IDEA GROUP PUBLISHING
01 E. Chocolate Avenue, Hershey PA 17033-1117, USA
el: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7515

Chapter 13

# A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views

Karine Zeitouni
PRiSM Laboratory
University of Versailles

*This chapter reviews the data mining methods that are combined with Geographic Information Systems (GIS) for carrying out spatial analysis of geographic data. We will first look at data mining functions as applied to such data and then highlight their specificity compared with their application to classical data. We will go on to describe the research that is currently going on in this area, pointing out that there are two approaches: the first comes from learning on spatial databases, while the second is based on spatial statistics. We will conclude by discussing the main differences between these two approaches and the elements they have in common.*

## INTRODUCTION

The growing production of maps is generating huge volumes of data that exceed people's capacity to analyze them. It thus seems appropriate to apply knowledge discovery methods like data mining to spatial data. This recent technology is an extension of the data mining applied to alphanumerical data on spatial data. The main difference is that spatial analysis must take into account spatial relations between objects.

The applications covered by spatial data mining are decisional ones, such as geomarketing, environmental studies, risk analysis, and so on. For example, in geomarketing, a store can establish its trade area, i.e., the spatial extent of its

Previously Published in *Challenges of Information Technology Management in the 21st Century*, edited by Mehdi Khosrow-Pour, Copyright © 2000, Idea Group Publishing.

This chapter appears in the book, Data Warehousing and Web Engineering by Shirley Becker.

customers, and then analyze the profile of those customers on the basis of both their properties and the properties related to the area where they live.

In our project, spatial data mining is applied to traffic risk analysis (Zeitouni, 1998). The risk estimation is based on the information on the previous injury accidents, combined to thematic data relating to the road network, population, buildings, and so on. The project aims at identifying regions with a high level of risk and analyzing and explaining those risks with respect to the geographic neighborhood. Spatial data mining technology specifically allows for those neighborhood relationships.

Nowadays, data analysis in geography is essentially based on traditional statistics and multidimensional data analysis and does not take account of spatial data (Sanders, 1989). Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. This constitutes the fundamental of a distinct scientific area called "spatial statistics" which, unlike traditional statistics, supposes inter-dependence of nearby observations. An abundant bibliography exists in this area, including well-known geostatistics, recent developments in Exploratory Spatial Data Analysis (ESDA) by Anselin and Geographical Analysis Machine (GAM) by Openshaw. For a summary, refer to Part 1.c of (Longley, Goodchild, Maguire, and Rhind, 1999). Multi-dimensional analytical methods have been extended to support contiguity (Lebart, 1984 & 1997). We maintain that spatial statistics is a part of spatial data mining, since it provides data-driven analyses. Some of those methods are now implemented in operational GIS or analysis tools.

In the field of databases, two main teams have contributed to developing data mining for spatial data analysis. The first one, DB Research Lab (Simon Fraser University, Vancouver), developed GeoMiner (Lu, Han and Ooi, 1993), which is an extension of DBMiner. The second one (Munich University) devised a structure-of-neighborhood graph (Ester, Kriegel, Sander and Xu, 1997), on which some algorithms are based. They have also worked on a clustering method based on a hierarchical partitioning (extension of DBSCAN with a R*Tree), classification (extension of ID3 and DBLearn), association rules (based upon an efficient spatial join), characterization and spatial trends. STING (University of California) uses a hierarchical grid to perform optimization on the clustering algorithm (Wang, Yang and Muntz, 1997). We might also mention work on Datawarehouse dedicated to spatial data (University of Laval) (Bedard, Lam, Proulx, Caron and Letourneau, 1997).

This chapter will describe data mining methods for Geographic Information Systems and highlight their value in performing spatial data analysis. It will survey both statistical approaches and those involving inference from databases.