MINING CENSUS AND GEOGRAPHIC DATA IN URBAN PLANNING ENVIRONMENTS

Donato Malerba, Francesca A. Lisi, Annalisa Appice and Francesco Sblendorio
Dipartimento di Informatica – Università degli Studi di Bari
Via Orabona 4 – 70126 BARI
{malerba | lisi | appice | sblendorio}@di.uniba.it

ABSTRACT

Urban planning is a knowledge-intensive activity. It acquires useful knowledge by analysing jointly socio-economic data and topographic maps. When supported by computers, this preliminary information gathering triggers a knowledge discovery process, which often consists of exploratory data analysis tasks. Advances in geo-referencing have caused - among the other things - a growing demand for more powerful exploratory data analysis techniques. We resort to the field of spatial data mining and propose the task of mining spatial association patterns/rules, i.e. frequent associations between spatial objects, as a means for data exploration. Strong points of the proposed technique for this task are the power of logics as a knowledge representation and reasoning means and the capability of discovering associations at multiple levels of description granularity. This enables new interesting exploratory tasks for spatial data. The technique has been implemented in SPADA and tested on census and geographic data of Stockport, one of the ten metropolitan districts of Greater Manchester, UK. We show results obtained by applying SPADA to a problem of urban accessibility of the Stepping Hill hospital in Stockport.

Key words: knowledge discovery, spatial databases, machine learning

1. INTRODUCTION

Urban planners, who typically work to integrate knowledge with action in the pursuit of more just, efficient and sustainable cities and environment, are used to acquire the necessary knowledge by analysing jointly socioeconomic data and topographic maps. This has been the practice in urban planning environments for centuries. Indeed population and economic census data overlapped to geographic data can be the key indicator of 'optimum' locations for public services, thus supporting a good public policy. For instance, population census data such as car availability and topographic maps such as public transport networks are core ingredients in studies of accessibility of public services.

Currently there is a growing demand for computer tools that can link population data to their spatial, or, more precisely, geographical distribution. Indeed geo-referencing have enabled the spatial representation of socio-economic phenomena as spatial objects in the sense of entities having both spatial location and spatially independent attribute characteristics (Martin, 1999). In the UK for instance the geo-referencing units for population census data are the areal objects ED (enumeration district), ward, district, and county, of which ED is the smallest unit for which data is published. Advances in geo-referencing have provided an added impetus to the development of Geographical Information Systems (GIS). GIS are increasingly used to store, manipulate and analyze physical, social and economic data of a geographic area in order to provide the information necessary for effective decision-making in urban planning (Han, Kim, 1989). As a toolbox, GIS allow

planners to perform spatial analysis using geo-processing functions such as map overlay or connectivity measurements. Yet GIS do not adequately support the spatial decision process because they lack of appropriate modeling capabilities (Densham, 1991; Keenan, 1998). To assist the work of public and private planners, the range of GIS applications can be greatly extended by adding some data analysis capabilities. Exploratory data analysis goes 'beyond the map' to use statistical tools in an informal, pattern-seeking vein (Anselin, 1993). More powerful data analysis techniques are actually needed to support knowledge-intensive activities such as urban planning.

In this paper we resort to the field of *spatial data mining* that investigates how implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial data can be extracted (Koperski et al., 1996). Knowledge discovered from spatial data can be in various forms including classification rules for the recognition of geographical objects (Koperski et al., 1998) or the interpretation of topographic maps (Malerba et al., 2001), clusters of spatial objects (Ng, Han, 1994; Sander et al., 1998), patterns describing spatial trends, that is, regular changes of one or more non-spatial attributes when moving away from a given start object (Ester et al., 1998), and subgroup patterns, which identify subgroups of spatial objects with an unusual, an unexpected, or a deviating distribution of a target variable (Klösgen, May, 2002). In this paper we propose *spatial association patterns/rules* as an advanced means for the exploration of geo-referenced data. In GIS terminology, a spatial pattern is a pattern showing the interaction of two or more spatially-referred objects or space-dependent attributes, according to a particular spacing or set of arrangements (DeMers, 2000). In the data mining context, a spatial association rule is a frequent association involving spatial relations between spatial objects playing the role of *reference objects* and spatial objects from different geographic layers playing the role of *task-relevant objects* (Koperski, Han, 1995). For instance, reference objects can be large towns and task-relevant objects can be spatial objects belonging to the map layers of road network and hydrography within a given geographic area. SPADA is a system for mining spatial association rules (Lisi, Malerba, 2002). It adopts logics as a knowledge representation and reasoning means and exploits possibly available spatial hierarchies to discover associations at multiple levels of description granularity. Results of previous experiments in urban planning environments are reported in (Malerba et al., 2002).

In this paper, we present an application of SPADA to a case study of urban accessibility of health care services. The area under analysis is Stockport, one of the ten metropolitan districts of Greater Manchester, UK. In particular, we study the access to Stepping Hill where a big hospital is located. We have chosen this case study for two reasons. First, it relies heavily on the aforementioned joint analysis of census and geographic data. Second, it is interesting from the twofold perspective of transportation planning and health care planning. Indeed accessibility of health care services is an issue for both planning activities. Several studies of this kind are reported in the literature. Love and Lindquist (1995) argue that crude access measures based on data for large areal units are no longer necessary given developments in GIS and spatially referenced data. In their study on the accessibility of centroid-referenced census blocks to point-referenced hospitals in Illinois (USA), they use simple accessibility measures based on straight line distances to the nearest five hospitals for each block. Jones and Bentham (1995) use GIS and logistic regression to test for a relationship between health outcomes and accessibility. Using grid-referenced police records on serious and fatal road traffic accidents and a digital road network for Norfolk, England, a routing procedure simulated the dispatch of ambulances from their stations to accident sites and then to hospitals, and estimated associated travel times. However all

these studies suffer from the limitations of the data analysis techniques they apply. Some of these limits are overcome in our approach.

The paper is organized as follows. In the next section, the system SPADA is briefly described. Section 3 illustrates the application of SPADA to Stockport data by giving details of the database, explaining the interface of SPADA with the database and reporting on the results of an accessibility study. Section 4 concludes the paper with remarks and future directions of work.

2. MINING SPATIAL ASSOCIATION RULES WITH SPADA

2.2. The method

The discovery of spatial association rules is a descriptive mining task aiming at the detection of associations between *reference objects* and some *task-relevant objects*. The former are the main subject of the description, while the latter are spatial objects that are relevant for the task in hand and are spatially related to the former. For instance, we may be interested in looking for spatial association rules that relate properties of some selected EDs (reference objects) with properties of other spatial objects, such as public transport stops (task-relevant objects) in order to investigate the socioeconomic phenomenon of deprivation of some urban areas.

In general, association rules are a class of regularities introduced by Agrawal et al., (1993) that can be expressed by the implication:

$$P \mathbin{®} Q \text{ (s, c)},$$

where P and Q are a set of literals, called *items*, such that $P \cap Q = \varnothing$, the parameter s, called *support*, estimates the probability $p(P \cup Q)$ of the underlying *pattern* $P \cup Q$, and the parameter c, called *confidence*, estimates the probability $p(Q|P)$.

We call an association rule $P \mathbin{®} Q$ *spatial*, if $P \cup Q$ is a *spatial pattern*, that is, it expresses a spatial relationship among spatial objects. Mining spatial association rules presents two main sources of complexity:

1. the implicit definition of spatial relations and
2. the granularity of the spatial objects.

The former is due to the fact that the location and the extension of spatial objects implicitly define spatial relations such as topological, distance and direction relations. The latter refers to the fact that spatial objects can be described at multiple levels of granularity. For instance, UK census data can be geo-referenced with respect to the hierarchy of areal objects:

$$\text{ED} \rightarrow \text{Ward} \rightarrow \text{District} \rightarrow \text{County},$$

based on the inside relationship between locations. Therefore, some kind of *taxonomic knowledge* of task-relevant geographic layers may also be taken into account to obtain descriptions at different granularity levels (*multiple-level association rules*).

The problem of mining association rules can be formally stated as follows:

Given

- a spatial database (SDB),
- a set of reference objects S,
- some sets R_k, $1 \le k \le m$, of task-relevant objects
- a background knowledge *BK* including some spatial hierarchies H_k on objects in R_k
- M granularity levels in the descriptions (1 is the highest while M is the lowest)
- a couple of thresholds *minsup*[l] and *minconf*[l] for each granularity level

Find strong multi-level spatial association rules.

Each R_k is typically a map layer and spatial hierarchies capture is-a relations among locations on the basis of their geometry. To deal with several spatial hierarchies at once in a uniform manner, objects in them are mapped to one or more of the M user-defined description granularity levels so that frequency of patterns as well as strength of rules depend on the level *l* of granularity with which patterns/rules describe data. To be more precise, a pattern *P* (*s*%) at level *l* is *frequent* if $s \geq minsup[l]$ and all ancestors of *P* with respect to H_k are frequent at their corresponding levels. An association rule $Q \rightarrow R$ *(s%, c%)* at level *l* is strong if the pattern $Q \cup R$ (s%) is frequent and $c \geq minconf[l]$.

The task of mining spatial association rules is supported by the module Geo-associator of the spatial data mining system GeoMiner (Han et al., 1997). However, the method implemented in Geo-associator suffers from severe limitations due to the expressive power of the language adopted for representing both data and patterns. For instance, Geo-associator can not discover associations involving two or more distinct objects belonging to the same map layer. This lack of expressivity is due to the so-called *single-table assumption* which is common to most data mining methods currently available in the literature. Conversely our system SPADA for mining spatial association rules (Malerba et al., 2002) follows the recently promoted (multi-)*relational* approach to data mining (Džeroski, Lavrac, 2001). Here associations are represented as conjunctive formulas of a logical language. An example of pattern is the following formula:

is_a(X,large_town), intersects(X,Y), is_a(Y,road),
intersects(X,Z), is_a(Z,road), Z¹ Y

to be read as "*X* is a large town **and** *X* intersects *Y* **and** *Y* is a road **and** *X* intersects *Z* **and** *Z* is a road **and** *Z* is distinct from *Y*". Note that the special-purpose language primitive ≠ allows object identity to be expressed. Suppose that this pattern is frequent with 91% support. Thus we can say that "91% of large towns intersect two distinct roads" in the given geographic area. From this pattern the following strong rule can be derived:

is_a(X, large_town), intersects(X,Y), is_a(Y, road)
→ intersects(X,Z), is_a(Z,road), Z¹ Y (91%, 100%)

It says that: "**If** a large town X intersects a road Y **then** X intersects a road Z distinct from Y **with** 91% support and 100% confidence". Since multiple concept levels are dealt with, finer-grained descriptions are also expected, such as:

is_a(X, large_town), intersects(X,Y), is_a(Y,regional_road),
intersects(X,Z), is_a(Z, main_trunk_road), Z¹ Y (75%)

which supplies more insight into the nature of the task-relevant objects *Y* and *Z*.

The choice of a logical language poses some issues of computational complexity. In particular, the number of possible patterns can be very large and it becomes necessary to limit their space of possible patterns by providing explicit constraints (*declarative bias*). SPADA requires one such specification as additional input. These constraints specify what relations should be involved in the patterns, how the relations may be interconnected and what other syntactic constraints the patterns have to obey.

Further details of the method underlying SPADA can be found in (Malerba, Lisi, 2001).

2.2. Distributed architecture of the system

SPADA is a component of a distributed client-server system named ARES (Association Rules Extractor from Spatial data), that incorporates, in addition to SPADA, a Graphical User Interface (GUI), a module for the extraction of spatial

features from maps (FEATEX), a module for the discretisation of numerical attributes (RUDE) and a SPADA server which allows multiple users to run concurrently SPADA on the server. The distributed architecture of ARES is showed in Figure 1.

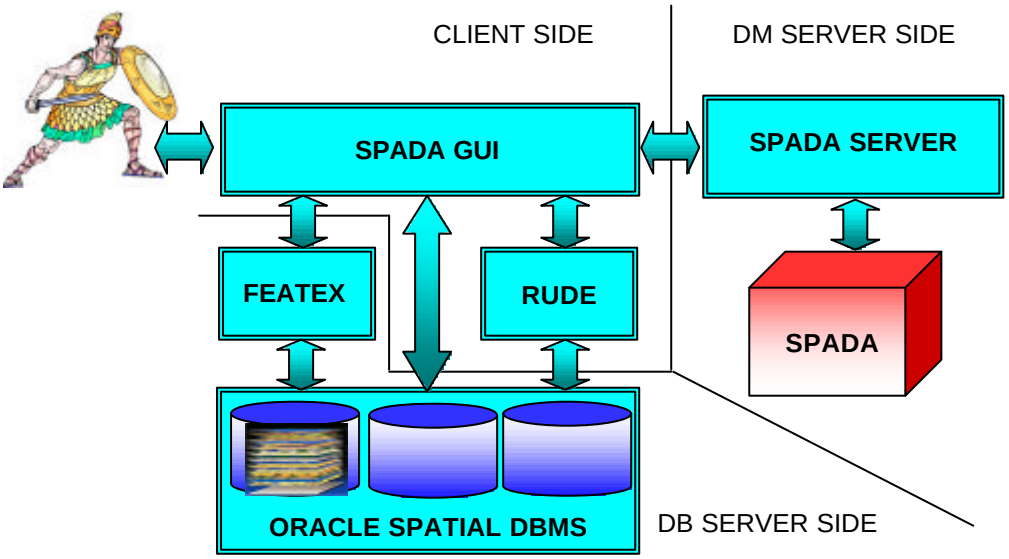

Figure 1. The software architecture of ARES

The GUI provides the user with facilities for controlling all parameters of the data mining process. More precisely, a wizard supports the user in the selection of layers (for spatial objects), tables (for aspatial properties) and attributes involved in the query to the SDB (Oracle Spatial). Conditions on both aspatial attributes (simple comparisons between two fields) and spatial features (simple comparisons with a field or a constant) can also be specified. Once the query is performed, the GUI allows the user to discretise some numerical attributes, to transform spatial relations and aspatial attributes into logic predicates, to specify the reference and the task relevant objects, to define the spatial hierarchies and their mappings into the M granularity levels, to specify the declarative bias, and finally to run SPADA on the server.

Many spatial features (relations and attributes) can be extracted from spatial objects stored in the SDB. The feature extraction requires complex data transformation processes to make spatial relations explicit. This function is partially supported by the SDBMS, which offers spatial data types in its data model and query language and supports them in its implementation, providing at least spatial indexing and efficient algorithms for spatial join (Güting, 1994). Therefore, a middle layer module (FEATEX) is required to make possible a loose coupling between SPADA and the SDB by generating features of spatial objects (points, lines, or regions). FEATEX is implemented as an Oracle package of procedures and functions, each of which implements a different feature (Appice et al., 2003).

Since SPADA, like many other association rule mining algorithms, cannot process numerical data properly, it is necessary to perform a discretization of numerical features with a relatively large domain. For this purpose we have implemented the relative unsupervised discretization algorithm RUDE (Ludl, Widmer, 2000) which discretizes an attribute of a relational database in the context defined by other attributes. At the end of this data preprocessing step, transformed data are stored in temporary Oracle Spatial tables. A PL-SQL function is responsible for transforming these data into a logic representation.

3. APPLICATION OF SPADA TO STOCKPORT DATA

In this section we describe a practical example that shows how it is possible to perform a spatial analysis on UK 1991 census data. The application has been developed in the context of the European project SPIN! (Spatial Mining for Data of Public Interest) (May, 2000). Data concern Stockport, one of the ten Metropolitan Districts of Great Manchester, UK. Stockport District is divided into twenty-two wards, each of which is decomposed into censual sections or enumeration districts (EDs), for a total of 589 EDs. Spatial analysis is enabled by the availability of vectorized boundaries of the 1991 census EDs as well as by other Ordnance Survey digital maps of the district, where several interesting layers are available, namely roads, bus priority lines, and so on. By joining UK 1991 census data available at the ED summarization level with some spatial objects (e.g. EDs, roads, and railways) it is possible to investigate socio-economic issues from a spatial viewpoint.

Both census data and digital maps are stored in an Oracle Spatial database, which is an object relational DBMS extended with spatial data handling facilities. Census data are distributed in 89 tables, each having 120 attributes on average, available for policy analysis. Census data are available only at ED level and are all numeric (in fact, integer-valued). They provide statistics on the population (resident at the census time, ethnic group, age, marital status, economic position, and so on), on the households in each ED (number of households with *n* children, number of households with *n* economically inactive people, number of households with two cars, and so on) as well as on some services available in each ED (e.g., number of schools).

For the application of our spatial association rule mining method we have focused our attention on transportation planning, which is one of the key issues in the Unitary Development Plan (UDP) of Stockport. In particular, we investigated the accessibility of the Stepping Hill Hospital. This is an inter-ED analysis, where spatial features concern relations between EDs, while aspatial features are mainly extracted from census data for all EDs.

The concept of “accessibility” appears initially in the context of geographical science and was progressively introduced in transport planning in the 1960’s and 1970’s. Many different definitions of accessibility and many ways to measure it can be found in the literature. In this work we are interested in urban accessibility, which refers to local (inner city) daily transport opportunities. A great effort has been made to define urban *accessibility indices*, which can be used to assess/compare transportation facilities within different regions of an urban area or between urban regions (Bhat et al., 2000). Accessibility is usually measured with respect to key activity locations for individuals (e.g., home, workplace) and evaluates the transportation services provided in these key locations to assess their relative advantages (Burns, 1979). In this work, we are interested in the accessibility “to” the Stepping Hill Hospital “from” the actual residence of people living within in the area served by the hospital. Since (micro) data on the actual residence of each involved household are not available, we study the accessibility at the ED level. Moreover, our study does not aim to synthesize a new accessibility index, but to discover human interpretable patterns that can also contribute to directing resources for facility improvement in areas with poor transport accessibility.

We decided to mine association rules relating five EDs close to the Stepping Hill Hospital (task relevant objects) with one-hundred and fifty-two EDs within a

distance of 10 Km of the hospital (reference objects). The goal is to understand which reference EDs have access to the task relevant EDs. To define the accessibility we used the Ordnance Survey data on transport network, namely the layers of roads, railways and bus priority lines (see Figure 2).

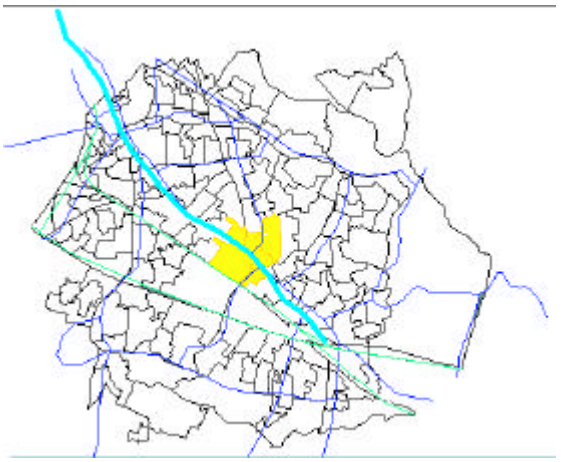

Figure 2. Stockport map around Stepping-Hill Hospital. Spatial objects are the following: one-hundred and fifty-two task relevant EDs (white regions), five task-relevant EDs (yellow regions), the only bus priority line (thick light-blue line), crossing roads (blue lines), crossing railways (green lines).

By using FEATEX we extracted facts concerning two topological relationships between EDs and the only bus priority line reported in the spatial database for that area. Two examples are the following:

external_touches_to(ed_03bsfq29,bus_priority_line_1).
comes_from(ed_03bsfl27,bus_priority_line_1).

The constants *ed_03bsfq29* and *ed_03bsfl27* denote two distinct EDs, while *bus_priority_line_1* is the only constant associated to a bus priority line. The topological relations *external_touches_to* and *comes_from* are computed according to the 9-intersection model (Egenhofer and Franzosa, 1991; Egenhofer and Herring,1994) and are schematized in Figure 3.

The set of topological relationships between EDs and roads is more varied. Some facts extracted by FEATEX are the following:

along(ed_03bsfk28,road_15329).
comes_from(ed_03bsfb23,road _12212).
crosses(ed_03bsfc13,road _12245).
external_ends_at(ed_03bsfc01,road_11501).
external_touches_to(ed_03bsfb22,road_ 15260).
external_comes_from(ed_03bsfc01,road_11502).
goes_out_of(ed_03bsfh01,road_ 10884).
inside(ed_03bsfc01,road_11494).
internal_ends_at(ed_03bsfc01,road_11500).
runs_along_boundary_ends_inside(ed_03bsfg22,road_ 10884).
runs_along_boundary(ed_03bsfc23,road _12312).

In this case constants *road_#* refer to roads crossing the interested area. It is noteworthy that the topological relations above are mutually exclusive, that is, it is impossible for two of them to concern the same pair of constants (ED, road).

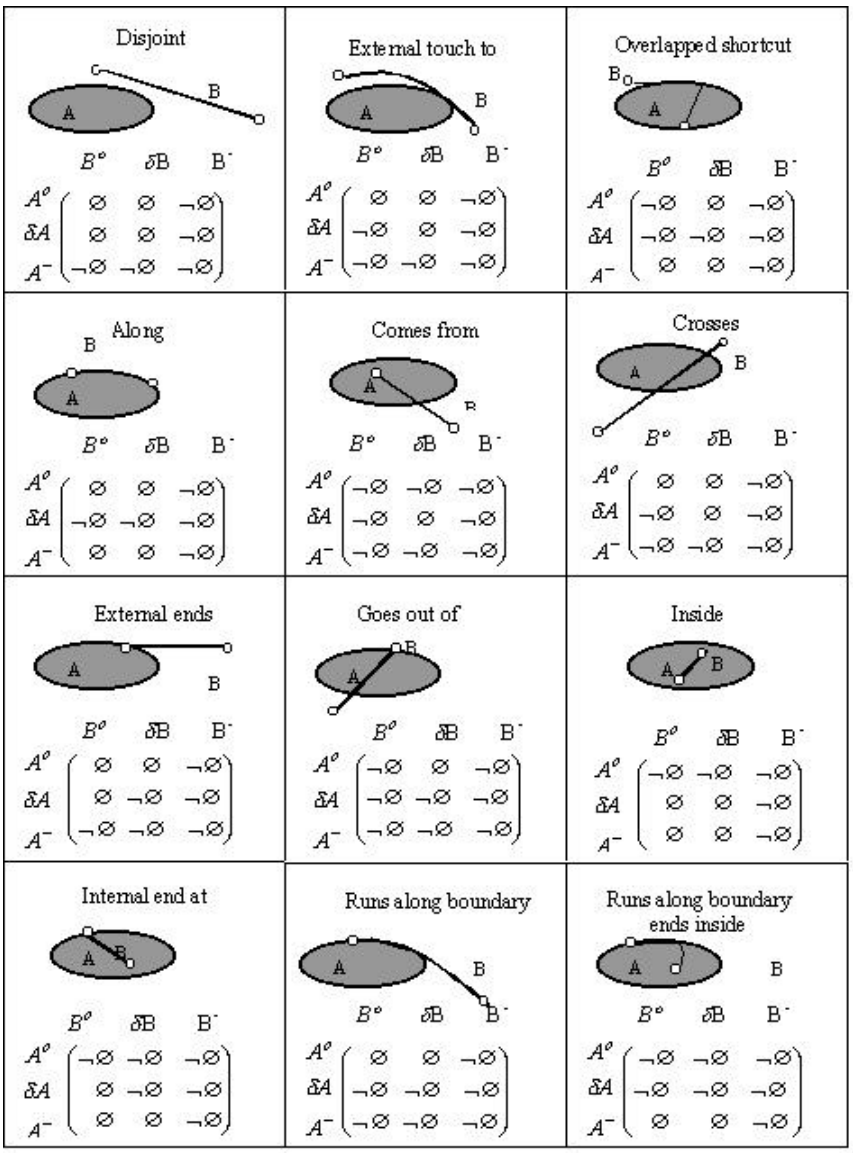

Figure 3. Twelve feasible relations between a region and a line according to the 9-intersection model.

Finally, we used FEATEX to extract spatial relationships between EDs and railways. Some examples of facts generated by the Oracle package are:

along(ed_ 03bsfg25,rail_ 2453).
comes_from(ed_ 03bsfc05,rail_ 2355).
crosses(ed_ 03bsfc13,rail_ 2391).
external_ends_at(ed_ 03bsfc05,rail_ 2389).
external_touches_to(ed_ 03bsfc20,rail_ 2389).
inside(ed_ 03bsfc01,rail_ 2341).
internal_ends_at(ed_ 03bsfc23,rail_ 2418).
overlapped_shortcut(ed_ 03bsfh12,rail_ 2487).
runs_along_boundary_ends_inside(ed_ 03bsfg11,rail_ 2429).
runs_along_boundary(ed_ 03bsfc10,rail_ 2391).

The total number of facts is 1,147. Despite the complexity of the spatial computation performed by FEATEX to extract these facts, the results are still not appropriate for the goals of our data analysis tasks. Indeed, we are interested in relationships between EDs, such as those stating that two EDs are 'connected' by the same bus priority line or the same road or the same railway. To solve this problem we specified the following rules to the domain specific knowledge:

1. crossed_by_bus_line(X) :- external_touches_to(X,bus_priority_line_1).
2. crossed_by_bus_line(X) :- comes_from(X,bus_priority_line_1).
3. connected_by_bus_line(X,Y) :- crossed_by_bus_line(X), crossed_by_bus_line(Y), X\=Y.
4. crossed_by_road(X,Z) :- along(X,Z), is_a(Z,road).
5. crossed_by_road(X,Z) :- comes_from(X,Z), is_a(Z,road).
6. crossed_by_road(X,Z) :- crosses(X,Z), is_a(Z,road).
7. crossed_by_road(X,Z) :- external_ends_at(X,Z), is_a(Z,road).
8. crossed_by_road(X,Z) :- external_touches_to(X,Z), is_a(Z,road).
9. crossed_by_road(X,Z) :- goes_out_of(X,Z), is_a(Z,road).
10. crossed_by_road(X,Z) :- inside(X,Z), is_a(Z,road).
11. crossed_by_road(X,Z) :- internal_ends_at(X,Z), is_a(Z,road).
12. crossed_by_road(X,Z) :- runs_along_boundary_ends_inside(X,Z), is_a(Z,road).
13. crossed_by_road(X,Z) :- runs_along_boundary(X,Z), is_a(Z,road).
14. connected_by_road(X,Y) :- crossed_by_road(X,Z), crossed_by_road(Y,Z), X\=Y.
15. crossed_by_rail(X,Z) :- along(X,Z), is_a(Z,rail).
16. crossed_by_rail(X,Z) :- comes_from(X,Z),is_a(Z,rail).
17. crossed_by_rail(X,Z) :- crosses(X,Z),is_a(Z,rail).
18. crossed_by_rail(X,Z) :- external_ends_at(X,Z),is_a(Z,rail).
19. crossed_by_rail(X,Z) :- external_touches_to(X,Z),is_a(Z,rail).
20. crossed_by_rail(X,Z) :- inside(X,Z),is_a(Z,rail).
21. crossed_by_rail(X,Z) :- internal_ends_at(X,Z),is_a(Z,rail).
22. crossed_by_rail(X,Z) :- overlapped_shortcut(X,Z),is_a(Z,rail).
23. crossed_by_rail(X,Z) :- runs_along_boundary_ends_inside(X,Z),is_a(Z,rail).
24. crossed_by_rail(X,Z) :- runs_along_boundary(X,Z),is_a(Z,rail).
25. connected_by_rail(X,Y) :- crossed_by_rail(X,Z), crossed_by_rail(Y,Z), X\=Y.

Here the use of the predicate is_a hides the fact that two hierarchies have been defined for spatial objects (see Figure 4). Both hierarchies have depth three and are straightforwardly mapped into three granularity levels.

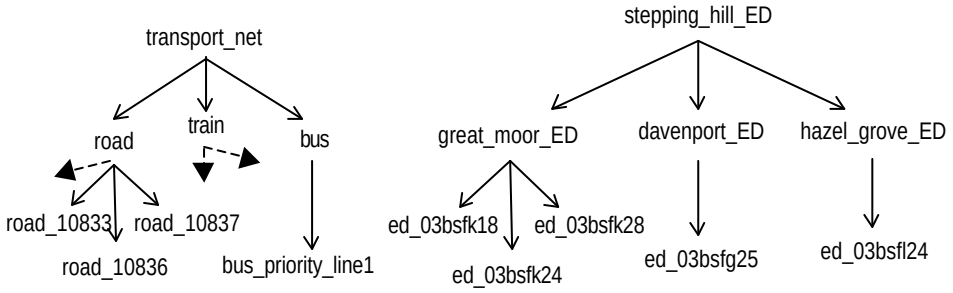

Figure 4. Two spatial hierarchies defined for the mining task concerning the accessibility of the Stepping Hill Hospital. They are mapped into three granularity levels.

The 'connection' predicates defined above express direct accessibility of an ED from another ED by means of only one road or railway or bus line. To express a more complex concept of accessibility, we added the following rules to the domain specific knowledge:

26. can_reach_by_road(X,Y) :- connected_by_road(X,Y).
27. can_reach_by_road(X,Y) :- connected_by_road(X,Z), can_reach_by_road(Z,Y), X\=Y.

28. can_reach_by_rail(X,Y) :- connected_by_rail(X,Y).
29. can_reach_by_rail(X,Y) :- connected_by_rail(X,Z), can_reach_by_rail(Z,Y), X\=Y.
30. can_reach_by_bus(X,Y) :- connected_by_bus_line(X,Y).
31. can_reach_by_bus(X,Y) :- connected_by_bus_line(X,Z), can_reach_by_bus_line(Z,Y), X\=Y.

These rules express a limited form of the transitivity property of 'connectedness'. Indeed, they state that an ED Y can be reached from another ED X if they are either directly connected by a road or a railway or a bus line, or if there is another "intermediate" ED Z, which is directly connected to X and can be reached from Y.

To complete the domain specific knowledge, we added the following rules on the accessibility by means of public transport:

32. can_reach_by_road_rail(X,Y) :- connected_by_road(X,Z), can_reach_by_rail(Z,Y), X\=Y.
33. can_reach_by_road_bus(X,Y) :- connected_by_road(X,Z), can_reach_by_bus (Z,Y), X\=Y.
34. can_reach_by_rail_bus(X,Y) :- connected_by_rail(X,Z), can_reach_by_bus (Z,Y), X\=Y.
35. can_reach_by_rail_bus(X,Y) :- connected_by_bus_line(X,Z), can_reach_by_rail(Z,Y), X\=Y.
36. can_reach_by_public_transport(X,Y) :- can_reach_by_bus(X,Y).
37. can_reach_by_public_transport(X,Y) :- can_reach_by_rail(X,Y).
38. can_reach_by_public_transport(X,Y) :- can_reach_by_rail_bus(X,Y).
and the complementary definition of accessibility by means of roads alone:
39. can_reach_only_by_road(X,Y) :- can_reach_by_road(X,Y), \+can_reach_by_public_transport(X,Y).

Until now, census data have not been used to define the accessibility of the Stepping Hill Hospital. All extracted data and user-defined background knowledge are purely spatial. However, we can observe that the accessibility of an area cannot be defined on the basis of the transport network alone. Even though some roads connect a reference ED X with a task relevant ED Y, people living in X might have problems reaching Y because they do not drive. This means that sociological data available in the census data tables can be profitably used to give an improved definition of accessibility. We selected four attributes on the percentage of households with zero, one, two, and three or more cars, we discretized them with RUDE and generated the following four binary predicates for SPADA: *no_car, one_car*, *two_cars*, *three_more_cars*. The first argument of the predicate refers to an ED, while the second argument is an interval returned by RUDE.

To complete the problem statement we specified a declarative bias both to constrain the search space and to filter out some uninteresting spatial association rules. In particular, we asked for rules containing only the following predicates: *can_reach_by_public_transport*, *can_reach_only_by_road*, *no_car*, *one_car*, *two_cars*, and *three_more_cars*. In this way, we ruled out all spatial relations directly extracted by means of FEATEX and all intermediate spatial relations that helped to define the two interesting ones, namely the accessibility by public transport and the accessibility only by roads. Moreover, the specification of the following filter:

pattern_constraint([no_car(_,_), one_car(_,_), two_cars(_,_), three_more_cars(_,_)], 1).

prevents the generation of association rules with purely spatial patterns, that is, patterns showing only spatial relations between spatial objects. Purely spatial patterns are indeed of no interest to the expert in transport planning, since it is very likely that they convey no additional information to what he/she already knows.

After some tuning of the parameters min_sup and min_conf for each granularity level, we decided to run the system with the following parameter values:

min_sup[1]=0.2 *min_conf[1]=0.5*
min_sup[2]=0.1 *min_conf[2]=0.4*
min_sup[3]=0.1 *min_conf[3]=0.3*

Despite the above constraints, SPADA generated 944 rules in 88 secs from a set of 39,830 extracted or inferred facts. More precisely, the system generated 28 rules in 38 secs at granularity level 1, 215 rules in 17 secs at level 2, and 701 rules in 33 secs at level 3. The output rules are stored in M×K XML files, where M is the number of granularity levels (3) and K is the maximum number of refinement steps (6). An additional index HTML file allows users to browse the output rules both by level and by refinement step (see Figure 5).

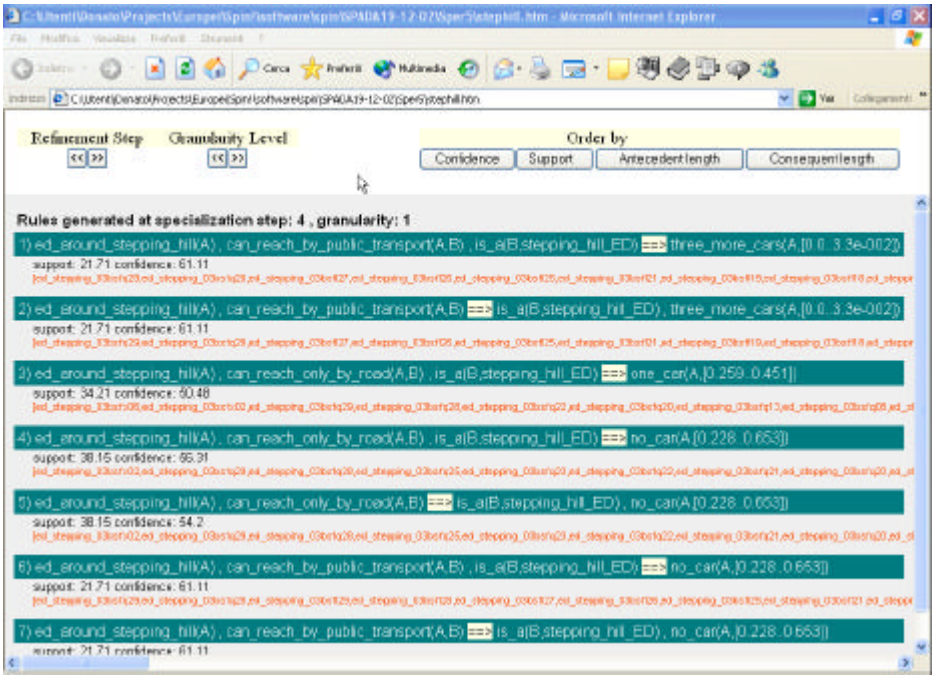

Figure 5. Browsing results of the SPADA system. The user can select the refinement (or specialization) step and the granularity level. Rules are reported in the order in which they are generated, but they can be sorted in a decreasing order by confidence, support, antecedent length and consequence length. In addition to usual confidence and support values, the name of the reference EDs supporting the rule is shown. Results can be easily shown on a map.

Two of the rules returned by SPADA at the first level are the following:

ed_around_stepping_hill(A), can_reach_only_by_road(A,B), is_a(B,stepping_hill_ED) ® no_car(A,[0.228..0.653]) *(38.15%, 56.31%)*

ed_around_stepping_hill(A), can_reach_by_public_transport(A,B), is_a(B,stepping_hill_ED) ® no_car(A,[0.228..0.653]) *(21.71%, 61.11%).*

The spatial pattern of the first rule occurs in fifty-eight distinct EDs. This means that from fifty-eight distinct EDs within a distance of 10Km from Stepping Hill Hospital, it is possible to reach the hospital only by road and the percentage of

households with no car is quite high (between 22.8% and 65.3%). Moreover, if from an ED A around Stepping Hill Hospital it is possible to reach one of the five task relevant EDs only by road, then the confidence that A has a high percentage of households with no car is 56.31%.

The spatial pattern of the second rule occurs in thirty-three distinct EDs. This means that from thirty-three reference EDs whose percentage of households with no car is quite high it is possible to reach the area of the Stepping Hill Hospital by public transport. The confidence in the second rule is a little higher than the first association rule.

At granularity level 2, SPADA specializes the task relevant object *stepping_hill_ED* considered at level 1. Only three specializations are possible for the five task relevant objects, namely *hazel_grove_ed, davenport_ed, great_moor_ed*, which correspond to three distinct wards (Hazel Grove, Davenport, Great Moor). The first rule above is specialized as follows:

ed_around_stepping_hill(A), can_reach_only_by_road(A,B),
is_a(B,great_moor_ED) ® no_car(A,[0.228..0.653]) (38.15%, 56.31%)
ed_around_stepping_hill(A), can_reach_only_by_road(A,B),
is_a(B,davenport_ED) ® no_car(A,[0.228..0.653]) (21.71%, 50.76%)
ed_around_stepping_hill(A), can_reach_only_by_road(A,B),
is_a(B,hazel_grove_ED) ® no_car(A,[0.228..0.653]) (21.71%, 50.76%).

As expected, the support of some rules have decreased. However, since both support and confidence are greater than the corresponding user-defined thresholds, all the three rules are output by SPADA. Similar considerations apply to granularity level 3, where specific task relevant EDs are reported.

Association rules found by SPADA in this application can be of interest to urban planners, since they relate data on the transport network with data on sociological factors. However, this study has three main limitations due to the nature of available data. First, we considered 1991 Census data, which are now obsolete. Second, the crossing of a railway does not necessarily mean that there is a station in an ED. Similar considerations can be made for bus priority lines and roads. Third, digital maps made available by the Ordnance Survey are devised for cartographic reproduction purposes and not for data analysis. Hence, a road may appear to be ‘blocked’ in the digital map, because it runs under a bridge.

4 CONCLUSIONS

In this paper an application of spatial data mining to geo-referenced census data and digitized topographic maps has been illustrated. The multi-relational approach and the usage of logic as representation and reasoning means is justified by the need to consider relationships implicitly defined between spatial objects. Spatial relationships and spatial reasoning rules can be easily represented by means of first-order logic clauses. The interface to a spatial database is another crucial issue in spatial data mining. In this work, a form of loose coupling between the rule mining system SPADA and the SDB has been presented. It is based on the implementation of an Oracle package for the extraction of a number of spatial and aspatial features initially represented as tuples and then translated into atoms. Advantages and drawbacks of this approach have been briefly discussed in the paper. The future implementation of a tight coupling will permit an experimental comparison of the two solutions.

The specific urban planning problem faced in this paper concerns the accessibility of an urban area. Unlike typical accessibility studies, no index has been developed in this application. Rather, we aimed at discovering human interpretable patterns that can also contribute to directing resources for facility improvement in areas with poor transport accessibility. Indeed, some of the discovered rules seem to convey new knowledge to urban planners, although the search for these "nuggets" requires a lot of tuning and effort on the part of the data analyst in order to constrain the search space properly and discard most of the obvious or totally useless patterns hidden in the data. One of the main limitations of our system, which is also a problem of many other relational data mining systems, is the requirement of some expertise in data and knowledge engineering. Indeed, the user should know how data are organized in the spatial database (e.g., layers and physical representation of objects), the semantics of spatial relations that can be extracted from digital maps, the meaning of some parameters used in the discretization process and in the generation of spatial association rules, as well as the correct and most efficient way to specify the domain knowledge and declarative bias. Also the system currently lacks of presentation modalities that make the discovered patterns immediately and intuitively understandable by end-users in urban planning environments, i.e. urban planners. For instance, a multi-modal interface has been presented in (De Carolis, Lisi, 2001). This and other solutions to usability problems should be investigated with the help of end-users themselves.

Acknowledgments

The authors thank Jim Petch, Keith Cole and Mohammed Islam (MIMAS, University of Manchester, England) and Chrissie Gibson (Department of Environmental and Geographical Sciences, Manchester Metropolitan University, England) for providing access to census data and digital OS maps of Stockport, Manchester. The work presented in this paper is in partial fulfillment of the research objectives set by the IST European project SPIN! (Spatial Mining for Data of Public Interest) and by the MURST COFIN-2001 project on "Methods for the extraction, validation and representation of statistical information in a decision context".

References

Agrawal R., Imielinski T., Swami A. (1993) Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Anselin L.G. (1993) Spatial statistical analysis and geographical information systems. In: Fischer M.M., Nijkamp P (Eds.) *Geographic information systems, spatial modelling, and policy evaluation*. Berlin, Springer, 35-49.

Appice A., Ceci M., Lanza A., Lisi F.A., Malerba D. (2003) Discovery of Spatial Association Rules in Georeferenced Census Data: A Relational Mining Approach, *Intelligent Data Analysis*, (in press).

Bhat C., Handy S., Kockelman K., Mahmassani H., Chen Q., Weston L. (2000) Urban accessibility index: literature review. *Technical Report No TX-01/7-4938-1*, Texas Dept. of Transportation, University of Texas at Austin.

Burns L.D. (1979) *Transportation, Temporal, and Spatial Components of Accessibility.* Lexington Books. Lexington, MA.

De Carolis B., Lisi F.A. (2001) A Multimodal Interface for supporting Urban and Land Planning. In G. Concilio & V. Monno (Eds.), *Proc. 2nd National Conference on Information Technology and Spatial Planning*, Dedalo Edizioni, Bari, Italy.

DeMers M.N. (2000) *Fundamentals of Geographic Information Systems*. *2nd ed.*, John Wiley & Sons, New York.

Densham P. (1991), Spatial Decision Support Systems. In: Maguire D.J., Goodchild M.F., Rhind D.W. (Eds.) Geographical Information Systems: principles and applications, John Wiley & Sons, New York, 403-412.

Dzeroski S., Lavrac N. (Eds.) (2001) *Relational Data Mining*. Springer-Verlag, Berlin.

Egenhofer M.J., Franzosa R. (1991) Point-Set Topological Spatial Relations, *International Journal of Geographical Information Systems*, 5(2), 161-174.

Egenhofer M.J, Herring J.R. (1994) Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. In:. Egenhofer, M.J, D.M. Mark, J.R. Herring (eds.): *The 9-intersection: Formalism and its Use for Natural-language Spatial Predicates*, Technical Report 94-1, Santa Barbara, NCGIA.

Ester M., Frommelt A., Kriegel H.-P., Sander J. (1998) Algorithms for Characterization and Trend Detection in Spatial Databases. *Proc. 4th Int. Conf.* on *Knowledge Discovery and Data Mining*, New York City, NY, 44-50.

Güting R.H. (1994) An introduction to spatial database systems*. VLDB Journal*, 4(3) 357-399.

Han S.Y., Kim T.J. (1989) Can expert systems help with planning? *Journal of the American Planning Association*, 55, 296-308.

Jones A.P., Bentham G. (1995) Emergency medical service accessibility and outcome from road traffic accidents, *Public Health*, 109, 169-177.

Keenan, P. (1998), Spatial decision support systems for vehicle routing. *Decision Support Systems,* 22, 65-71.

Klösgen W., May M. (2002) Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.): *Principles of Data Mining and Knowledge Discovery (PKDD), 6th European Conference*, LNAI 2431, Springer-Verlag, Berlin, 275-286.

Koperski K., Han J. (1995) Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer, M.J., Herring, J.R. (eds.): *Advances in Spatial Databases*. LNCS 951, Springer-Verlag, Berlin, 47-66.

Koperski K., Adhikary J., Han, J.: Knowledge discovery in spatial databases: progress and challenges, *Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, (1996).

Koperski K. , Han J., Stefanovic N. (1998) An Efficient Two-Step Method for Classification of Spatial Data. *Proc. Symposium on Spatial Data Handling (SDH '98)*, Vancouver, Canada, 45-54.

Lisi F.A., Malerba D. (2002) SPADA: A spatial association discovery system. In A. Zanasi, C. A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.) *Data Mining III*, Series Management Information Systems, Vol 6, WIT Press, Southampton, 157-166.

Love D., Lindquist P. (1995) The geographical accessibility of hospitals to the aged: a geographical information systems analysis within Illinois, *Health Services Research*, 29, 627-651.

Ludl M.C., Widmer, G. (2000) Relative Unsupervised Discretization for Association Rule Mining. In: Zighed D.A., H.J. Komorowski, J.M. Zytkow (Eds.): *Principles of Data Mining and Knowledge Discovery*, LNCS 1910, Springer-Verlag, 148-158.

Malerba D., Lisi F.A. (2001) An ILP method for spatial association rule mining. Working notes of the *First Workshop on Multi-Relational Data Mining*, Freiburg, Germany, 18-29.

Malerba D., Esposito F., Lisi F.A., Appice A. (2002) Mining Spatial Association Rules in Census Data. *Research in Official Statistics*, 5(1), 19-44.

Malerba D., Esposito F., Lanza A., Lisi F.A. (2001) First-order Rule Induction for the Recognition of Morphological Patterns in Topographic Maps. In: Perner, P. (ed.): *Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Artificial Intelligence, vol. 2123, Springer-Verlag Berlin Heidelberg, Germany, 88-101.

Martin D. (1999) Spatial representation: the social scientist's perspective. In: P. Langley, M. Goodchild, D. Maguire, and D. Rhind (eds.): *Geographical Information Systems, vol. 1, Principles and Technical Issues, 2nd edition*, Vol.1, John Wiley and Sons, pp. 71-80.

May M. (2000) Spatial Knowledge Discovery: The SPIN! System. *Proceedings of the 6th EC-GIS Workshop*, Lyon, ed. Fullerton, K., JRC, Ispra.

Mitchell T.M. (1999) Machine Learning and Data Mining. *Communications of the ACM*, 42(11), 30-36.

Ng R., Han J. (1994) Efficient and effective clustering method for spatial data mining. Proceedings of the *International Conference VLDB*, Santiago, Chile, September (1994) 124-155.

Sander J., Ester M., Kriegel H.P., Xu X. (1998) Density-Based Clustering in Spatial Databases: A New Algorithm and its Applications. *Data Mining and Knowledge Discovery, an International Journal*, Kluwer Academic Publishers, 2(2), 169-194.