

Research Article

Knowledge discovery from mining the spatial associations between cancer mortality and socioeconomic characteristics.

SRINIVAS VINNAKOTA and NINA S.-N LAM*

Department of Geography and Anthropology, 227 Howe-Russell Geoscience Complex,
Louisiana State University, Baton Rouge, LA 70803, USA.

Correspondence: *Nina S.-N. Lam. Email: nlam@lsu.edu

Abstract. Cancer is the second leading cause of death in the United States of America, with a rather uneven distribution among the population. Revealing the associations between socioeconomic characteristics and the spatial distributions of cancer mortality will help in explaining patterns of health disparity, as well as in generating hypotheses for cancer etiology. This paper demonstrates the use of association rule mining to extract associations between selected socioeconomic variables and the four most common cancer types in the United States (lung, colorectal, breast, and prostate) using mortality data at the Health Service Area (HSA) level. A geographic information system is used to integrate data of different spatial resolutions, and to visualize and analyze the results from the association rule mining process. Results show that areas with high rates of low education levels, high unemployment rates, and low paying jobs tend to have higher rates of cancer mortality.

Keywords: Spatial association rule mining; Knowledge discovery; Cancer mortality; Socioeconomic inequality;

1 Introduction

From the application of a paper based point-map by John Snow to investigate the outbreak of cholera epidemic in 1854 to the current day practices of using digital maps, the medical and public health profession has greatly benefited from the use of spatial information as manifested by maps. In this regard contemporary Geographic Information Systems (GIS) not only provide an excellent platform in which digital maps and data can be manipulated to extract valuable information, but also serve as an excellent medium for visualizing geographic phenomenon. Currently available geographic information system software also provides a capability in which data can be subject to statistical procedures for analysis. The ability of a GIS to bring together data from different sources underpinned by a common spatial unit and its ability to manipulate such data at various scales makes it an ideal choice for analyzing spatial data. But one inherent limitation of existing commercial software is the limited choice of analytical functions. In the field of public health and medical geography, geographic information systems have previously been used to analyze disease clusters (Nkhoma et al. 2004, Gregorio et al. 2005), predicting disease outbreaks (Lam et al. 1996, Moonan et al. 2004), accessibility to health care facilities (McLafferty 2003), environment - human health interactions (Cuthe et al. 1992, Ruiz et al. 2004, Jacquez and Greiling 2003a), and spatial distribution of disease (Jacquez and Greiling 2003b,

Pickle et al. 1996). One such area has been the study of spatial distribution of cancer incidence and mortality (Bithell and Vincent 2000).

Cancer is a family of diseases and there is no one cause or cure for cancer (Meade and Earickson 2000). In the year 2002 cancer had been the second leading cause of death in the United States of America (American Cancer Society 2005). Some kinds of cancer have long been suspected of being genetically linked (Reiss-Starr et al. 1998) while some are more prevalent in certain population groups than others (Lam 1986). There has been evidence that some cancers affect men more than women while there are other forms of cancer that affect people of a particular race and ethnicity more than others. In this regard Parker et al. (1998) argues, “Because race and ethnicity are so strongly correlated with socioeconomic status, some of the differences in cancer incidence and mortality rates that exist among racial and ethnic groups are probably the result of socioeconomic status rather than genetic and cultural aspects of race and ethnicity”. Tobacco consumption and environmental pollution are widely perceived to be the major cause for different types of cancer that are prevalent today. Acquavella (1999), Weinrich et al (1999), and Morrison et al (1993) report associations between occupational exposure and prostate cancer while a study by Briggs et al (2003) found a higher incidence of cancer among African-American workers reflecting racial disparities in levels of exposure to occupational carcinogens. In a study of breast cancer mortality among US women, Wagener and Schatzkin (1994) have found that mortality rates decreased in areas of higher socioeconomic status while during the same time-period it increased in areas of lower socioeconomic status.

From a public health point of view it then becomes imperative that one has to assess the differences in socioeconomic, environment and disease characteristics to find explanations. Some of the methodological problems associated with analyzing cancer mortality using GIS have previously been investigated (Pickle et al. 2005, Lam 1986, Teppo 1998). These studies urge not to overtly interpret area based associations of cancer mortality with socioeconomic or environmental variables but call for exercising caution in doing so.

The objective of this study is to demonstrate the use of association rule mining in uncovering patterns of association between cancer mortality and socioeconomic characteristics at a national scale. Although the unit of analysis in this study is not at the county level, this study will provide useful information on whether health disparity exists at a national level. With few exceptions (Devesa et al. 1999, Singh and Siahpush 2002, Singh et al. 2002a, Singh et al. 2002b), reports on the spatial distribution of socioeconomic inequalities in cancer mortality at a national level have seldom been made. Information of this nature could also be useful to generate hypotheses regarding cancer etiology, and its control and prevention. Such efforts are needed to reach areas that exhibit high cancer mortality rates and to understand the underlying differences of its occurrence across a diverse spectrum of the population.

2 Knowledge Discovery and Association Rule Mining

The concept of knowledge discovery is to extract implicit information in a dataset. Knowledge extracted from a dataset refers to a set of rules that are implicit, valid, novel, potentially useful, and those that are easily comprehensible by humans (Fayyad et al. 1996, Frawley et al. 1992).



Figure 1. Flowchart of the knowledge discovery process.

The process of extracting knowledge is an interactive and iterative procedure involving many tasks (Fayyad et al. 1996, Han and Kamber 2001) as illustrated in figure 1.

Association analysis is one of the most widely researched topics in data mining (Hipp et al. 2000). The main focus of association rule mining is to generate hypothesis rather than to test them as is commonly achieved using statistical techniques. Association rule mining, first conceived by Agrawal et al. (1993) was used for analyzing market-basket data to *mine* customer shopping patterns. The idea was to find relations among the items (termed as predicates) purchased so that the customers could then be targeted for marketing specific products. An association rule typically consists of 3 parts – an antecedent (X%), a consequent (Y%), and a measure of the interestingness of the rule (*support%*, *confidence%*, *lift*) and is represented in the form shown in equation 1.

$$X \rightarrow Y (\text{support}\%, \text{confidence}\%, \text{lift}) \quad \text{-- Equation 1.}$$

The antecedent and the consequent are a set of one or more predicates. The support of a rule measures the frequency of occurrence of all the predicates of a rule in the dataset, the confidence measures the frequency of occurrence of the consequent predicates of the rule given the occurrence of the antecedent predicates of the rule, while the lift of a rule measures the correlation among the antecedent and consequent predicates of a rule. A lift value of 1 would indicate that the predicates expressed in the rule are independent of each other while a value greater than 1 would indicate a positive correlation among the antecedent and consequent predicates, and a value less than 1 would indicate a negative correlation among the antecedent and consequent predicates of the rule.

For example, consider a dataset composed of 100 records of cancer mortality and median income. Let 60 records indicate high rates of mortality, 50 records indicate low median income, and 40 records be comprised of both high rates of mortality and low median income values. Assuming the association analysis produces a rule linking the low median income and high cancer mortality, the rule would be of the form shown in equation 2:

$$\text{Low median income} \rightarrow \text{high cancer mortality} (40\%, 80\%, 1.33) \quad \text{-- Equation 2.}$$

The support of a rule is calculated as the ratio of the number of records containing both antecedent and consequent predicates of the rule to the total number of records in the dataset expressed as a percentage. In this case support of the rule is the ratio of number of records containing high cancer mortality (consequent predicate) and low median income (antecedent predicate) to the total number of records in the dataset i.e., $(40/100)*100 = 40\%$. The confidence of a rule is the ratio of the number of records containing both antecedent and consequent predicates of a rule to the number of records that contain the antecedent predicates of the rule, and is expressed as a percentage. The confidence in this case is calculated as $(40/50)*100 = 80\%$. The lift is measured as the ratio of the probability of all the predicates in a rule occurring together to the probability of each predicate of the rule occurring independently. In this case the lift is calculated as

$$\text{lift is calculated as } \frac{P(\text{lowMedianIncome} \cup \text{highCancerMortality})}{P(\text{lowMedianIncome})P(\text{highCancerMortality})} = \frac{40/100}{(60/100)*(50/100)} =$$

1.33. The lift value of 1.33 (greater than 1) indicates a positive correlation among low median income and high cancer mortality.

Based on the market-basket analysis, a spatial extension of association rule mining for analyzing geographic datasets was developed by Koperski and Han (1995). A spatial association rule contains a spatial predicate in either (or both) the antecedent or the consequent part of the rule. Ester et al. (2001) proposed a neighbourhood graph based mining of association rules, Klösgen and May (2002) used subgroup mining to analyze dependencies between a target and a

from three different sources: U.S. Census Bureau, U.S. Bureau of Economic Analysis, and National Atlas of the United States of America. The county-based socioeconomic data was spatially aggregated at the Health Service Area level so that it can be linked with the cancer mortality dataset. A Health Service Area (HSA) is a geographic area comprising of one or more contiguous counties delineated for the purposes of health planning and treatment. In all there were a total of 805 HSAs encompassing 3,141 counties in the United States of America. The following sections give a brief overview of the data variables and their spatial resolution.

The cancer mortality dataset used in this study was extracted from Pickle et al. (1996). The dataset included gender and race specific mortality rates for the 4 most common types of cancer - colorectal cancer, lung cancer, breast cancer for women, and prostate cancer for men aggregated over the years 1988 – 1992 at the HSA level and age-adjusted based on the US standard million population from the 1940 census population count. The mortality data was classified according to the International Classification of Diseases (ICD) 9 classification scheme.

The socioeconomic data used in this study comprised five categories: family composition, education, housing, economic conditions, and occupation. A complete list of the variables used in this study is summarized in table 1. These variables were chosen based on their ability to describe the living conditions of the general population and in part on having been used in area-based statistical studies involving the use of socioeconomic characteristics to describe public health (Devesa et al. 1999, Singh and Siahpush 2002, Turrel and Mathers 2001). Direct income measures were not used in the current study because the Census Bureau provides median income values of a county and these values cannot be aggregated to the HSA level, and per capita income tends to be a biased measure because of the unequal distribution of wealth among the population. As such measures of the availability of household plumbing facilities and households with no cars were used as surrogate measures of income and poverty levels.

4. Methodology

Firstly, all the datasets had to be converted to a common spatial unit, which in this study was a Health Service Area. This was because the cancer mortality dataset was made available at the HSA level, whereas the socioeconomic variables provided by the US Census Bureau and the Bureau of Economic Analysis were at a county level. The aggregation process involved loading a shapefile containing HSAs into a GIS along with a county shapefile. Using the geoprocessing wizard of ESRI® ArcGIS 8.3 a database was created that contained a list of all HSAs and their associated counties. This list was used to aggregate the county based socioeconomic variables to the HSA level. In all 3,141 counties were aggregated into 805 HSAs. The socioeconomic variables were then normalized based on aggregated values obtained at the HSA level.

For association rule mining the Classification Based Association (CBA ver2.0) software developed by Liu et al. (1999) was used. Since association rule mining requires categorical input data, each variable in the dataset was independently discretized in to 5 quintile groups with each group containing about 20% of the records in the dataset. The five groups were labeled low, medium-low, medium, medium-high, and high respectively. As the association rules are independent of the number of variables contained in the dataset no attempt was made to reduce

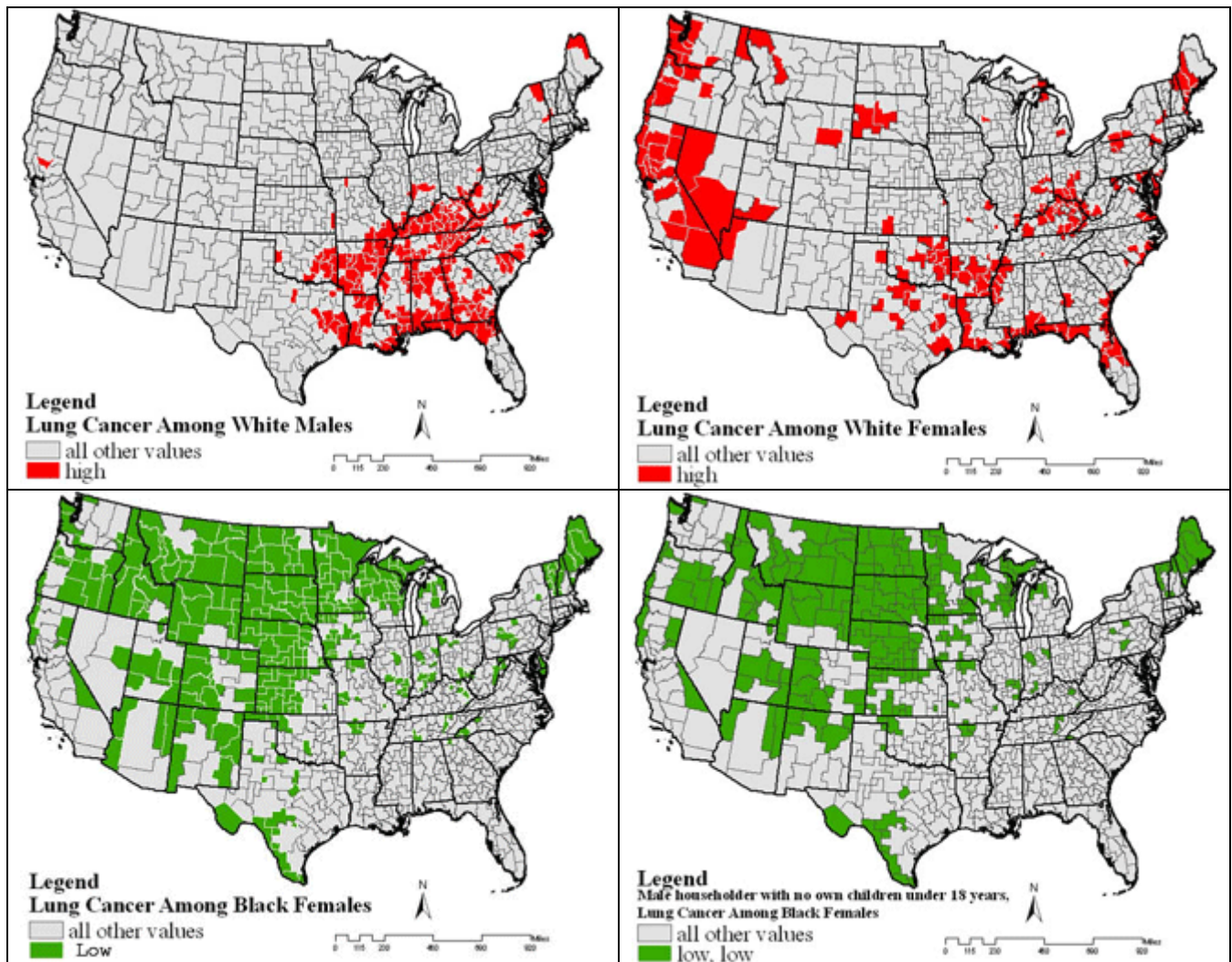


Figure 2. Spatial distribution of lung cancer mortality rate among (a) white males (b) white females (c) black females, and (d) areas with low rate of lung cancer mortality among black females and low percentage of households with male householder with no spouse and having no children under 18 years.

the number of variables included in the study. A minimum support value of 10 generated large number of association rules either among the socioeconomic variable themselves or among the cancer mortality rates. Since the intent of the study was to find associations between the socioeconomic variables and cancer mortalities the support value was iteratively changed (decreased) until a suitable number of rules consisting of such associations were generated. After several iterations, the minimum support level was set at 7% and a minimum confidence level of 40%, in other words the probability that a particular rule occurs in the dataset was set to ≥ 0.07 and the minimum probability that a rule occurs given its antecedent was set to ≥ 0.4 . While the probability values may seem low, consider the following fact: two variables are independently discretized in to 5 classes with each class containing (roughly) 161 of a total of 805 records in the dataset. The probability that a class occurs in the dataset is 0.2. Since each variable can take one of the 5 class values, the two variables could possibly have 25 different combinations (5C_5) in the dataset. The least probable scenario in which only 1 record contains a probable combination could occur with a probability of 0.0012 while at the other extreme a

particular combination may occur at most 161 times with a probability of 0.2. An association with a low probable value might not be interesting while an association with a high probable value might not occur in the dataset. Since the concept of a minimum support value is to eliminate only those rules with a probability less than the specified value from being outputted, the minimum value in this study was set to 0.07, i.e. all associations with a probability of 0.07 or greater are displayed.

5. Results and Discussion

The following section describes three different categories of association analysis. The first part describes the associations with the highest support and confidence levels, the second part describes the associations that are interesting although their support levels are not as high as the rules described in the first part, and the third part describes rules that have relatively low support values compared to the first two parts but are interesting. A high support level of 7% i.e., an occurrence in a minimum of 56 of 805 HSAs and a minimum confidence level of 40% were specified to generate rules that describe the general trend of cancer mortality in the United States. In all 211,499 such rules were generated based on the above specified criteria.

Table 2 lists the top three association rules having the highest support values. Association rule 2a through 2c correlate density of households with black male householder living without the presence of spouse and with no own children under the age of eighteen years with breast cancer among black females, colorectal cancer among black females, and lung cancer among black females. These rules also exhibit high positive correlation as indicated by their respective lift values (greater than 2). Even though the above rules indicate high degree of interestingness as dictated by their respective measures they are neither novel nor provide any useful information. The strongest rules generated are most likely dominated by the most commonly occurring cases such as HSA's with average or low mortality rates. Figures 2a and 2b show the distribution of lung cancer mortality among white males and females respectively. Among white males HSAs that have the highest lung cancer mortality are Madison, MO (119.29), Scott, TN (115.03), Pike, KY – Logan, WV (100.69), Johnson, AR (100.51), Decatur, GA – Seminole, GA (96.43). Among white females highest rates of lung cancer mortality are relatively lower compared to white males and the areas that have the highest rates are Northampton, VA – Accomack, VA (49.36), Pike, KY – Logan, WV (44.82), Kent, MD (43.93), Coos, OR – Del Norte, CA (43.13), Floyd, KY – Johnson, KY (40.21). Figure 2c shows the distribution of high lung cancer mortality rate among black females while figure 2d shows a map of the association rule 2c. For cancer mortality studies, rules associated with high mortality rates, while less frequent, are more useful. The second set of rules selected for discussion here is

	Association Rule	Support%, Confidence%, lift (number of HSAs in the rule)
a	Percent of households with black male householder with no spouse and with no own children below 18 years (low) → breast cancer mortality among black females (low)	27.95%, 89.29%, 2.49 (225)
b	Percent of households with black male householder with no spouse and with no own children below 18 years (low) → colorectal cancer mortality among black females (low)	27.70%, 88.49%, 2.39 (223)
c	Percent of households with black male householder with no spouse and with no own children below 18 years (low) → lung cancer mortality among black females (low)	27.20%, 86.90%, 2.48 (219)

based on rules associated with high mortality rates, and in this case lung cancer mortality among white males and females, though their support levels are not as high as the first set (Table 2).

From table 3, association rule 3a describes areas with high percentage of 3 person households as areas that have a high rate of lung cancer mortality among white men. This rule has a support of 11.06% and confidence of 55.28% i.e., of 161 HSAs that have a high density of 3 person households 89 areas have had a high rate of lung cancer mortality among men. Based on the lift value (2.76) the predicates of this rule can be said to exhibit a strong positive correlation. Association rules 3b and 3c describe HSAs that have a high density of whites below poverty and HSAs that have a high density of whites who had low education attainment respectively as areas where the lung cancer mortality rate among white males is high. Association rule 3d describes the lung cancer mortality rate among white females based on the marital status. Health Service Areas with a higher percentage of female divorcees tend to have a high rate of lung cancer mortality among white females. Figure 3 a-b map the association rules 3b and 3d described above.

Table 3: Association rules generated using high minimum support level (7%) and high minimum confidence level (40%).		
	Association Rule	Support%, Confidence% (number of HSAs in the rule)
a	Areas with 3 person household density is high → lung cancer among white males is high	11.06%, 55.28%, 2.76 (89)
b	Areas with high density of whites below poverty → lung cancer among white males is high	8.70%, 43.48%, 2.17 (70)
c	Areas with high density of low education among whites → lung cancer among white males is high	9.07%, 45.34%, 2.26 (73)
d	Areas with high density of female divorcees → lung cancer among white females is high	8.44%, 42.24%, 2.11 (68)

The association rules generated using a high support and confidence values describe the socioeconomic characteristics of cancer mortality of the majority of the health service areas. Alternately, it might be probable that a particular socioeconomic characteristic might occur infrequently in the dataset but when this occurs the probability that it is associated with cancer mortality is high. In order to extract such rules one has to consider low minimum support value and a high minimum confidence value, which becomes the third set of rules selected for discussion below.

In all, 4,979 association rules were extracted using a support value of 0.5% and a minimum confidence value of 50% and examples of these rules are listed in table 4. Association rule 4a and 4b correlate areas that have high proportion of whites with low education, and areas with high density of workers in transportation and low density of workers in services industry respectively, with areas of high rate of lung cancer mortality among white men. Rule 4c specifies that areas with low density of whites having higher education and high unemployment rates for white males have a tendency to have high rates of lung cancer mortality among white men. Association rules 4d and 4f correlate low educational standards and white households with no vehicles, and households with no plumbing and white households with no vehicles with high rate of lung cancer mortality among white males. The last rule 4g associates areas with low density of whites with higher education and density of people born in the south but living in a

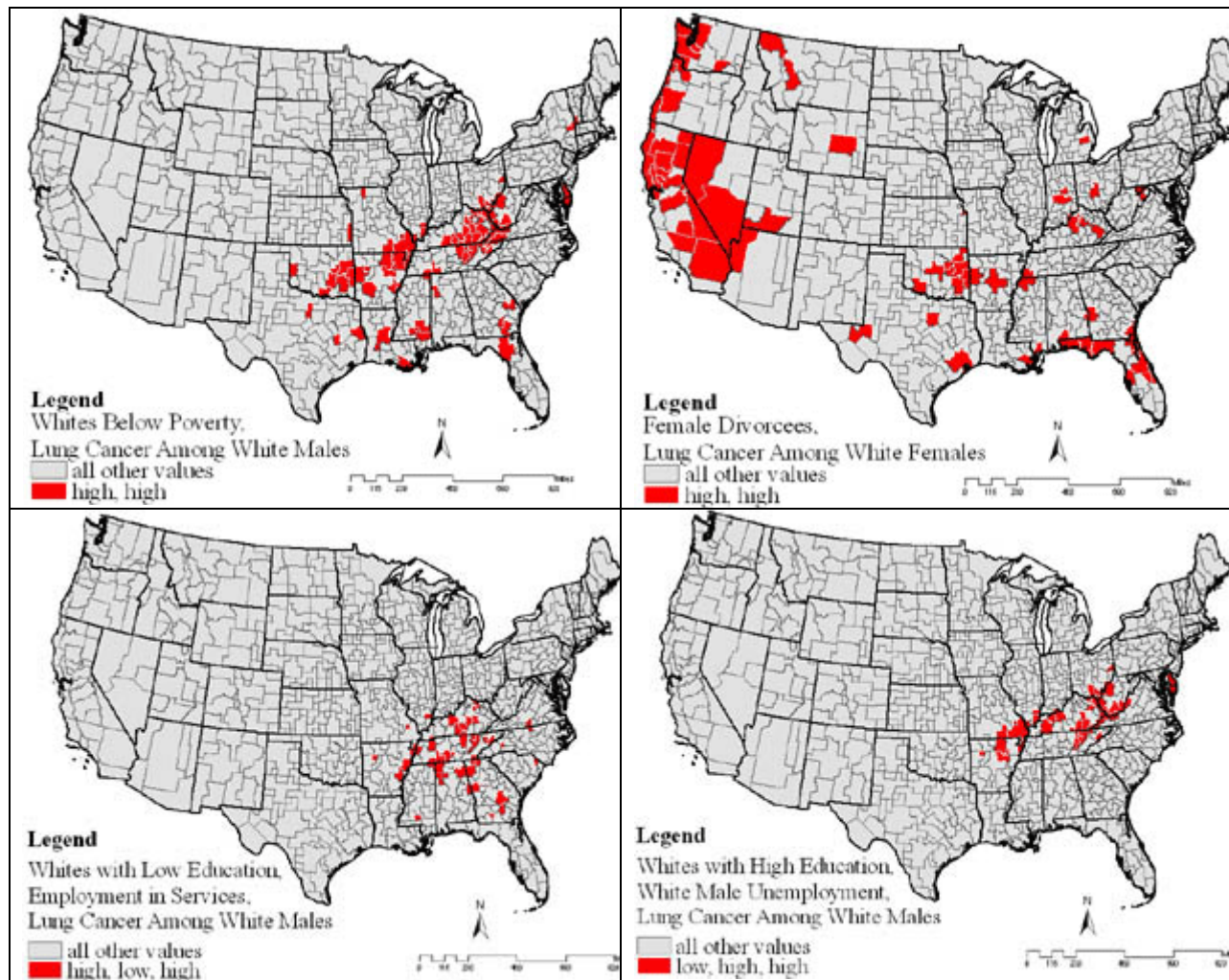


Figure 3. Spatial distribution of (a) high lung cancer mortality rate among white men and percentage of whites living below poverty level, (b) high lung cancer mortality rate among white females and percentage of female divorcees, (c) high rate of lung cancer mortality among white men and areas with high proportion of whites with low education and low proportion of employment in services industry, and (d) high rate of lung cancer mortality among white men and low density of whites with high education attainment and high rate of white male unemployment.

different region in US as areas that have a high rate of lung cancer mortality among white men. Figure 3 c-d map the association rules 4 b-c respectively obtained from using a low support and high confidence values.

The results presented previously are area based measures of socioeconomic characteristics. As such these should be interpreted accordingly. While an association rule stating that areas with high unemployment among white males are also the areas that are more likely to contain high rate of lung cancer mortality among white males, it certainly does not imply that lung cancer affects white men with no jobs. It only states that there is a higher probability that in an area of high unemployment rate one could find a high rate of mortality due to lung cancer among white men. This could be either because the areas lack adequate health facilities, screening centers, or the population in these areas is exposed to more polluted environment than others. Since lung cancer is highly associated with tobacco consumption it could also be that white men in these areas tend to smoke more. Persons who are separated from their family often times lead a stressful life (Singh et al. 2002b). So, areas with high percentage

	Association Rule	Support%, Confidence% (number of HSAs in the rule)
a	Density of whites with low education is high AND percentage of workers employed in transportation industry is high → lung cancer among white males is high	2.23%, 64.29%, 3.21 (18)
b	Density of whites with low education is high AND percentage of workers employed in services industry is low → lung cancer among white males is high	5.093%, 53.95%, 2.69 (41)
c	Density of whites with higher education is low AND unemployment rate among white males is high → lung cancer among white males is high	4.22%, 52.31%, 2.61 (34)
d	Density of whites with low education is high AND number of white households with no vehicles is high → lung cancer mortality among white males is high.	4.96%, 51.95%, 2.59 (40)
e	Percentage of workers employed in construction industry is medium-high AND number of houses with no plumbing facilities is medium-high → lung cancer among white males is high	1.61%, 50.00%, 2.5 (13)
f	Density of households with no plumbing is high AND number of white households with no vehicles is high → lung cancer mortality among white males is high	3.97%, 52.46%, 2.62 (32)
g	Density of whites with higher education is low AND density of people born in the south is high → lung cancer mortality among white males is high.	2.48%, 60.61%, 3.03 (20)

of female divorcees would indicate a breakdown in the family structure and these areas tend to have high number of persons leading a stressful life. Gee and Payne-Sturges (2004) suggest that increased levels of stress might imbalance the functioning of body systems thereby leading to chronic illness. Low educational attainment would indicate lower levels of income and employment in blue collared jobs while households with no plumbing and no cars would indicate poverty. In general areas with high rates of cancer mortality were associated with low educational attainment, high rates of unemployment, higher percentage of the population employed in construction, mining, transportation, and agricultural industries. These areas tend to have high density of households with no plumbing and no vehicles.

Knowledge discovery using association analysis is an iterative procedure at times requiring expert domain knowledge. Since the intent of the analysis is to extract implicit patterns from among the variables in the dataset, different combinations of support and confidence values have to be tested. As had been previously mentioned, sometimes association rules might have strong measures of support, confidence, and lift values but still not provide any useful information. Some basic domain knowledge regarding the dataset might help overcome this limitation. With the exception of correlation as measured by lift, the main drawback of association analysis is that it does not provide a measure of statistical significance, for the rules. On the other hand the same argument can be used in favor of association analysis, in that it makes no assumption of the data being independent and identically distributed as required by statistics. Since the process of association rule mining is data driven in which a pattern is induced based on the available data while making no assumption about the extracted pattern, this makes it an exploratory approach. The patterns extracted using association rule mining could be used to generate a hypothesis that could then be tested using statistical techniques.

6 Future work

One of the critical issues in association rule mining is that it requires categorical data as its input. Since the discrete class combinations depend on the method of discretizing continuous variables and on the number of classes chosen to represent the data, it would be interesting to study the

effects of these class interval selections on the results of association rule mining. Also, it is generally accepted that cancer mortality varies significantly with environmental conditions (Brody 2003). As such a future study would incorporate environmental data in addition to socioeconomic data to study the distribution of cancer mortality across the United States. Finally, using more detailed level data (e.g., county-level data at the national level) will also help, as more variables can be included (e.g. median income, whether the county is urban or rural) to better reveal the associations.

References

- ACQUAVELLA, J. F. (1999) Farming and prostate cancer. *Epidemiology*, 10, 349-51.
- AGRAWAL, R., IMIELINSKI, T. & SWAMI, A. (1993) Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD International Conference on Management of Data*. Washington D. C., USA.
- AMERICAN CANCER SOCIETY. <http://www.cancer.org>. Last accessed May 23rd 2005.
- BITHELL, J. F. & VINCENT, T. J. (2000) Geographical Variations in Childhood Leukaemia. IN ELLIOT, P. & WAKEFIELD, J. C. (Eds.) *Spatial Epidemiology: Methods and Applications*. Oxford, Oxford University Press.
- BRIGSS, N. C., LEVINE, R. S., HALL, H. I., COSBY, O., BRANN, E. A. & HENNEKENS, C. H. (2003) Occupational Risk Factors for Selected Cancers among African-American and White Men in the United States. *American Journal of Public Health*, 93, 1748 - 1752.
- BROSSETTE, S. E., SPRAGUE, A. P., HARDIN, J. M., WAITES, K. B., JONES, W. T. & MOSER, S. A. (1998) Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, 5, 373 - 381.
- BRODY, J. G. & RUDEL, R. A. (2003) Environmental Pollutants and Breast Cancer. *Environmental Health Perspectives*, 111, 1007 - 1019.
- CLASSIFICATION BASED ASSOCIATION. <http://www.comp.nus.edu.sg/~dm2>. Last accessed May 23rd 2005.
- CUTHE, W. G., TUCKER, R. K., MURPHY, E. A., ENGLAND, R., STEVENSON, E. & LUCKARDT, J. C. (1992) Reassessment of Lead Exposure in New Jersey using GIS Technology. *Environmental Research*, 59, 318 - 325.
- DEVESA, S. S., GRAUMAN, D. J., BLOT, W. J. & FRAUMENI, J. F. (1999) Cancer Surveillance Series: Changing Geographic Patterns of Lung Cancer Mortality in the United States, 1950 Through 1994. *Journal of the National Cancer Institute*, 91, 1040 - 1050.
- ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. <http://www.esri.com>. Last accessed May 23rd 2005.
- ESTER, M., KRIEGEL, H.-P. & SANDER, J. (2001) Algorithms and Applications of Spatial Data Mining. IN MILLER, H. J. & HAN, J. (Eds.) *Geographic Data Mining and Knowledge Discovery*. New York, USA, Taylor and Francis.
- FAYYAD, U. M., PIATETSKY-SHAPIRO, G. & SMYTH, P. (1996) From Data Mining to Knowledge Discovery: An Overview. IN FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P. & UTHURUSAMY, R. (Eds.) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California, USA, AAAI Press.
- FRAWLEY, W., PIATETSKY-SHAPIRO, G. & MATHEUS, C. (1992) Knowledge Discovery in Databases: An Overview. *AI Magazine*, 14, 57 - 70.

- GREGORIO, D. I., DECHELLO, L. M., SAMOCIUK, H. & KULLDORFF, M. (2005) Lumping or splitting: seeking the preferred areal unit for health geography studies. *International Journal of Health Geographics*, 4.
- GEE, G. C. & PAYNE-STURGES, D. C. (2004) Environmental Health Disparities: A Framework Integrating Psychosocial and Environmental Concepts. *Environmental Health Perspectives*, 112, 1645 - 1653.
- HAN, J. & KAMBER, M. (2001) *Data Mining Concepts and Techniques*, San Diego, USA, Academic Press.
- HIPP, J., GUNTZER, U. & NAKHAEIZADEH, G. (2000) Algorithms for Association Rule Mining - A General Survey and Comparison. *SIGKDD Explorations*, 2, 58 - 64.
- JACQUEZ, G. M. & GREILING, D. A. (2003a) Geographic Boundaries in Breast, Lung and Colorectal Cancers in Relation to Exposure to Air Toxics in Long Island, New York. *International Journal of Health Geographics*, 2.
- JACQUEZ, G. M. & GREILING, D. A. (2003b) Local Clustering in Breast, Lung and Colorectal Cancer in Long Island, New York. *International Journal of Health Geographics*, 2.
- KLÖSGEN, W. & MAY, M. (2002) Spatial Subgroup Mining. 6th European Symposium on Principles of Knowledge Discovery in Databases. Berlin, Germany.
- KOPERSKI, K. & HAN, J. (1995) Discovery of Spatial Association Rules in Geographic Information Databases. IN EGENHOFER, M. J. & HERRING, J. R. (Eds.) *Advances in Spatial Databases*. New York, USA, Springer-Verlag.
- LAM, N. S.-N. (1986) Geographic Patterns of Cancer Mortality in China. *Social Science and Medicine*, 23, 241 - 147.
- LAM, N. S.-N., FAN, M. & LIU, K. B. (1996) Spatial-temporal spread of the AIDS epidemic: A correlogram analysis of four regions of the United States. *Geographical Analysis*, 28, 93 - 107.
- LIU, B., HSU, W., MA, Y. & S.CHEN (1999) Mining Interesting Knowledge using DM-II. *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, California, United States, ACM Press, New York, USA.
- MALERBA, D. & LISI, F. A. (2001) Discovering Associations between Spatial Objects: An ILP Application. IN ROUVEIROL, C. & SEBAG, M. (Eds.) *Inductive Logic Programming*. Berlin, Germany, Springer - Verlag.
- MALERBA, D., ESPOSITO, F., LISI, A. F. & APPICE, A. (2002) Mining Spatial Association Rules in Census Data. *International Journal for Research in Official Statistics*, 5, 19 - 44.
- MCLAFFERTY, S. L. (2003) GIS and Health Care. *Annual Review of Public Health*, 24, 25 - 42.
- MEADE, M. S. & EARICKSON, R. J. (2000) *Medical Geography*, New York, USA, The Guilford Press.
- MENNIS, J. & LIU, J. W. (2005) Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. *Transactions in GIS*, 9, 5 - 17.
- MOONAN, P. K., BAYONA, M., QUITUGUA, T. N., OPPONG, J., DUNBAR, D., JOST, K. C., BURGESS, G., SINGH, K. P. & WEIS, S. E. (2004) Using GIS Technology to Identify Areas of Tuberculosis Transmission and Incidence. *International Journal of Health Geographics*, 3.

- MORRISON, H., SAVITZ, D., SEMENCIW, R., HULKA, B., MAO, Y., MORISON, D. & WIGLE, D. (1993) Farming and prostate cancer mortality. *American Journal of Epidemiology*, 137, 270-80.
- NATIONAL ATLAS OF THE UNITED STATES OF AMERICA. <http://nationalatlas.gov>. Last accessed May 23rd 2005.
- NKHOMA, E. T., HSU, C. E., HUNT, V. I. & HARRIS, A. M. (2004) Detecting Spatiotemporal Clusters of Accidental Poisoning Mortality Among Texas Counties., 1980 - 2001. *International Journal of Health Geographics*, 3.
- ORDONEZ, C., OMIECINSKI, E., BRAAL, L. D., SANTANA, C. A., EZQUERRA, N., TAABOADA, J. A., COOKE, D., KRAWCZYNSKA, E. & GARCIA, E. V. (2001) Mining Constrained Association Rules to Predict Heart Disease. *Proceedings of the IEEE International Conference on Data Mining*, 433 - 440.
- ORDONEZ, C., SANTANA, C. A. & BRAAL, L. D. (2000) Discovering Interesting Association Rules in Medical Data. *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 78 - 85.
- PARKER, S. L., DAVIS, K. J., WINGO, P. A., RIES, L. A. & HEATH, C. W. (1998) Cancer Statistics by Race and Ethnicity. *CA: A Cancer Journal for Clinicians*, 48, 31 - 48.
- PICKLE, L. W., MUNGIOLE, M., JONES, G. K. & WHITE, A. A. (1996) Atlas of the United States Mortality. IN SERVICES, D. O. H. A. H. (Ed.), National Center for Health Statistics.
- PICKLE, L. W., WALLER, L. A. & LAWSON, A. B. (2005) Current Practices in Cancer Spatial Data Analysis: A Call for Guidance. *International Journal of Health Geographics*, 4.
- REISS-STARR, C. A., WEINRICH, S. P., CREANGA, D. & WEINRICH, M. (1998) The Association of Family History and Participation in Free Prostate Cancer Screening. *American Journal of Health Studies*, 14, 95 - 105.
- RUIZ, M. O., C.TEDESCO, MCTIGHE, T. J., AUSTIN, C. & KITRON, U. (2004) Environmental and Social Determinants of Human Risk During a West Nile Virus Outbreak in the Greater Chicago Area, 2002. *International Journal of Health Geographics*, 3.
- SINGH, G. K., MILLER, B. A. & HANKEY, B. F. (2002a) Changing Area Socioeconomic Patterns in U.S. Cancer Mortality, 1950 - 1998: Part II - Lung and Colorectal Cancers. *Journal of the National Cancer Institute*, 94, 916 - 925.
- SINGH, G. K., MILLER, B. A., HANKEY, B. F., FEUER, E. J. & PICKLE, L. W. (2002b) Changing Area Socioeconomic Patterns in U.S. Cancer Mortality, 1950 - 1998: Part I - All Cancers Among Men. *Journal of the National Cancer Institute*, 94, 904 - 915.
- SINGH, G. K. & SIAHPUSH, M. (2002) Increasing Inequalities in All-Cause and Cardiovascular Mortality among US Adults aged 25 - 64 years by Area Socioeconomic Status, 1969 - 1998. *International Journal of Epidemiology*, 31, 600 - 613.
- TEPPO, L. (1998) Problems and Possibilities in the Use of Cancer Data by GIS - Experience in Finland. IN GATRELL, A. C. & LOYTONEN, M. (Eds.) *GIS and Health*. Philadelphia, Taylor and Francis.
- TURREL, G. & MATHERS, C. (2001) Socioeconomic Inequalities in All-Cause and Specific-Cause Mortality in Australia: 1985-1987 and 1995-1997. *International Journal of Epidemiology*, 20, 231 - 239.
- UNITED STATES CENSUS BUREAU. <http://www.census.gov>. Last accessed May 23rd 2005.

UNITED STATES BUREAU OF ECONOMIC ANALYSIS. <http://www.bea.gov>. Last accessed May 23rd 2005.

WAGENER, D. K. & SCHATZKIN, A. (1994) Temporal trends in the socioeconomic gradient for breast cancer mortality among US women. *American Journal of Public Health*, 84, 1003-1006.

WEINRICH, S., WALLER, J., P. GREENWALD, WEINRICH, M. & ARONSON, K. (1999) Occupational Exposures and Abnormal Prostate Cancer Screening Results in Black and White Men. *American Journal of Health Studies*, 15, 113 - 120.