Mining Spatial Association Rules in Census Data: A Relational Approach

Donato Malerba, Francesca A. Lisi, Annalisa Appice, Francesco Sblendorio

Dipartimento di Informatica, Università degli Studi di Bari,
via Orabona 4, 70126 Bari, Italy
{malerba, lisi, appice, sblendorio}@di.uniba.it

Abstract. In this paper we propose a method for the discovery of spatial association rules, that is, association rules involving spatial relations among (spatial) objects. The method is based on a multi-relational data mining approach and takes advantages of the representation and reasoning techniques developed in the field of Inductive Logic Programming (ILP). In particular, the expressive power of predicate logic is profitably used to represent spatial relations and background knowledge (such as spatial hierarchies and rules for spatial qualitative reasoning) in a very elegant, natural way. The integration of computational logics with efficient spatial database indexing and querying procedures permits applications that cannot be tackled by traditional statistical techniques in spatial data analysis. The proposed method has been implemented in the ILP system SPADA (Spatial Pattern Discovery Algorithm). We report the preliminary results on the application of SPADA to Stockport census data.

1 Introduction

Censuses make a huge variety of general statistical information on society available to both researchers and the general public. Population and economic census information is of great value in planning public services (education, funds allocation, public transportation) as well as in private businesses (locating new factories, shopping malls, or banks, as well as marketing particular products).

The application of data mining techniques to census data, and more generally, to official data, has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society [22]. Nevertheless, it is not straightforward and requires challenging methodological research, which is still in the initial stage.

One of the research issues related to mining census data is geo-referenciation. The practice of attaching socio-economic data to specific locations has increasingly spread over the last few decades. In the UK, for instance, household expenditure data are provided for each enumeration district (ED), the smallest areal unit for which census data are published. At the same time, vectorized boundaries of the 1991 census EDs

enable the investigation of socio-economic phenomena in association with the geographical location of EDs. These advances cause a growing demand for more powerful data analysis techniques that can link population data to their spatial distribution.

Spatial data mining methods and techniques have been proposed for *the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases* [13]. In this paper we focus our attention on the specific task of discovering *spatial association rules*, that is, association rules involving spatial objects and relations.

The problem has already been tackled by [12], who implemented the module Geo-associator of the spatial data mining system GeoMiner [10]. This method, however, suffers from severe limitations due to the restrictive data representation formalism, known as *single-table assumption*. More specifically, it is assumed that data to be mined are represented in a single table (or relation) of a relational database, such that each row (or tuple) represents an independent unit of the sample population and columns correspond to properties of units.

In spatial data mining applications this assumption turns out to be a great limitation. Indeed, different geographical objects may have different properties, which can be properly modeled by as many data tables as the number of object types. In addition, attributes of the neighbors of some spatial object of interest may influence the object itself, hence the need for representing object interactions. From a database perspective, this means that two relations are required, one for the *reference* EDs, that is, the EDs whose socio-economical factors are the subject of investigation, and one for the neighboring EDs, which are considered *task relevant*, because they are spatially adjacent to some reference EDs.

The recently promoted *relational* approach to data mining [6], looks for patterns that involve multiple relations of a relational database. Thus data taken as input by these approaches typically consists of several tables and not just a single one, as is the case in most existing data mining approaches. Patterns found by these approaches are called *relational* and are typically stated in a more expressive language than patterns defined in a single data table.

The following is an example of a *relational association rule*:

$$male\text{-}full\text{-}time\text{-}employee\%(X,low) \wedge male\text{-}part\text{-}time\text{-}employee\%(X,low) \wedge neighbor(X,Y) \wedge comm\text{-}activities(Y,high) \rightarrow male\text{-}self\text{-}employed\%(X,high) \quad (32\%,70\%)$$

which states that in 70% of the cases, the low percentage of full-time and part-time male employees in some reference ED *X*, adjacent to another task relevant ED *Y*, with many commercial activities, implies a high percentage of self-employed males in *X*. The relational *pattern*

$$male\text{-}full\text{-}time\text{-}employee\%(X,low) \wedge male\text{-}part\text{-}time\text{-}employee\%(X,low) \wedge neighbor(X,Y) \wedge comm\text{-}activities(Y,high) \wedge male\text{-}self\text{-}employed\%(X,high)$$

occurs in 32% of reference EDs.

It is noteworthy that in this example, and more generally in relational association rules, the items are first-order logic *atoms*, that is, *n*-ary *predicates* applied to *n terms*.

In this example terms can be either *variables*, such as *X* and *Y*, or *constants*, such as *low* or *high*. In other words, subsets of *first-order logic*, which is also called predicate calculus or relational logic, are used to express relational patterns and relational association rules.

Considering this strong link with logics, it is not surprising that many algorithms for multi-relational data mining originate from the field of *inductive logic programming* (ILP) [19, 5, 14, 20]. Extending a single table data mining algorithm to a relational one is not trivial. Efficiency is also very important, as even testing a given relational pattern for validity is often computationally expensive. Moreover, for relational pattern languages, the number of possible patterns can be very large and it becomes necessary to limit their space by providing explicit constraints (*declarative bias*).

However, mining *spatial* association rules is a more complex task than mining *relational* association rules, whose solutions have already been reported in the literature [4]. Two further degrees of complexity are:

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* defines spatial relations such as topological, distance and direction relations. Therefore, complex data transformation processes are required to make spatial relations explicit (see the application of machine learning techniques to topographic map interpretation [16]).

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 following hierarchy:

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

based on the *inside* relationship between locations. Interesting rules are more likely to be discovered at low granularity levels (ED and ward) than at the county level. On the other hand, large support is more likely to exist at higher granularity levels (District and County) rather than at low levels.

In the next section, a new algorithm for mining spatial association rules is reported. The algorithm, named SPADA (Spatial Pattern Discovery Algorithm), is based on an ILP approach to relational data mining and permits the extraction of multi-level association rules, that is, association rules involving spatial objects at different granularity levels. SPADA has been implemented in Sictus Prolog and is interfaced to an Oracle8i™ database, empowered by an Oracle Spatial cartridge, which enables spatial data to be stored, accessed, and analyzed quickly and efficiently. The system also performs the appropriate data transformation by extracting spatial features (FEATEX module) and by discretizing numerical attributes (RUDE module). The application of SPADA to two data mining tasks involving UK census data is reported in Section 3.

2 Mining spatial association rules with SPADA

The discovery of spatial association rules is a descriptive mining task aiming to detect associations between *reference objects* and some *task-relevant objects*. The former

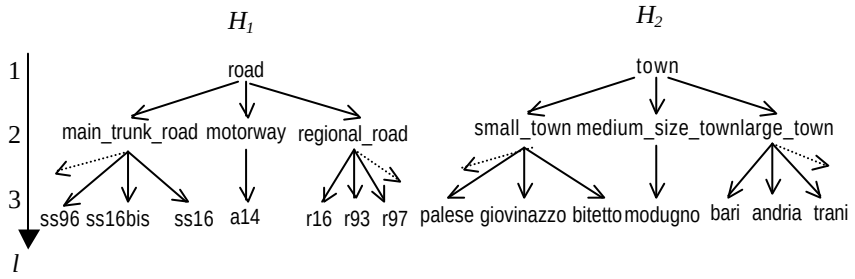

Fig. 1. Two spatial hierarchies and their association to three granularity levels (*l*).

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 describing a given area by finding associations between large towns (reference objects) and spatial objects in the road network, hydrography, and administrative boundary layers (task-relevant objects). The following is an example of spatial association rule that can be generated:

$$is\_a(X,large\_town) \wedge intersects(X,Y) \wedge is\_a(Y,road) \rightarrow$$
$$intersects(X,Z) \wedge is\_a(Z, road) \wedge Z^1 Y \qquad (91\%, 85\%).$$

It states 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 some kind of taxonomic knowledge on task-relevant objects may also be taken into account to obtain descriptions at different granularity levels (*multiple-level association rules*), finer-grained answers to the above query are also expected, such as:

$$is\_a(X,large\_town) \wedge intersects(X,Y) \wedge is\_a(Y,regional\_road) \rightarrow$$
$$intersects(X,Z) \wedge is\_a(Z, main\_trunk\_road) \wedge Z^1 Y \qquad (45\%, 90\%)$$

which provides more insight into the nature of the task relevant objects *Y* and *Z*, according to the spatial hierarchy reported in Fig. 1. It is noteworthy that the support and the confidence of the last rule changed. Generally, the lower the granularity level, the lower the support of association rules. Therefore, we follow Han and Fu's [9] proposal to use different thresholds of support and confidence for different granularity levels.

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, $1 \leq k \leq m$, of task-relevant objects
- some spatial hierarchies H_k involving objects in R_k
- *M* granularity levels in the descriptions (1 is the highest while M is the lowest) (see Fig. 1)
- a set of granularity assignments ψ_k which associate each object in H_k with a granularity level
- a *domain specific knowledge DK*
- a *declarative bias DC*

- a couple of thresholds *minsup*[*l*] and *minconf*[*l*] for each granularity level

Find strong multi-level spatial association rules.

An ILP approach to mining spatial association rules has already been reported in [17]. Representation problems, and algorithmic issues related to the application of our logic-based computational method are discussed in the next two sub-sections.

2.1 The representation

The basic idea in our proposal is that a spatial database boils down to a deductive relational database (DDB) once the spatial relationships between reference objects and task-relevant objects have been extracted. The expressive power of first-order logic in databases also allows us to specify background knowledge (BK), such as spatial hierarchies and domain specific knowledge expressed as sets of *rules*, which are stored in the intensional part of the DDB and can support, amongst other things, spatial qualitative reasoning.

Henceforth, we denote the DDB in hand $D(S)$ to mean that it is obtained by adding the spatial relations extracted from SDB regarding the set of reference objects S to the previously supplied *BK*. The ground facts in $D(S)$ can be grouped into distinct subsets: Each group, uniquely identified by the corresponding reference object $s \in S$, is called *spatial observation* and denoted $O[s]$. It is given by:

$$O[s] = O[s|s] \cup \{ O[r \mid s] \mid \text{a spatial relation } q(s,r) \text{ exists in } D(S) \}$$

It contains not only spatial relations between s and some task-relevant object $r \in R_k$ but also spatial relations between r and some $s' \in S$. It is noteworthy that a spatial observation refers to one and only one reference object $s \in S$. The unique reference object associated to a spatial observation allows us to define the support and the confidence of a spatial association rule (see definition below).

Let $A=\{a_1, a_2, \ldots, a_t\}$ be a set of atoms whose terms are either variables or constants (Datalog atoms [2]). Predicate symbols used for A are all those permitted by the user-specified declarative bias, while the constants are only those defined in DDB. Conjunctions of atoms on A are called *atomsets* [3] like the itemsets in classical association rules. In our framework, a language of patterns $L[l]$ at the granularity level l is a set of well-formed atomsets generated on A. Necessary conditions for an atomset P to be in $L[l]$ are the presence of the *key atom* defining a reference object ω at level l, the linkedness [11], and safety. To a pattern P we assign an existentially quantified conjunctive formula $eqc(P)$ obtained by turning P into a Datalog query.

Definition A pattern P covers an observation $O[s]$ if $eqc(P)$ is true in $O[s] \cup BK$.

Definition Let O be the set of spatial observations in $D(S)$ and O_P denote the subset of O containing the spatial observations covered by the pattern P. The *support* of P is defined as $\sigma(P) = |O_P| / |O|$.

Definition A spatial association rule in $D(S)$ at the granularity level l is an implication of the form

$$P \rightarrow Q \ (s\%, c\%)$$

where $P \cup Q \in L[l]$, $P \cap Q = \emptyset$, P includes the key atom and at least one spatial relationship is in $P \cup Q$. The percentages s% and c% are respectively called the support and the confidence of the rule, meaning that s% of spatial observations in $D(S)$ is covered by $P \cup Q$ and c% of spatial observations in $D(S)$ that are covered by P is also covered by $P \cup Q$.

Definition The support and the confidence of a spatial association rule $P \rightarrow Q$ are given by $s = \sigma(P \cup Q)$ and $c = \varphi(Q|P) = \sigma(P \cup R) / \sigma(P)$.

In multi-level association rule mining, an *ancestor* relation between two patterns at different granularity levels $PL[l]$ and $P' \in L[l']$, $l < l'$, exists if and only if P' can be obtained from P by replacing each spatial object $h \in H_k$ at granularity level $l = \psi_k(h)$ with a spatial object $h' < h$ in H_k, which is associated with the granularity level $l' = \psi_k(h')$.

The frequency of a pattern depends on the granularity level of task-relevant spatial objects.

Definition Let $minsup[l]$ and $minconf[l]$ be two thresholds setting the minimum support and the minimum confidence respectively at granularity level l. A pattern P is *large* (or frequent) at level l if $\sigma(P) \geq minsup[l]$ and all ancestors of P with respect to the hierarchies H_k are large at their corresponding levels. The confidence of a spatial association rule $P \rightarrow Q$ is high at level l if $\varphi(Q|P) \geq minconf[l]$. A spatial association rule $P \rightarrow Q$ is *strong* at level l if $P \cup Q$ is large and the confidence is high at level l.

2.2 Method

The task of mining spatial association rules itself can be split into two sub-subtasks:

1. Find large (or frequent) spatial patterns;
2. Generate highly-confident spatial association rules.

Algorithm design for frequent pattern discovery has turned out to be a popular topic in data mining. The blueprint for most algorithms proposed in the literature is the levelwise method [18], which is based on a breadth-first search in the lattice spanned by a generality order $\geq$ between patterns. The space is searched one level at a time, starting from the most general patterns and iterating between the candidate generation and candidate evaluation phases. The algorithm SPADA implements the aforementioned levelwise method.

The pattern space is structured according to the θ-subsumption [21]. Many ILP systems adopt θ-subsumption as the generality order for clause spaces. In this context we need to adapt the framework to the case of atomsets. More precisely, the restriction of θ-subsumption to *Datalog queries* (i.e. existentially quantified conjunctions of Datalog atoms) is of particular interest.

Definition Let Q_1 and Q_2 be two queries. Then Q_1 *q-subsumes* Q_2 if and only if there exists a substitution θ such that $Q_1 \supseteq Q_2\theta$.

We can now introduce the generality order adopted in SPADA.

Definition Let P_1 and P_2 be two patterns. Then P_1 is more general than P_2 under θ-subsumption, denoted as $P_1 \geq_\theta P_2$, if and only if P_2 θ-subsumes P_1.

It is noteworthy that $\geq_\theta$ on patterns represented as Datalog queries is monotone with respect to support, which is the criterion for candidate evaluation in SPADA. The quasi-ordered set spanned by $\geq_\theta$ can be searched by a *refinement operator*, namely a function which computes a set of refinements of a pattern. In particular, we need a refinement operator under θ-subsumption that enables the bottom-up search of the pattern space from the most specific to the most general patterns.

Definition Let $\langle G, \geq_\theta \rangle$ be a pattern space ordered according to $\geq_\theta$. An *upward refinement operator under q-subsumption* is a function ρ such that $\rho(P) \subseteq \{Q \mid Q \geq_\theta P\}$.

Such a refinement operator drives the search towards patterns with decreasing support, therefore all refinements $\rho(P)$ of an infrequent pattern P are infrequent. This is the first-order counterpart of one of the properties holding in the family of the Apriori-like algorithms [1], on which the pruning criterion is based.

For each granularity level ℓ(), SPADA generates and evaluates candidates by searching the pattern space. The *candidate generation* phase consists of a refinement step followed by a pruning step. The former applies the refinement operator under θ-subsumption to patterns previously found to be frequent by preserving the property of linkedness [11]. The latter mainly involves verifying that candidate patterns do not θ-subsume any infrequent pattern. Further pruning criteria have been implemented in SPADA. In particular, the system checks that candidates are not alphabetic variants of previously discovered patterns. The complexity of this test is $O(n^2)$, where n is the number of atoms in the two patterns to be compared. The *candidate evaluation* phase is performed by comparing the support of the candidate pattern with the minimum support threshold set for the level being explored. If the pattern turns out not to be a large one, it is rejected.

2.3 Integrating SPADA with other software components

The application of the ILP approach to spatial databases is made possible by a middle-layer module for feature extraction, as shown in Fig. 2. This layer is essential to cope with one of the main issues of spatial data mining, namely the requirement of complex data transformation processes to make spatial relations explicit.

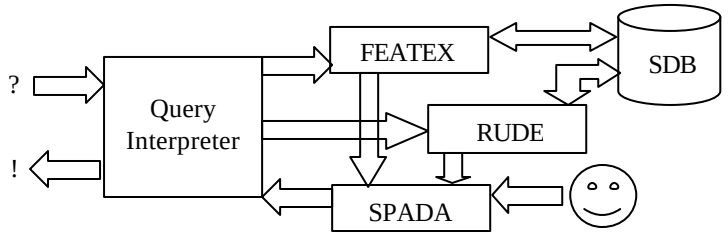

Fig. 2. Integration of SPADA with other software modules which support spatial feature extraction (FEATEX) and discretization of numerical features (RUDE). Additional input to SPADA, such as declarative bias and background knowledge, is directly provided by the user.

This function is partially supported by the spatial database (SDB), 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 [8]. Thus spatial databases supply an adequate representation of both single objects and spatially related collections of objects. In particular, the abstraction primitives for spatial objects are point, line and region. Among the operations defined on spatial objects, spatial relationships are the most important because they make it possible, e.g., to ask for all objects in a given relationship with a query object. The Oracle Spatial cartridge implements the 9-intersection model [7] to support the computation of some topological relations.

Many spatial features (relations and attributes) can be extracted from spatial objects stored in SDB. They can be categorized as follows:

1. *geometric*, that is, based on the principles of Euclidean geometry;
2. *directional*, that is, regarding relative spatial orientation in 2 or 3D;
3. *topological*, that is, binary relations that preserve themselves under topological transformations such as translation, rotation, and scaling;
4. *hybrid*, that is, features which merge properties of two or more of the previous three categories.

This variety requires the development of a feature extractor module, named FEATEX, which also enables the coupling of SPADA with the SDB. FEATEX is implemented as an Oracle package of procedures and functions implemented in the PL-SQL language. In this way, it is possible to formulate complex SQL queries involving both spatial and aspatial data (e.g., census data). The set of spatial features that can be extracted by this module is reported in Table 1.

Table 1. Spatial features extracted by the feature extractor module.

Feature	Meaning	Type	Values
almost_parallel(Y, Z)	Parallelism relation between Y and Z	Hybrid relation	{true, false}
almost_perpendicular(Y,Z)	Perpendicularity relation between Y and Z	Hybrid relation	{true, false}
density(Y, Z)	AREA(Y)/AREA(Z)	Hybrid relation	Real
direction(Y)	Geographic direction of object Y	Directional attribute	{north, east, north_west, north_east}
distance(Y,Z)	Distance between Y and Z	Geometrical relation	Real
layer_name(Y)	Object Y type	Aspatial attribute	Layer name
line_shape(Y)	Object Y shape	Geometrical attribute	{Straight, curvilinear}
relate(Y,Z)	Topological Relation between Y and Z	Topological attribute	Type of topological relation

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 [15] which proves to be suitable for dealing with numerical data in the context of association rule mining. At the end of all this data processing, query results are stored in temporary database tables. An ad-hoc PL-SQL function transforms these tuples into ground Datalog facts of *D(S)*.

3 Application to Stockport census data

In the context of the SPIN! project we investigated the application of spatial data mining techniques to some issues reported in the Unitary Development Plans (UDP) of Stockport, one of the ten Metropolitan Districts of Greater Manchester, UK.

3.1 The data

Spatial analysis is made possible by the use of the Ordnance Survey's digital maps of the district, where several interesting layers are available, namely ED/ward/district boundaries, roads, bus priority lines, and so on. In particular, Stockport is divided into twenty-two wards for a total of 589 EDs. By joining UK 1991 census data available at the ED summarization level with ED spatial objects it is possible to investigate socio-economic issues from a spatial viewpoint. In total 89 tables, each having 120 attributes on average, have been made available for policy analysis. Census attributes 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 UDP.

3.2 Characterizing the area crossed by the M63 motorway

One of the problems is a decision-making process concerning the M63 motorway. More precisely, we are asked to describe the area of Stockport served by the M63 (i.e. the wards of Brinnington, Cheadle, Edgeley, Heaton Mersey, South Reddish) from the sociological viewpoint, in order to provide some hints for transport planners. The data considered in this analysis concerns census statistics on commuters. The description of the area is expressed by some spatial association rules at two levels of granularity. A hierarchy for the Stockport ED layer has been obtained by grouping EDs on the basis of the ward they belong to (see Fig. 3) and expressed as Datalog facts in BK.

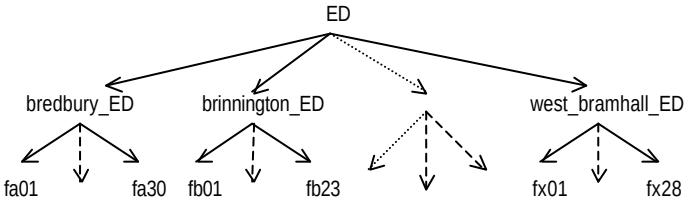

Fig. 3. An is-a hierarchy for the Stockport ED layer

Spatial association rules should relate EDs crossed by the M63 (reference objects) to EDs in the area served by the M63 (task relevant objects). The relations of intersection (EDs-motorways) and adjacency (EDs-EDs) have been extracted for the area of interest and transformed into Datalog facts of $D(S)$. The following census attributes have selected for this experiment:

- *s820161*, persons who work outside the district of usual residence and drive to work;
- *s820213*, employees and self-employed workers who reside in households with 3 or more cars and drive to work;
- *s820221*, employees and self-employed workers who reside in households with 3 or more cars and work outside the district of usual residence.

Since they refer to residents aged 16 and over, they have been normalized with respect to the total number of residents aged 16 and over (s820001). Moreover, they have been discretized by RUDE, since they are all numeric (more precisely, integer valued). At the end of this transformation process, each ED is described by three ground atoms in $D(S)$, namely *dr_out*(X, [a..b]), *cars3_dr*(X, [a..b]), *cars3_out*(X, [a..b]), where X denotes an ED, while [a..b] is one of the intervals returned by RUDE.

The key atom defining the reference objects in S is *ed_on_M63*(X), which is intensionally defined in the BK by means of the following rule:

```
ed_on_M63(X) :- intersect(X, m63).
```

The BK also includes the declarative specification of some rules for spatial qualitative reasoning, namely

```
can_reach(X, Y) :- intersect(X, m63), intersect(Y, m63), Y\=X.
close_to(X, Y) :- adjacent_to(X, Z), adjacent_to(Z, Y), Y\=X.
```

Finally, the following thresholds for support and confidence were defined: *min_sup*[1]=0.7 and *min_conf*[1]=0.9 at the first level, and *min_sup*[2]=0.5 and *min_conf*[2]=0.8 at the second level.

SPADA was run on the $D(S)$ obtained. The runtime was 331 secs for association rules at granularity level 1, and 310 secs for level 2 (data refers to a PC Pentium III 1GHz with 256 Mb RAM).

Initially, the system returned 12,925 frequent patterns out of 74,338 candidate patterns, for a total of 12,466 strong rules. By analyzing them we observed that some were actually useless, since they did not relate spatial data to census data. In other words, some association rules were pure spatial patterns, such as the following:

ed_on_M63(X) , can_reach(X,Y) → *is_a(Y,ward_on_m63_ED)* (90.0%, 100.0%)

which states that if an ED (*Y*) in the area served by the M63 can be reached from an ED crossed by the M63, then that ED is certainly (100% confidence) an ED of a ward crossed by the M63. Despite the high support and confidence, this pure spatial pattern is of no interest for transport planners.

In a second run, we decided to declare a bias for patterns containing at least one of the census attributes *dr_out*(X, [a..b]), *cars3_dr*(X, [a..b]) and *cars3_out*(X, [a..b]). The system generated 10,513 strong association rules in 1520 secs (time increased because of constraint checking for each generated pattern). Some of them have a very high support and confidence and provide the expert with some hints on the habits of commuters, such as the following association rule discovered at level 2:

(rule 27) *ed_on_M63(X), close_to(X,Y), is_a(Y,Bedgeley_ED)* →
cars3_out(X,[0.0..0.037]) , cars3_dr(X,[0.0..0.037]) *(100%,100%)*

which states that "if an ED crossed by the M63 (*X*) is close to another ED of the ward of Bedgeley (*Y*), then in that ED the percentage of people living in households with 3 or more cars and going/driving out of the district to work is very low (less than 4%)". It is important to point out that this is simply an association and does not define any kind of cause-effect relationship between the place where people live and their social habits. Another interesting spatial association rule at the same granularity level is the following:

(rule 177) *ed_on_M63(X), can_reach(X,Y)* → *is_a(Y,heaton_mersey_ED) ,*
dr_out(Y,[0.2857..0.4782]) , cars3_out(Y,[0.0..0,037]) *(80.0%,88.88%)*

which states that "if an ED *Y* in the M63 area can be reached from another one crossed by the M63 motorway (*X*), then it is in the Heaton Mersey ward and has quite a high percentage of people that drive to work but don't live in households with 3 ore more cars".

Finally, we decided to constrain the search space further, by asking only for those spatial patterns involving EDs where people have the same commuting habits. This time SPADA found only 345 strong rules (79 for level 1 and 266 for level 2) in about 833 secs. The following is an example of association found by the system at the granularity level 2:

(rule 76) *ed_on_M63(A)* → *can_reach(A,B), is_a(B,cheadle_ED), can_reach(A,C),*
C ¹ B, is_a(C,edgeley_ED), cars3_dr(C,[0.0..0.037]), cars3_dr(B,[0.0..0.037])
(90%, 90%)

which states that from an ED crossed by the M63 it is possible to reach (by the same motorway) two EDs, one in Cheadle and one in Edgley, with the same low percentage of people living in families with three or more cars and driving out of the district to work.

3.3 Accessibility of the Stepping Hill Hospital

Another problem concerning transport planning is the accessibility of the Stepping Hill Hospital in Stockport. To study this problem we decided to mine association rules relating five EDs close to the Stepping Hill Hospital (*task relevant* objects) with EDs

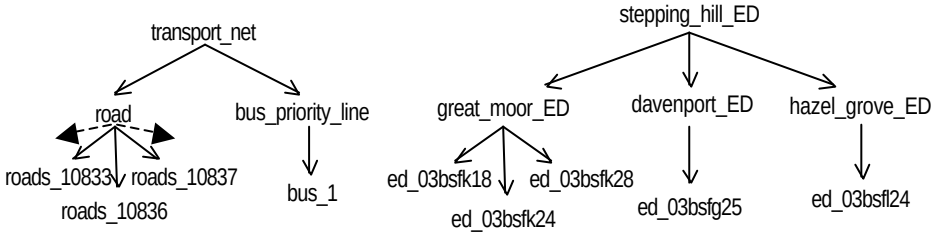

Fig. 4. Two spatial hierarchies defined for the mining task concerning the accessibility of the Stepping Hill Hospital.

within a distance of 10 Km from the hospital (reference objects). The goal is that of understanding which reference EDs have direct access to the task relevant EDs. To define the accessibility we used the Ordnance Survey data on transport network (roads and bus priority line). In the domain knowledge we defined a predicate *can_reach(X,Y)* stating that ED Y can be reached from ED X if one of the two following conditions hold:

1. Both are crossed by the same road or bus priority line;
2. From X it is possible to reach Z and from Z it is possible to reach Y (transitivity property)

This is the only spatial relation used in the spatial association rules. Our observation is 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 leaving in X might have problems to reach 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.

In this task we have two spatial hierarchies mapped into three granularity levels (Fig. 4). The declarative bias requires that the spatial association rules contain at least one of the four predicates above. SPADA generated 63 rules in 12 secs. Two of the rules returned by SPADA are the following:

ed_around_stepping_hill(A), can_reach(A,B), is_a(B,stepping_hill_ED) → *two_cars(A,[9.0e-003..0.179])*

(11.84%, 66.66%)

ed_around_stepping_hill(A), can_reach(A,B), is_a(B,stepping_hill_ED) ® *no_car(A,[0.266..0.653])*

(13.15%, 74.07 %)

They state that if from an ED it is possible to reach the area of the Stepping Hill Hospital, then the percentage of households without car can be between 26.6% and 65.3% while the percentage of households with two cars is between 9% and 17.9%. These association rules are interesting for urban planners, since they relate data on the transport network with data on sociological factors. In the future work, this task will be more deeply investigated.

4 Conclusions

In the above application we have seen that some of the discovered rules actually convey new knowledge, however the search for these "nuggets" requires a lot of tuning and efforts by 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. This is typical of exploratory data analysis, and SPADA can be considered one of the most advanced tools that data analysts currently use in their iterative knowledge discovery process.

One of the main limitations of SPADA, which is also a problem of many other relational data mining algorithms, 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. Finally, in future work, we will investigate some "interestingess measures" of rules for presentation purposes, so that the user can browse the output XML file of spatial association rules as simply as possible.

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". Thanks to Lynn Rudd for her help in reading the paper.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of the Twentieth VLDB Conference, Santiago, Chile (1994)
2. Ceri, S., Gottlob, G., Tanca, L.: What you Always Wanted to Know About Datalog (And Never Dared to Ask). IEEE Transactions on Knowledge and Data Engineering 1,1, (1989) 146-166
3. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations. In: Lavrac, N., Dzeroski, S. (eds.): Inductive Logic Programming. LNCS 1297, Springer-Verlag, Berlin

(1997) 125-132
4. Dehaspe, L., Toivonen, H.: Discovery of frequent Datalog patterns. Data Mining and Knowledge Discovery 3(1) (1999) 7-36
5. De Raedt, L.: Interactive Theory Revision. Academic Press, London (1992)
6. Dzeroski, S., Lavrac, N. (eds.): Relational Data Mining, Springer-Verlag, Berlin (2001)
7. Egenhofer, M.J, Herring, J.R.: Categorizing binary topological relations between regions, lines, and points in geographic databases. In M.J. Egenhofer, D.M. Mark, and J.R. Herring (eds.): The 9-intersection: Formalism and ist Use for Natural-language Spatial Predicates, pages (1994) 183-271
8. Güting, R.H.: An introduction to spatial database systems. VLDB Journal, 3,4 (1994) 357-399.
9. Han, J., Fu, Y.: Discovery of multiple-level association rules from large databases. In U. Dayal, P.M.D. Gray, S. Nishio (eds.): VLDB'95, Proceedings of the 21st International Conference on Very Large Data Bases, Morgan-Kaufmann (1995) 420-431.
10. Han, J., Koperski, K., Stefanovic, N.: GeoMiner: A System Prototype for Spatial Data Mining. In Peckham, J. (ed.): SIGMOD 1997, Proceedings of the ACM-SIGMOD International Conference on Management of Data. SIGMOD Record 26, 2 (1997) 553-556.
11. Helft, N.: Inductive generalization: a logical framework. In: Bratko, I., Lavrac, N. (eds): Progress in Machine Learning. Sigma Press (1987) 149-157
12. Koperski, K., Han, J.: 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 (1995) 47-66.
13. Koperski, K., Adhikary, J., Han, J.: Spatial Data Mining: Progress and Challenges. In Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada (1996)
14. Lavrac, N., Džeroski, S.: Inductive Logic Programming: techniques and applications. Ellis Horwood, Chichester (1994)
15. Ludl, M.-C., Widmer, G.: Relative Unsupervised Discretization for Association Rule Mining. In D.A. Zighed, H.J. Komorowski, J.M. Zytkow (Eds.): Principles of Data Mining and Knowledge Discovery, LNCS 1910, Springer-Verlag (2000) 148-158.
16. Malerba, D., Esposito, F., Lanza, A., Lisi, F.A.: Machine learning for information extraction from topographic maps. In H. J. Miller & J. Han (Eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, London, UK, (2001) 291-314
17. Malerba, D., Lisi, F.A.: An ILP method for spatial association rule mining. Working notes of the First Workshop on Multi-Relational Data Mining, Freiburg, Germany (2001) 18-29.
18. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3) (1997) 259-289
19. Muggleton, S. (ed): Inductive Logic Programming, Academic Press, London (1992)
20. Nienhuys-Cheng, S.-H., deWolf, R.: Foundations of inductive logic programming. Springer, Heidelberg, Germany (1997)
21. Plotkin, G.: A note on inductive generalization. Machine Intelligence, 5 (1970) 153-163
22. Saporta, G.: Data Mining and Official Statistics. Atti della Quinta Conferenza Nazionale di Statistica, Rome (2000) 15-17