii) determining the shortest distance between the two regions; or iii) determining the furthest distance between the two regions.

direction relations, for the cone-shaped system with eight acceptance regions, the quantitative intervals adopted are: [337.5, 22.5), [22.5, 67.5), [67.5, 112.5), [112.5, 157.5), [157.5, 202.5), [202.5, 247.5), [247.5, 292.5), [292.5, 337.5) from N to NW respectively. In the case of distances, there should exist a constant ratio ($\text{ratio} = \text{length}(\text{dist}_i) / \text{length}(\text{dist}_{i-1})$) relationship between the lengths of two neighbouring intervals (Hong, 1994). For example, for a ratio 4^2 the obtained intervals are:

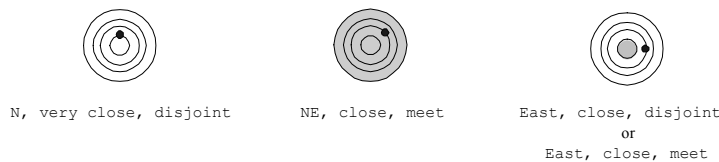
Ratio	dist ₀	dist ₁	dist ₂	dist ₃
4	(0, 1]	(1, 5]	(5, 21]	(21, 85]

Analysing the furthest distance that can exist between two regions of the addressed geographic domain, the ratio 4 intervals need to be magnified³ by a factor of 10, resulting in the validity intervals: (0, 10], (10, 50], (50, 210] and (210, 850] corresponding from very close to very far respectively.

Qualitative spatial reasoning with the adopted identifiers requires the construction of a composition table that aggregates a set of rules, used in the inference of new spatial relations. Qualitative rules can be constructed using quantitative methods (Hong, 1994) or manipulating qualitatively the set of identifiers adopted (Sharma, 1996) (through the definition of axioms and properties for the spatial domain).

The adoption of a mixed approach (Santos, 2000) allowed the integration of direction, distance and topological relations, under the principles of qualitative spatial reasoning. The composition table is represented recurring to graphical symbols, like the presented in Figure 5. In order to exemplify the set of rules stored in the final composition table, Table 1 exhibits a subset of the rules contained in it.

Figure 5: Graphic symbols used in the integration of direction, distance and topological relations



The set of rules stored in the composition table (Table 1) can be used in the inference of new spatial relations needed in the knowledge discovery process. For example, knowing the facts A Northeast, very close, meet B and B East, close, disjoint C, it can be inferred that A East, close, disjoint C, as can be verified in Figure 6.

² Others validity intervals, for different ratios, can be found in (Hong, 1994) or (Santos and Amaral, 2000).

³ Since the same scale magnifies all intervals and quantitative distances, qualitative inferences will remain the same (Hong, 1994).

Table 1: Excerpt of the final composition table




























































































































































































































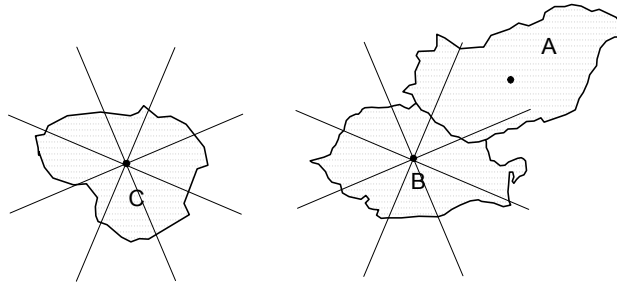
																			
																			
																			
																			
																			
																			
																			
																			
																			
																			
																			

Figure 6: Example of the inference process



THE PADRÃO SYSTEM

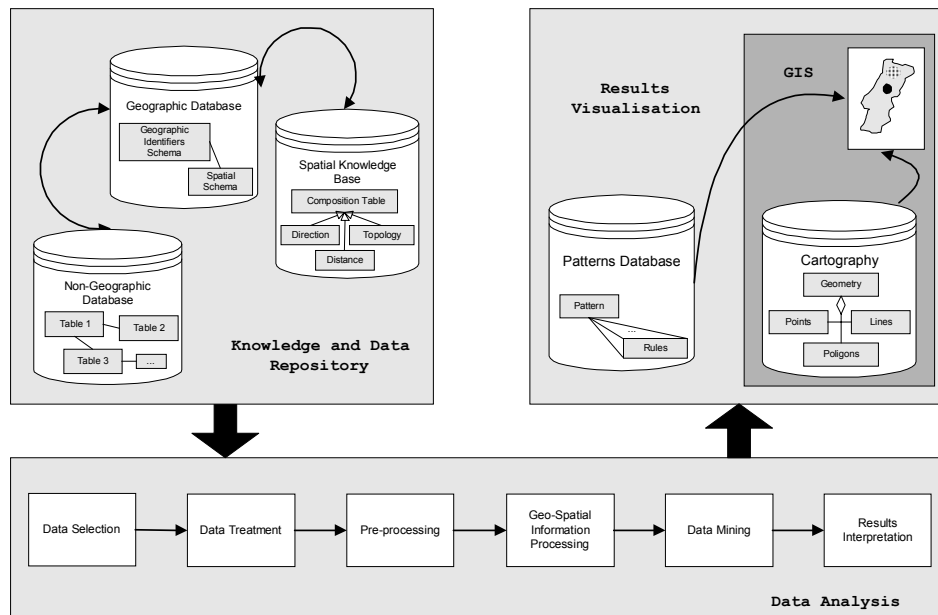
PADRÃO is a system for KDSD based in qualitative spatial reasoning. This section presents its architecture and gives some technical details about its implementation.

The PADRÃO's architecture

The architecture of PADRÃO (Figure 7) aggregates three main components: Knowledge and Data Repository, Data Analysis and Results Visualisation. The Knowledge and Data Repository component stores the data and knowledge needed in the knowledge discovery process. This process is implemented in the Data Analysis component, which allows the discovery of patterns or others relationships implicit in the analysed geographic and non-geographic data. The discovered

patterns can be visualised in a map using the Results Visualisation component. These components are afterwards described.

Figure 7: PADRÃO's Architecture



The **Knowledge and Data Repository** component group three central databases:

1. A **Geographic Database (GDB)** constructed under the principles established by the European Committee for Normalisation in the CEN TC 287 standard for Geographic Information. Following its recommendations was possible to implement a GDB in which the positional aspects of geographic data are provided by a *geographic identifiers system* (CEN/TC-287, 1998). This system characterises the administrative subdivisions of Portugal at the municipality and district level. Also includes a geographic gazetteer with the several geographic identifiers used and the concept hierarchies existing between them. The geographic identifiers system was integrated with a spatial schema (CEN/TC-287, 1996) allowing the definition of the *direction*, *distance* and *topological* spatial relations that exist between the adjacent regions of the municipality level.
2. A **Spatial Knowledge Base (SKB)** that stores all the qualitative rules needed in the inference of new spatial relations. The knowledge available in this database aggregates the composition table constructed, the set of used identifiers and the validity interval of them. This knowledge base is used in conjunction with the GDB in the inference of implicit spatial relations.
3. A **non-Geographic Database (nGDB)** that is integrated with the GDB and analysed in the Data Analysis component. This procedure enables the discovery of implicit relationships that exist between the geographic and non-geographic data.

The **Data Analysis** component is implemented through the knowledge discovery module, and is characterised by six main steps:

1. **Data Selection.** This step allows the selection of the relevant non-geographic and geo-spatial⁴ data needed for the execution of a defined data mining task. In this phase must be evaluated what is the minimal sub-set of data to be selected, the size of the sample needed and the period of time to be considered.
2. **Data Treatment.** This phase is concerned with the cleaning of the selected data, allowing the corrupt data treatment and the definition of strategies for dealing with missing data fields.
3. **Data Pre-Processing.** This step allows the reduction of the sample set to be analysed. Two tasks can here be effected: i) reduction of the number of rows or, ii) reduction of the number of columns. In the reduction of the number of rows, data can be generalised attending to the domain's hierarchies or attributes with continuous values can be transformed into discreet values attending to the defined classes. The reduction of the number of columns attends to verify if any of the selected attributes can afterwards be omitted.
4. **Geo-Spatial Information Processing.** This step verifies if the geo-spatial information needed is available in the GDB. In many situations it implicitly exists, due to the properties of the spatial schema implemented. In this case, and to ensure that all geo-spatial knowledge is available for the data mining algorithms, these implicit relations are transformed into explicit relations through the inference rules stored in the SKB.
5. **Data Mining.** Several algorithms can be used for the execution of a given data mining task. In this step, the several available algorithms are evaluated in order to identify the most appropriate for the defined task. The selected one is applied to the relevant non-geographic and geo-spatial data, in order to find implicit relationships or other interesting patterns that exist between them.
6. **Results Interpretation.** The interpretation of the discovered patterns aims to evaluate their utility and importance to the application domain. In this step it may be realised that relevant attributes were ignored in the analysis, suggesting that the process should be repeated. The relevant discoveries can be stored in the database of patterns, allowing its subsequent use in further analyses (meta-rules construction) or its visualisation in a map.

The **Results Visualisation** component is responsible for the management of the discovered patterns and its visualisation in a map. For that, PADRÃO uses a Geographic Information System (GIS), integrating the discovered patterns with the cartography of the analysed region. This component aggregates two main databases:

1. The **Patterns Database** (PDB) that stores all relevant discoveries. In this database, each discovery is catalogued and associated with the set of rules that represents the discoveries made in a given data mining task.

⁴ The term *geo-spatial* is used to emphasise that geographic data has associated spatial data. Traditionally, geographic data is associated with a location, positioning an object or fact in space, while spatial data defines the characteristics of that localisation, namely its geometry and topology.

2. A **Cartographic Database** (CDB) with the cartography of the region. It aggregates a set of points, lines and polygons with the geometry of the geographical objects.

The PADRÃO's implementation

PADRÃO has been implemented using *Microsoft Access*, the relational database system; *Clementine* (SPSS, 1999), the knowledge discovery tool, and *Geomedia Professional* (Intergraph, 1999), the GIS used for the graphical representation of results.

The overall set of databases that integrates the Knowledge and Data Repository and the Results Visualisation components were implemented in *Access*. The data stored in them are available to the Data Analysis component, or from it, through ODBC (Open Database Connectivity) connections.

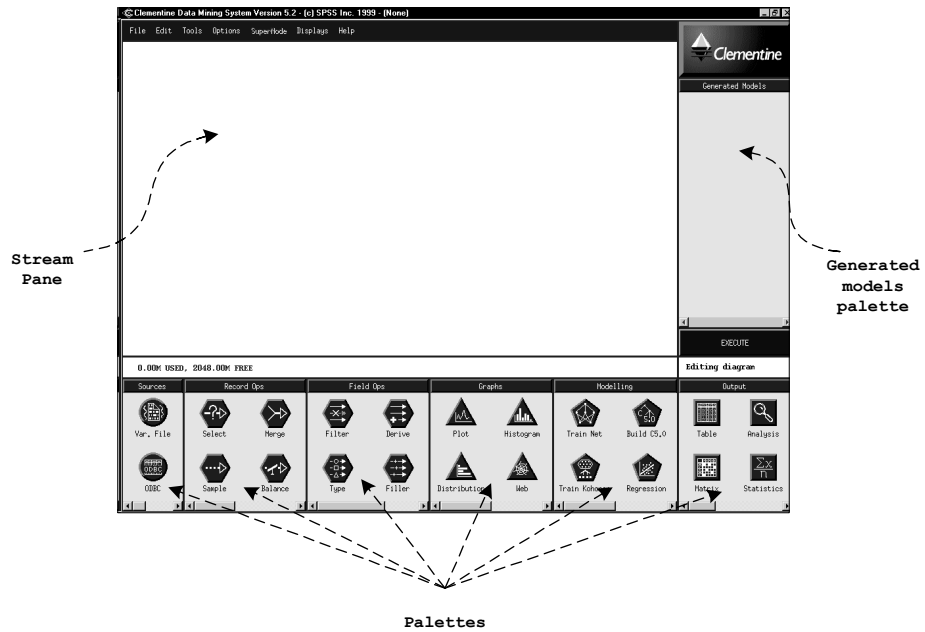
Clementine is a data mining toolkit based on visual programming⁵, which include machine learning technologies like rule induction, neural networks, association rule discovery and clustering. The knowledge discovery process is defined in *Clementine* through the construction of a *stream*, in which each operation on data is represented by a *node*.

Figure 8 presents a view of the workspace of *Clementine*, in which it is possible to see several sets of icons, grouped into palettes according to their functions: links to sources of information, operations on data (rows or columns), visual facilities and modelling techniques (data mining algorithms). The figure already points out the space used to the streams construction (stream pane, the main work area) and the generated models palette.

The PADRÃO's Data Analysis component recurs to the construction of several streams that implement the knowledge discovery module. The several models obtained in the data mining phase represent knowledge about the analysed data, and can be saved or reused in others streams. In PADRÃO, these models can be exported through an ODBC connection to the PDB. Its integration with the CDB allows its visualisation in a map.

⁵ Visual programming involves placing and manipulating icons representing processing nodes.

Figure 8: Clementine's workspace



THE KNOWLEDGE DISCOVERY PROCESS IN PADRÃO

The nGDB, of the Knowledge and Data Repository component, used in this application of PADRÃO is a Demographic Database (DDB) that stores the parish registers dated between 1690 and 1990 in the *Aveiro* district. This database collects attributes like birth date, birthplace, death date, death place, occupation, number of descendants, number of marriages, etc. The several attributes related to places allow the integration of the DDB with the GDB, providing the geo-spatial data needed in the knowledge discovery module.

Table 2 presents some records of the `individual` table, in which is possible to see the existence of some missing data fields and attributes with continuous values, requiring its transformation into discreet values. At the geographical level, the demographic data will be generalised at the municipality level. Following some suggestions (Rodrigues et al., 1999) of possible concept hierarchies for demographic data, Table 3 systematises the several hierarchies used, and the classes defined for the transformation of continuous attributes into discreet values.

Table 2: Some records of the individual table

Num	Name	S	Birth date	Birth place	Died	Died place	Occupation:	M	Ch
6224	JOAO ANTONIO	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 Pica	Oleira	1	8
6233	ANTONIO DA M		24-05-1790	Quinta do Pica	16-09-1864	Quinta do Pica	Oleiro	2	10
6235	JOSE FRANC	M	28-05-1790	Quinta do Pica	05-10-1849	Verdemilho	Lavrador	1	10
6239	MANUEL FR	M	03-08-1790	Quinta do Pica	01-08-1830	Quinta do Pica	Lavrador	1	7
6241	ROSA DOS S	F	25-08-1790	Bom Sucesso	27-08-1830	Quinta do Pica	Lavadora	1	7
6249	MANUEL JO	M	25-09-1790	Verdemilho	20-03-1841	Verdemilho	Lavrador	1	10
6250	ANTONIO SIM	M	21-09-1790	Arada	07-03-1874	Arada	Lavrador	1	9
6253	JOANA MARI	F	31-10-1790	Verdemilho	06-03-1863	Verdemilho	Lavadora	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	Lavadora	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	Lavadora	1	8
6267	PERPETUA	F	06-03-1791	Verdemilho	10-12-1855	Verdemilho	Lavadora	1	10
6288	MARIA DE J	F	06-06-1791	Arada	24-03-1877	Arada	Jornaleira	0	0
6299	JOSEFA DE	F		Quinta do Pica	30-03-1836	Quinta do Pica	Lavadora	1	9
6314	JOANA TERE	F	05-11-1791	Arada	17-04-1870	Arada	Mendicante	1	4
6331	MAURICIO F	M	09-06-1792	Quinta do Pica	27-02-1858	Quinta do Pica	Lavrador	1	7
6335	MARIA ROSA	F	07-09-1792	Quinta do Pica	20-05-1870	Quinta do Pica	Lavadora	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	Lavadora	1	11

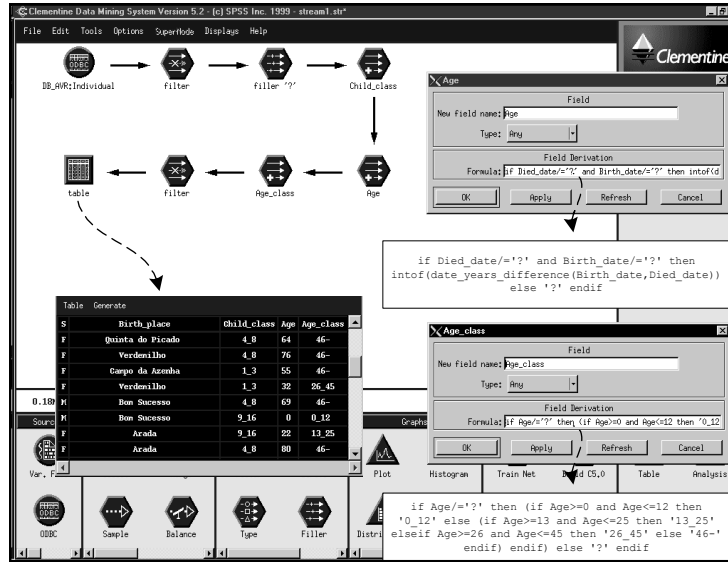
Table 3: Hierarchies and classes for data reduction

Attributes	Hierarchies / Classes for 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-	

The definition of an objective for the discovery requires its transformation in a data mining objective. In the example presented here, the objective was to find a model that characterise the age at death and the number of children attributes for the analysed geographic region. Transforming this objective in a data mining objective begins with the *selection*, *treatment* and *pre-processing* of the relevant data. After the data selection, all missing data fields were marked as unknown ('?') and the continuous attributes were transformed into discreet values attending the classes previously described. Figure 9 presents the stream charged with the execution of these three steps. By its analysis it is possible to verify that the first node (DB_AVR:Individual) makes the data available through an ODBC connection. After that, the filter node permits the selection of the relevant attributes and the filler node substitutes all missing data fields by a '?'. The Child_class node assigns the correspondent class to the number of Children attribute; the Age node calculates the age at death of the individuals, based on the attributes Birth_date and Died_date. The result is used by the Age_class node, which assigns to each value the respective class. The figure also presents the

CLEM⁶ code used in two of the nodes, the Age and Age_class nodes, and a table with a set of the resulting records.

Figure 9: Selection, treatment and pre-processing of data



The next step is concerned with the *geo-spatial information processing*. As the GDB only stores spatial relations for adjacent regions, and we want to verify the geographical distribution of the age at death and number of children attributes, all the others relations, existing between non-adjacent regions and needed in the data mining step, must be inferred. First, Clementine must learn the composition table stored in the SKB and that allows the inference of new spatial relations. In Clementine, a rule induction⁷ algorithm learnt the inference rules available in the composition table that allows integrated qualitative spatial reasoning. Figure 10 presents the stream constructed for the learning process and some of the rules of the decision trees.

The created models, *infDir*, *infDis* and *infTop*, can now be used in the inference process. For that, another stream was constructed. This stream, presented in Figure 11, is cyclically executed until no more inferences have to be done. The available spatial relations are gathered from the *Regions* table and are combined recurring to the *Inflation*⁸ node, ensuring that new inferences can be performed.

⁶ CLEM, the *Clementine Language for Expression Manipulation*, is a language for manipulating the data that flows along Clementine streams.

⁷ A rule induction algorithm creates a decision tree aggregating a set of rules for classify the data into different outcomes. This technique only includes in its rules the factors that really matter in decision making.

⁸ This node represents an external program (an executable module, .exe) that was implemented for the combination of the existing spatial relations. It combines non-adjacent regions, linked through some existing third region, for which the spatial relations will be inferred. This procedure was necessary since array's manipulation capabilities were needed and at this moment not available in the used

The obtained inferences are stored in the GDB, upgrading the available geographical knowledge. Figure 11 also presents the Clementine's script that allows the stream cyclical execution.

Figure 10: Learning process of the spatial inferences rules

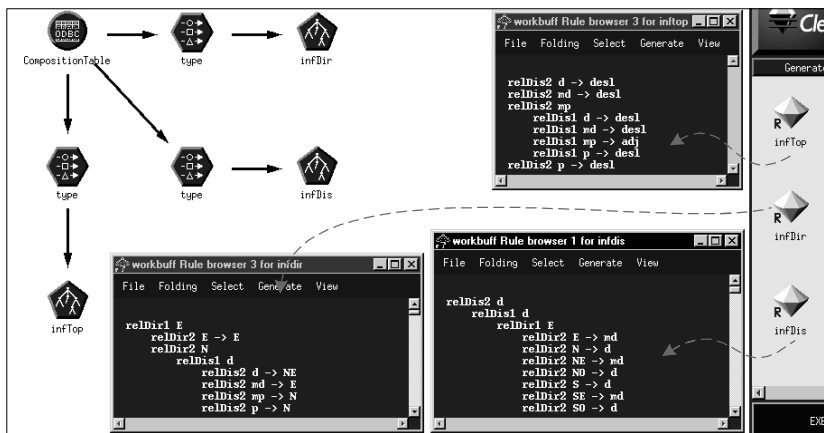
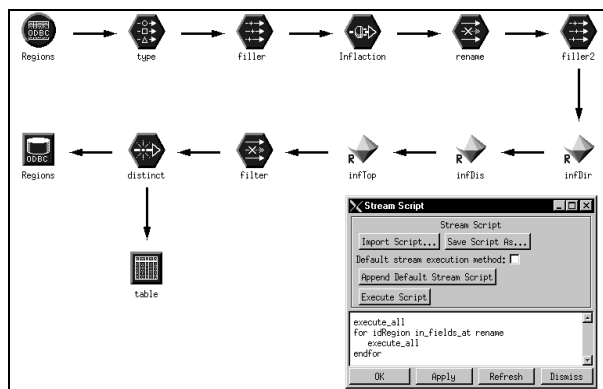


Figure 11: Cyclically inference process of new spatial relations



The knowledge discovery process proceeds with the construction of the geographical model of the region. This model describes the localisation of each municipality in the *Aveiro* district. After that, it will be used by the *data mining* algorithm, in order to integrate the geographic component in the analysis. Figure 12 presents the two streams constructed for this propose. The stream located at the right side of the figure selects, from the *Regions* table, all the geographic information related with the *Aveiro* district. The selected records are analysed by the C5.0⁹ algorithm, constructing the district geographical model (*geo_AVR*). The obtained

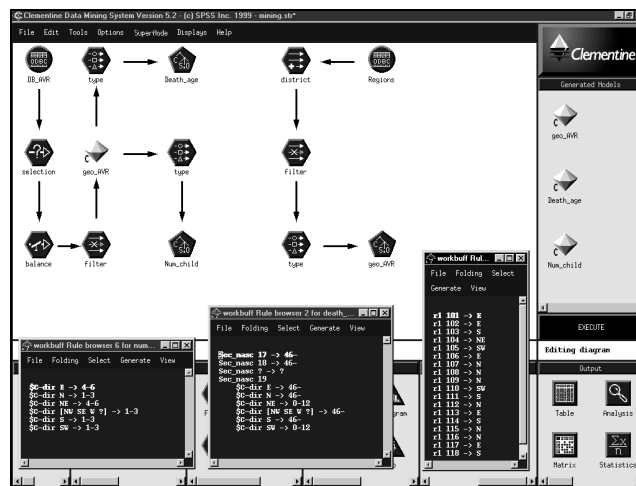
Clementine's version. The Inflation node was constructed recurring to a Clementine's *specification file*, which makes it available as one of the nodes of the *Record Ops* palette.

⁹ The C5.0 algorithm is a rule induction algorithm that generates decision trees or rule sets, predicting the value of an output field. It can only be used with symbolic outputs.

model (browsed in the tabular table at the right of the figure) was afterwards used in the other stream, allowing the geographical characterisation of the number of children and the age at death attributes.

Analysing the obtained results for the age at death attribute (tabular table at middle of the figure), only in the XIX century exist a pattern 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 number of children attribute (tabular table at left of the figure) point out that regions with a high birth rate fall in the Northeast and East of the analysed district.

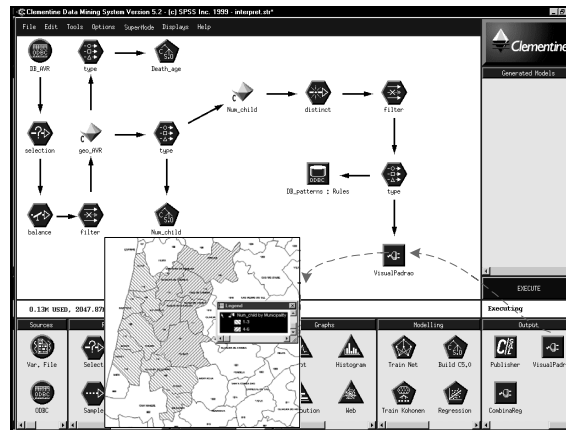
Figure 12: Geographical characterisation of the age at death and number of children attributes



The last step of the Data Analysis component is related with the *interpretation of the discovered patterns*, verifying the relevance of each one for the application domain. The desired patterns can be stored in the PDB, allowing its visualisation in a map. The transference process of the discovered rules only requires that the desired pattern (the generated model) be linked to the stream that originated it and that be connected to an output ODBC node, in order to export the rules to the PDB. To visualise the model in the GIS, the user only needs to use the VisualPadrao¹⁰ node available in the Output palette and the rules are automatically represented in the map of the analysed region.

¹⁰ This node was again created through a Clementine's specification file, and is available in the Output palette. It receives the rules contained in the generated model and opens the GIS to its visualisation.

Figure 13: Storage of the discovered rules in the PDB and its visualisation in a map



SUMMARY

This paper presented the PADRÃO system, a system for knowledge discovery in spatial databases based on qualitative spatial reasoning. PADRÃO represents a new approach to this particular case of knowledge discovery, in which the positional aspects of geographic data are provided by a spatial reference, given by a geographic identifier (in the described case, the municipalities' name).

The PADRÃO's *geographic database* and *spatial knowledge base* store the spatial semantic knowledge and the qualitative spatial reasoning principles needed for the inference of new spatial relations required in the knowledge discovery process. The analysis of a *demographic database* with PADRÃO allowed the discovery of implicit relationships that exist between the analysed demographic and geo-spatial data.

The data analysis techniques described along this document constitute a special case of *Spatial Analysis* and are useful in the *Urban and Regional Research*. This application domain analyses huge amounts of data that are related with some place on the Earth's surface. PADRÃO has the capability to analyse databases of great dimension. These databases usually store geo-referenced data, which spatial component is included in the spatial analysis process of PADRÃO using qualitative spatial reasoning strategies.

Acknowledgements

This work has been partially supported by a Portuguese grant from PRODEP II (*acção 5.2, Concurso n°3/98 Doutoramentos*). Also, our acknowledgements to NEPS (*Núcleo de Estudos da População e Sociedade*), of University of Minho, who provide us the original demographic data.

References

- Abdelmoty, A. I. and El-Geresy, B. A. (1995): *A General Method for Spatial Reasoning in Spatial Databases*, Proceedings 4th International Conference on Information and Knowledge Management, Baltimore, 312-317.
- CEN/TC-287 (1996): *Geographic Information: Data Description, Spatial Schema*, European Committee for Normalisation, European pre-standard prENV 12160.

- CEN/TC-287 (1998): Geographic Information: Referencing, Geographic Identifiers, European Committee for Normalisation, European pre-standard prENV 12661.
- Egenhofer, M. J. (1994): Deriving the Composition of Binary Topological Relations, *Journal of Visual Languages and Computing*, **5**, 133-149.
- Ester, M., Frommelt, A., Kriegel, H.-P. and Sander, J. (1998): *Algorithms for Characterization and Trend Detection in Spatial Databases*, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.
- Ester, M., Kriegel, H.-P. and Sander, J. (1997): *Spatial Data Mining: A Database Approach*, Proceedings of the 5th International Symposium on Large Spatial Databases, Berlin, Germany.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a): The KDD process for extracting useful knowledge from volumes of data *Communications of the ACM*, **39**, 27-34.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.) (1996b) *Advances in Knowledge Discovery and Data Mining*, The MIT Press, Massachusetts.
- Frank, A. U. (1996): Qualitative Spatial Reasoning: cardinal directions as an example, *International Journal of Geographical Information Systems*, **10**, 269-290.
- Freksa, C. (1992): *Using Orientation Information for Qualitative Spatial Reasoning* In *Theories and Methods of Spatio-Temporal Reasoning in Geographic space*, Lectures Notes in Computer Science 639 (Eds, Frank, A. U., Campari, I. and Formentini, U.), Springer-Verlag, Berlin.
- Hernández, D., Clementini, E. and Felice, P. D. (1995): *Qualitative Distances* In *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. and Kuhn, W.) Springer-Verlag, Austria, 45-57.
- Hong, J.-H. (1994): *Qualitative Distance and Direction Reasoning in Geographic Space*, PhD thesis, University of Maine.
- Intergraph (1999): *Geomedia Professional v3, Reference Manual*, Intergraph Corporation.
- Koperski, K. and Han, J. (1995): *Discovery of Spatial Association Rules in Geographic Information Systems*, Proc. 4th International Symposium on Large Spatial Databases (SSD95), Maine, 47-66.
- Koperski, K., Han, J. and Stefanovic, N. (1998): *An Efficient Two-Step Method for Classification of Spatial Data*, Proceedings of the International Symposium on Spatial Data Handling (SDH'98), Canada.
- Lu, W., Han, J. and Ooi, B. C. (1993): *Discovery of General Knowledge in Large Spatial Databases*, Proc. of the 1993 Far East Workshop on Geographic Information Systems, Singapore, 275-289.
- Papadias, D. and Sellis, T. (1994): On the Qualitative Representation of Spatial Knowledge in 2D Space *Very Large Databases Journal, Special Issue on Spatial Databases*, **3**, 479-516.
- Rodrigues, M. F., Ramos, C. and Henriques, P. R. (1999): *An Intelligent System to Study Demographic Evolution*, Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology, Orlando, Florida, 161-170.
- Santos, M. (2000): *Padrão: Um sistema para a descoberta de conhecimento em bases de dados georeferenciadas (in Portuguese)*, PhD thesis (in finalisation), University of Minho.
- Santos, M. and Amaral, L. (2000): *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, Manchester, 73-88.
- Sharma, J. (1996): *Integrated Spatial Reasoning in Geographic Information Systems: Combining Topology and Direction*, PhD thesis, University of Maine.
- SPSS (1999): *Clementine, User Guide, Version 5.2*, SPSS Inc.
- Zimmermann, K. (1995): *Measuring without Measures: The Δ -Calculus* In *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. and Kuhn, W.) Springer-Verlag, Austria, 59-67.