

## Research Article

# Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change

Jeremy Mennis  
*Department of Geography and  
Urban Studies  
Temple University*

Jun Wei Liu  
*Department of Geography  
University of Colorado*

### Abstract

This research demonstrates the application of association rule mining to spatio-temporal data. Association rule mining seeks to discover associations among transactions encoded in a database. An association rule takes the form  $A \rightarrow B$  where  $A$  (the antecedent) and  $B$  (the consequent) are sets of predicates. A spatio-temporal association rule occurs when there is a spatio-temporal relationship in the antecedent or consequent of the rule. As a case study, association rule mining is used to explore the spatial and temporal relationships among a set of variables that characterize socioeconomic and land cover change in the Denver, Colorado, USA region from 1970–1990. Geographic Information Systems (GIS)-based data pre-processing is used to integrate diverse data sets, extract spatio-temporal relationships, classify numeric data into ordinal categories, and encode spatio-temporal relationship data in tabular format for use by conventional (non-spatio-temporal) association rule mining software. Multiple level association rule mining is supported by the development of a hierarchical classification scheme (concept hierarchy) for each variable. Further research in spatio-temporal association rule mining should address issues of data integration, data classification, the representation and calculation of spatial relationships, and strategies for finding ‘interesting’ rules.

## 1 Introduction

Spatio-temporal data mining is an emerging research area dedicated to the development and application of novel computational techniques for the analysis of very large,

**Address for Corresponding Author:** Jeremy Mennis, Department of Geography and Urban Studies, Temple University, 1115 West Berks Street, 309 Gladfelter Hall, Philadelphia, PA 19122, USA. E-mail: [jmennis@temple.edu](mailto:jmennis@temple.edu)

spatio-temporal databases (Buttenfield et al. 2001, Koperski et al. 1996). Data mining techniques are typically inductive, as opposed to deductive, in that they are not used to prove or disprove pre-existing hypotheses but rather to identify patterns embedded within data, and thereby support hypothesis generation. Most research in spatial, temporal, and spatio-temporal data mining has sought to adapt 'classical' data mining algorithms intended to operate on more conventional data types (cf. Ladner et al. 2002, Roddick and Hornsby 2000). Spatio-temporal data mining presents a number of challenges due to the complexity of geographic domains, the mapping of all data values into a spatial and temporal framework, and the spatial and temporal autocorrelation exhibited in most spatio-temporal data sets (Miller and Han 2001).

The purpose of this research is to demonstrate the application of a certain type of data mining technique, association rule mining, to spatio-temporal data. As a case study, we use association rule mining to explore the spatial and temporal relationships among geographic data that characterize socioeconomic and land cover change in the Denver, Colorado, USA region. This case study is intended to elicit associations among processes of socioeconomic change and urban growth. Strategies for data integration and pre-processing to support spatio-temporal association rule mining are discussed.

## 2 Association Rule Mining

Association rule mining seeks to discover associations among transactions encoded within a database (Agrawal et al. 1993). An association rule takes the form  $A \rightarrow B$  where  $A$  (the antecedent) and  $B$  (consequent) are sets of predicates. For example, consider a database that encodes transactions made at a supermarket. An association rule may state that 'customers that purchase bagels also purchase cream cheese'. This statement may be expressed as:

$$Is\_a(x, bagel) \wedge is\_purchased(x) \rightarrow is\_a(y, cream\_cheese) \wedge is\_purchased(y) \quad (1)$$

Association rule mining uses the concepts of support and confidence to identify rules that are particularly interesting or unexpected. The support is the probability of a record in the database satisfying the set of predicates contained in both the antecedent and consequent, for instance the probability that a record in the database contains the purchase of a bagel and cream cheese in the example above. The confidence is the probability that a record that contains the antecedent also contains the consequent. The support and confidence of a rule are typically reported in support-first order in parentheses following the rule, i.e. '(support%, confidence%)'. Thresholds of support and confidence can be set to weed out rules that are not of interest. A spatial association rule occurs when a predicate in either the antecedent or the consequent contains a spatial relationship (Koperski and Han 1995). Likewise, a spatio-temporal association rule contains a spatio-temporal relationship.

Many databases to which data mining is applied are arranged in a concept hierarchy, a hierarchical classification (Koperski et al. 1996). For example, in the supermarket example above, a bagel may be considered a kind of baked good. In this case, sales data are stored at the level of the individual product (i.e. the total sales of bagels) and also aggregated at 'higher' levels of the concept hierarchy (i.e. the total sales of baked goods). Association rule mining of data arranged in a concept hierarchy is called multiple level association rule mining, and is supported by mining rules at varying levels of the concept hierarchy to find the hierarchy resolution that best captures the rule (Han and Fu 1995).

A number of authors have noted that there are problematic issues in applying association rule mining to spatial data, and, analogously, spatio-temporal data. One issue in spatial association rule mining is that whereas non-spatial association rule mining seeks to find associations among transactions that are encoded explicitly in a database, spatial association rule mining seeks to find patterns in spatial relationships that are typically not encoded in a database but are rather embedded within the spatial framework of the georeferenced data (Shekhar and Chawla 2003). These spatial relationships must be extracted from the data prior to the actual association rule mining. There is therefore a trade-off between pre-processing spatial relationships among geographic objects and computing those relationships on-the-fly (Klosgen and May 2002). Pre-processing improves performance, but massive data volumes associated with encoding spatial relationships for all combinations of geographic objects prohibits the storage of all spatial relationships. A number of approaches have been developed to address this issue, including the use of  $R^*$  trees and minimum bounding rectangles for fast computation of spatial relationships (Kopersky and Han 1995), the use of spatial relationship indices (Ester et al. 2000, Zeitouni et al. 2001), and the encoding of spatial relationships among certain target sets of geographic objects prior to data mining (Malerba et al. 2002).

Another issue with spatial association rule mining is that conventional association rule mining is designed to work with categorical data, not numeric data such as metric distance. One approach to this problem is to discretize numeric data into ordinal categories and then mine those ordinal data for association rules (Piatetsky-Shapiro 1991, Srikant and Agrawal 1996). For example, data such as metric distance may be parsed into categories of ‘near’ and ‘far.’ The choice of interval breaks in the conversion from numeric to categorical data type impacts the results of the rule mining, however, and can be particularly problematic if the discretization is too coarse to capture an interesting rule. As an approach to this problem, some researchers have proposed methods for optimizing the discretization of numeric data for association rule mining, for instance using cluster analysis (Zhang et al. 1997), computational geometry (Fukada et al. 1999), or a heuristic method (Wang et al. 1998). Others have suggested approaches for association rule mining of numeric data that are oriented toward finding statistical differences among subsets of the entire data set (Aumann and Lindell 2003), although such approaches have generally not been extended to spatial association rule mining.

### 3 Case Study: Urban Growth in Denver, Colorado, USA

#### 3.1 Overview

As a case study, we focused on mining association rules in data relating to urban growth in the Denver, Colorado, USA region from 1970 to 1990. The objective of the case study is to identify patterns within a database containing socioeconomic and land cover change data, and thus support hypothesis generation regarding the relationship between socioeconomic change and urban growth. U.S. Bureau of the Census data for 1970 and 1990 were acquired at the tract level using the Geolytics, Inc. Neighborhood Change Database CD, which maps a select set of 1970–2000 Census data variables to Census 2000 tract boundaries (Geolytics 2001). Data on limited and unlimited access highways were acquired from the Environmental Systems Research Institute (ESRI) Streets Database. Land cover data for the 1970s and 1990s were acquired from the U.S. Geological Survey’s (USGS) Front Range Infrastructure Resources Project (FRIRP). These vector

**Table 1** Sample of land cover classes contained in the three level hierarchical land cover classification

Level 1	Level 2	Level 3
1 Water	...	...
2 Developed	21 Residential	211 Single-Family Residential 212 Multi-Family Residential
	22 Non-Residential	...
3 Bare	...	...
4 Vegetated	41 Woody Vegetation	411 Forested 412 Shrubland
	42 Herbaceous Vegetation	...

polygon data were generated from historic aerial photography, USGS digital orthophotographic quadrangles (DOQs), and ancillary data such as wetlands inventories (Stier 1999). Each polygon is classified using a hierarchical three level classification of land cover using the modified Anderson land cover classification scheme (Anderson et al. 1976) (Table 1).

### 3.2 *Data Integration and Pre-Processing*

As noted above, there is a trade-off in spatio-temporal association rule mining involving pre-processing spatial and temporal relationships versus calculating those relationships on-the-fly. Here, we used a geographic information system (GIS) to pre-process these relationships and encode them in a single table, referred to hereafter as the ‘mining table,’ so that they may be mined using association rule mining software developed for conventional (i.e. non-spatio-temporal) data. Compare our approach to previous approaches which would, say, mine spatial relationships among tracts and land cover polygons by finding overlapping tracts and land cover polygons on-the-fly, by using a spatial join index to retrieve indexed tract-land cover polygon spatial relationships, or by pre-processing the spatial coincidence of a particular target set of tracts and land cover polygons. Our approach is rather to integrate the tract and land cover data to a common spatial unit prior to association rule mining, so that spatio-temporal relationships among different tracts and land cover polygons are encoded as attributes of those common spatial units within a single table. While this approach incurs an upfront performance cost in calculating the spatial and temporal relationships, as well as in data storage, once the pre-processing is complete this strategy allows for fast association rule mining without significant customization of either GIS or conventional association rule mining software.

The Census and land cover data were integrated within a GIS and processed to produce a mining table with 24,258 records, each of which represented a polygon that was homogeneous in both socioeconomic character and land cover change from 1970 to 1990. The following variables were then calculated for each polygon (Figure 1):

- **Land Cover Change** Change in land cover, 1970–1990
- **Change in Minority** Change in percent minority (non-white or Hispanic), 1970–1990

**Table 2** Descriptive statistics for the quantitative variables used in the analysis

Variable	Minimum	Maximum	Mean	St. Deviation
Change in Minority (%)	-18	68	6	9
Change in Poverty (%)	-41	37	1	7
Urban Density (cells)	1	3,409	1,299	1,086
Distance to Highway (m)	0	6,937	959	1,085

*N* = 24,258 for all variables.

- **Change in Poverty** Change in percent of population living below the poverty line, 1970–1990
- **Urban Density** Mean density of developed land in 1970
- **Distance to Highway** Minimum distance to limited and unlimited access highways

Descriptive statistics for the quantitative variables are presented in Table 2. Note that Urban Density is the density of land classified as developed in 1970, created by first generating a 30 m resolution binary grid of developed/not developed land cover, then creating a second grid that encodes the number of developed grid cells within one km of each grid cell, and then finally calculating the mean grid cell value within each polygon. The units of Urban Density are the number of developed cells per unit 1 km radius circle area. Distance to Highway was calculated by generating a 30 m resolution grid in which each grid cell encoded the distance to the nearest limited or unlimited access highway. The minimum grid cell value within each tract was then calculated.

As noted above, conventional association rule mining works only with categorical data. We therefore converted each of the numeric variables to an ordinal value. After experimenting with quantile and equal interval classification schemes, we settled on the natural breaks classification available in the GIS software package, which is based generally on the Jenks optimal classification algorithm (Jenks and Coulson 1963). Natural breaks was used to create an eight class scheme for each of the variables; class breaks for each variable are reported in Table 3 (column ‘Level 3’ reports class IDs for the eight classes).

We also developed a strategy to support multiple level association rule mining in which rules are mined at multiple levels of a concept hierarchy. As demonstrated in Table 1, the land cover data were already arranged in a concept hierarchy via their hierarchical classification scheme. We adapted this hierarchical classification to create a three level concept hierarchy for the Land Cover Change variable in which level 1 encoded the change (or absence of change) from the 1970 to 1990 land cover at level 1 of the land cover hierarchy (e.g. from vegetated to developed), level 2 encoded the change in land cover at level 2 (e.g. from woody vegetation to residential), and level 3 encoded the change in land cover at level 3 (e.g. from forested to single family residential).

A three level concept hierarchy, denoted levels 1, 2, and 3, was also created for each of the other (non-land cover change) variables through data classification. Level 3 of the concept hierarchy for each variable was defined using the eight class natural breaks classification scheme reported in Table 3 (column ‘Level 3’). Level 2 and level 1 map the eight classes given in level 3 into four classes and two classes, respectively (Table 3). Figure 1 shows three of the four non-land cover change variables mapped according to level 2 of its concept hierarchy.

**Table 3** Class breaks for the eight class natural breaks classification of the non-land cover change variables used in the association rule mining. The level numbers indicate a three level concept hierarchy based on data classification for each variable

Level			Change in % Minority	Change in % Poverty	Urban Density (cells)	Distance to Highway (m)
1	2	3				
0	0	0	-18-6	-41-25	1-328	0-273
0	0	1	-5-01	-24-5	329-697	274-702
0	1	2	2-7	-4-0	698-1,138	703-1,228
0	1	3	8-13	1-5	1,139-1,622	1,229-1,852
1	2	4	14-20	6-10	1,623-2,126	1,853-2,562
1	2	5	21-28	11-15	2,127-2,626	2,563-3,384
1	3	6	29-37	16-22	2,627-3,071	3,385-4,366
1	3	7	38-68	23-37	3,072-3,409	4,367-6,937

### 3.3 Software and Methods

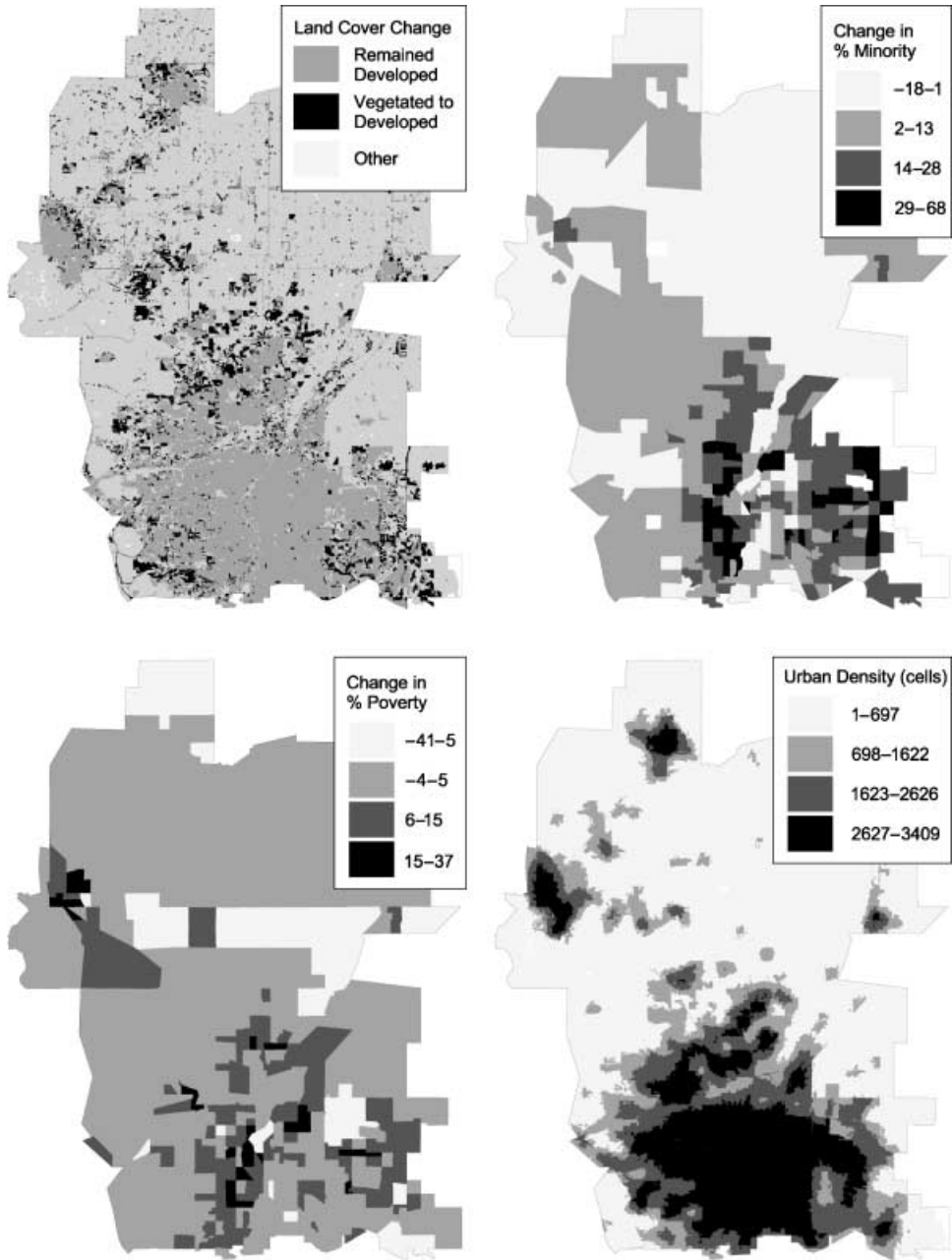
We used the association rule mining software CBA (version 2.1) (Classification Based on Associations) (Liu et al. 1998), which uses the well-known Apriori algorithm for finding association rules (Agrawal and Srikant 1994). We structured our multiple level association rule mining by first investigating rules among variables at level 1 (the coarsest level) of the concept hierarchies. We sought 'interesting' rules among the socioeconomic variables, then added the land cover change, urban density, and distance to highway variables. Within the process of adding different variables, if an interesting rule was found, we proceeded to investigate that rule using finer levels of the concept hierarchy. Note that there are an enormous number of rules that may be generated, and 'paths' of investigation to explore them. We describe just a few interesting rules here for demonstration purposes by focusing on the relationship between changes in percent minority and poverty. We found that the most useful approach to finding interesting rules was to compare sets of similar rules. By holding the values of certain variables constant within each rule, and varying the values of other variables, the effects of an individual variable on a rule may be determined.

### 3.4 Results

We began by mining rules on only the level 1 Change in Minority and Change in Poverty variables. This run generated four rules, including the following (note that we use a slightly easier-to-follow rule format here than that presented in Equation 1):

$$\begin{aligned} &\text{Change in Minority} > 13\% \\ \rightarrow &\text{Change in Poverty} > 5\% \quad (9.53\%, 63.25\%) \end{aligned} \quad (2)$$

This statement indicates that in 63.25% of the polygons in which there is an increase of greater than 13% in percent minority, there is also an increase greater than 5% in percent living below the poverty line. When the level 1 Change in Minority variable is replaced with the level 2 and level 3 Change in Minority variables, the following rules are generated:



**Figure 1** The Denver, Colorado, USA study region and four of the variables used in the association rule mining: Change in Land Cover, Change in Minority, Change in Poverty, and Urban Density

Change in Minority > 28%  
 → Change in Poverty > 5% (2.02%, 80.16%) (3)

Change in Minority > 37%  
 → Change in Poverty > 5% (0.34%, 100.00%) (4)

As the Change in Minority variable is parsed into categories at finer granularities, the percentage of the polygons which are increasing in poverty in the highest percent minority class increases. Note that all of the polygons with a 37% or greater increase in percent minority also increased at least 5% in the percent living below the poverty line (Equation 4), although the support for this rule is relatively low (0.34%).

We investigated further by next incorporating the level 1 land cover change variable. We included Change in Minority (level 2) in order to keep the support at a reasonable level. Of the generated rules, the following two allow for a comparison of Change in Poverty between those polygons that remained developed and those polygons that changed from vegetated to developed, among polygons strongly increasing in percent minority:

$$\begin{array}{l} \text{Land Cover Change from Developed to Developed} \quad \text{and} \\ \text{Change in Minority (level 2) > 28\%} \quad \text{and} \\ \rightarrow \text{ Change in Poverty (level 1) > 5\% (1.27\%, 77.78\%)} \end{array} \quad (5)$$

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \quad \text{and} \\ \text{Change in Minority (level 2) > 28\%} \quad \text{and} \\ \rightarrow \text{ Change in Poverty (level 1) > 5\% (0.30\%, 82.02\%)} \end{array} \quad (6)$$

Note that the confidence of the rules expressed in Equations 5 and 6 have similar values. We may conclude that the change in land cover of a polygon does not appear to make a significant difference in the relationship between strongly increasing percent minority and increasing poverty rate. Strongly increasing percent minority is associated with increasing poverty in polygons that remained developed as well in those polygons undergoing development.

In the next mining run, level 1 Urban Density was incorporated to investigate whether these socioeconomic and land cover changes were associated with areas that were primarily urban or rural in 1970. Note that while the land cover data indicate whether a polygon is developed in 1970, it does not capture whether that polygon is in primarily urban or rural surroundings. The Urban Density variable addresses this by measuring the density of developed area within the surrounding region. Thus, a polygon with a developed land cover and a low Urban Density value indicates a developed area surrounded by undeveloped land, such as a small town in a rural area. Because the actual values of Urban Density (the mean number of grid cells within 1 km of a developed area) are unintuitive, the rules refer only to the classes of Urban Density as shown in Table 2 (e.g. classes 0–7 for level 3, ranging from very rural to very urban, respectively). We also replaced level 2 Change in Minority with level 1 Change in Minority to ensure adequate support values, which naturally decrease because of the combinatorial nature of calculating the support when adding predicates within a rule. The following two rules were generated which compares Change in Poverty among polygons undergoing development in rural versus urban areas:

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \quad \text{and} \\ \text{Urban Density = 1} \quad \text{and} \\ \text{Change in Minority > 13\%} \\ \rightarrow \text{ Change in Poverty > 5\% (0.90\%, 55.87\%)} \end{array} \quad (7)$$

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \quad \text{and} \\ \text{Urban Density = 0} \quad \text{and} \\ \text{Change in Minority > 13\%} \\ \rightarrow \text{ Change in Poverty > 5\% (0.64\%, 59.23\%)} \end{array} \quad (8)$$

The rules given in Equations 7 and 8 suggest that there is not a difference between urban (Urban Density = 1) and rural (Urban Density = 0) areas in terms of the association between increasing percent minority and increasing percent poverty in developing areas. We further investigated this pattern by generating similar rules to those given in Equations 7 and 8, but replacing Urban Density (level 1) with levels 2 and 3 in two subsequent mining runs (Equations 9 and 10, and 11 and 12, respectively). The following two pairs of rules were generated, one pair for each run, again comparing urban versus rural areas:

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \\ \text{Urban Density} = 3 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (9)$$

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \\ \text{Urban Density} = 0 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (10)$$

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \\ \text{Urban Density} = 7 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (11)$$

$$\begin{array}{l} \text{Land Cover Change from Vegetated to Developed} \\ \text{Urban Density} = 0 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (12)$$

Interestingly, as the attribute resolution of the Urban Density variable becomes finer, a pattern emerges. Of those polygons that are very urban (Urban Density = 7), changed from vegetated to developed, and are strongly increasing in percent minority, 84.21% are increasing in percent poverty (Equation 11). In contrast, of those polygons that are very rural (Urban Density = 0), changed from vegetated to developed, and are strongly increasing in percent minority, only 23.33% are increasing in poverty (Equation 12).

Similar rules to those generated in Equations 11 and 12, except focusing on those polygons that remained developed, are shown below in Equations 13 and 14. Note that the confidence values for these two rules are approximately the same (76.95% and 80.00%). In contrast to polygons undergoing development, for polygons that remained developed there is not a significant difference in the percent minority/poverty relationship between urban and rural areas.

$$\begin{array}{l} \text{Land Cover Change from Developed to Developed} \\ \text{Urban Density} = 7 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (13)$$

$$\begin{array}{l} \text{Land Cover Change from Developed to Developed} \\ \text{Urban Density} = 0 \\ \text{Change in Minority} > 13\% \\ \rightarrow \text{Change in Poverty} > 5\% \end{array} \begin{array}{l} \text{and} \\ \text{and} \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \end{array} \quad (14)$$

It is also interesting to explore the impact of distance to highways on the relationship between increasing minority and poverty in developing areas. Equations 15 and 16 below indicate that polygons undergoing development in rural areas that are nearby highways are much more likely to have a positive relationship between increasing percent minority and percent poverty, as compared to those lands that are far from a highway. However, it should be noted that there are very few rural polygons that are undergoing development, increasing in both percent minority and percent poverty, and are far from a highway.

$$\begin{array}{ll}
 \text{Land Cover Change from Vegetated to Developed} & \text{and} \\
 \text{Urban Density} = 0 & \text{and} \\
 \text{Distance to Highway} \leq 1902 \text{ m} & \text{and} \\
 \text{Change in Minority} > 13\% & \\
 \rightarrow \text{Change in Poverty (level 1)} = 1 \text{ (0.89\%, 71.65\%)} & (15)
 \end{array}$$

$$\begin{array}{ll}
 \text{Land Cover Change from Vegetated to Developed} & \text{and} \\
 \text{Urban Density} = 0 & \text{and} \\
 \text{Distance to Highway} > 1902 \text{ m} & \text{and} \\
 \text{Change in Minority} > 13\% & \\
 \rightarrow \text{Change in Poverty (level 1)} = 1 \text{ (0.004\%, 2.17\%)} & (16)
 \end{array}$$

In sum, the association rules generated here indicate the following. First, increasing percent minority is associated with increasing poverty in general, and this relationship holds both for polygons that underwent development and for those that remained developed. However, in very urban areas that underwent development, increasing percent minority is associated with increasing poverty; in very rural areas that underwent development, increasing percent minority is not associated with increasing poverty. The exception is the unusual situation in which a rural polygon undergoing development is located nearby a highway, in which case increasing minority is associated with increasing poverty. In contrast to developing areas, the urban versus rural factor does not alter the strongly positive percent minority/poverty relationship in areas that remained developed. Clearly, the dynamic relationship between factors of race and poverty in developing areas is controlled in large measure by the urban versus rural setting where that land cover change occurs.

#### 4 Conclusions

The case study concerning urban growth in the Denver region has demonstrated how association rule mining may be applied to spatio-temporal data. Data pre-processing within GIS can be used to facilitate spatio-temporal association rule mining by integrating diverse data sets and extracting spatio-temporal relationships embedded in databases. These spatio-temporal relationships may then be encoded in tabular format for use by conventional association rule mining software intended for non-spatial data. The development of concept hierarchies through data classification demonstrates a methodology to support multiple level spatio-temporal association rule mining and thereby explore the effect of attribute resolution on the generation of interesting rules.

While this case study shows association rule mining to be a promising analytical tool for spatio-temporal data analysis, there are a number of issues that warrant further

investigation. First, the integration of diverse data sets can be problematic. Data integration in this case demanded areal interpolation, the transformation of spatial data from one set of areal units to another. We chose to assume a homogeneous distribution of the Census tract data and use a simple areal weighting technique to apportion the tract data to homogeneous land cover change units. A related problem is that the tracts range widely in size. Thus, the confidence of a rule, expressed in terms of the percentage of the number of tracts in the database, may not reflect the actual percentage in area.

A second issue for further research is the impact of scale of analysis and spatial resolution on association rule mining results. It is well-established that the results of statistical analyses of spatial data are sensitive to the scale of spatial data aggregation (e.g. the use of Census tracts versus block groups) as well as to the spatial partitioning scheme at any one scale (Fotheringham and Wong 1991). This problem, referred to as the modifiable areal unit problem (MAUP), likely impacts the results of many spatial data mining techniques as well, including spatial association rule mining.

A related issue is the impact of data classification techniques on spatio-temporal association rule mining. We briefly experimented with quantile and equal interval classification schemes before settling on the natural breaks method. As noted above, other researchers have proposed novel methods for optimizing the classification of numeric data to find interesting association rules. However, the impact of using different classification schemes on the results of spatio-temporal association rule mining is currently unknown. We suspect that impact is significant in many situations.

Another issue concerns the representation of spatial relationships. The case study focused on relationships of spatial coincidence (a topologic relationship) and distance (a metric relationship). However, there are other types of spatial relationships that may be used in spatio-temporal association rule mining, such as formal topologic relationships as defined by Egenhofer and Franzosa (1991), directional relationships (Frank 1996), and spatial relationships as specified in natural language (Shariff et al. 1998).

It should also be acknowledged that the data integration approach to encoding spatial relationships described here explicitly encodes spatial relationships only for the individual common spatial units that result from the data integration pre-processing; and those relationships are encoded only for other entire classes of geographic objects. In the case study mining table, for example, each record (polygon) records only the distance to the nearest highway and not the distance to each individual highway. And there is no encoding of distance between, say, an individual land cover polygon and an individual highway. Nonetheless, the data integration approach described here offers an efficient way to support the representation of topologic, metric, and other types of spatial relationships for spatial association rule mining.

Finally, and perhaps most importantly, there is the issue of how to find interesting rules among the multitude of rules that are generated from even moderately sized input data sets. Although CBA does provide a graphic user interface that supports rule exploration, we struggled to find rules that were unexpected, and thus, of particular interest. A number of authors have suggested metrics beyond support and confidence for measuring the 'interestingness' of association rules (e.g. Tan et al. 2002), and there are commercial association rule mining packages that incorporate many of these. For example, the commercial package *Magnum Opus* (Rulequest Inc.) reports the 'lift' metric, which quantifies how much more likely the consequent occurs when it is associated with the antecedent than one would expect given a random distribution of the consequent throughout the data. Preliminary applications of *Magnum Opus* for rule mining in

geographic data show the lift metric to be particularly useful in evaluating the interestingness of rules, particularly when combined with the support metric.

Other authors have addressed the issue of finding interesting patterns in data mining results (meta-mining) by pre-specifying expected rules in a rule ‘template’ to guide association rule mining (Fu and Han 1995), by tracking association rules that change over time (Spiliopoulou and Roddick 2000), and by developing database support for storing and refining discovered geographic knowledge (Mennis and Peuquet 2003). In future research, we intend to investigate how more sophisticated interestingness measures and meta-mining approaches may be used to improve the utility and efficiency of applying association rule mining to spatio-temporal data.

## Acknowledgements

The authors would like to thank Manish Salian and Supriya Ramdasi for assistance with the data pre-processing. This research was supported by a NASA New Investigator Program grant.

## References

- Agrawal R, Imielinski T, and Swami A 1993 Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 207–16
- Agrawal R and Srikant R 1994 Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Databases*: 487–99
- Anderson I R, Hardy E E, Roach J T, and Witmer R E 1976 *A Land Use and Land Cover Classification System for Use With Remote Sensor Data*. Reston, VA, U.S. Geological Survey Professional Paper No 964
- Aumann Y and Lindell Y 2003 A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems* 20: 255–83
- Buttenfield B, Gahegan M, Miller H, and Yuan M 2001 *Geospatial Data Mining and Knowledge Discovery*. Washington D.C., University Consortium for Geographic Information Science White Paper on Emerging Research Themes (available at <http://www.ucgis.org/emerging/gkd.pdf>)
- Egenhofer M J and Franzosa R D 1991 Point-set topological spatial relations. *International Journal of Geographical Information Systems* 5: 161–74
- Ester M, Frommelt A, Kriegel H-P, and Sander J 2000 Spatial data mining: database primitives, algorithms and efficient DBMS support. *Data Mining and Knowledge Discovery* 4: 193–216
- Fotheringham A S and Wong D W S 1991 The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23: 1025–44
- Frank A U 1996 Qualitative spatial reasoning: Cardinal directions as an example. *International Journal Geographical Information Science* 10: 269–90
- Fu Y and Han J 1995 Meta-rule-guided mining of association rules in relational databases. In *Proceedings of the International Workshop on the Integration of Knowledge Discovery with Deductive and Object-Oriented Databases*: 39–46
- Fukada, T, Morimoto Y, Morishita S, and Tokuyama T 1999 Mining optimized rules for numeric attributes. *Journal of Computer and System Sciences* 58: 1–12
- Geolytics Inc. 2001 *Appendix J: Description of Tract Remapping Methodology*. East Brunswick, NJ, Geolytics Inc. Census Neighborhood Change Database, 1970–2000 Census Tracts (CD)
- Han J and Fu Y 1995 Discovery of multiple-level association rules from large databases. In *Proceedings of the International Conference on Very Large Databases*: 420–31

- Jenks G F and Coulson M R 1963 Class intervals for statistical maps. *International Yearbook of Cartography* 3: 119–34
- Klosgen W and May M 2002 Spatio-temporal subgroup discovery. In Ladner R, Shaw K, and Abdelguerfi M (eds) *Mining Spatio-Temporal Information Systems*. Boston, MA, Kluwer: 149–68
- Koperski K, Adhikary J, and Han J 1996 Spatial data mining: Progress and challenges survey paper. In *Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*: 55–70
- Koperski K and Han J 1995 Discovery of spatial association rules in geographic information databases. In *Proceedings of the Fourth International Symposium on Large Spatial Databases*: 47–66
- Ladner R, Shaw K, and Abdelguerfi M (eds) 2002 *Mining Spatio-Temporal Information Systems*. Boston, MA, Kluwer
- Liu B, Hsu W, and Ma W 1998 Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*: 80–6
- Malerba D, Lisi F A, Appice A, and Sblendorio F 2002 Mining spatial association rules in census data: A relational approach. In *Notes of the ECML/PKDD Workshop on Mining Official Data*: 80–93
- Mennis J and Pequet D J 2003 The role of knowledge representation in geographic knowledge discovery: A case study. *Transactions in GIS* 7: 371–91
- Miller H J and Han J 2001 Geographic data mining and knowledge discovery: An overview. In Miller H J and Han J (eds) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis: 3–32
- Piatetsky-Shapiro G 1991 Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro G and Fawley W J (eds) *Knowledge Discovery in Databases*. Menlo Park, CA, AAAI/MIT Press: 229–248
- Roddick J F and Hornsby K (eds) 2001 *Temporal, Spatial, and Spatio-Temporal Data Mining*. Berlin, Springer
- Shariff A R, Egenhofer M J, and Mark D M 1998 Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms. *International Journal of Geographical Information Science* 12: 215–46
- Shekhar S and Chawla S 2003 *Spatial Databases: A Tour*. Upper Saddle River, NJ, Prentice Hall
- Spiropoulou M and Roddick J F 2000 Higher order mining: modelling and mining the results of knowledge discovery. In *Data Mining II: Proceedings of the Second International Conference on Data Mining Methods and Databases*: 309–20
- Srikant R and Agrawal R 1996 Mining quantitative association rules in large relational tables. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 1–12
- Stier M 1999 Temporal Land Use and Land Cover Mapping. WWW document, <http://rockyweb.cr.usgs.gov/frontrange/land/templanduse/apsabs.htm>
- Tan P-N, Kumar V, and Srivastava J 2002 Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*: 32–41
- Wang K, Tay S H W, and Liu B 1998 Interestingness-based interval merger for numeric association rules. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*: 121–27
- Zeitouni K, Yeh L, and Augaure M-A 2001 Join indices as a tool for spatial data mining. In Roddick J F and Hornsby K (eds) *Temporal, Spatial, and Spatio-Temporal Data Mining*. Berlin, Springer: 105–16
- Zhang Z, Lu Y, and Zhang B 1997 An effective partitioning-combining algorithm for discovering quantitative association rules. In *Proceedings of the First Pacific Asia Conference on Knowledge Discovery and Data Mining*: 241–51