

CONFERENCE OF EUROPEAN STATISTICIANS

**Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of
Statistics and Geography**

(Tallinn, Estonia, 25-28 September 2001)

Topic (ii): New technological solutions, including those based on online data access

**PROJECT "SPIN!"
INTEGRATING DATA MINING AND GEOGRAPHIC INFORMATION SYSTEMS**

Submitted by the Fraunhofer Institute for Autonomous Intelligent Systems¹

Invited paper

I. INTRODUCTION

1. Geographic Information Systems (GIS) are widely used for analysing and visualising geo-referenced data. In the last few years, a new generation of Geographic Information Systems has emerged that extends the interactivity of dynamically generated maps, greatly enhancing visual exploratory data analysis ([1], [3], [6], [13]). While being an exciting development for automating cartography, these systems have limited capabilities to visualise attribute interaction on a map having more than a few dimensions. Hence, complex multi-variate dependencies are easily overlooked.

2. Searching for multi-variate dependencies is where data mining promises great benefits. Data mining is the partially automated search for hidden patterns in typically large and multi-dimensional databases. It draws on results in machine learning, statistics and database theory. Some data mining methods, such as k-nearest neighbor, are extensions of statistical techniques known for a long time. Others, especially from the area of machine learning and inductive logic programming (ILP), are essentially new (cf. [9]). These techniques have been packaged in data mining platforms, which are software environments providing support for the application of one or more data-mining algorithms.

3. So far Data Mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualisation and data analysis. Recently, the task of integrating these two technologies has become highly actual, especially as various public and private sector organisations possessing huge databases with thematic and geographically referenced data have begun to realise the huge potential of information hidden there. Among those organisations are:

- statistical offices wanting to analyse or disseminate geo-referenced statistical data;
- public health services searching for explanations of disease clusters;
- environmental agencies assessing the impact of changing land use patterns on climate change;
- geo-marketing companies doing customer segmentation based on spatial location.

4. As a response to this demand a prototype has been developed [2] which demonstrates the potential of combining data mining and GIS. This initial prototype encouraged the formation of the SPIN! project. The project is funded by the European Commission under IST-10536-SPIN!. The overall

¹ Prepared by Michael May.

objective of the SPIN! project [5,12] consists in developing a web-based spatial data mining system by integrating state of the art Geographic Information System (GIS) and data mining functionality in a closely coupled open and extensible system architecture. The new generation SPIN! system pays special attention to such features as scalability, security, multi-user access, robustness, platform independence and adherence to standards. In this paper, we describe the general architecture of the SPIN! data mining platform.

II. COMBINING DATA MINING AND GIS

5. What benefits does data mining offer for the GIS user? Data mining and geographical information systems are best seen as complementary tools for describing and analyzing data. Whereas in GIS the user guides the search and generates hypotheses, data mining partially delegates this task to the computer, pre-selecting and presenting to the analyst only those patterns deemed most interesting (according to some measure of quality). Whereas GIS relies on visualization in geographical space, data-mining searches for patterns in multi-dimensional abstract space. Both techniques are essentially exploratory, leaving the final decision of whether a hypothesis is an important new finding (a “nugget” in data mining language) or just an artifact to the analyst.

6. How are spatial data handled in usual data mining systems? Although many data-mining applications deal at least implicitly with spatial data, they essentially ignore the spatial dimension of the data, treating them as non-spatial. This has ramifications both for the analysis of data and for their visualisation. First, one of the basic tasks of exploratory data analysis is to present the salient features of a data set in a format understandable to humans. It is well known that visualisation in geographical space is much easier to understand than visualisation in abstract space. Secondly, results of a data mining analysis may be sub-optimal or may even be distorted if unique features of spatial data, such as spatial autocorrelation ([7]), are ignored.

7. In sum, convergence of GIS and data mining in an Internet enabled spatial data mining system is a logical progression for spatial data analysis technology. Related work in this direction has been done by Koperski and Han, Ester et al. [4,10].

III. SPIN!: THE ELEMENTS

8. To describe the functionality of the SPIN!-system, it is useful to distinguish several levels of functionality.

III.1 Level 1: Data access and management

9. The basic functionality provides data access to heterogeneous data sources, data transformation capabilities, and facilities for organizing and documenting analysis tasks.

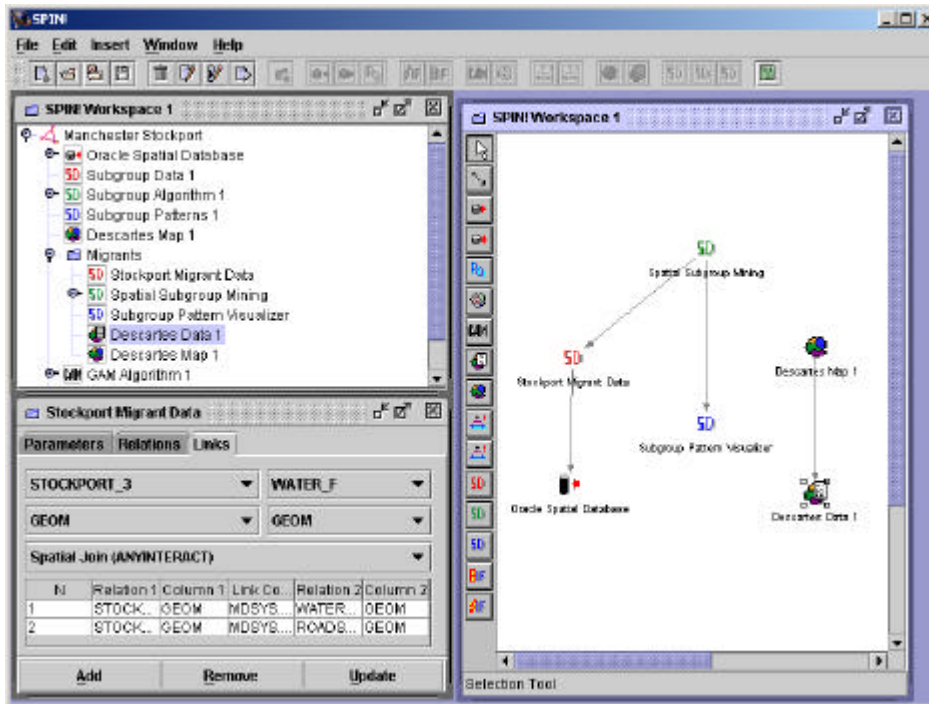


Fig. 1. SPIN! workspace with workspace tree (upper left) for organizing analysis tasks, a graph editor for constructing analysis tasks in a visual manner (left) and property editors (bottom left) for configuring data sources, algorithms, and visual output

III.2 Level 2: Interactive thematic mapping for visualizing statistical data

10. For visual exploratory spatial analysis the Descartes module for interactive manipulation of statistical maps is used ([1]). It supports basic GIS operations such as zooming, panning, querying features and changing visual appearance. Yet its real strength lies in its capabilities for interactive visual exploration of statistical data. Descartes automates map design by incorporating the knowledge of thematic cartography in the form of generic, domain-independent rules, taking into account data characteristics and relations among data components or attributes. The automation of map generation releases the user from the necessity of thinking about how to present the data and from the routine work of map building; instead he/she can concentrate on the analysis of his/her data. Among Descartes features are linked displays, interactive classification, box-plots, scatter-plots and a module for temporal visualisation.

11. One application area in the SPIN! project is urban planning for the area of Stockport, a district in greater Manchester, UK. UK 1991 census data, available on the level of enumeration districts, is the data source used. Also available are detailed geographical layers, among them streets, rivers, buildings, railway lines, shopping areas. Data are provided for the project by the project partners Manchester University and Manchester Metropolitan University. As an example, Figure 2 shows areas with a high rate of long term illness in Manchester Stockport. The map has been produced using Descartes' interactive classification tool. Areas above a user-defined threshold are coloured in shades of red, those below the threshold are coloured in green. We see a cluster of districts having a high rate of long-term illnesses in the north of the centre, and several other districts scattered around the map, among them a large one in the west.

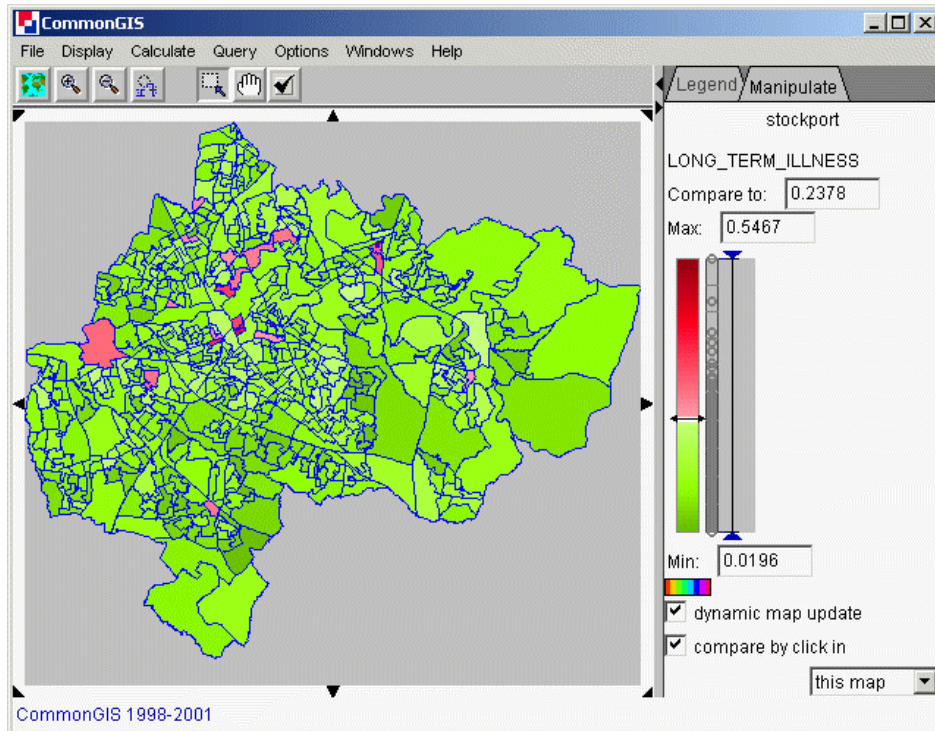


Fig. 2. Areas with high-rate of long-term illness in Manchester Stockport (red areas)

III.3 Level 3: Spatial cluster detection

12. Descartes can be used for interactive, visual identification of spatial clusters. Yet the SPIN!-system also contains modules for performing this search automatically. The objective of the Geographical Analysis Machine GAM [13] is to look for local spatial clusters without knowing in advance where to look. GAM works by examining a large number of overlapping circles of varying sizes that completely cover a region of interest, retaining cycles with a statistically significant deviation in distribution.

III.4 Level 4: Explaining clusters and spatial phenomena

13. Assume we are interested in Stockport enumeration districts with a high migration rate. We ask ourselves how those enumeration districts are characterised and what might distinguish them from other enumeration districts not having a high migration rate. Spatial subgroup discovery is a data mining approach that helps to answer this kind of question. Subgroup discovery is searching the hypothesis space for interesting deviation patterns with respect to an attribute of interest.

Example : “Migration is high in enumeration districts with high unemployment”

14. The attribute of interest or target attribute T is “high migration rate”, the concept C is “enumeration districts with high unemployment” and the subgroup is the conjunction of both, i.e. “Migration is high in enumeration districts with high unemployment”. The deviation pattern is that the proportion of districts satisfying the target T is much higher in districts that satisfy pattern C than in the overall population ($p(T|C) > p(T)$).

Type	Abbreviation	Description
Concept	C	“enumeration districts with high unemployment”
Target	T	“Migration rate is high”
Subgroup	S	„Migration rate is high in enumeration districts with high unemployment“
Attribute	A	Migration rate
Value	v	high, low, medium

Table 1. Terminology for subgroup discovery.

15. The hypothesis space is searched in a top-down manner from more general to more specific; e.g. the description “area with high unemployment and a large number of medical establishments” is more specific than the description “area with high unemployment”. A beam search is performed so that only the best n hypothesis found so far are expanded at each level of search. Hypotheses are ranked by an evaluation function, given by the formula:

$$\sqrt{\frac{n}{p_0(1-p_0)}}(p-p_0),$$

where

- n is the concept size, $|C|$;
- p_0 is the relative frequency of T in the population P ($|T|/|P|$);
- p is the relative frequency of the subgroup with respect to the concept ($|T\&C|/|C|$).

The search terminates when the algorithm has searched the hypothesis space up to a user-defined search depth. The best n hypotheses are reported to the user.

16. The key to spatial data mining is to make proper use of spatial information inherent in the data by extending the representational capabilities of data mining algorithms. While traditional attribute-value based learning methods have difficulties in expressing topological features such as `being_inside`, `adjacent_to`, etc. in a natural and general way, they can be easily expressed in first-order-logic. This makes inductive logic programming (ILP), which uses a first-order representation, a natural and promising approach to many forms of spatial data mining. In the SPIN! project we investigate spatial association rules and subgroup discovery ([8], [11], [14]). We have translated an ILP multirelational subgroup discovery algorithm to relational algebra/SQL, so that a search can be carried out directly in the database. The details of this are, however, beyond the scope of this paper.

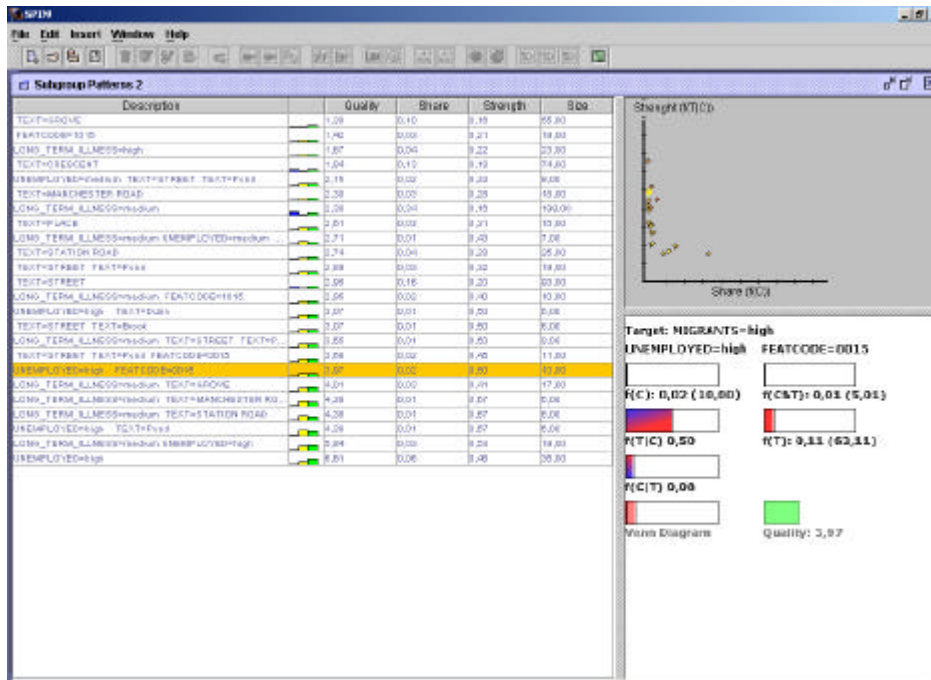


Fig 3. Overview of subgroups found by the subgroup discovery algorithm (left) showing the subgroup description. Bottom right side shows a detailed view for the overlap of the concept C (e.g. located near a railway line) and the target attribute T (high unemployment rate). The larger the difference between $p(T)$ and $p(T|C)$ and the larger the subgroup C, the higher the quality of a subgroup. The window on the right top plots $p(T|C)$ against $p(C)$ for all subgroups.

17. The way data mining results are presented to the user is crucial for their appropriate interpretation. We use a combination of cartographic and non-cartographic displays linked together through simultaneous dynamic highlighting of the corresponding parts. The user navigates in the list of subgroups, which are dynamically highlighted in the map window.

18. Figures 3 and 4 show an example for the migrant scenario, where the subgroup discovery method reports a relation between districts with high migration rate and high-unemployment. The conditional probability ($p(T|C)$) becomes even larger (though the subgroup C becomes smaller) if those districts are located near a railway line (shown in brown on the map). This is an example for a subgroup combining spatial features (being crossed by a railway line) and non-spatial features (unemployment rate).

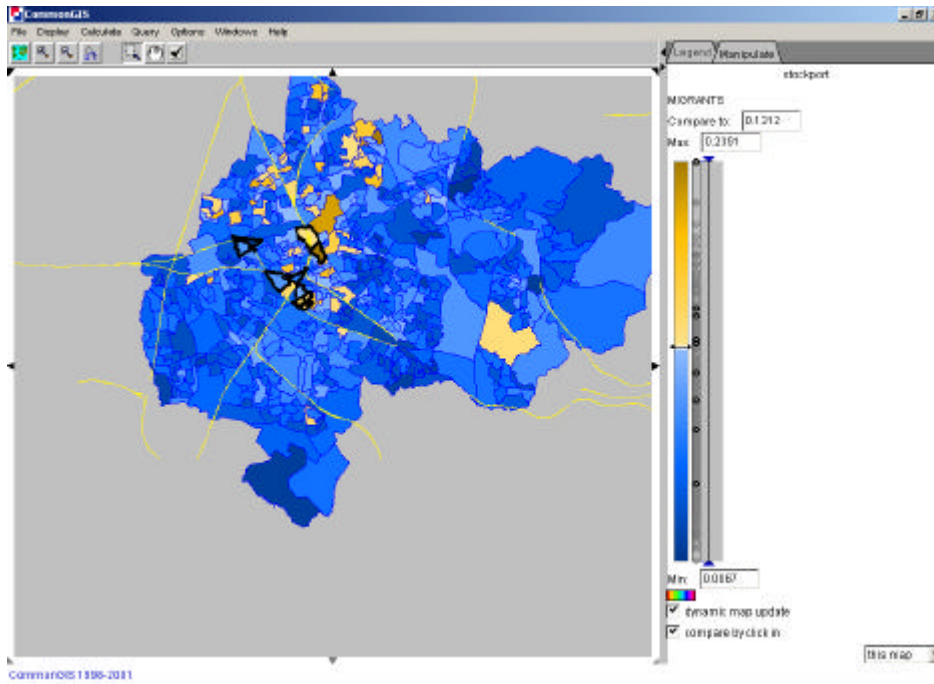


Figure 4. Enumeration districts satisfying the subgroup description C (high unemployment rate and crossed by a railway line) are highlighted with a thicker black line. Enumeration districts also satisfying the target (high migration rate) are displayed in brown.

IV. N-TIER EJB-BASED ARCHITECTURE

19. The general SPIN! architecture is shown in Fig. 5. It is an n-tier Client/Server-architecture based on Enterprise Java Beans for the server side components. A major advantage of using Enterprise Java Beans is that tasks such as controlling and maintaining user access rights, handling multi-user access, pooling of connections, caching, handling persistency and transaction management are delegated to the EJB container. The architecture is comprised of the following major subsystems: client, application server with one or more EJB containers, one or more database servers and optionally compute servers.

20. The client is a GUI Java application or applet. It always creates one server side representative in the form of session bean, the methods of which are accessed either directly through the corresponding remote reference (Java RMI or CORBA IIOP protocol) or indirectly by means of servlets (HTTP protocol). The client session bean executes various server side tasks on behalf of the client. In particular, it may load/save workspace objects from/in its persistent state.

21. The application server is an Enterprise Java Bean container. It manages the client workspace, analysis and visualisation tasks, data access and persistency. There may be more than one simultaneously running container on one or more servers so that, for example, different algorithms and other tasks can be executed on different computers under different restrictions.

22. User data are stored in primary data storage, which is a relational database system (it may be the same machine as the application server). There may be one or more optional secondary databases for analysing data. In addition, data can be loaded from other sources – databases, ASCII files in the file system or Excel files. Analysis tasks can run on one or more compute servers (possibly the same machine(s) as the application server).

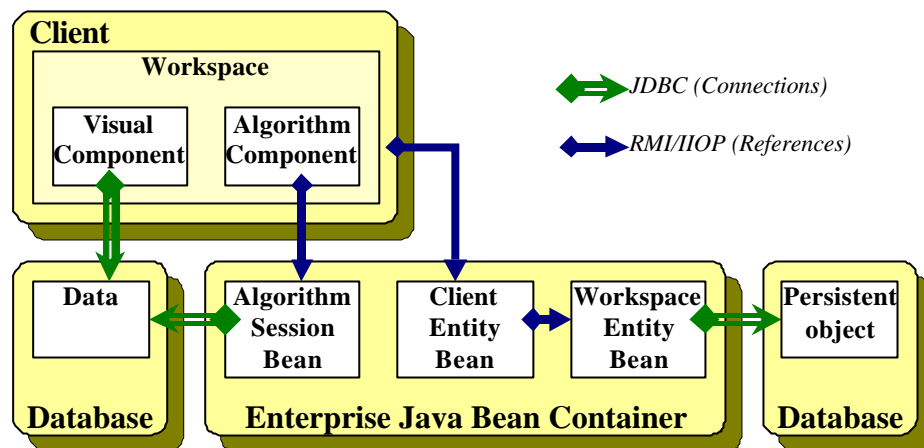


Fig. 5. SPIN! architecture. Main components are a Java-based client, an Enterprise Java Beans Container and one or more databases serving spatial a non-spatial data

23. The client creates one remote object for each analysis task to be run so that data is transferred directly from the database to the algorithm. After the analysis is finished its result is transferred to the client for visualisation. A connector machine, which is a Java Virtual Machine running on the application server, is used for accessing non-Java analysis tasks. These may run on additional compute servers.

V. RUNNING DATA MINING ALGORITHMS

24. The developed architecture supposes that all algorithms are executed on compute servers. For each running algorithm, a separate session bean is created which implements high-level methods for controlling its behaviour, particularly, starting/stopping the execution, obtaining/setting parameters, setting the data to process, and obtaining the result. The session bean then is responsible for the methods implementation. There are several ways this can be done.

- A clean and very convenient but in some cases inefficient approach is using Java for implementing the complete algorithm directly within the corresponding EJB, loading all data via JDBC into the work-space.
- A second approach divides the labour between the EJB container and the relational database. An example is the multi-relational spatial subgroup-mining algorithm that does most of the analysis work (especially the spatial analysis) directly in the database. The EJB part retrieves summary statistics, manages hypotheses and controls the search.
- A third approach consists in implementing computationally intensive methods in native code wrapped into shared library by means of Java Native Interface (JNI).
- A fourth option is that the algorithm session bean directly calls an external executable module with a set of parameters to carry out its procession task.
- And finally other remote objects (e.g. CORBA) can be used to execute the task.

25. The algorithm parameters are formed in the client and transferred to the algorithm EJB as a workspace component before the execution. In particular, data to be processed by the algorithm have to be specified. It is important that only a data description is specified and not the complete data set transferred. In other words, the algorithm bean states where and how to take data and what kind of restrictions to use. Thus when the algorithm starts the data is directly retrieved by the algorithm EJB rather than passing through the client.

VI. CONCLUSION

26. We have described the general architecture of the SPIN! spatial data mining platform and have provided an overview of its components. It integrates GIS and data mining algorithms that have been adapted to spatial data. The choice of EJB technology allows us to meet requirements such as security, scalability and platform independence in a principled manner. The system is tightly integrated with a RDBMS and can serve as a data access and transformation tool for spatial and non-spatial data.

Acknowledgement: Work on this paper has been partially funded by the European Commission under IST-1999-10536-SPIN!

REFERENCES

- [1] Andrienko, G.; Andrienko, N.. "Interactive Maps for Visual Data Exploration", *International Journal of Geographical Information Science* 13(5), 355-374, 1999
- [2] Andrienko, N., G. Andrienko, A. Savinov, and D. Wettschereck, "Descartes and Kepler for Spatial Data Mining", *ERCIM News*, No. 40, January 2000, 44-45.
- [3] Dykes, J., "Exploring spatial data exploration with dynamic graphics", *Computers and Geosciences*, 23, 345-370, 1997
- [4] Ester, M., Frommelt, A., Kriegel, H.P, Sander, J., "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support", in *Data Mining and Knowledge Discovery, an International Journal*, 1999
- [5] European IST SPIN! project web site, <http://www.ccg.leeds.ac.uk/spin/>
- [6] Gitis V., Dovgyallo A., Osher B., Gergely T., "GeoNet: an information technology for WWW on-line intelligent Geodata analysis", *Abstracts of 4th EC-GIS Workshop, Hungary*, 1998
- [7] Haining, R. *Spatial data analysis in the social and environmental sciences*, Cambridge Univ. Press, 1991
- [8] Klös gen, W., "Deviation and association patterns for subgroup mining in temporal, spatial, and textual data bases", In: Polkowski, L., Skowron, A. (eds): *Rough sets and current trends in computing*, 1-18, New York, Springer, 1998
- [9] Klös gen, W., Zyt kow, J. (eds.), *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, to appear 2001
- [10] Koperski, K., Han, J. "GeoMiner: A System Prototype for Spatial Mining", *Proceedings ACM-SIGMOD, Arizona*, 1997
- [11] Malerba, D.; Esposito, F., Lisi, F., "A logical framework for frequent pattern discovery in spatial data", *Proceedngs of the 14th International FLAIRS Conference*, accepted, 2001
- [12] May, M.: *Spatial Knowledge Discovery: The SPIN! System*. Fullerton, K. (ed.) *Proceedings of the 6th EC-GIS Workshop, Lyon, 28-30th June*, European Commission, JRC, Ispra.
- [13] 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*, Taylor and Francis, London, 1999
- [14] Wrobel, S. "Scalability Issues in Inductive Logic Programming", In *Proc. 9th Int. Workshop on Algorithmic Learning Theory (ALT-98)*, Berlin, Springer, 1998