Spatial Data Mining and University Courses Marketing

Hong Tang
School of Environmental and Information Science
Charles Sturt University
htang@csu.edu.au

Simon McDonald
Spatial Data Analysis Network
Charles Sturt University
smcdonald@csu.edu.au

Abstract. Movement of university admission is not random. Student admission data can be used to define the likely source of students and movement of the source. Thus, it can help the university improve its courses marketing strategy. Standard database and statistical methods do not work well with interrelated spatial data. The ongoing research presented in this paper attempts to use Geographic Information System (GIS), spatial statistical techniques, and spatial data mining to explore the relationships between the source of students (such as area), its spatial components (such as connectivity and distance), and the attribute data (such as family income and education backgrounds). Multiple level spatial classifications and association rules are adopted to identify the patterns and to predict the trend of incoming student source. The preliminary results of this research confirm that sources of students tend to be located in the particular spatial areas and with certain geographical/non-geographical settings. This paper also discusses the limitations of the adopted approach and the directions for future research.

Keywords: Geographic Information System (GIS), spatial data mining, spatial statistics, trend prediction.

1. INTRODUCTION

The tertiary education market has become very competitive. There is a battle for universities to gain student market share. As a relatively young regional university, the success of Charles Sturt University (CSU) depends more on the quantity, quality and diversity of the student source then other universities in Australia.

Many efforts have been taken to develop a suitable student recruitment marketing strategy for the better use of the limited university marketing resources. Among those are the collection of the student admission and enrollment data and the classification and visualization of students' preference categories (Blackmore and Yang, 2000). Blackmore and Yang (2000) use two courses from CSU as the study case. Their study reveals the insight of student's course preference pattern by categorising their preferences into home-based, campus-based, course-based and random selection. The "hot spot" of each student's incoming source are identified. Drawbacks of this approach, among others, are the highly abstractive and ambiguous nature of such categorization.

Marble et al. (1996, 1997) apply a series of filters, in particular, demographic filter, geographic filter and institutional filters to identify the areas of which have higher population of potential students, thus the target marketing zone. Their study finds a substantial amount of stability in each different year. Nevertheless the use of such filters is subjective and displays a lack of consideration of many other factors which may influence the students' choice.

Many companies offer target marketing service to universities. A number of commercial tools for academic marketing are now existed. These tools utilise statistical methodologies and Geo-demographic analysis techniques. STA (Target Statistical Analysis from Target Marketing Inc, 2001) profiles current enrollment and rates prospective students accordingly therefore prioritising the inquiry pool. It identifies only the potential students who had already inquired about the courses. They also use MicroVision (Claritas Inc, 2001), a Geo-demographic tool, and attempt to pinpoint lifestyle clusters.

Today's university course markers need a reasonably clear indication of student admission patterns and the future trends of student source movement. This indication should come from two aspects: marketer's knowledge of admission matters and, knowledge mined from the database.

Stevenson et al. (2000) study the effect of campus proximity and socio-economic status on university participation rates in different regions of Australia. Our studies further analyse it by considering a particular course in a university rather than all courses in the university. Course marketers believe that same socio-economic characteristics may have relatively different degrees of significance and play different roles in the different courses even in the same campus in same university.

Focused on the methodology, this research is aiming to use Geographic Information System (GIS), Spatial Statistics and Spatial Data Mining techniques to study the students' admission data, analyse the admission pattern and predict the trend of "hot spot" movement. The CSU course chosen for this analysis is the Bachelor of Applied Science (Park Recreation and Heritage), due to its decreasing enrolment numbers in recent years. The main objective of this case study is to present those responsible for marketing with information they can use to improve their course marketing strategy.

2. GIS AND SPATIAL DATA MINING

Increasing availability of large datasets from different agents (in our case, large datasets from New South Wales University Admissions Center, Victorian Tertiary Admissions Center, CSU's Planning and Development Division and Communication and International Relations Division, Australian Bureau of Statistics and State Department of Infrastructure) 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., (eds) 1996). Data mining involves the fields of database systems, data visualization, statistics, machine learning and information theory (Koperski et al., 1996).

Traditional data mining methods assume independence among studied objects. These methods lack the ability to handle the inter-relational nature of spatial data.

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. Spatial data mining methods can be used to understand spatial data, discover relationships between spatial and non-spatial data, detect the distribution pattern of certain phenomena, predict the future movement of such patterns, etc. Foundations of spatial data mining include spatial statistics and data mining.

Han and Kamber (2000) group data mining tasks into two broad categories: descriptive data mining and predictive data mining. This categorization also fits the spatial data mining tasks. Descriptive spatial data mining describes spatial data and spatial phenomena. It also explores the relationships among spatial or non-spatial data, and identifies the pattern of spatial distribution. Predictive spatial data mining, on the other hand, based on the current state of spatial data, attempts to develop a model to predict the future state of the spatial data and forecasts the trend of pattern change.

Spatial data mining techniques include, but are not limited to, visual interpretation and analysis, characterization and classification, defining association rules, clustering, and spatial regression.

Koperski et al (1996, 1998) present an architecture for spatial data mining process and state that in every step of knowledge discovery, there are the need of user's control and background knowledge input.

2.2 GIS And Spatial Data Mining

Data related to students' admission involves all feature types of point, line and polygon, including the location of town and city, extent of the road network and subdivided area. It also contains a rich set of attribute data, including students' background and other social economic data. The unique capacity of spatial data handling makes GIS the ideal database tool for such data. It provides the functions to store, manipulate, analyse and display spatial data.

GIS has a long history of being used as a tool to visualise spatial and statistical data. Such uses of GIS include choropleth mapping, dasymetric mapping, phcophylactic interpolation, and trend surface analysis. The availability of functions such as classification, spatial and non-spatial query, network analysis and map creation make GIS a useful tool for spatial data mining. GIS can be also used to aid visual analysis of spatial data and detect the distribution of certain features and their patterns. Visualisation through a tool such as GIS gives the user ability to spot spatial errors that may otherwise have gone unnoticed by analysing raw data.

Spatial data are interrelated. Distribution of one feature is not only depended on its own characteristics, but also related to other features in some degree. To reveal the hidden patterns from large databases, it's desirable to identify the association rule between dependent and independent variables. GIS alone does not have algorithms to carry out such task.

3. EXPLORING SPATIAL ASSOCIATIONS

To identify the "hot spot" or the potential markets of incoming student sources and extend it to the future, we need to find out the reasons of forming such "hot spot", that is, the rules that associate students' admission with their own characteristics and other objects and features.

3.1 Association Rules

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. For examples:

- Is_a (X, origin of CSU student) --> close to (X, railway stations) (70%), this rule states that 60% of CSU students originally live close to highway.

- Is_a (X, CSU student) --> from (X, middle education level areas), (80%), this rule states that 80% of CSU students are from the area which have middle level of education.

For a large database where there are a large set of objects and attributes, there may exist a large number of associations between them (Koperski, 1998). Some rules may only apply to a small number of objects, for example, less than 5% of CSU's Park Recreation and Heritage students are associated with a disability code, therefore it may not be of interest for the further study. While 75% of students live within 10 km of railway stations, therefore it attracted further study. A minimum support threshold needs to be specified along with minimum confidence threshold to filter out the uninterested associations.

3.2 Methods Of Mining Multi-level Association Rules

Han et al. (1999) urge that for many applications there is the need of mining association rules in multiple levels of data abstraction. For example, after we find the CSU students who are close to highway (within 10 km of distance), we may need to find that Parks Recreation and Heritage students who are close to an interstate highway, or in particular, within 5 km of interstate highway. Or we may need to analyse the percentage of students from middle income areas and from the areas which have more specified ranges of incomes (eg. $500-$699 weekly, $700-$999 weekly, etc). Different levels of abstraction are applied here for both spatial and non-spatial data to find out more specific and concrete associations.

Mining multi-level association rules needs to have algorithms to abstract data into different levels and, to have methods to explore the associations within the same level or across the different levels.

Classification can be used to arrange data into different levels of abstraction. A basic function of common GIS is spatial or attribute data classification. Expert knowledge needs to be used to provide a meaningful yet useful grouping of data to fit the application, for example, classify students into different courses, different residential status, different distance to the university and to the main road etc.

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 the method of applying different minimum support thresholds and different minimum confidence thresholds for mining associations in different levels of abstraction. Knowledge of the studied matter needs to be applied here to define those thresholds in different levels for different associations.

3.3 Mining Specific Associations From Lower Levels Of Abstraction

Our study is aiming to find out relatively low abstraction level's association between our students and their spatial and non-spatial characteristics, e.g. students and distance to railway stations < 10 km instead of distance to railway stations <50 km, or students with prior TAFE qualification instead of prior qualifications. The lower levels of abstraction reveal more specific associations and are therefore more informative.

An assumption of mining associations into low level of abstraction is that, we need only to study the sub-classes of data if their ascendants are strongly associated, e.g. if association of students and distance to railway stations <50 km is less then defined minimum confidence threshold (thus is not significant), then we do not need to explore further the association between students and distance to stations < 10 km or less.

To effectively find out the association in a lower level of data abstraction, we need to set a high threshold at high abstraction level and progressively reducing it to the lower levels of abstraction. By doing this, we can only filter out the uninteresting object/character in each level and will not miss out the useful information. For this refinement process, it is important here for users to apply the prior knowledge to define and control the thresholds at different levels.

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

Using GIS scripting language, for example, Avenue (ESRI, 2000), a purpose built algorithm/process can be integrated into GIS based on its query and statistical functions to implement the algorithm and to automate the processes. 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. S-Plus commands can also run from Avenue script.

4. INTERPRETATION AND DISCUSSION OF RESULT

There are massive amounts of data stored in databases related to students, their course preferences and enrollment. Our task is to mine valuable information from such data. Those data come from different sources and are found in different formats. Table 1 lists the data which are used in our study.

4.1 Exploratory Data Analysis

After careful consultation with course marketing staffs, a set of variable were selected, which included students' home location and their background. Those data were then classified to study students' accessibility to transports, their proximity to CSU campus and other universities, their ethnic backgrounds and the area's socio-economic status (represented by average family income and education level). Figure 1 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.

Table 1: List of data sets and their attribute

Dataset	Brief description
CSU administration enrolment file	Data file contains students' information such as DOB, home location, country of birth, prior education, etc
Census data 1996	Contains various level's aggregated data in different spatial unit: census district, postcode area, location government area, etc
Reduced output spatial datasets	Contains data related to high ways, main road and other classes of road networks, location of railway stations
Location of other university campus	Contains location of other universities' campus
CSU 1999, 2000 UAC course analysis	Contains student's course preference analysis, school leaver statistic, competitor analysis
Other data	White pages, AUSLIG (both for location reference)

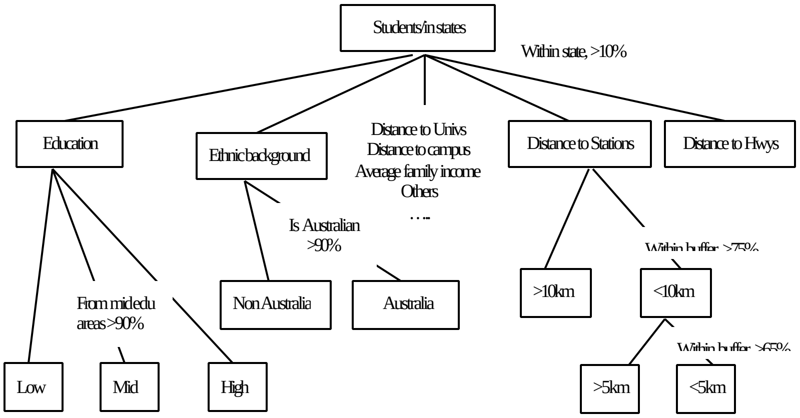

Figure 1: Classification of the relevant data items

4.2 The Iterative Process

Following is the suede code for defining the algorithm and carrying out the process.

```
Open enrolment Dataset
Match coordinates to each student's records
% breakdown based on variable 1 (within states)
        set user defined %
        if greater then  %  extract
        calculate summary statistics
        map over postcode area and other spatial unit
For extracted dataset
        open dataset (variable 2)
        set user defined distance and %
        buffer dataset
        extract records
        calculate summary statistics
        map over different spatial units
        accept/reject extract based on %
        loop until satisfied
        rule defined
For extracted dataset
        open dataset (variable 3)
        set user defined constrain and %
        extract record
        calculate summary statistics
        map over different spatial units
        accept/reject extract based on %
        loop until satisfied
        rule defined
        …
repeated above process for each variable
        …
        result in a set of rule
map those rules and define the area
```

4.3 Discussion Of The Results

There are different steps involved in this analysis. After all students' home locations were plotted, we first calculated the number of student from each state. Since CSU's major campuses are located in NSW (with one on the NSW/VIC border), over 80% of students are from these two states, therefore we focus our further analysis in NSW and VIC. Following figures (figure 2a, 2b and 2c) illustrate the distribution of student sources for Park Recreation and Heritage and are mapped over two different spatial units (postcode area and census district) for comparison.

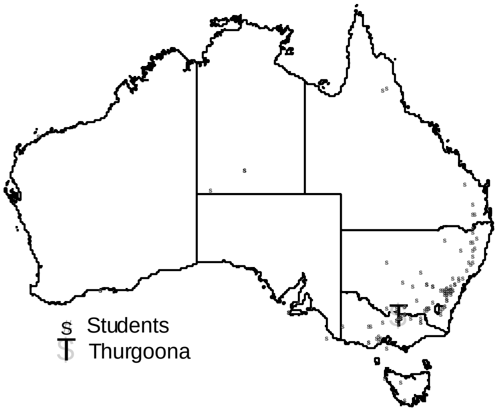

Figure 2a: All students in Australia

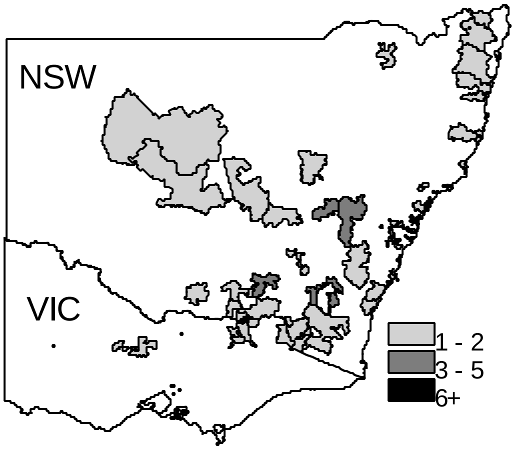

Figure 2b: Numbers of student in each postcode area

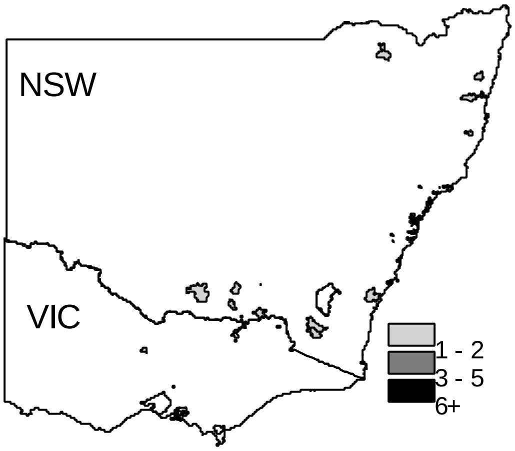

Figure 2c: Numbers of student in each census district

Accessibilities

A simple measure of student's accessibility would be their distance to railway stations and highway (access to public transportation) and distance to the campus via road network (access using private car).

While nearly 100% of students live less than 50 km from a station, our analyses classified the closeness to the station into different levels (eg. 10 km and 5 km from the stations) and studied their relations to student's enrolments. Figure 3a, 3b and 4a, 4b show the results of different levels of query/classification. Same processes are carried out for other accessibility factors, thus cross level classifications of different variables (for accessibility) can be combined together to study their association with student enrolments (figure 5).

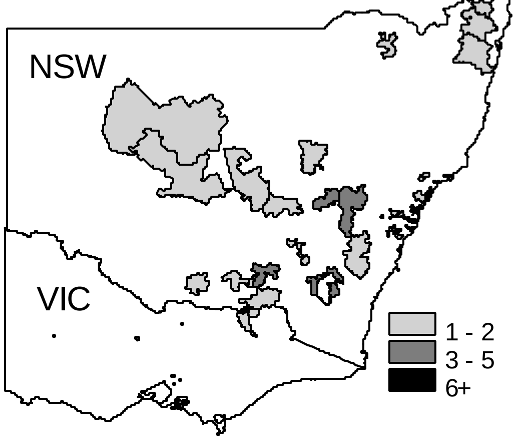

Figure 3a: Number of students that live within 10 km of a station in each postcode area (higher level of classification). 75% of the total students number.

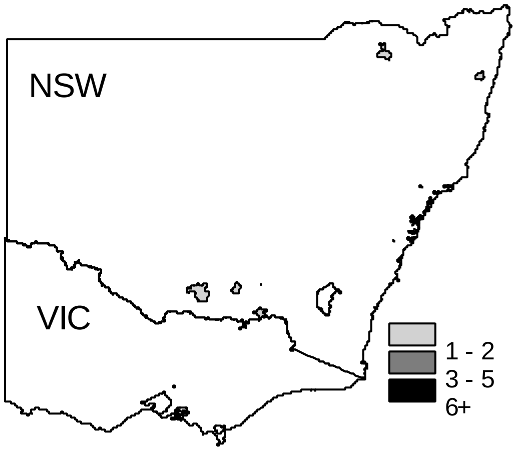

Figure 3b: Number of students that live within 10 km of a station in each census district.

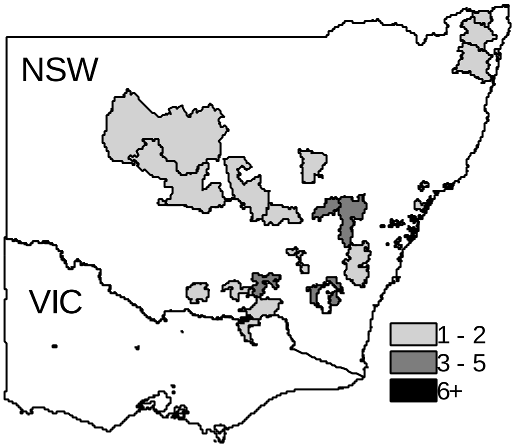

Figure 4a: Number of students that live within 5 km of a station in each postcode area (lower level of classification). 65% of the total student number.

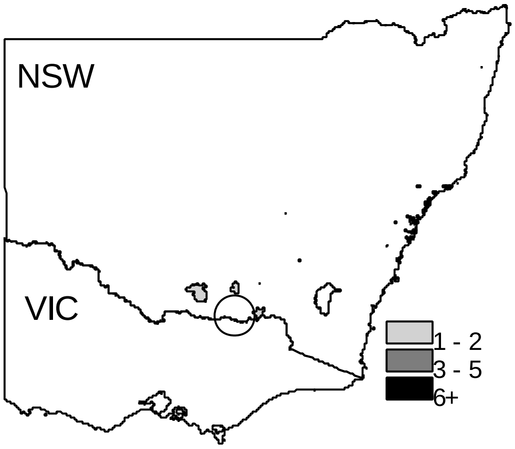

Figure 4b: Number of students that live within 5 km of a station in each census district.

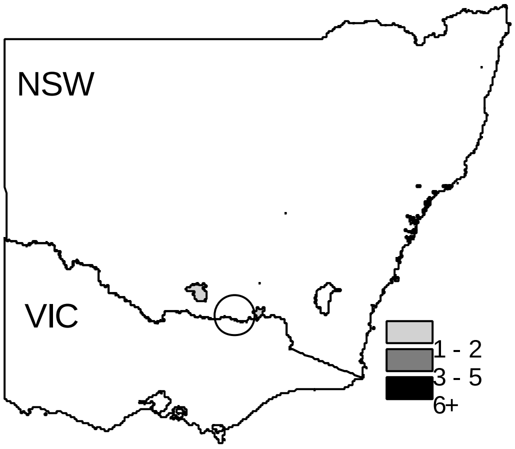

Figure 5: Number of students that live within 5km of a station and within 5 km of a highway. 50% of the total students number.

Proximity to the universities and campus

Study of student's proximity to other universities was found to have little significant impact on the enrolment pattern, as the regions which have higher number of students are clustered around each of the CSU's campuses. This may be due to their exposure to CSU and easy access to CSU's facilities.

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

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 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 forms of undergraduate education.

The enrolment pattern also had strong association with average family weekly income, as over 75% of students are from the middle income area (with weekly income of AUS$500-999). This result can be seen as the collateral evidence for the previous finding.

Defining the potential markets

A set of characteristics which strongly associated with Park Recreation and Heritage student's enrolments may be plotted against postcode area or census district to find out the possible 'hot spots' for future students sources. Those hot spots are as a collection of postcode areas and census districts. Figure 5 shows such hot spots in the unit of census district.

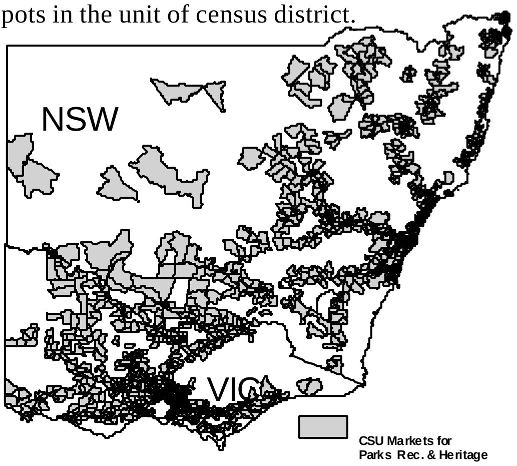

Figure 5 Possible market area.

Our results supported the findings of research conducted by Stevenson et al. (1999) which also demonstrated that variation of student characteristics (accessibilities, socio-economic, etc) across regional area are much smaller than that of metropolitan regions. While access is significant in making their enrolment decision, socio-economic status is playing more important role.

5. LIMITATIONS

While our findings explain the enrollment pattern for CSU's Bachelor of Park Recreation and Heritage in some extents, this research suffers from some limitations.

Many rural region/town areas are very large and the locations of population centers are unlikely to correspond to the regions' centroids. When numbers of student assigned to each town as points and further expended to the area, there will be mismatch and misinterpretation.

Using postcode areas as the spatial analysis unit causes the 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 units. In each district, homogeneity among the households can be assumed. Nevertheless, the use of such a spatial unit has its difficulties in the visualisation of data and implementation of analysis results. To overcome these problems, data need to be aggregated to a more suitable spatial unit without loosing its integrity.

Differences in aggregation and categorisation of social data, on the other hand, can cause different inference about a student's socio-economic status. In our case, integers are 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.

Same processes need to be carried out using historical data to check the validation of the algorithm and variable selection. The differing "hot spot" locations of each year may be detected and explained thus the trend and future locations may be derived. These trends must be evaluated against previous, current and future enrollment data.

6. FUTURE DIRECTIONS

Sources of students tend to be located in particular spatial area and geographical settings. These locations (hot spots) are not static. To predict the movement of such hot spots, many factors need to be considered, including the economic growth of the regions, trend of employment, etc.

There are a very large number of ways to define a set of area units for the collection or reporting of statistics. The choice can have important consequences for the later analysis of the data. Significantly different spatial patterns may be emerged as the level of spatial analysis is shifted. Defining spatial analysis units becomes the critical research task.

There are many spatial statistical techniques that need to be explored for spatial modeling. Those techniques include local spatial association, spatial autocorrelation, spatial regression and spatial cluster analysis.

For tools which assist the non-GIS and statistics specialists in the handling and analysis of spatial data, a range of graphical and numerical facilities need to be carefully designed and integrated to enhance the application of their special knowledge of the subject matters.

REFERENCES

Anselin, L. and Bao, S., 1997, "Exploratory Spatial Data Analysis Linking SpaceStat and Arc View", 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. L and Yang, X., 2000, "GIS in University Admissions: Analysis and Visualization of Students Flows", in *Proceeding of The 28th Annual Conference of AURISA*, Coolum, QLD, Australia, 20 - 24 November 2000.

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, pp 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), September 1999.

Han, J. and Kamber, M., 2000, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

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

Koperski, K., Han, J. and Adhikary, J., 1998, "Mining Knowledge in Geographic Data", in *Communications of the ACM*, 1998.

Marble, D., Mora, V. and Herries, J., 1995, "Applying GIS Technology to the Freshman Admissions Process at a Large University", in *Proceeding of ERSI User Conference*, 1995.

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", in *Proceeding of ERSI User Conference*, 1997.

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, Canberra, Higher Education Division, Department of Education, Training and Youth Affairs, 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, Canberra, Higher Education Division, Department of Education, Training and Youth Affairs, 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]