Mining Association Rules in Spatio-Temporal Data

J. Mennis and J.W. Liu

Department of Geography, University of Colorado,
Boulder, Colorado 80309-0260, USA.
Telephone: (303) 492 4794
FAX: (303) 492 7501
Email: jeremy@colorado.edu

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* ? *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, U.S.A. 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, 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, U.S.A. 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* ? *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:

$$is\_a(x, bagel) \wedge is\_purchased(x) \; ? \; is\_a(y, cream\_cheese) \wedge is\_purchased(y) \qquad (1)$$

Association rule mining uses the concepts of support and confidence. 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%)'. 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).

Shekhar and Chawla (2003) also point out that association rule mining is designed to work with nominal and ordinal data, not numeric data such as a metric distance. Thus, spatial data that are of a numeric type must be discretized into categories such as 'near' and 'far' for use in spatial association rule mining. Of course, 'near' and 'far' are relative and fuzzy terms that carry no intrinsic metric distance information. The determination of what distances may be defined as 'near' versus 'far' may have a large impact on the association rule mining results.

3. Case Study: Urban Growth in Denver, Colorado, U.S.A.

3.1. Overview

As a case study, we focused on mining association rules in data relating to urban growth in the Denver, Colorado, U.S.A. region from 1970 to 1990 (Figure 1). The study region extends from the Denver County boundary in the south to the cities of Boulder and Longmont in the north, and from the Rocky Mountain foothills in the east to the current edge of development on the Great Plains in the west. 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.

3.2. Data Sources

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. These data are generated through an areal interpolation methodology that uses sub-tract resolution Census data (e.g. block groups), and a weighting scheme based on the presence of local streets and other indicators of population concentration, to apportion Census data from previous years to Census 2000 tract boundaries (Geolytics, 2001).

Land cover data for the 1970s and 1990s were acquired from the U.S. Geological Survey's (USGS) Front Range Infrastructure Resources Project (FRIRP, http://rockyweb.cr.usgs.gov/frontrange/). These vector polygon data were generated from historic aerial photography, USGS digital orthophotographic quadrangles (DOQs), and ancillary data such as wetlands inventory (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).

3.3. 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 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. 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.

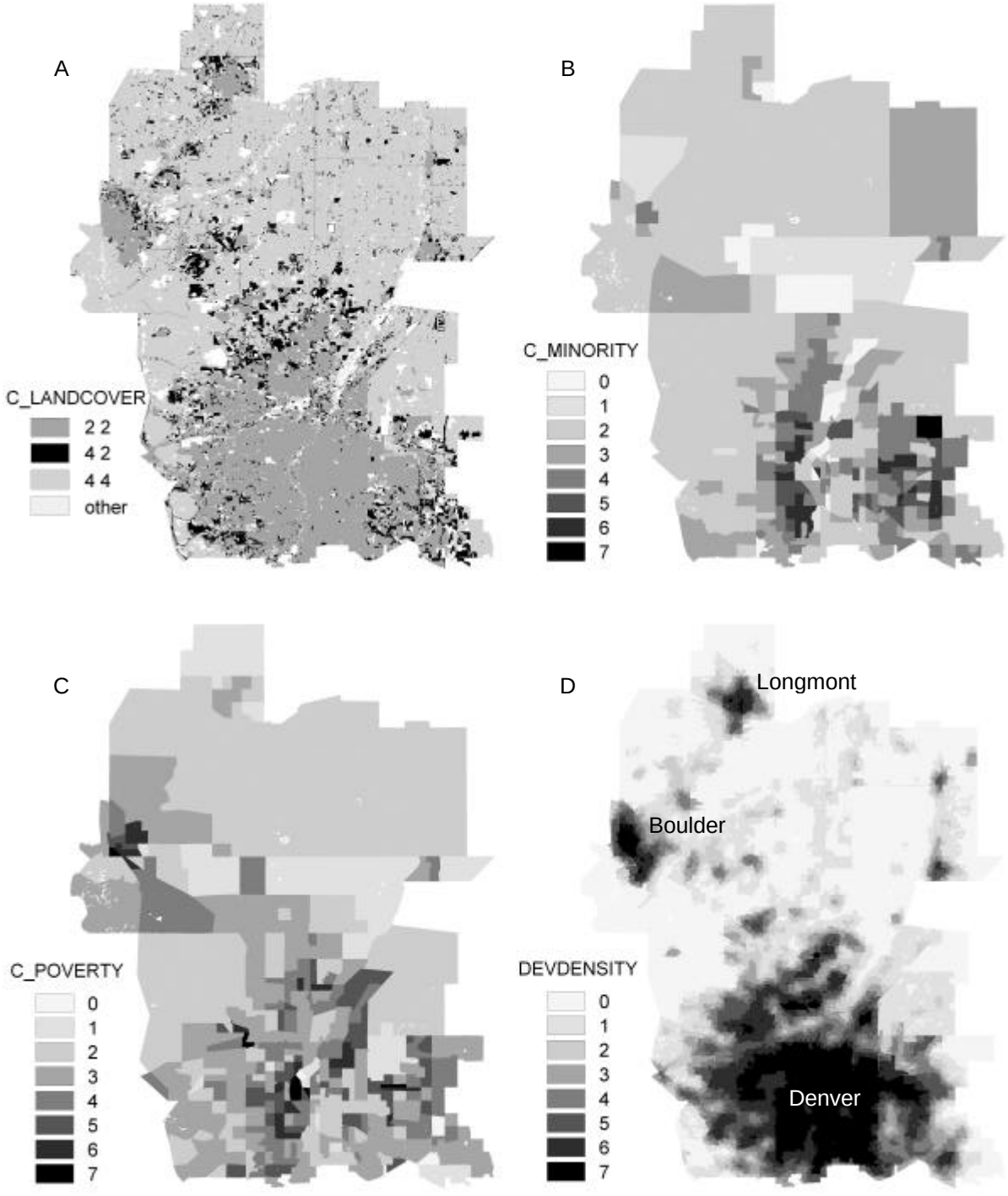

Figure 1. The Denver, Colorado, U.S.A. study region and the four variables used in the association rule mining: C_LANDCOVER (A), C_MINORITY (B), C_POVERTY (C), and DEVDENSITY (D). Refer to Tables 1 and 2 (column 'Level 3') for the meanings of the numeric codes shown in the legend for each variable.

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	…

The Census and land cover data were integrated within ArcGIS (Environmental Systems Research Institute, Inc.) and processed to produce a mining table in which each record represented a polygon that was homogeneous in both socioeconomic character and land cover change from 1970 – 1990. This table was created by intersecting the 1970 and 1990 land cover layers and then intersecting the resulting layer with the tract layer. After removing 'sliver' polygons and other artefacts of the polygon overlay operations, this approach produced a data layer with 24,258 polygons and an equal number of records in the polygon attribute table. We then calculated a number of variables that encode spatial and temporal relationships and characteristics for these polygons. We focus on four of these variables here (Figure 1):

- C_LANDCOVER Change in land cover, 1970 – 1990
- C_MINORITY Change in percent minority, 1970 – 1990
- C_POVERTY Change in percent of population living below the poverty line, 1970 – 1990
- DEVDENSITY Mean density of developed land in 1970

C_LANDCOVER stores a single value indicating the polygon's land cover in 1970 and its subsequent land cover in 1990. C_MINORITY and C_POVERTY were calculated by subtracting the 1990 value of the appropriate Census variable for the host tract by the 1970 value. Note that 'minority' (used to calculated the C_MINORITY variable) is defined here as anyone who self-identifies as either non-white or Hispanic. DEVDENSITY is the density of land classified as developed in 1970, created by first generating a 30 meter resolution binary grid of developed/not developed land cover, then creating a second grid that encodes the number of developed grid cells within 1 kilometer of each grid cell, and then finally calculating the mean grid cell value within each polygon.

As noted above, association rule mining works only with categorical data. We therefore converted each of the variables (with the exception of the nominal variable C_LANDCOVER) to an ordinal value. After experimenting with quantile and equal interval classification schemes, we settled on the natural breaks classification available in ArcGIS, which is based 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 2 (column 'Level 3' reports class IDs for the eight classes).

Table 2. 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			C_MINORITY (%)	C_POVERTY (%)	DEVDENSITY (mean # dev. cells w/in 1 km)
1	2	3			
0	0	0	-18 - -6	-41 - -25	1 - 328
0	0	1	-5 - 01	-24 - -5	329 - 697
0	1	2	2 - 7	-4 - 0	698 - 1138
0	1	3	8 - 13	1 - 5	1139 - 1622
1	2	4	14 - 20	6 - 10	1623 - 2126
1	2	5	21 - 28	11 - 15	2127 - 2626
1	3	6	29 - 37	16 - 22	2627 - 3071
1	3	7	38 - 68	23 - 37	3072 - 3409

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 C_LANDCOVER 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). The actual values of C_LANDCOVER are encoded as a text string which contains the 1970 numeric land cover code (e.g. '2' for 'developed'), followed by a space, and then the 1990 land cover code.

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 2 (column 'Level 3'). Figure 1 shows each of the three non-land cover change variables mapped according to level 3 of its concept hierarchy. Level 2 and level 1 map the eight classes given in level 3 into four classes and two classes, respectively (Table 2). In the case study, we indicate the level of a variable by appending '_L1,' '_L2,' or '_L3' onto the variable name (e.g. C_MINORITY_L3 for level 3).

The resulting mining table was composed of 13 fields, one field for each of the three levels of the concept hierarchy for each of the four variables, in addition to one field that represented a unique identifier for each record. The table had 24,258 records, each record representing one polygon.

3.4. Software and Methods

We used the association rule mining software CBA (v.2.1) (Classification Based on Associations) (Liu et al., 1998), developed at the School of Computing, National University of Singapore and available for free download (http://www.cs.uic.edu/~liub/). CBA uses the well known Apriori algorithm for finding association rules (Agrawal and Srikant, 1994). CBA provides two graphic user interfaces for rule exploration. The first interface sorts all rules by support or confidence; the second supports the exploration of rules composed of particular combinations of variables (figure 2). One can therefore choose to look only at rules composed of a particular set of predicates in the antecedent or consequent, and then sort those rules by support or confidence. We used multiple level association rule mining by generating rules at multiple levels of the three level hierarchy for each of the variables.

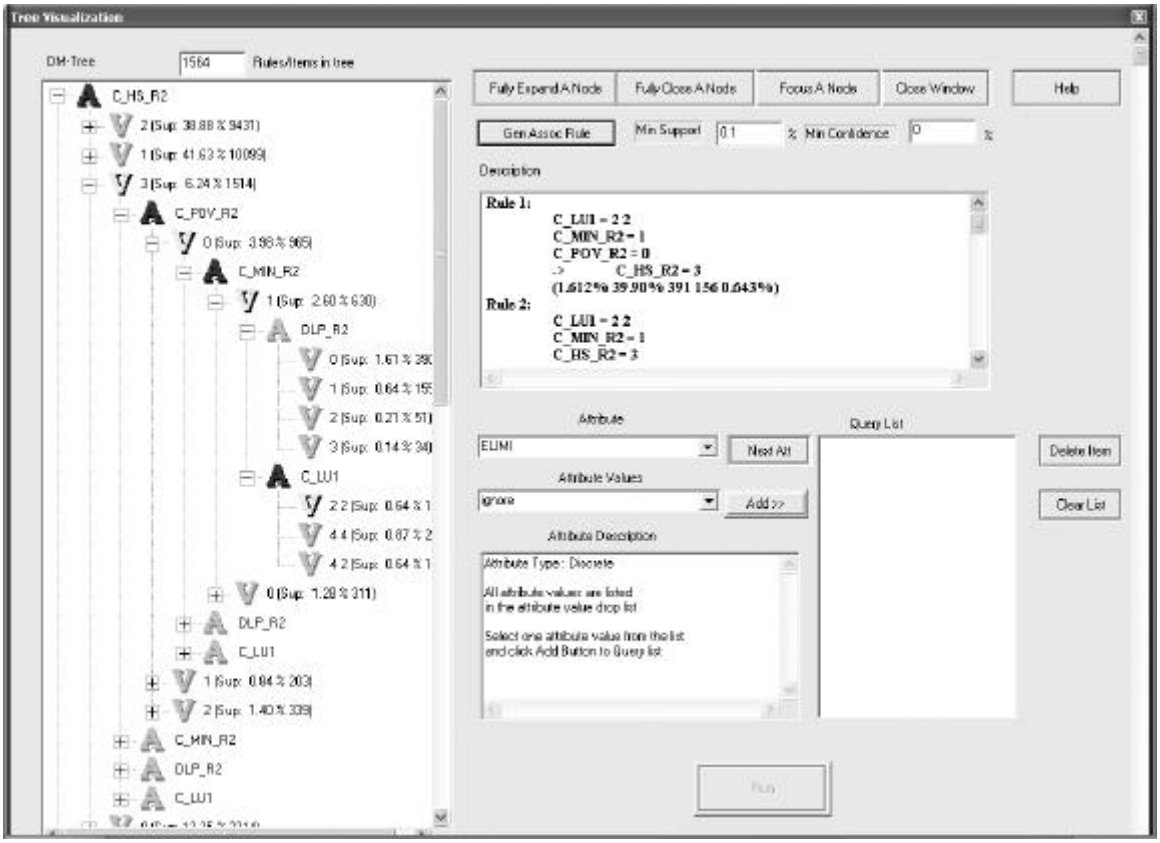

Figure 2. Screen capture of the CBA graphic user interface that supports the generation of association rules on a select set of variables using a hierarchical 'drill down' menu.

3.5. Results

The first association rule mining run encompassed all fields in the mining table, including all variables and all levels of the concept hierarchies, resulting in the generation of 122,970 rules. The run finished in under one minute on a Dell Dimension 4550 computer running Windows XP with a Pentium 4 2.40 GHz CPU and 512 Mb of RAM. Many of the generated rules are obvious or uninteresting, however. For instance, inclusion of all three levels of the hierarchy within one run produces rules that simply associate one level of the hierarchy with another for a given variable, e.g. 100% of those polygons that changed from 'woody vegetation' to 'residential' also changed from 'vegetated' to 'developed.' Consequently, 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 variable, then added the urban density variable. 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. Clearly, 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.

In the second mining run, only the level 1 C_MINORITY_L1 and C_POVERTY_L1 variables were included. For simplicity in discussing the rules, we refer to polygons where C_MINORITY_L1 = 1 (0) to be increasing (decreasing) in percent minority, even though the class break does not fall precisely on 0% (refer to Table 2 for the class breaks for each variable). We follow the same form of reference when discussing other variables. This run generated four rules, including the following:

$$(C\_MINORITY\_L1 = 1) \; ? \quad (C\_POVERTY\_L1 = 1) \; (9.53\%, 63.25\%) \qquad (2)$$

This statement indicates that in 63.25% of the polygons in which there is an increase in percent minority, there is also an increase in poverty. An important consideration in evaluating this rule is the rate of occurrence for the consequent out of all the records in the mining table. Here, out of all 24,258 polygons represented in the mining table, only 20.52% have C_POVERTY_L1 = 1. A comparison of 20.52% with the confidence of the rule (63.25%) gives an indication of the significance of the rule; it is a comparison of the rate of occurrence of the consequent in the rule to the consequent's rate of occurrence in the data set as a whole. In this case, it is a relatively unusual occurrence for poverty to be increasing (it happened in only 20.52% of all polygons); but it is approximately three times (63.25%/20.52%) as likely to occur when there is an increase in percent minority.

Further investigation of the rule given in Equation 3 is facilitated by choropleth mapping of the C_MINORITY_L1 and C_POVERTY_L1 variables (Figure 3). Figures 3A and 3B show that increases in both percent minority and percent poverty are concentrated around Denver, although an increase in poverty can also be seen to occur in the region around Boulder. Figure 3C shows the 63.25% of the C_MINORITY_L1 = 1 area where C_POVERTY_L1 = 1. For contrast, Figure 3D shows the areas where C_MINORITY_L1 = 0, with different values of C_POVERTY_L1.

We continued to explore the nature of the relationship between changes in percent minority and poverty by removing C_MINORITY_L1 and using the level 2 and level 3 variables for C_MINORITY in two subsequent mining runs, generating the following rules:

$$(C\_MINORITY\_L2 = 3) \; ? \quad (C\_POVERTY\_L1 = 1) \; (2.02\%, 80.16\%) \qquad (3)$$

$$(C\_MINORITY\_L3 = 7) \; ? \quad (C\_POVERTY\_L1 = 1) \; (0.34\%, 100.00\%) \qquad (4)$$

As the C_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 at level 3 (C_MINORITY_L3), all of the polygons with a C_MINORITY_L3 value of '7' (the highest percent minority class) also have a

C_POVERTY_L1 value of ‘1’ (Equation 4), although the support for this rule is quite low (0.34%); the combination of C_MINORITY_L3 = 7 and C_POVERTY_L1 = 1 occurs in only 82 polygons.

Figure 3. Choropleth maps of the variables given in the rule expressed in Equation 2. Shown are maps of C_MINORITY_L1 (A), C_POVERTY_L1 (B), and different combinations of C_MINORITY_L1 and C_POVERTY_L1 (C and D).

We investigated further by next incorporating the level 1 land cover change variable. We included C_MINORITY_L2 (level 2 of the concept hierarchy) 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 (C_LANDCOVER = 2 2) and those polygons that changed from vegetated to developed (C_LANDCOVER = 4 2) (refer to Table 1 to see the numeric land cover codes), among polygons strongly increasing in percent minority:

(C_LANDCOVER_L1 = 2 2) ^ (C_MINORITY_L2 = 3) ? (C_POVERTY_L1 = 1) (5)
(1.27%, 77.78%)

(C_LANDCOVER_L1 = 4 2) ^ (C_MINORITY_L2 = 3) ? (C_POVERTY_L1 = 1) (6)
(0.30%, 82.02%)

Interestingly, 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. Strongly increasing percent minority is associated with increasing poverty in polygons that remained developed as well in those polygons that changed from vegetated to developed.

In the next mining run, DEVDENSITY_L1 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 DEVDENSITY 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 DEVDENSITY value indicates a developed area surrounded by undeveloped land, such as a small town in a rural area. We also replaced C_LANDCOVER_L2 with C_LANDCOVER_L1, and C_MINORITY_L2 with C_MINORITY_L1 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 compare changes in poverty among polygons changing from vegetated to developed land covers in rural versus urban areas:

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L1 = 1) ^ (C_MINORITY_L1 = 1) ? (7)
(C_POVERTY_L1 = 1) (0.90%, 55.87%)

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L1 = 0) ^ (C_MINORITY_L1 = 1) ? (8)
(C_POVERTY_L1 = 1) (0.64%, 59.23%)

The rules given in Equations 7 and 8 suggest that there is not a difference between urban (DEVDENSITY_L1 = 1) and rural (DEVDENSITY_L1 = 0) areas in terms of the association between increasing percent minority and increasing percent poverty in areas that changed from vegetated to developed. We further investigated this pattern by generating similar rules to those given in Equations 7 and 8, but replacing DEVDENSITY_L1 with levels 2 and 3 for the DEVDENSITY variable in two subsequent mining runs. The following two pairs of rules were generated (Equations 9 and 10, and Equations 11 and 12), one pair for each run, again comparing urban versus rural areas:

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L2 = 3) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (0.26%, 69.57%) (9)

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L2 = 0) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (0.14%, 47.22%) (10)

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L3 = 7) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (0.13%, 84.21%) (11)

(C_LANDCOVER_L1 = 4 2) ^ (DEVDENSITY_L3 = 0) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (0.03%, 23.33%) (12)

Interestingly, as the attribute resolution of the DEVDENSITY variable becomes finer, a pattern emerges. Of those polygons that have very high urban density (DEVDENSITY_L3 = 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 heavily rural (DEVDENSITY_L3 = 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 that changed from vegetated to developed, for polygons that remained developed there is not a significant difference in the percent minority/poverty relationship between urban and rural areas.

(C_LANDCOVER_L1 = 2 2) ^ (DEVDENSITY_L3 = 7) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (2.56%, 76.95%) (13)

(C_LANDCOVER_L1 = 2 2) ^ (DEVDENSITY_L3 = 0) ^ (C_MINORITY_L1 = 1) ? (C_POVERTY_L1 = 1) (0.03%, 80.00%) (14)

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 the set of polygons that remained developed and for the set that changed from vegetated to developed. However, in areas that changed from vegetated to developed, the general percent minority/poverty relationship changes depending on whether the development occurred in a very urban versus very rural area. In very urban areas that underwent development, increasing percent minority is associated with increasing poverty. However, in very rural areas that underwent development, increasing percent minority is not associated with increasing poverty. Although, it should be noted that it is relatively rare for a polygon to be increasing in minority and undergoing development in a very rural area. In contrast, the urban versus rural factor does not alter the percent minority/poverty relationship in areas that remained developed. One may thus hypothesize that the dynamic relationship between factors of race and poverty is controlled in part by the type of land cover change (or lack of change), but even more so by the urban versus rural setting where that land cover change occurs.

4. Conclusion

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. This research also demonstrates how generated rules may be transformed into GIS queries that produce choropleth maps to support visual data exploration.

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 (Goodchild and Lam, 1980) to apportion the tract data to homogeneous land cover change units. While we experimented with other approaches, such as aggregating the land cover change data to tracts, we found these other approaches to be awkward and cumbersome. A related problem is that the land cover change polygons range widely in size. Thus, the confidence of a rule, expressed in terms of the percentage of the number of polygons in the database, may not reflect the actual percentage in area. These issues are, of course, not restricted to association rule mining, but are applicable to many types of spatial statistical analysis. Nonetheless, the nature of these problems in the context of association rule mining has yet to be determined.

A second issue for further research is the impact of data classification on spatio-temporal association rule mining. We briefly experimented with quantile and equal interval classification schemes before settling on the natural breaks method. Other researchers have proposed novel methods for optimizing the classification of numeric data to find interesting association rules (e.g Wang et al., 1998). 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.

A third issue concerns the representation of spatial relationships. In the case study we focused on relationships of spatial coincidence and distance. However, there are other types of spatial relationships that may be used in spatio-temporal association rule mining, such as directional and topologic relationships (Egenhofer and Franzosa, 1991; Peuquet and Zhan, 1987). A structured experiment comparing different spatial relationship types in association rule mining would illuminate the impact of the choice of spatial relationship type on the mining results.

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 approaches beyond support and confidence for measuring the

'interestingness' of association rules (e.g. Tan et al., 2002). Others have suggested approaches for analyzing data mining results (meta-mining), for instance 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.

5. 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 (#02-0000-0028).

6. 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*, pp. 207-216.

AGRAWAL, R. and SRIKANT, R., 1994, Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Databases*, pp. 487-499.

ANDERSON, J.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, U.S. Geological Survey Professional Paper 964* (Reston: U.S. Geological Survey).

BUTTENFIELD, B., GAHEGAN, M., MILLER, H., and YUAN, M., 2001, *Geospatial Data Mining and Knowledge Discovery*. White Paper on Emerging Research Themes, University Consortium for Geographic Information Science (WWW document, http://www.ucgis.org/emerging/gkd.pdf, last accessed February 21, 2003).

EGENHOFER, M.J. and FRANZOSA, R.D., 1991, Point-set topological spatial relations. *International Journal of Geographical Information Systems*, **5**, 161-174.

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.

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*, pp. 39-46.

GEOLYTICS, 2001, *Appendix J: Description of Tract Remapping Methodology*. CensusCD Neighborhood Change Database: 1970 – 2000 Census Tracts. Geolytics, Inc. (CD).

GOODCHILD, M. and LAM, S.-N., 1980, Areal interpolation: a variant on the traditional spatial problem. *Geo-processing*, **1**, 297-312.

HAN, J. and FU, Y., 1995, Discovery of multiple-level association rules from large databases. In *Proceedings of the Interational Conference on Very Large Databases*, pp. 420-431

JENKS, G.F. and COULSON, M.R., 1963, Class intervals for statistical maps. *International Yearbook of Cartography*, **3**, 119-134.

KLOSGEN, W. and MAY, M., 2002, Spatio-temporal subgroup discovery. In (R. Ladner, K. Shaw, and M. Abdelguerfi, Eds.) *Mining Spatio-Temporal Information Systems*

(Boston: Kluwer Academic Publishers), pp. 149-168.

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*, pp. 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*, pp. 47-66.

LADNER, R., SHAW, K., and ABDELGUERFI, M. (Eds.), 2002. *Mining Spatio-Temporal Information Systems* (Boston: Kluwer Academic Press).

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*, pp. 80-86.

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 2002 Workshop on Mining Official Data*, pp. 80-93.

MENNIS, J. and PEUQUET, D.J., 2003, The role of knowledge representation in geographic knowledge discovery: a case study. *Transactions in GIS*, **7**, 371-391.

MILLER, H.J. and HAN, J., 2001, Geographic data mining and knowledge discovery: an overview. In (H.J. Miller and J. Han, Eds.) *Geographic Data Mining and Knowledge Discovery* (London: Taylor and Francis), pp. 3-32.

PEUQUET, D. and ZHAN, C.-X., 1987, An algorithm to determine the directional relationship between arbitrarily shaped polygons in the plane. *Pattern Recognition*, **20**, 65-74.

RODDICK, J.F. and HORNSBY, K. (Eds.), 2001. *Temporal, Spatial, and Spatio-Temporal Data Mining* (Berlin: Springer).

SHEKHAR, S. and CHAWLA, S., 2003. *Spatial Databases: A Tour* (Upper Saddle River, New Jersey: Prentice Hall).

SPILIOPOULOU, 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*, pp. 309-320.

STIER, M., 1999. Temporal land use and land cover mapping. Paper presented at the American Society for Photogrammetry and Remote Sensing (ASPRS) Conference, Portland, Oregon (WWW document, http://rockyweb.cr.usgs.gov/frontrange/land/templanduse/apsabs.htm, last accessed June 11, 2003).

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*, pp. 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*, pp. 121-127.

ZEITOUNI, K., YEH, L, and AUGAURE, M.-A., 2001, Join indices as a tool for spatial data mining. In (J.F. Roddick and K. Hornsby, Eds.) *Temporal, Spatial, and Spatio-Temporal Data Mining* (Berlin: Springer), pp. 105-116.