INTEGRATING GIS AND SPATIAL DATA MINING TECHNIQUE FOR TARGET MARKETING OF UNIVERSITY COURSES

Hong Tang
School of Environmental and Information Science
Charles Sturt University
P O Box 789, Albury, NSW 2640 Australia
Tel: 61 2 6051 9617
Fax: 61 2 6051 9897
htang@csu.edu.au

Simon McDonald
Spatial Data Analysis Network
Charles Sturt University
P O Box 789, Albury, NSW 2640 Australia
Tel: 61 2 6051 9922
Fax: 61 2 6051 9919
smcdonald@csu.edu.au

ABSTRACT

Student populations tend to be located in particular spatial areas and within specific demographic settings. Spatial and non-spatial data such as the university's admission records, census data, transport network data, and university campus location data can be used to describe, explain, and predict university student admission patterns, thus provide information to assist the university with improving its strategy in course marketing. Standard database and statistical methods do not work well with interrelated spatial data. Our study attempts to use Geographic Information System (GIS), spatial statistics, and spatial data mining techniques to explore the associations between the students of a specified course and their demographic characteristics (such as accessibility and proximity to university campus, ethnic background, and socio-economic status). A method of integrating experts' knowledge for mining multi-level association rules iteratively between spatial and non-spatial data is proposed to identify the pattern and predict the spatial trend of student admission, and hence, the potential market areas of students. Historical data are used to evaluate the results and validate the method. This paper also discusses the limitations of the adopted approach and the directions for future study.

Keywords: Geographic Information System (GIS), university admission, spatial data mining, spatial statistics, potential market areas.

1. Introduction

Tertiary education has evolved into a very complex and competitive environment. As a relatively young regional university, Charles Sturt University (CSU) is recruiting students primarily from regional Australia, particularly from regional Victoria and New South Wales. The success of CSU depends more on the quantity, quality, and diversity of the students than other universities in Australia. Therefore, it is critical for CSU to explore more potential student markets.

Many efforts have been taken to develop a suitable student recruitment and marketing strategy to maximise university marketing resources. Blackmore and Yang (2000) used the collections and statistical reports of student application and admission data to classify and visualise students' preferences for two courses in CSU. Their study revealed student's course preference pattern by categorising their preferences into home-based, campus-based, course-based, and random selection. The flow of students from different regions was presented. Drawbacks to this approach and other similar methods, include the highly abstractive and ambiguous nature of such categorization.

Marble, et al. (1996, 1997) applied a series of filters such as demographic, geographic, and institutional to identify the areas which have higher population of potential students, thus the target marketing zone. Their study revealed a substantial

amount of stability in the results from filtering over several years. Nevertheless, the use of such filters is subjective. It also lacks the consideration of other factors that may influence a student's' choice. Furthermore, such filters do not consider the characteristics of particular courses.

Many companies offer target marketing services to businesses. Presently, there are a number of commercial tools for academic marketing. These tools utilize statistical methodologies and Geo-demographic analysis techniques and adopt the filter approach. TSA (Target Statistical Analysis from Target Marketing Inc, 2001) profiles current enrollment and rates prospective students accordingly, therefore, prioritizing the inquiry pool. It identifies only the potential students who have already inquired about the courses hence it does not provide a pro-active means of marketing.

Today's university course marketers need a clear indication of student admission patterns for a special course and the prediction of future trends of student distribution. This indication and prediction should come from two aspects: marketer's knowledge of admission matters; and knowledge mined from the database. Course marketers believe (1) there is a pattern for student enrollment, and (2) that the same socio-economic characteristics have relatively different degrees of significance and play different roles in different courses even in the same campus of the same university.

The Bachelor degree of Applied Science (Park Recreation and Heritage (PR&H)) which is offered from Albury/Woodonga campus of Charles Sturt University (Australia), is a unique program in its field. It offers a very flexible admission program which includes internal, distance, full time, part time, and associate modes of study. It is chosen as a study case for this analysis due to its decrease of enrollment numbers in recent years. The main objective of this study is to provide the university with additional student marketing information thereby assisting in an improved course marketing strategy.

Focused on methodology, this study aims to use Geographic Information System (GIS), Spatial Statistics, and Spatial Data Mining techniques to mine knowledge from various data, and integrate marketer's expert knowledge to analyze admission patterns and predict potential market areas.

2. GIS and spatial data mining

Increasing availability of large datasets from different agents creates the necessity of knowledge / information discovery from data, which leads to an emerging field of data mining or knowledge discovery in databases (KDD) (Fayyad, et al., 1996). Data mining involves the fields of database systems, data visualization, statistics, machine learning, and information theory (Koperski, et al., 1996). It is an exploratory process aimed at discovering hidden features in the database, testing the hypothesis and building the model.

2.1 Spatial Data Mining

Recent widespread use of spatial databases has lead to the studies of Spatial Data Mining (SDM), Spatial Knowledge Discovery (SKD), and the development of spatial data mining techniques. Traditional data mining methods assume independence among studied objects and lack the ability to handle the inter-relational nature of spatial data. Spatial data mining methods can be used to understand spatial data, discover relationships between spatial and non-spatial variables, detect the spatial distribution patterns of certain phenomena, and predict the trend of such patterns. Foundations of spatial data mining include spatial statistics and data mining.

Spatial data mining tasks can be grouped into description, exploration, and prediction. To understand the data, spatial data and spatial phenomena have to be first described and analyzed; and hidden patterns and relationships among spatial or non-spatial variables have to be explored. Based on the current pattern of spatial distribution and the understanding of spatial relationships, future state and trend of the spatial pattern and spatial distribution can be predicted.

Spatial data mining techniques include, but are not limited to, visual interpretation and analysis, spatial and attribute query and selection, characterization, generalization and classification, detection of spatial and non-spatial association rules, clustering analysis, and spatial regression.

2.2 The Integration of GIS And Spatial Data Mining

Data related to students' admission includes all feature types of point, line and polygon, such as the locations of campuses, towns, and cities, extent of the road networks, and subdivided areas. It also contains a rich set of attribute data such as

students' background and other social economic dataset. The unique capacity of spatial data handling makes GIS the ideal database tool for such data. It provides functions to store, manipulate, analyze, and display spatial data.

GIS has a long history of being used as a tool to visualize spatial, attribute, and statistical data. Such uses of GIS include choropleth mapping, dasymetric mapping, and trend surface analysis. The availability of functions such as spatial and non-spatial query and selection, classification, map overlay, network analysis, and map creation, make GIS a useful tool for spatial data mining. Visualization through GIS gives the user the ability to spot spatial errors that can be otherwise unnoticed by analyzing raw data, and to aid visual analysis and detection of certain features' distribution and their patterns.

Even though, by using GIS, students' locations can be plotted and 'hot areas', which have a majority of student sources, can be spotted, GIS alone does not have the capacity of finding the relationship between such 'hot areas' and their spatial and attribute characteristics. Therefore, the areas which have similar characteristics but as yet, have no enrolled students, can not be identified.

3. Exploring spatial associations

To identify the potential student markets from the current 'hot areas', we need to find out the reasons for forming such 'hot areas', that is, finding the underlying common characteristics of the students and their areas, and their association to admission. Data related to student's admission are voluminous. They contain a range of different variables. To find out such associations from the data, academic marketers' expert knowledge needs to be integrated with spatial data mining techniques.

3.1 Association Rules And Minimum Confidence Threshold

Spatial and non-spatial association rules are in the form of X-->Y(c%), where X and Y are sets of spatial or non-spatial predicates and c% is the confidence of the rule which indicates the strength of the association. For example:

- Is_a (X, origin of CSU PR&H student) --> close to (Y, railway stations) (70%), this rule states that 70% of CSU PR&H students enrol from a location close to a highway. In this case, distance threshold has to be specified for the 'close to' predicate.

- Is_a (X, CSU PR&H student) --> from (Y, middle family income areas) (80%), this rule states that 80% of CSU PR&H students are from areas which have a middle level of family income. A set of income thresholds can be specified for the 'middle income'.

For a large database where there are a large set of variables, there may exist a large number of associations between them (Koperski, 1998). Some rules may have a very low level of confidence (c%), for example, less than 5% of CSU's PR&H students are associated with a disability code, therefore it is not significant and may not be of interest for further study. Conversely, 75% of students lived within 10 km of railway stations, therefore this rule is significant and attracted further examination. Expert knowledge is applied here to define the minimum confidence threshold for each rule and filter out the uninterested associations.

3.2 Methods Of Mining Multi-level Association

Han, et al. (1999) suggest that for many applications there is the need of mining association rules in multiple levels of data abstraction. For example, after we analyze the association between students and proximity to highway (within distance of 10 km), we may need to find the association between students and proximity to an interstate highway, or in particular, within 5 km of an interstate highway. In the case of family income level, we may need to analyze the association of students with middle income areas first then more specified ranges of incomes (eg. $500-$699 weekly, $700-$999 weekly, etc). For each variable, different thresholds are defined which result in the different levels of data abstraction. More specific and concrete associations can then be found. Expert knowledge is applied here for defining spatial hierarchy and setting thresholds in different level for both spatial and non-spatial data abstractions.

Mining association is a well-studied area in the field of Data Mining. There are many methods to explore the multi-level association rules after the classification. Han, et al. (1999) suggest a method of applying different thresholds and different

minimum confidence thresholds for mining associations of different levels of abstraction. Basic functions of common GIS such as spatial generalization and attribute classification can be used to arrange data into different levels of abstraction.

3.3 Mining Specific Associations From Lower Levels Of Abstraction

Our study is aiming to find out relatively low abstraction level's association between students' admission and their spatial and non-spatial characteristics, e.g. students and distance to railway stations < 5 km instead of distance to railway stations < 20 km, or students with prior TAFE qualification instead of prior qualifications. The lower levels of abstractions can reveal more specific associations and are therefore more informative and useful.

An assumption of mining associations for lower levels of abstraction using this methodology is that, we only need to study the rule in sub-classes of data if a strong association is found for their ascendants. For example, if the association between student enrollments and the distance to railway stations < 20 km is less than a defined minimum confidence (thus is, not significant), then we do not need to explore further association between students and distance to stations < 10 km or less.

To effectively find out the association in lower levels of data abstraction, we need to set a high minimum confidence threshold at higher levels of abstraction and progressively reduce it while sequencing to lower levels of abstraction. By doing this, we can only filter out the uninteresting variables in each level and will not miss out the useful information. In this process, it is important for users to apply their prior knowledge to define and control the thresholds at different levels.

This process will identify relationships between student admission and their spatial and demographic characters at a defined lowest abstraction level of each variable. The potential student market areas or the potential 'hot areas' can then be defined. These are the areas where all the significant rules are present.

GIS does not have build-in functions and algorithms to effectively perform the statistical tasks and mine the associations in multiple levels. Haining, et al (1996) reviewed different means by which extra analytical functions could be added to GIS. Anselin and Bao (1997) demonstrated an ArcView extension for the visualization of results from SpaceStat.

Using GIS scripting language Avenue (ESRI, 2000), for example, an algorithm and procedure can be integrated into GIS. Alternatively, in our case, the ArcView extension of S-Plus for ArcView GIS (MathSoft, 1998) can be used to take advantage of its powerful statistical functions. A set of S-Plus commands can be run from an Avenue script which carries out query, selection, buffering, generalization, and classification tasks.

Apart from the various technical considerations, it is more important to have continuous interaction between the data miner (researchers) and subject matter experts (university marketers in this case) throughout the processes, from database selection, data exploration, variable selection, modeling and prediction, to model testing and evaluation.

4. Method of predicting potential market areas

There are massive amounts of data stored in databases related to student's admission. Data useful to this study include student's admission records, census data, road network data, and many others. Data comes from many sources in different formats. After the process of selection, cleaning and transformation, our task is to mine valuable information from such data and use that information to support marketer's decision-making. Following datasets are used in this study:

- 1999, 2000, and 2001's CSU student admission data (with information such as date of birth, home location, country of birth, prior education, etc);
- 1996 Census data;
- reduced output spatial datasets (location of highways, main road and other classes of road networks, and location of railway stations);
- location of other university campuses;
- CSU 1999 and 2000 University Admission Centre (UAC)'s course analysis and statistical reports;
- Other datasets including White pages and The Australian Surveying and Land Information Group (AUSLIG) - both for location reference.

4.1 Exploratory Data Analysis

All PR&H students' home locations were plotted and the percentage of students from each state were calculated. As CSU's major campuses are located in NSW (with one on the NSW/VIC border), for each year from 1999 - 2001, over 85% of students are from these two states, therefore we focus our further analysis in NSW and VIC. Following figures (1a, 1b, and 1c) illustrate the distribution of 1999's PR&H student sources. The data was mapped over three different spatial units, postcode area, local government area, and census district for comparison.

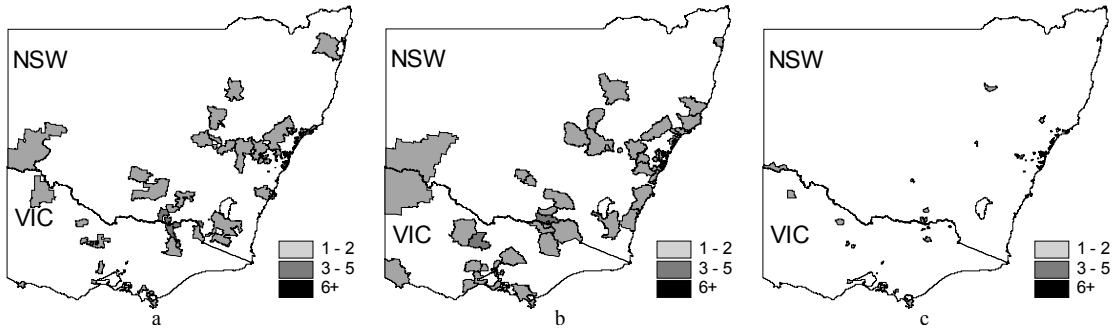

Figure 1 Numbers of student in each postcode area (1a), local government area (1b) and census district (1c) (1999 enrolment data)

Only a limited number of variables that are critical to student's admission are considered here for each year due to the nature of this study and the availability of the resource. Those selected variables were classified to study the relationships between students' admission and their accessibility to transports (public and private transport), their ethnic backgrounds, and the area's socio-economic status (represented by average family income and education level). Selecting, adding and removing variables can be a challenging process.

Figure 2 shows the method used to classify relevant attribute items based on different levels of abstraction and different thresholds. The process for each variable is iterative. The number of iteration depends on the minimum confidence threshold defined by the user and the actual degree of association mined from the data. Each further iteration provides more specific information. Algorithms can be developed by using Avenue scripts to carry out the iterative process.

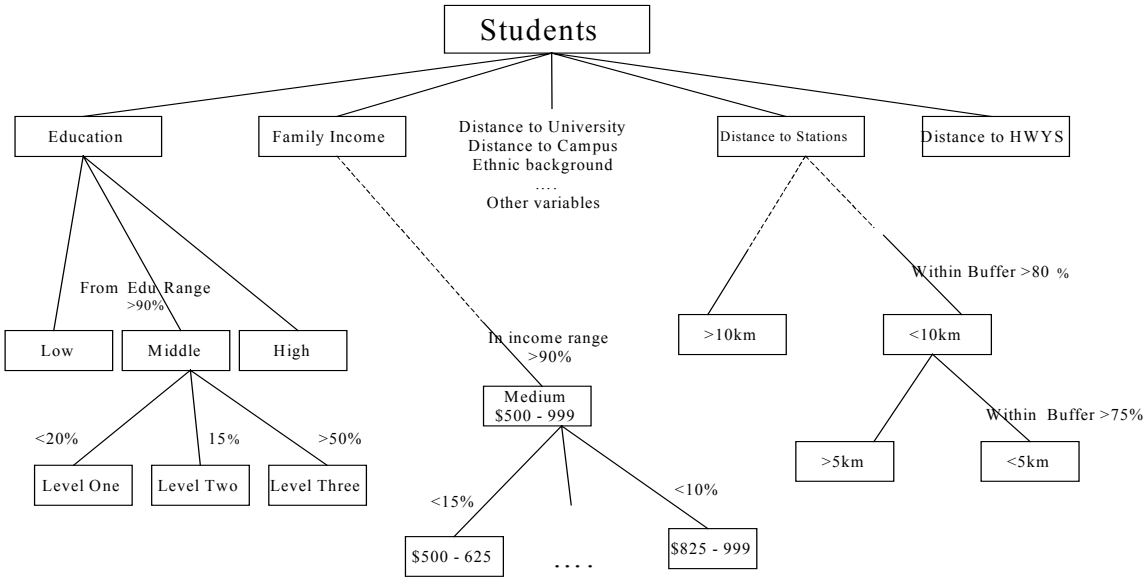

Figure 2 Generalization and classification of relevant data items into different levels

4.1.1. Accessibilities

An example of simple measurement of student's accessibility to public and private transport would be their distance to railway stations, distance to highway, and distance to campus via road network.

While nearly 100% of 1999's PR&H students live less than 50 km from a station and a highway, our analyses classified the closeness to the station and highway into different levels (eg. 10 km and 5 km from the stations or highways) and studied their significance to student's enrollments. Figure 3a and 3b demonstrate the results of different levels of classification for the distance to highways. The same processes can be carried out for other accessibility factors (such as distance to railway stations), thus cross level classifications of different variables (for accessibility) can be combined together to study their association with student enrollments. Our study indicated that the majority of 1999's PR&H students have easy access to public and private transports (e. g., within 5 km to the station and 5 km to highway).

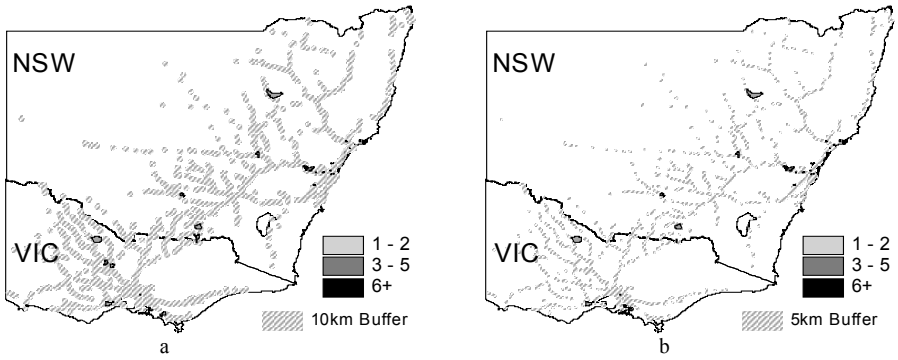

Figure 3 Over 80% of students live within 10 km of a station (higher level of abstraction, 3a). Over 75% of students live within 5 km of a station (lower level of classification, 3b).
(Both maps and figures are derived from 1999 students' admission data.)

Distance via road networks (shortest path) to the particular campus (Thurgoona, Albury NSW) had an impact on enrollments, but is less significant than expected.

4.1.2 Socio-economic status

Student home location's socio-economic status (chosen variables are average weekly family income and education level) tended to be unique. Over 90% of 1999' PR&H students came from an area where the average education level was classified as 'middle' - with a majority of people having either an associate diploma or having completed some level of undergraduate education.

The enrollment pattern also had a strong association with average family weekly income. Nearly 90% of 1999's PR&H students are from the middle income area (with weekly income of AUS$500-999). The results from the above studies can be seen as the collateral evidence for the previous knowledge from the university's marketing officers.

4.2 Defining The Potential Markets

A set of characteristics which are strongly associated with 1999's PR&H student's enrollments can be identified. When areas, where all these characteristics are present, are plotted, that is, maps overlaid and the areas which have these characteristics intersected, the potential market for 2000 can be defined. This potential market is the collection of postcode areas, local government areas or census districts. Figure 4 illustrates the potential market areas of 2000 defined by using 1999's student enrollment data in the unit of census district.

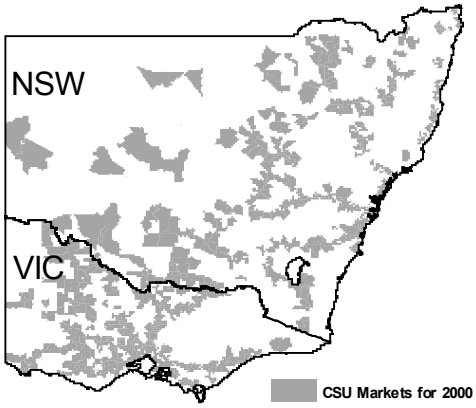

Figure 4 2000's market areas for 2000 predicted by using 1999's enrollment data.

4.3 Evaluations

Historical data can be used to check the stability of the prediction result, and consequently the employed methodology. Year 2000's actual student home locations were plotted against the potential market defined by using the previous year data. Approximately 50% of students were found to be within the predicted areas. When 2001's actual student home locations were plotted against the potential market defined by 2000's, 56% of students were found to be within the predicted areas.

Figure 5a and 5b show the actual student admission data plotted against the potential market areas defined by using the previous year's data. Based on these figures, we can say that the predictions are close to the reality and they provide a reasonably accurate indication of the potential student sources. Our methodology has therefore been proven to be able to predict a considerable proportion of student enrollments.

Further to these results, it was discovered that around 50% of PR&H students did not come from the areas that have had enrolled students from the previous year and yet are still predicted by the potential market. This indicates that the origins of students are dynamic however, using our methodology, they can still be predicted. This methodology has the potential to be used to study the dynamic nature of student enrollment and to predict future trends.

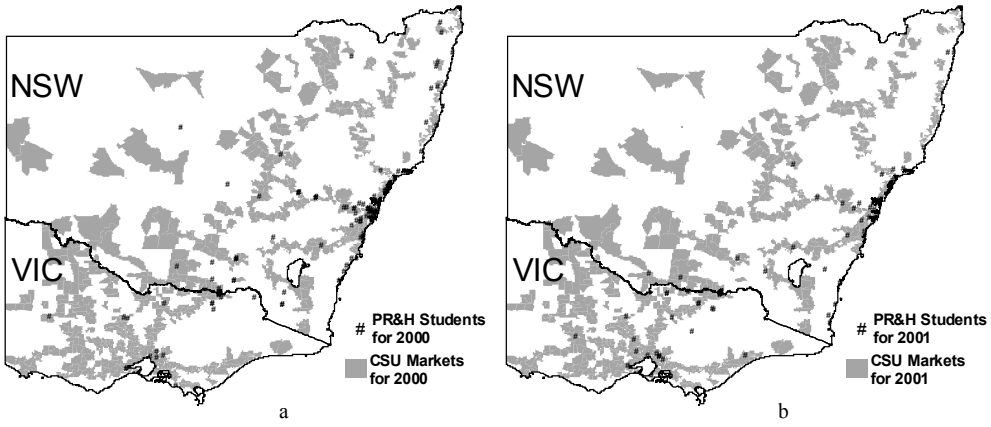

Figure 5a Actual admission data 2000 Vs 'potential market areas' defined previous year
Figure 5b Actual admission data 2001 Vs 'potential market areas' defined previous year

5. Conclusions, limitations and future directions

Sources of students for a specified course in the university tend to be located in particular spatial areas with certain demographical settings. Different courses in the university may have different enrollment patterns. When the association between admission and students' spatial and non-spatial characteristics are analyzed, the potential student market areas can be defined. This process involves a series of complex iteration involving data miners and subject matter experts. This study proposed a methodology and mechanism that integrate data mining techniques and expert's knowledge of student

admission. It can be used to: 1) mine such associations in different levels of data abstraction; 2) define potential areas of incoming students; and 3) predict the trend of admission pattern. Case study results showed that reasonably accurate potential areas can be defined by using the proposed methodology.

While our findings explain the enrollment pattern for CSU's Bachelor of Applied Science (Park Recreation and Heritage) program and in large extents also provide the methodology which can define the potential student market with reasonable accuracy, this research suffers from some limitations thus leads to the following research directions:

1) Many rural regions/town areas are very large and the locations of population centers are unlikely to correspond to the regions' centroids. When numbers of students are assigned to each town as points and further expanded to the area, there will have some degrees of mismatch and misinterpretation. For example, students or the general population are most likely to be clustered around town centres instead of spreading evenly through out the whole region.

2) Using postcode areas and local government areas as the spatial analysis unit causes boundary problems. Areas are different in size and in some case it is discontiguous. Postcode areas also do not respect administrative boundaries or other descriptive geography such as urban/rural splits and contain a substantial amount of internal heterogeneity with respect to many socio-economic characteristics (Marble, et al. 1997).

There are many advantages of using census districts as our smallest spatial analysis unit. In each district, homogeneity among the households can be assumed. Nevertheless, the use of such a spatial unit has difficulties in the visualization of data and implementation of analysis results. To overcome these problems, data need to be aggregated to a more suitable spatial unit such as University catchment area. The choice of spatial analysis unit can have important impact in the research. Significantly different spatial patterns may be emerged as the unit of spatial analysis is shifted. Defining spatial analysis units becomes the critical research task of its own.

3) Differences in aggregation and categorization of social data, on the other hand, can cause different inferences concerning a student's socio-economic status. In our case, integers were assigned to different qualification levels and in each census district those figures are aggregated, averaged, and reclassified. The result in some cases may not represent the reality of those areas.

4) The reasons for a student to select a particular course are very complex. Selection of variables and the study of their relationship determine the outcome of this research. There are already some known factors which we have not yet considered due to the unavailability of the dataset. These factors may include: the competition of similar course in other universities, the competition amount the different campuses in same university, region's proximity to other universities and campuses, and the region's employment situation, economic growth.

5) We need also to consider the differences between distance education students and internal students in the same course. Attempting to develop one model for these two quite different cohort types may be less than satisfactory. More accurate predictions may be yielded by separating these types of enrollment.

6) A student survey is necessary for such a study to confirm the result of data mining. A carefully structured questionnaire can provide invaluable information and can be used as ground truth for checking the analysis results. The newly available New Student Survey which includes student profile and other useful student demographic data from CSU Division of Marketing and Communications may be utilized in the future research.

7) It is our aim to create a tool to assist the non-GIS and statistics specialists in handling and analysing spatial data. A range of graphical and numerical facilities will need to be carefully designed and integrated to enhance the application of their special knowledge. An algorithm needs to be interfaced with GIS in an user-friendly and menu-driven manner for marketers to: 1) apply their knowledge to select as many variables for analysis as appropriate, 2) define spatial and attribute hierarchy, and 3) define and control thresholds in different levels of abstraction for different characteristic.

REFERENCES

Anselin L and Bao S (1997) Exploratory Spatial Data Analysis Linking SpaceStat and ArcView. In M. Fischer and A.Getis (eds) Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling And Neuro-computing, Berlin: Springer-Verlag, pp 35-59.

Blackmore K and Yang X (2000) GIS in University Admissions: Analysis and Visualization of Students Flows. Proceedings of 28th Annual Conference of AURISA, November 2000, Coolum, QLD, Australia, 20 - 24.

Haining R, Wise S and Ma J (1996) The Design of a Software System for the Interactive Spatial Statistical Analysis Linked to a GIS. Computational Statistics**,** (11), 449-466.

Han J and Fu Y (1999) Mining Multiple-Level Association Rules from Large Databases. IEEE Transactions on Knowledge and Data Engineering, 11(5), 798-805.

Han J and Kamber M (2000) Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers.

Koperski K, Adhikary J and Han, J. (1996) Spatial Data Minding: Progress and Challenges - Survey Paper. Proceedings of Workshop on Research Issues on Data Mining and Knowledge Discovery, June 1996, Montreal, QB, 55-70.

Koperski K, Han J and Adhikary J (1998) Mining Knowledge in Geographic Data. Communications of the ACM, 41(12), 47-66.

Marble D, Mora V and Herries J (1995) Applying GIS Technology to the Freshman Admissions Process at a Large University. Proceeding of ERSI User Conference, 22-25 May, 1995, Palm Springs, CA.

Marble D, Mora V and Cranados M (1997) Applying GIS Technology and Geodemographics to College and University Admissions Planning: Some Results From The Ohio State University. Proceeding of ERSI User Conference, 8-11 July,1997, San Diego, CA.

Stevenson S, Maclachlan M and Karmel T (1999) Regional Participation in Higher Education and Distribution of Higher Education Resources Across Regions. Occasional Paper Series, No 99-B, Higher Education Division, Department of Education, Training and Youth Affairs, Canberra, Australia.

Stevenson S, Evans C, Maclachlan M, Karmel T and Blakers R (2000) Access - Effect of Campus Proximity and Socio-economic Status on University Participation Rates in Regions. Occasional Paper Series, No 00-D, Higher Education Division, Department of Education, Training and Youth Affairs, Canberra, Australia.

Target Marketing Inc (2001) A Proven Tool for Strategic Academic Marketing. http://targetusa.com /educational/about_edu/about_tsa.html. [Last access 29 July 2001]