

Census Data Mining – An Application

Willi Klösgen and Michael May

Fraunhofer Institute for Autonomous Intelligent Systems
Knowledge Discovery Team
D-53757 Sankt Augustin, Germany
{kloesgen, may}@ais.fhg.de

Abstract. Because of data privacy regulations, census data are available for analysis only in aggregated form. Primary data (responses of persons) are aggregated in many cross tabulations for small geographical units. Thus the target objects of secondary analysis are small areas (enumeration districts or wards). Any cell or marginal of a cross tabulation can be used as variable on these target objects. The target objects can be linked with other spatial objects (e.g. rivers, roads, railway lines) for spatial analyses. In this paper we discuss the special requirements that occur for this type of aggregate data mining including spatial analyses. We show an application of SubgroupMiner, which is an advanced subgroup mining system supporting multirelational hypotheses, efficient data base integration, discovery of causal subgroup structures, and visualization based interaction options.

1 Introduction: Mining Spatial Subgroups

The goal of spatio-temporal data mining is to discover attributive, spatial and temporal patterns and to analyse their potential interactions. The patterns describe hypotheses about spatially and timely referenced data. Spatial patterns additionally include variables that do not only refer to properties of the analysis objects themselves (attributive patterns), but also to spatially neighbored objects and their properties. Temporal patterns include analyzing change and trend. In this paper we focus on spatial patterns from the perspective of the subgroup mining paradigm.

Subgroup Mining [Klösgen 1991, 1996, 2002] is used to analyze dependencies between a target and a large number of explanatory variables. The approach can be applied for exploration, classification, or optimization. Interesting subgroups with some designated type of deviation, change, or trend pattern are searched, e.g. subgroups with an over proportionally high target share (mean) for a value of a discrete (continuous) target, or subgroups for which the target share (mean) has significantly changed between two time points, or shows a trend pattern for a sequence of time points. Subgroups are subsets of analysis objects described by selection expressions of a query language, e.g. simple conjunctive attributive selections, or multirelational selections joining several tables representing different (spatial) objects. Interestingness aspects include statistical significance, interpretability, and non-redundancy of subgroups.

A spatial query language that includes operations on the spatial references of objects describes spatial subgroups. A spatial subgroup, for instance, consists of the young children that live near a nuclear power plant of type boiling water reactor. A spatial predicate (nearby) operates on the coordinates of the spatially referenced objects persons and power plants. Further some attributive selectors (age = young, type = boiling_water_reactor) define which objects belong to the subgroup.

While the spatial dimension is covered by spatial description languages for subgroups, the temporal dimension is represented by change or trend patterns that determine the evaluation criteria for an interesting or statistically significant subgroup.

The subgroup-mining paradigm provides the main components for these approaches: description languages for subgroups, search strategies in hypothesis spaces, hypothesis evaluation, scaling, visualization, and causality analysis.

This paper describes an application of *SubgroupMiner* on census data. The goal of the system is to provide a spatial and temporal mining tool. The system improves all stages of the knowledge discovery cycle:

- Data Access: Subgroup Mining is partially embedded in a spatial database, where analysis is performed. No data transformation is necessary and the same data is used for analysis and mapping in a GIS. This is important for the applicability of the system since pre-processing of spatial data is error-prone and complex.
- Pre-processing and analysis: SubgroupMiner handles both numeric and nominal target attributes. For numeric explanatory variables on-the-fly discretization is performed. Spatial and non-spatial joins are executed dynamically.
- Post-processing and Interpretation: Similar subgroups are clustered according to degree of overlap of instances to identify multicollinearities. A Bayesian network between subgroups can be inferred to support causal analysis.
- Visualisation. SubgroupMiner is dynamically linked to a GIS, so that spatial subgroups are visualized on a map. This allows to bring in background knowledge into the exploration process, performing several forms of interactive sensitivity analysis and exploring relations to further variables and spatial features.

The paper is organized as follows. Section 2 introduces the context of census data. In section 3, the representation of spatial data and spatial subgroups is discussed. The analysis framework is presented in section 4.

2 Census Data

We discuss an application example to illustrate the special requirements of census data mining and especially show the interaction between spatial subgroup mining and GIS mapping. The UK Census, undertaken every ten years, collects population and other statistics essential to those who have to plan and allocate resources. Major customers include departments of national and local government, and providers of services such as health and education.

In the example, we analyse UK 1991 census data for North West England, one of the twelve regions in UK. The basic geographical units used in our analyses are the 1011 wards situated in the 43 local authorities of NW England. Deprivation indices that are the focus of our analysis are given for these wards. The next geographical level below wards are enumeration districts.

Census data can be aggregated to any level of spatial unit. The appropriate level for an analysis depends of the problem and especially the available secondary data (e.g. on deprivation). Lower levels ensure a higher homogeneity of aggregated variables thus providing a higher potential to identify and evaluate hypotheses on individuals (persons). On the other side, lower levels require scalable methods, since the number of the main analysis objects can get very large when the overall region (as North West England) is not strongly limited.

For the 2001 Census England and Wales had 116,895 EDs with an average size close to 200 households (450 people). Census data are available as aggregated cross tabulations for each geographical unit (wards). Table 1 is one (small) cross tabulation of the about 100 tabulations that are provided for different dimensions (economic position, ethnic group, gender in Table 1). Each of the cells of the cross tabulations (e.g. 54 cells of Table 1) can be used as a variable on the geographical units. Thus some 10.000 variables are available for the main analysis objects (wards). Typically a small subset of these variables is selected for a special analysis.

Table S09 Economic position and ethnic group: Residents aged 16 and over						
Economic position	Ethnic group					
	TOTAL PERSONS	White	Black and other groups	Indian Pakistani and Bangladeshi	Chinese and other groups	Persons born in Ireland
TOTAL PERSONS	1	2	3	4	5	6
Males 16 and over	7	8	9	10	11	12
Economically active	13	14	15	16	17	18
Unemployed	19	20	21	22	23	24
Economically inactive	25	26	27	28	29	30
Females 16 and over	31	32	33	34	35	36
Economically active	37	38	39	40	41	42
Unemployed	43	44	45	46	47	48
Economically inactive	49	50	51	52	53	54

Table 1: An aggregated cross tabulation available e.g. for all wards

Also available are detailed geographical layers, among them streets, rivers, buildings, railway lines, shopping areas. Table 2 shows these layers including subtypes for some layers. These layers have own attributes such as featcode (indicating the subtype of the spatial object) or length (of line). Only a few of the many point layers on sports and tourist facilities are included in our analyses, because most of them seem not relevant for the selected target variables.

Layer name	Description	Type	Objects
Motorway	Motorway	Line	494
PrimRoad	Motorway (over), Motorway tunnel Primary route, dual carriageway Primary route, dual carriageway (over) Primary route, single carriageway Primary route, single carriageway (over) Primary route, narrow Primary route, narrow (over) Primary route tunnel	Line	3945
A_Road	A road, dual carriageway Other subtypes: see PrimRoad	Line	3882
B_Road	B road, dual carriageway Other subtypes: see PrimRoad	Line	4368
Mnr_Rd4o	Minor road over 4 meters wide Minor road over 4 meters wide (over) Minor road over 4 meters wide tunnel	Line	9705
Mnr_Rd4u	Minor road under 4 meters wide / over / tunnel	Line	8756
Railway	Railway, standard gauge Railway, standard gauge (over) Railway, narrow gauge / narrow gauge (over) Railway tunnel / Railway station	Line	4231
UrbAreaL	Large Urban Area (outer limit) Large Urban Area (inner limit)	Line	384
UrbAreaS	Small Urban Area (outer limit) / (inner limit)	Line	2235
Water	Inland water (inner limit) Inland water (outer limit)	Line	438
River	River (primary), source / middle / lower River (secondary), source / middle / lower River (other and drains)	Line	12103
Canal	Canal Canal tunnel / Canal (over)	Line	968
Wood	Wood/Forest (inner limit) Wood/Forest (outer limit)	Line	859
Foreshor	Foreshore (sand, inner limit) Foreshore (other) and offshore rocks (il) Foreshore (sand, outer limit) Foreshore (other) and offshore rocks (ol)	Line	209
National	National boundary	Line	12
County	County boundary	Line	88
District	District boundary	Line	61
Park	National park/forest park	Line	11
CampCara	Camping and caravanning combined sites	Point	212
...			

Table 2: Geographic Layers (spatial objects of type line / point)

Deprivation indices are selected as target variables, i.e. the analysis goal is to gain some information on attributive and spatial dependencies of these variables and their interactions. Information from the Census (sometimes in combination with other variables) is often combined into a single index score (Table 3) to show the level of deprivation in an area. Over the years a number of different such indices have been developed for different applications. In general, these measures show a strong correlation between the level of deprivation and a variety of health indicators.

Variable	Jarman	Townsend	Carstairs	DoE
Unemployment	X	X	X(males)	X
Low social class	X		X	
Overcrowded households	X	X	X	X
Households lacking basic amenities				X
Single parent	X			
Under age 5	X			
Lone pensioner	X			
residents who have changed address in the previous year	X			
head of household born in the new commonwealth	X			
Households with no car		X	X	X
Not owner occupied		X		
Children living in flats				X
Children in low earning households				X
Low educational participation				X
Low educational attainment				X
Standard Mortality Ratios				X
Male long term unemployment				X
Income Support recipients				X
Home Insurance Weightings				X

Table 3: Variables used in the calculation of four deprivation indices

Individual variables are usually weighted before they are combined. One of the simplest approaches is to normalize the scores around a mean of zero and express individual components in terms of the number of standard deviations. As a result, the measures are ordinal, hence they are often accompanied by ranking.

3 Representation of Spatial Data and of Spatial Subgroups

Census data, deprivation indices, and the data for the other geographic layers are loaded into a spatial database system (Oracle Spatial). Before analysing the data, a special view is constructed by selecting a subset of the very many census variables and their normalization.

Most modern Geographic Information Systems (GIS) use an underlying Database Management System (DBMS) for data storage and retrieval. In object-relational databases spatial data is represented as follows:

A **spatial data base S** is a set of relations R_1, \dots, R_n such that each relation R_i in S has a geometry attribute G_i or an identifier A_i such that R_i can be linked (joined) to a relation R_k in S having a geometry attribute G_k .

A **geometry attributes G_i** consists of ordered sets of x,y -pairs defining points, lines, or polygons.

Different types of spatial objects are organized in different relations R_i , e.g. roads, rivers, enumeration districts or wards, buildings. Each such relation is called a **geographical layer**. Each layer can have its own set of attributes A_1, \dots, A_n , called **thematic data**, and at most one geometry attribute G . The attributes A_1, \dots, A_n are the usual numeric or nominal attributes found in a relational database.

For querying multirelational spatial data, a major extension a spatial database adds is the efficient implementation of a **spatial join**. A spatial join links two relations each having a geometry attribute based on distance or topological relations (inside, covers, adjacent, touches). For supporting spatial joins efficiently, special purpose indexes like KD-trees or Quadrees are used.

Preprocessing vs. dynamic approaches

The above description shows that a GIS representation is *multi-relational*. A **relation graph** is shown in Figure 1 for seven tables. A link in this graph connects two tables. Foreign keys are simple links: e.g. from diagnoses to persons. Implicit spatial links are given by the spatial references of objects. E.g. a spatial predicate relates persons and industrial plants: a person lives near a plant (either precalculated and materialized as in Figure 1, or dynamically calculated during analysis).

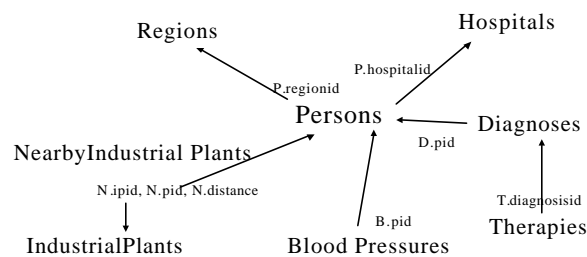


Figure 1: Object classes (tables) of a multirelational spatial application

While the relation graph of Figure 1 has a maximal depth of 3 (e.g. persons, diagnoses, therapies), the relation graph for the census application is a simple star shaped graph. The spatial objects (table 2) are arranged around the target analysis units (wards) such that the maximal depth is only 2.

There are different strategies to deal with multi-relational data in data mining. One possibility is to **preprocess** the data and join the relevant variables from secondary

tables to a target table. The resulting table has a rectangular form and can be analysed using standard methods like decision trees or regression. Multirelational analysis approaches are not necessary in this case.

However the preprocessing approach has disadvantages. First, the set of relations S and the set of possible joins L between tables in S constrain the hypothesis space. Each type of object can spatially interact with any other type of spatial object in numerous ways according to the topological relations. When all these joins are meaningful, the set L is prohibitively large. It would be desirable to set up the problem in such a way that at least in principle all hypotheses are in the search space of an algorithm, or, at least, that it is not the preprocessing that prevents this.

The extended target table generated by preprocessing will include only a small part of the information available in the original tables. But often, it is not known before, in which parts of the data the interesting results can be found. Thus it may be difficult to select the potentially relevant variables and aggregations. When e.g. the number of diagnoses of a person is aggregated from the diagnoses table to the persons table, the correlation of this number with other attributes from the diagnoses table is lost (e.g. number of diagnoses of a special type). Thus it would be necessary to aggregate the diagnosis numbers for each value of (some) attributes of the diagnoses table and possibly also for combinations of values. Thus the number of derived attributes will easily explode for complex relation graphs.

Thus secondly, there is an obvious trade-off between the computing time needed for pre-processing, and the required space for redundant storage on one hand, and the computational complexity of the analysis run. Preprocessing may take a long time, and much of the preprocessing may turn out to be unnecessary since a certain part of the hypothesis space will be pruned away by the data-mining algorithm anyway. It would be desirable to perform expensive spatial joins only for that part of the hypothesis space that is really explored during search.

Thirdly, a further disadvantage occurs in applications where the data can change, either due to adding, deleting or updating. Since pre-processing leads to redundant data storage, we suffer the usual problems of non-normalized data storage, thoroughly investigated in the database literature.

An advantage of preprocessing with respect to a dynamic approach is that once the data is preprocessed the calculation has not to be done again. But here a dynamic approach could cache search results to improve efficiency.

The general approach we apply for multi-relational data mining relies on dynamically joining the tables. The joins that are arranged during search follow paths in a prespecified relation graph. The relation graph includes the edges between table nodes. In the multirelational model we have k object classes O_1, \dots, O_k that are represented by tables. They are the nodes of the relation graph. Further we have a set of links where each link is a relation between two object classes and is represented by a prespecified link condition that defines a subset of the product of the two object classes. These links are the edges in the relation graph.

When deciding on dynamic versus static spatial joins for the census application, the following characteristics are important: structure of the relation graph, number

of thematic attributes in geographical layers, size of the relations, and data dynamics. The number of attributes that can be induced for the primary table (e.g. wards) depends on the depth of the relation graph and the number of thematic attributes. E.g. for discrete thematic attributes, an own attribute can be induced for each value and for each combination of values of several attributes (when combinations of attributes are included). Such an attribute could hold the information that a ward is intersected by a road of type A (or of type A and length L). The number of potential induced attributes exponentially grows with the depth of the relation graph. With each additional layer joined, the number of induced attributes is multiplied by the number of combinations of attribute values of the additional layer. Since the structure of the relation graph is simple (depth = 2; only joins between wards and geographic layers and no joins between geographic layers) and the number of thematic attributes is small, the number of induced attributes is manageable for this application.

Also the data dynamics are extremely low. The UK census is undertaken every ten years and also the other layers are fairly stable, such that no data update problems occur. Therefore the generation of an overall ward relation extended by all the possible induced attributes from the geographical layers would be preferable, because efficiency (computation time) of analyses will be higher avoiding expensive joins during analyses. This would especially be necessary for finer levels of target objects resulting in large tables (enumeration districts instead of wards). We apply a dynamic join approach (no precalculated universal join) for the analyses (section 4), because the joins are performed for these table sizes (1011 wards and e.g. 9705 roads) within some few seconds such that an interactive analysis is still possible.

In general, an extended dynamic strategy could be useful. This strategy would not require an universal target relation constructed in a preprocessing step, but would dynamically store the induced attributes, which are generated during a multirelation search by joining several tables, into the target table. In a subsequent analysis, it would not be necessary any more to construct the join again, but the stored induced attributes could be accessed from the target table.

Spatial subgroups

A multirelational subgroup is a subset of target objects that is defined by conditions on variables including variables induced from secondary tables. These conditions are described by a query that consists of conjunctive selectors. The query language of SubgroupMiner is described in (Klösigen and May 2002) and is only summarized here referring to the main options of a multirelational subgroup language.

The first option determines which links (joins) between the various (spatial) object classes are selected, i.e. which links are used to construct a (next) conjunctive selector. SubgroupMiner exploits a predefined Relation Graph (Figure 1), that includes the possible links and their details (which attributes and aggregations).

As a basic aggregation option, SubgroupMiner uses the existential quantifier, e.g. the subgroup *Wards.male=high and Rivers.type=primary* is a condensed description of the set of wards with a high share of males and intersected by at least one primary river.

A next option includes aggregate functions such as count, average, max, min (Knobbe et al 2001). The subgroup *Wards.area=large and Rivers.max(length)<l* describes wards with a large area and only intersected by rivers with a limited length.

Another option includes variables to distinguish several objects of one class for applying a predicate on these objects, e.g. *wards situated near two industrial plants with special conditions* (e.g. distance between two plants is small). Such selections are typically included in ILP approaches such as Malerba and Lisi (2001).

The type of refinement is another option. There are two possibilities how a further thematic attribute can refine a subgroup. E.g. the *wards intersected by at least one primary river* can be refined (introducing an additional conjunctive condition) by *wards intersected by at least one primary river and intersected by at least one polluted river*. Another refinement are *wards intersected by at least one primary and polluted river*. The type of refinement is important for aggregation functions.

Details on how these options are applied (e.g. which aggregation functions on which variables, the number of objects to be distinguished and the predicate(s) to be applied on the objects) are prespecified in the Relation Graph.

4 Applying Subgroup Mining to Deprivation Indices

After loading census, deprivation, and geographical data, an Oracle Spatial database holds a table for each census cross tabulation and each geographical layer. As a next preprocessing step, a tool is used to select variables from the very many census tables and to normalize them. Generally we select variables from the margins of the cross tabulations and not so often the inner cells (e.g. for cross table 1: *total Chinese persons* and not *Chinese unemployed males*). Especially for cross tabulations with very deep classifications, the cells are correlated providing (too much) redundancy. Normalization is necessary to adjust the different sizes of wards, thus not the number of Chinese persons, but the number of Chinese persons divided by the total number of persons is included in the resulting ward table. Several normalization options can often be useful, e.g. unemployed males wrt males or total persons. The selection and normalization tool will typically be used many times during an analysis process to include additional variables or to modify normalizations.

With this preprocessing step performed, we can analyse a target table including numerical variables derived from the census (shares such as white persons in a ward related to all residents) and join the target table with geographic layers. Selectors of subgroup descriptions (section 3) need discretizations for numerical variables. Subgroup Miner can automatically discretize the numerical variables during an analysis or rely on predefined discretizations. We at first use the simplest automatic option that generates only two selectors for a numerical variable, e.g. *Wards.males=high* and *Ward.males=low* comparing the percentage of males in a ward to the average percentage over all wards.

In a first experiment, we select *carstairsidx* as target variable and include all selected census variables as well as the other deprivation variables into search to build subgroups. The target variable is numeric and the system uses the mean pattern as a

default. Thus subgroups are searched for which the mean of the target variable is significantly higher than the total mean (over all wards). The found subgroup *low_social=high and married=low and unempl_male=high* (subset of wards with above average value of low_social and below average value of percentage of married persons and above average value of unempl_male) has e.g. an average value for the Carstairs index of 6,24 compared to the overall average of 0,94.

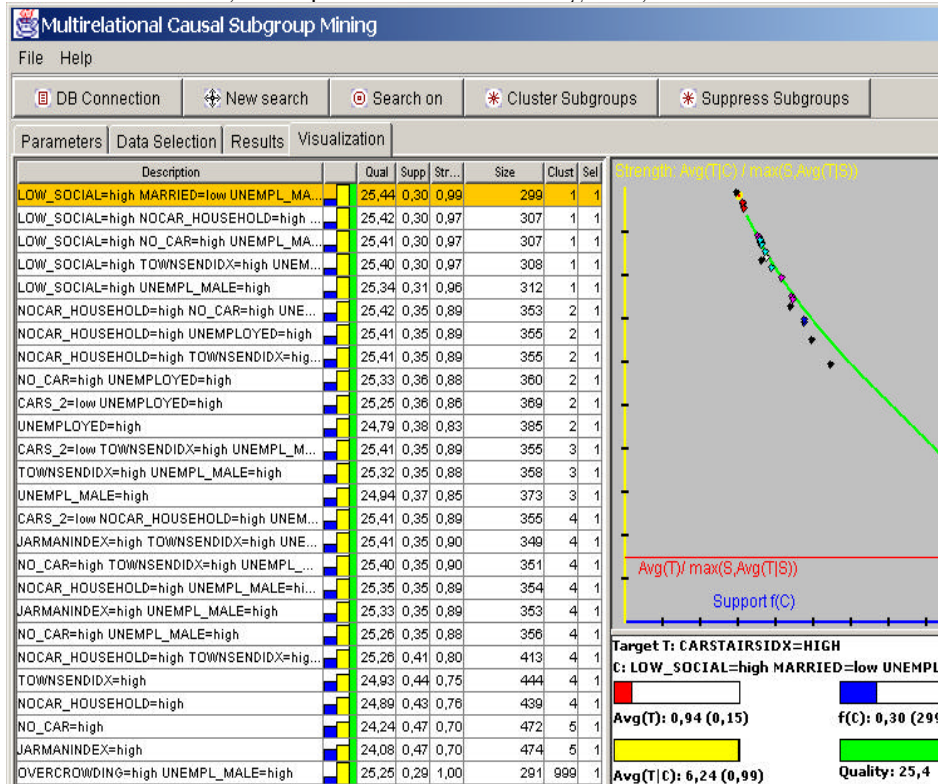


Figure 2: Subgroups with high Carstairs index (all selected census attributes included)

The found subgroups (shown in Figure 2) are ordered by clusters after applying a clustering option (complete linkage method, similarity measure for pair of subgroups based on their overlapping). Five clusters and five remaining single element clusters (999) have been identified. The first cluster includes subgroups with refinements of *low_social=high* and the second cluster subgroups with *no-car_household=high*. The third and fourth cluster include subgroups with *unempl_males=high* (third cluster consists of refinements of fourth cluster). These clusters fairly well reproduce the definition of the Carstairs index (compare section 2). The fourth variable included in the definition of the Carstairs index (overcrowding) occurs in two single element clusters. The results also show the high correlation between the four deprivation indices.

This first experiment has been performed to check the validity of subgroup results. Although some very simple default options have been used such as discretization by average value and mean pattern for ordinal target variable (compare section 2), the results reproduce the definition of the target variable (similar results are found for the negative index, i.e. subgroups with a significantly low average value of the index). More detailed subgroup analyses (not shown here) study the relative importance of the contributing variables of the indices (an index is constructed as a weighted sum of not independent variables that strongly correlate). Other analyses compare the four deprivation indices (e.g. subgroups with a difference of indices).

Since the distribution of many census variables is very skew for the wards, we next apply a more profound discretization based on clustering. Then dense discretization intervals are constructed. When there are e.g. very many small values of a variable and some middle and a few high values, two or three homogeneous intervals are identified based on optimizing the boundary points where e.g. the first interval includes all the small values. This clustering method can especially exclude variables that are not useful to build subgroups (e.g. one cluster includes nearly all values due to the extremely skew distribution of the variable).

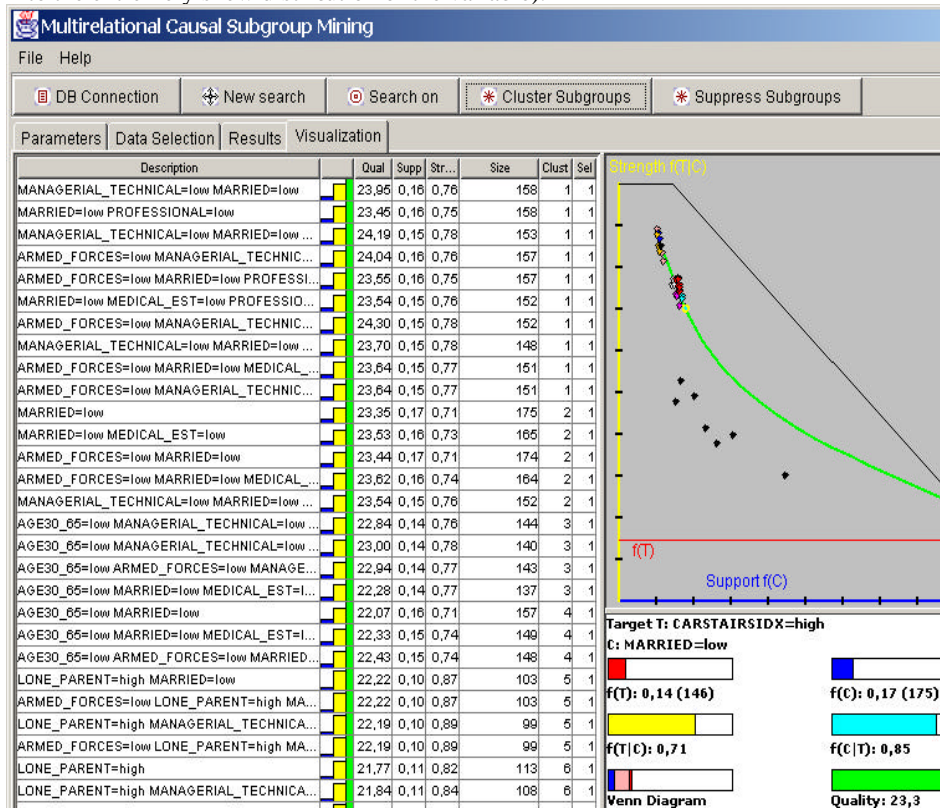


Figure 3: Subgroups dependent of high Carstairs index

Further we exclude all variables that are included in the definition of the Carstairs index from search and also the other deprivation indices. To avoid problems of the mean pattern with only ordinally scaled target variables, we select as target variable the binary variable *carstairsidx=high* (high interval identified with discretization).

Seven subgroup clusters and some single subgroups are identified by the subgroup clustering method (Figure 3). To reduce these results, a redundancy elimination algorithm is run suppressing subgroups that are conditionally independent of the target group given another subgroup. This Bayesian Network based causality approach (Klösger 2002) suppresses 24 of the 40 subgroups resulting in the following “causal” subgroups (in the ordering of Figure 3).

Causal Subgroups: 5 7 13 14 23 25 27 28 30 31 33 34 36 37 38 40

A summary of the main “causal” factors (including the dual problem: wards with low Carstairs index) is shown in Table 4.

Variable	Category	Subgroup Size	TargetRate T in Subgroup T = High Carstairs index (14 % of all wards)	TargetRate T in Subgroup T = Low Carstairs index (52 % of all wards)
lone_parent	high	11 %	82 %	
	low	50 %		86 %
age0-4	high	19 %	50 %	
	low	35 %		87 %
unskilled	high	16 %	53 %	
	low	45 %		84 %
long_term_illness	high	22 %	42 %	
	low	34 %		90 %
partly-skilled	high	15 %	49 %	
	low	43 %		89 %
married	low	17 %	71 %	
	high	46 %		91 %
managerial_technical	low	44 %	31 %	
age6-29	low	34 %		88 %
cohabit	low	41 %		75 %

Table 4: Main single factors causing high / low Carstairs index

Underprivileged wards (e.g. high Carstairs index defined by a combined high rate of unemployed males, low social status, overcrowded households, households with no car) tend to be populated by lone parents, families with young children, unskilled and partly skilled persons, long term ill persons, unmarried persons. The dual properties characterize privileged wards. Since data are only given as aggregates thus characterizing wards and not individual persons, it can not be concluded that these subgroups (e.g. lone parents or unmarried persons) hold the Carstairs properties on the individual level. Lone parents have not necessarily a low social status, but tend to live in areas with a high rate of persons with a low social status. Using a lower aggregation level (enumeration units or the still more homogeneous output areas) will increase the possibility to infer individual hypotheses. Discussing these problems of aggregate data analysis are beyond the scope of this paper.

Next we analyse the dependence of the Carstairs index of the spatial objects. We include all line objects listed in Table 2 and two thematic attributes for each spatial object class (*featcode* represents subtypes and *length* is a numerical attribute holding the length of the line object). Table 5 summarizes the results.

Subgroups with high average Carstairs index overall carstairs ave rage for all wards = 0.94	Quality (significance)	Support (wards#)	Carstairs Average
DISTRICT.DISTRICT_ID=6	6.23	36	5.32
DISTRICT.DISTRICT_ID=22	3.99	35	3.79
DISTRICT.LENGTH=high	3.48	174	1.97
DISTRICT.ALL	3.16	240	1.71
COUNTY.COUNTY_ID=5	4.38	12	6.34
RIVER.ALL	3.26	857	1.13
MNR_RD4U.ALL	1.22	734	1.05
PARK.PARK_ID=2	1.20	76	1.51
Subgroups with low average Carstairs index			
WOOD.LENGTH=high	6.20	215	-0.67
WOOD.FEAT=inner limit	5.90	48	-2.62
WOOD.ALL	5.74	344	-0.13
WATER.LENGTH=high	6.06	128	-1.20
WATER.ALL	4.70	263	-0.12
WATER.FEAT=inner limit	2.23	6	-2.95
PRIMROAD.FEAT=dual carriageway, over other feature	5.14	44	-2.31
PRIMROAD.FEAT= dual carriageway	4.74	152	-0.58
RIVER.FEAT=secondary, source	4.99	301	-0.09
RIVER.FEAT=secondary,middle	4.81	160	-0.55
RIVER.FEAT=primary,lower	2.80	35	-1.05
RIVER.FEAT= primary,source	2.33	45	-0.51
MOTORWAY.LENGTH=low	4.77	145	-0.66
MOTORWAY.ALL	4.60	210	-0.27
MOTORWAY.FEAT=over other feature	4.35	125	-0.62
RAILWAY.FEAT=standard gauge, over other feature	4.68	248	-0.16
RAILWAY.FEAT=tunnel	4.09	34	-2.02
B_ROAD.FEAT=single carriageway, over other feature	4.24	162	-0.36
MNR_RD4O.FEAT=over other feature	4.12	220	-0.11
URBAREAL.LENGTH=low	4.07	247	-0.02
URBAREAL.FEAT=large , inner limit	2.99	79	-0.44
NATIONAL.ALL	3.94	21	-2.71
CANAL.FEAT=over other feature	3.77	38	-1.63
CANAL.LENGTH=high	2.86	141	-0.01
COUNTY.LENGTH=low	3.65	87	-0.66
COUNTY.COUNTY_ID=31	3.02	8	-3.62
PARK.PARK_ID=1	3.54	26	-1.99
DISTRICT.DISTRICT.ID=2	3.47	22	-2.19
MNR_RD4U.FEAT=over other feature	3.47	134	-0.25
A_ROAD.FEAT=single carriageway, over other feature	3.13	105	-0.29
A_ROAD.FEAT=dual carriageway	2.41	102	-0.02

Table 5: Spatial Subgroups with high / low average Carstairs index

(Underprivileged) wards with a high Carstairs index are situated near (large) district boundaries or boundaries of special single districts or counties, near rivers, and

near main roads under 4 m wide. However, wards situated near the source or middle part of a secondary river or near the source or lower part of a primary river have a low Carstairs index. Also main roads under 4 m wide with this feature dominating another feature (over other) have a low Carstairs index.

There are more spatial characteristics for wards with a low Carstairs index (privileged wards). They are e.g. located near woods (especially large woods or inner areas of woods), near waters (especially large waters or inner areas), near dual carriageway primeroads, motorways, tunnels of railways, inner parts of large urban areas, national boundaries, long canals.

The way these data mining results are presented to the user is essential for their appropriate interpretation. We use a combination of cartographic and non-cartographic displays linked together through simultaneous dynamic highlighting of the corresponding parts.

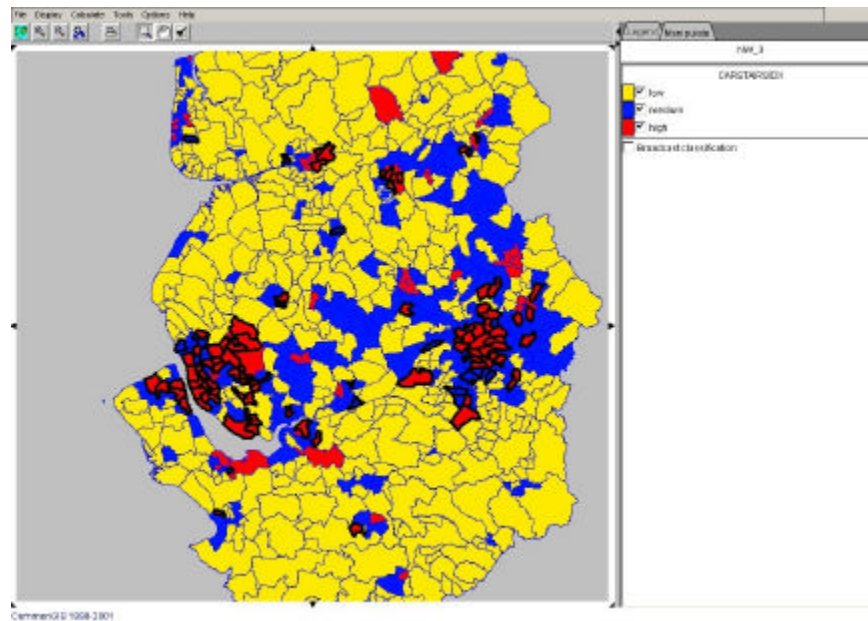


Fig. 5. Wards satisfying the subgroup description C (`lone_parent=high`) are highlighted with a thicker black line. Wards also satisfying the target (high Carstairs index) are in a lighter color.

The user navigates in the list of subgroups (Fig. 3), which are dynamically highlighted in the map window (Fig. 5). As mapping tool, the the CommonGIS system [Andrienko and Andrienko 1999] is integrated, whose strengths lie in the dynamic manipulation of spatial statistical data.

The application has been developed within the IST-SPIN!-project, that integrates a variety of spatial analysis tools into a spatial data mining platform based on Enter-

prise Java Beans [May and Savinov 2001]. Besides Subgroup Mining these are Spatial Association Rules [Malerba and Lisi 2001], Bayesian Markov Chain Monte Carlo and the Geographical Analysis Machine GAM [Openshaw et al. 1999]. Data are provided by the partners Manchester University and Metropolitan University.

Conclusion and Future Work

Two-layer database integration of multirelational subgroup-mining search strategies has proven as an efficient and easy portable architecture. Scalability of subgroup mining for large datasets has been realized for single relational and multi-relational applications with a not complex relation graph. The complexity of a multirelational application mainly depends of the number of links, the number of secondary attributes to be selected, the depth of the relation graph, and the aggregation operations. Scalability is also a problem, when several tables are very large. Some spatial predicates are expensive to calculate. Then sometimes a grid for approximate (quick) spatial operations can be selected that is sufficiently accurate for data mining purposes. When several large tables are spatially joined, it is advantageous to precalculate the spatial operations. We are currently investigating options to combine static and dynamic links; links can e.g. be declared as static in the relation graph definition. The specification of textual link conditions and predicates in the relation graph that are then embedded into a complex SQL query has proven as a powerful tool to construct multirelational spatial applications. While the analyses of deprivation indices described in this paper treat very general problems with fairly obvious results, a more detailed study on the differences and problems of the various indices is performed as a pilot application within the SPIN! Project.

Acknowledgments

Work was partly supported by SPIN! – Spatial Mining for Data of Public Interest (IST Program, IST-1999-10563, 2000-2002).

References

- G. Andrienko, N. Andrienko, N. Interactive Maps for Visual Data Exploration, *International Journal of Geographical Information Science* 13(5), 355-374, 1999
- W. Klösgen 1991. Visualization and Adaptivity in the Statistics Interpreter EXPLORA. In *Proceedings of the 1991 Workshop on KDD*, ed. Piatetsky-Shapiro, G., pp. 25-34.
- W. Klösgen 1996. Explora: A Multipattern and Multistrategy Discovery Assistant. *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Cambridge, MA: MIT Press, 249–271.
- W. Klösgen 2002. Subgroup Discovery. Chapter 16.3 in: *Handbook of Data Mining and Knowledge Discovery*, eds. W. Klösgen and J. Zytkow, Oxford University Press, New York.
- W. Klösgen 2002. Causal Subgroup Mining. To appear.
- W. Klösgen, M. May 2002. Spatial Subgroup Mining. In *Proceedings of Sixth European Symposium on Principles of KDD (PKDD 2002)*, Berlin:Springer.
- A. Knobbe, M. de Haas, A. Siebes 2001. Propositionalisation and Aggregates. In *Proceedings of Fifth European Symposium on Principles of KDD (PKDD 2001)*, Berlin:Springer.
- D. Malerba, F. Lisi 2001. Discovering Associations between Spatial Objects: An ILP Application. *ILP 2001*, LNAI 2157, Berlin: Springer, 156-163.
- M. May, Savinov, A. An Architecture for the SPIN! Spatial Data Mining Platform, *Proc. New Techniques and Technologies for Statistics, NTTS 2001*, 467-472, Eurostat, 2001
- Openshaw, S., Turton, I., Macgill, J. and Davy, J. Putting the Geographical Analysis Machine on the Internet, in Gittings, B. (ed.) *Innovations in GIS 6*, 1999