1999

Aggregation and site selection

Rebecca Marie Tumicki
San Jose State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

AGGREGATION AND SITE SELECTION

A Thesis

Presented to

The Faculty of the Department of Geography

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Rebecca Marie Tumicki

August 1999

APPROVED FOR THE DEPARTMENT OF GEOGRAPHY

/Richard Taketa 7/14/99

Dr. Richard Taketa, Chair

Earl D Bossard

Dr. Earl Bossard

Carlos Quilez

Carlos Quilez

APPROVED FOR THE UNIVERSITY

William Fisher

ABSTRACT

AGGREGATION AND SITE SELECTION

by Rebecca Tumicki

Previous research in the field of spatial analysis has shown geographical problems exist when using geographical data in spatial analysis. The problems range from aggregation bias in correlation and regression analysis, the sensitivity of analytical results to the definition of geographical units for which data are collected, and identifying zonal boundaries for optimal performance of any model or process at a given scale. The objective of this study was to evaluate these issues when testing the hypothesis that an optimum level of geography can be identified to assist in targeting appropriate high schools for recruiting college freshmen.

ACKNOWLEDGEMENTS

I wish to thank Dr. Richard Taketa for his guidance throughout the preparation of this thesis. I am also grateful to Carlos Quilez for providing background information and data for the case study. Also to Marica Holstrom for responding to desperate pleas of assistance in obtaining San Jose related data while I was living in Colorado. Special thanks are extended to Gary Zaepfel for his words of encouragement and for his thorough review of the document. Finally, to all my family and friends who have morally supported me for the three years while completing the thesis.

TABLE OF CONTENTS

PAGE

ACKNOWLEDGEMENTS v

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF MAPS xi

CHAPTER

I. INTRODUCTION 1

II. RELATED RESEARCH 3

Increasing Correlation Coefficients 3

Different Levels of Aggregation Produce Different Results 5

Outcome of Grouping Data at a Given Scale 6

III. CASE STUDY 8

Potential Pools of First Time Freshmen 8

Study Area 9

Levels of Data Aggregation 12

Zone Design 13

IV. METHODS 15

Selection Criteria 15

CB and FTF Data 16

Socioeconomic and Independent Variables 17

Digital Boundaries 18

Datasets 19

Regression Residuals 19

Spearman's Rank Correlation Coefficient 21

V. RESULTS AND DISCUSSION 23

Residual Results Comparison 24

Effects of Zone Design Evaluation 28

Information Loss Evaluation 29

Base Level Evaluation 34

Spearman's Rank Correlation Coefficient Evaluation 34

Future Research 36

VI. CONCLUSION 38

REFERENCE LIST 40

APPENDICES

A. SCHOOL NUMBERS BY HIGH SCHOOL AREA 42

B. SCHOOL NUMBERS BY ZIP CODE 43

C. DATA USED IN THE STUDY

Table of High School Data by District 45

Table of % First Time Freshmen and Known Potential Pool 46

Rank Order of Known Potential Pool 47

Case 1a Dataset 48

Case 1b Dataset 49

Case 1c Dataset 50

Case 2 Dataset 51

Case 3 Dataset 52

Case 1a Residual Ranking and Rank Order Results 53

Case 1b Residual Ranking and Rank Order Results 54

Case 1c Residual Ranking and Rank Order Results 55

Case 2 Residual Ranking and Rank Order Results 56

Case 3 Residual Ranking and Rank Order Results 57

LIST OF TABLES

TABLE		PAGE
4.1	Variables and Data Sources	18
5.1	Residual Output for the Three Levels of Aggregation and Zone Design	26
5.2	Schools Identified as having Potential Pools of FTF	29
5.3	Results of Rank Correlation	35

LIST OF FIGURES

FIGURE		PAGE
2.1	Example of Zone Design	7
3.1	Pie Chart of Relationship between CB, FTF and Potential Pool	9
3.2	Three Levels of Aggregation	12
3.3	Zone Designs for School #10	14

LIST OF MAPS

MAP		PAGE
3.1	High School Locations	11
5.1	Residual Output from CB Model by High School Service Area	30
5.2	Residual Output from CB Model by Zip Code	31
5.3	Residual Output for County Level using High School Service Area Boundaries	32

CHAPTER I

INTRODUCTION

San Jose State University (SJSU) is experiencing a decline in freshmen enrollment numbers. This causes concern for the university because its budget allotments from the state of California are based on enrollment numbers, especially freshmen numbers. SJSU wants to maintain a stable budgetary environment; this is accomplished by maintaining enrollment figures. In order to do this SJSU needs to increase these numbers by recruiting more effectively.

One way for SJSU to recruit more effectively is by selecting high schools to target for recruitment. The option of selecting all high schools in SJSU's service area is cost prohibitive. Therefore, the university must target high schools that provide a high return of enrolled freshmen students. Thus, SJSU needs to identify high schools with the largest potential pools of college bound students to make the recruitment process cost effective.

Deciding which schools to target for recruiting requires a variety of data. A profile of a typical first time freshman might include family income, ethnic origin, and the area where the student comes from. Demographic data provides the income levels and the ethnic make-up. SJSU possesses data by high school indicating the number of students who enrolled at the university for a particular semester. Also, the California Department of Education Research, Evaluation, and Technology Division supplies data

on the number of college bound students by high school district. SJSU has access to many sources of data to create profiles of first time freshmen.

The necessary data for selecting high schools are available at many levels of geographic aggregation. For example, the United States Bureau of the Census has demographic data available by census tract, Zip Code, county, and state levels. The California Department of Education Research, Evaluation, and Technology Division has data by high school district and high school service area levels. SJSU has data available by the individual high school. SJSU needs to determine the most useful level of aggregated data from the many levels available.

This study attempts to identify the appropriate level of data aggregation for selecting such high schools by using spatial analysis. The definition of appropriate for this study includes a level requiring minimal data manipulation, and a level that best indicates potential pools of college bound students. Possible levels to evaluate would be census block, census tract, high school service area, ZIP Code, and county. This study tests the hypothesis than an optimum level of aggregation can be identified by comparing several appropriate levels of data aggregation using spatial analysis to determine the best level at which SJSU should focus recruiters efforts.

CHAPTER II

RELATED RESEARCH

Research related to this study focuses on the aggregation problem common to spatial analysis. In its most general form the aggregation problem can be defined as the information loss which occurs in the substitution of aggregate data for individual data (Clark and Avery, 1976). More specifically, the side effects that relate to this research include increasing correlation coefficients as levels of data aggregation are grouped together; different levels of aggregation produce different results; and the outcome of grouping data at a given scale may or may not differ. These three aggregation effects apply to this research because they impact the task of determining the better level of data aggregation. The following sections describe these three aggregation effects in more detail.

Increasing Correlation Coefficients

The teams of Gehkle and Biehl and Kendall and Yule have conducted research in the area of increasing correlation coefficients as units of observation are grouped together. Gehkle and Biehl (1934) reported the earliest concerns with aggregation problems in the social sciences. They noted as smaller non-homogeneous areas, such as census tracts, are aggregated, the resulting unit shows more similarity or correlation with other units. This observation was also noted in work performed by Kendall and Yule

(1950). They showed that by using the same data and by changing the level of geography, reducing the number and increasing the size (areas to states to regions), a correlation coefficient between potatoes and wheat yields varied from .2189 to .9902.

Another researcher, W. S. Robinson, termed this phenomenon as "ecological correlation." In 1950 he gave one of the most well known commentaries on the relationship between micro level correlation and macro level correlation. Using race and illiteracy as his variables, Robinson correlated them at several levels: geographic divisions, states, and individuals. The results were .946, .773, and .203, respectively. Robinson explained that the increasing magnitude of correlation coefficients results from increasing size of the unit by demonstrating that aggregation reduces the between unit variation in a variable making the variable seem more homogeneous. This has the effect of increasing the magnitude of the correlation coefficient.

Other studies have noted that with increasing level of aggregation more and more spatial frequencies may be lost and so obscure important processes in the system (Curry, 1966). There is also loss of information when units are aggregated prior to fitting probability distributions (Cliff and Ord, 1973).

From the previous examples, one could conclude that out of the many levels of aggregation used in a study, the level with the largest unit size would show more similarity. However, the information loss that is inherent when using aggregated data in spatial analysis should not be overlooked. This study suggested that even though the level of aggregation with the largest unit size had the highest correlation coefficient,

making it the most relevant from a statistical standpoint, it may not be the better level of aggregation because of information loss.

Different Levels of Aggregation Produce Different Results

One of the most stubborn problems related to the use of areal data in spatial analysis is what is commonly referred to as the modifiable areal unit problem (MAUP) (Fotheringham and Wong, 1990). MAUP is defined as the sensitivity of analytical results to the definition of units for which data are collected. Fotheringham and Wong observed that using data gathered and reported at different resolutions for analysis can derive significantly different analytical results. They explored the sensitivity of the calibration results of two multivariate models to variations in scale and zoning systems. A data set at the block-group level for the Buffalo Metropolitan Area was used to calibrate the models. Their results showed that in an analysis of the spatial distribution of mean family income, an 800-zone data system yields the result that an increase of 0.1 in the proportion of elderly would create a decrease in the predicted mean family income of only $308. When these data were aggregated to only 25 zones, the results suggested that the same increase would produce a decrease in the predicted mean family income of $2654. An analysis of goodness of fit in the same model showed that there was an alarming increase in the coefficient of multiple determination as data were increasingly aggregated. The results suggested that it is possible to find any desired level of accuracy simply by aggregating data sufficiently.

The effects of the MAUP in multivariate analysis are essentially unpredictable. Even in the simplest multiple regression containing only two independent variables, the

interaction between changes in variance (the spread of the data) and the covariance (the degree of association) cannot be anticipated. Therefore, given that a large number of geographical studies make use of such data (perhaps the most common being analysis of aggregated census data) caution should be used when analyzing aggregate data.

This study acknowledged that different levels of aggregation produce different results and investigated these results. A comparison was made among the schools identified as having potential pools of students and the size of these pools for the levels of aggregation. This inquiry answered the question, how different were the results?

Outcome of Grouping Data at a Given Scale

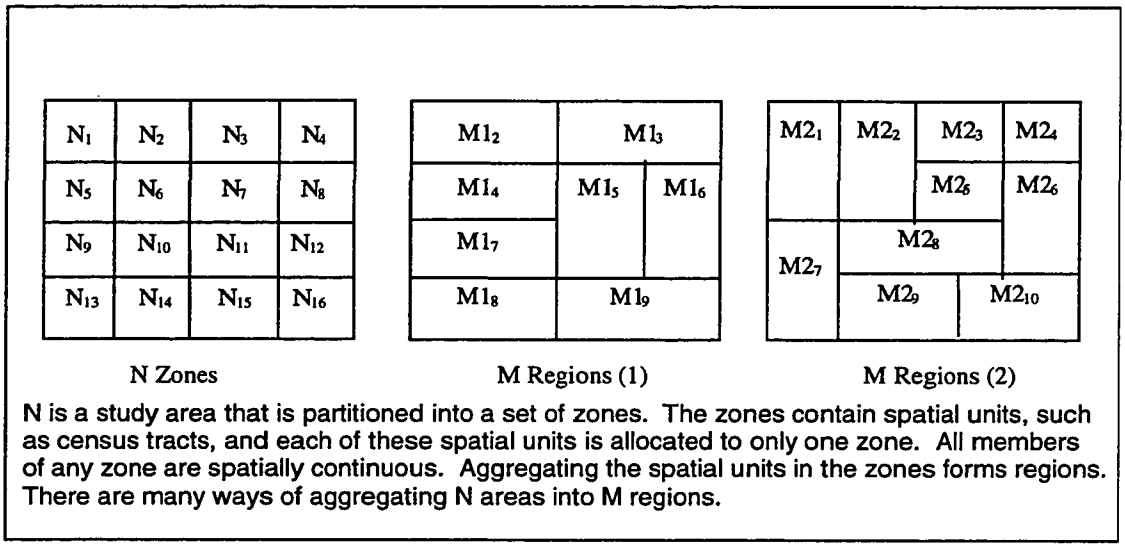
Grouping of data at a given scale is referred to as zone design (Openshaw, 1976). Zone design addresses the MAUP issue by identifying zonal boundaries which result in the optimal performance of any model of pattern or process in which spatial data is involved (Openshaw, 1976). Openshaw and Rao (1991) described the zone design process as the aggregation of spatial data for N zones into M regions where M is less than N. The M output regions should be comprised of internally connected, contiguous subzones. Provided N is not too small, and M is not large in relation to N; aggregating N areas into M regions can result in several possible configurations (see Figure 2.1).

Zones are mainly based on considerations of convenience and the existence of readily available data. Openshaw and Rao suggested starting with data at one scale and then re-aggregating it to create a new set of regions designed to be suitable for a specific purpose. Different zoning of the same data may or may not provide deviant results. Zone

design was applied to one of the levels used in the case study to determine if the results did differ among the zone designs.

FIGURE 2.1

EXAMPLE OF ZONE DESIGN



N is a study area that is partitioned into a set of zones. The zones contain spatial units, such as census tracts, and each of these spatial units is allocated to only one zone. All members of any zone are spatially continuous. Aggregating the spatial units in the zones forms regions. There are many ways of aggregating N areas into M regions.

Past research on spatial aggregation suggests researchers performing spatial analysis should heed warnings on two main points. The first is to be wary of inflated correlation coefficients as aggregation levels increase. The second point is that analytical results are sensitive to the definition of units for which data are collected. The zone design process is a way to ameliorate the situation. With these warnings in mind, a purpose of this study was to determine if our ability to identify high schools is inhibited as the level of geography changes.

CHAPTER III

CASE STUDY

Several issues were addressed before testing the hypothesis that a relevant level of aggregated data can be identified from several appropriate levels. The issues included defining the potential pools of first-time freshmen, selecting a study area, choosing the levels of data aggregation, and applying zone design to one of the levels of aggregation. Following is a short discussion of each research issue.

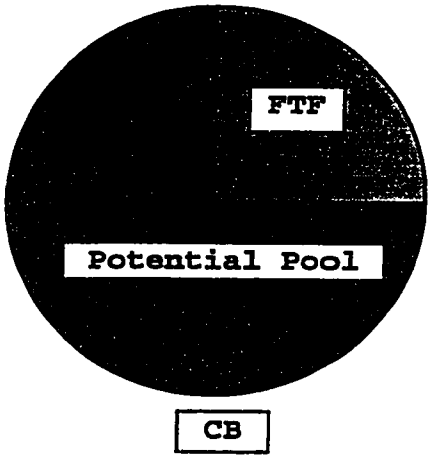
Potential Pools of First Time Freshmen

Identifying potential pools of first-time freshmen (FTF) is important to SJSU because the university can become more efficient at recruiting if they can predict the size and location of these pools before the start of the next school year. This study predicts these potential pools and location by using spatial analysis. Pertinent data for predicting these pools of students and schools are the number of eligible seniors by high school and the number of seniors who did enroll at the SJSU by high school. For this study, the number of eligible seniors by high school is labeled "college bound" (CB). Figure 3.1 depicts the relationship between CB and FTF. In this example, roughly 25% of the CB students enrolled at the university as FTF. This means the remaining 75% of the CB students represent the potential pool of FTF. By using CB and FTF data in spatial

FIGURE 3.1

PIE CHART OF RELATIONSHIP BETWEEN

CB, FTF, AND POTENTIAL POOL



Data for College Bound (CB) students and First Time Freshman (FTF) aid in determining the potential pools of high school seniors.

Study Area

analysis, SJUS can predict the potential pools of college bound students for an upcoming school year.

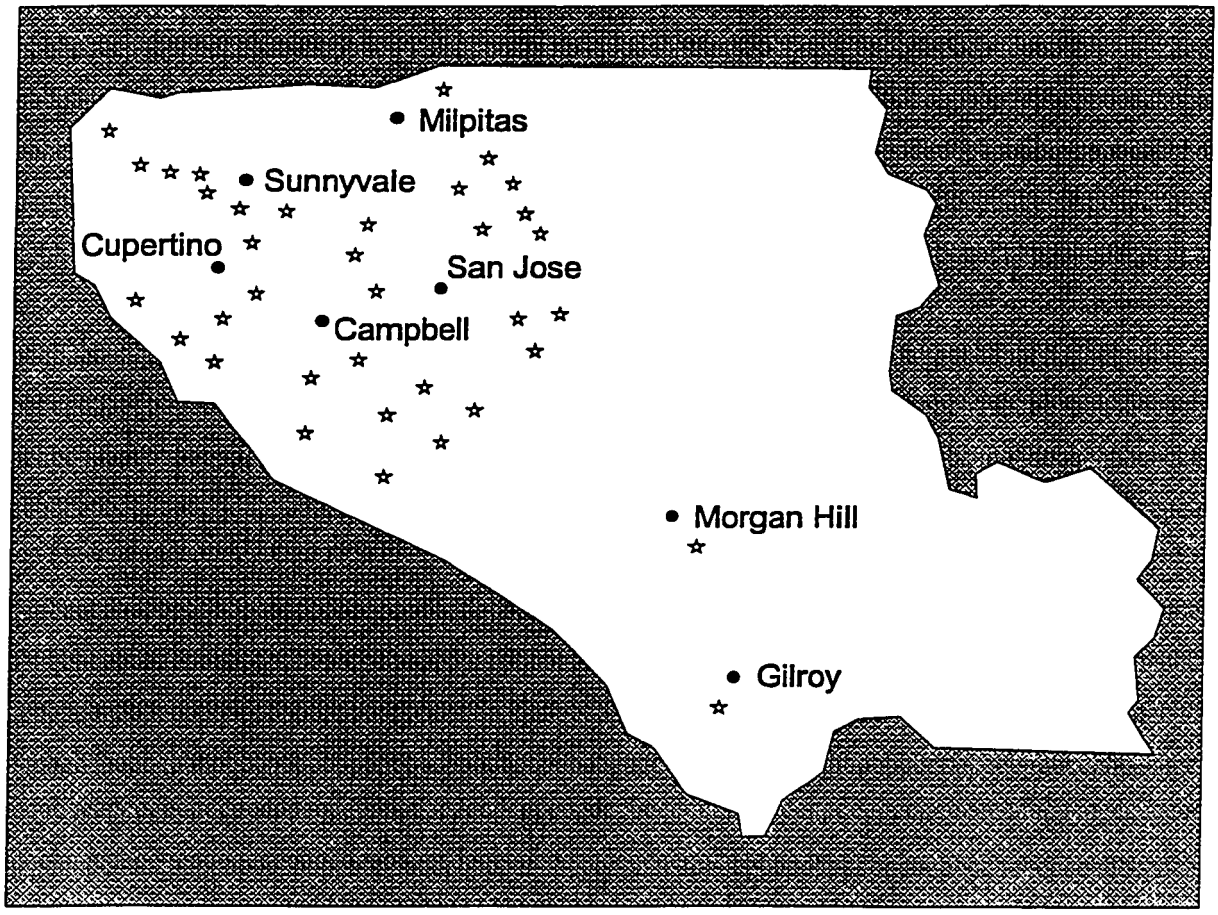
The study area was Santa Clara County, California. Two factors made this area an appropriate target: decrease in FTF enrollment and the majority of students that enroll at SJSU reside in this county. According to Leon Washington, Director of Student Outreach at SJSU, the freshmen applications to SJSU have declined (Spartan Daily 9/1/95 Vol. 105, No.3). In the past, the California State University system possessed an abundant

number of students enrolling at its campuses. A recent drop in college-age students and increased competition with other universities and colleges throughout the nation has reduced the pool of students (Spartan Daily 11/6/95 Vol. 105, No. 47). In addition, SJSU has seen overall enrollment drop from 30,338 in 1990 to 25,997 in the Fall of 1995 (Spartan Daily 11/6/95 Vol. 105, No. 47).

Another important observation is the fact that the majority of the students enrolling at SJSU live within, roughly, a twenty-mile radius of the institution. This means the recruiting effort can concentrate in a well-defined geographic area, which is Santa Clara County. Within Santa Clara County there are thirty-six public high schools to target for recruiting first-time freshmen students. Map 3.1 provides the locations of the thirty-six public high schools that were used in this case study.

MAP 3.1

HIGH SCHOOL LOCATIONS



Levels of Data Aggregation

The geographic extent for the case study data was aggregated into three levels: High School Service Area, containing 36 units; ZIP Code, containing 30 units; and Santa Clara County with one unit of aggregation. The three divisions of the study area are shown in Figure 3.2.

FIGURE 3.2

THREE LEVELS OF AGGREGATION

High School Service Area ZIP Code County

The High School Service Area represented the finest level of resolution or base level for this study. Base level is the level that best describes the geographic areas in answering the question of which high schools to target for recruiting. Other levels are more generalized, meaning they have units that are fewer in number but cover more area per unit. The High School Service Area level was the finest level of geography that should yield "the best" results. However, this level also demanded more time to prepare for analysis. Of pertinent consideration is whether this level of geography provided more useful information to outweigh the factors of time, cost and data analysis.

Zone Design

Zone design was applied to the High School Service Area because this level had to be aggregated by grouping census tracts. The three groupings are referred to as Case 1a, Case 1b, and Case 1c:

Case 1a. Aggregated only the census tracts that were completely inside the High School Service Area. This case covers less than the entire High School Service Area.

Case 1b. Aggregated all census tracts that lay within and overlap with the High School Service Area. This case covers more area than the High School Service Area.

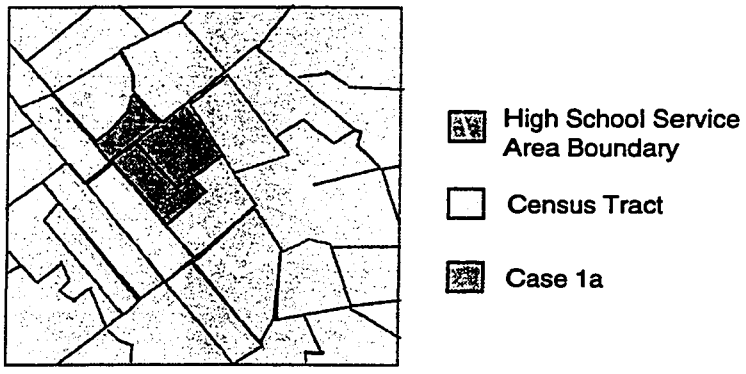
Case 1c. Aggregated only the census tracts that were 100% included and added the percentage of census tracts that overlapped with that High School Service Area. This case represents the closest approximation of the High School Service Area.

Figure 3.3 shows the three cases in relation to the actual High School Service Area boundary.

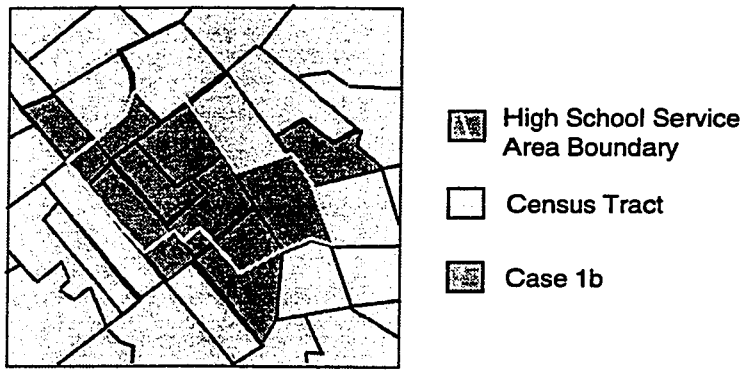
After determining the appropriate levels of data aggregation and geographic extent, the case study focused on identifying the necessary data required to aid in the recruitment process. With the case study established, methods were then defined to drive the research. The next chapter elaborates on these methods.

FIGURE 3.3

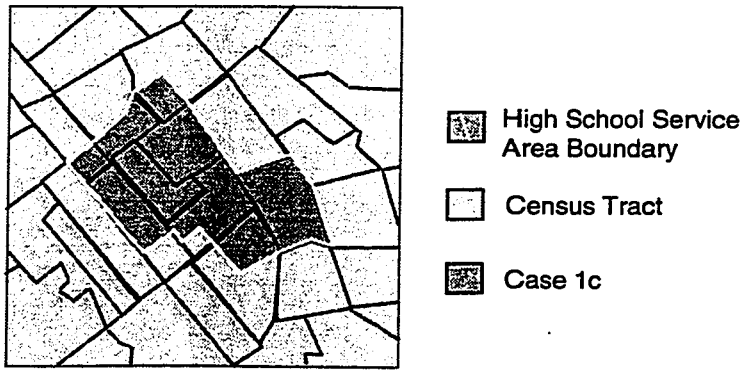
ZONE DESIGNS FOR SCHOOL # 10



Case 1a – Aggregating census tracts that fall within the High School Service Area Boundary



Case 1b – Aggregating census tracts that overlap with the High School Service Area Boundary



Case 1c – Aggregating % of census tracts that fall within and overlap with the HSSA Boundary

CHAPTER IV

METHODS

This study followed a particular methodology in order to accomplish the research objectives. To the best of the author's knowledge, this chosen methodology has not been used previously. The methods in their simplest form consisted of defining the selection criteria, collecting the data, and producing residuals. This chapter is useful to those who wish to expand on this research or create a similar case study in another geographic area.

Selection Criteria

To determine the optimum level of data aggregation among the several appropriate levels, it was necessary to define selection criteria. The selection criteria for the optimum level were defined as a level that:

1. maximized the number of high schools with the largest potential pools
2. was statistically acceptable when correlated with similar data from SJSU
3. minimized time and cost for data collection and manipulation

These selection criteria required things such as CB and FTF data, socioeconomic and independent variables, digital boundaries for the enumeration units for each level, datasets, regression residuals to represent potential pools of college bound students, and the Spearman's Rank Correlation Coefficient. The following sections describe the methods in detail.

CB and FTF Data

This study defined potential pools of students for SJSU to use in their recruitment process. The Case Study chapter defined the potential pool as the difference between the number of college bound students and the number of students that enrolled at SJSU (CB-FTF). The scope of the data for CB and FTF was limited to the 1989-1990 school year and to the 36 public schools identified in the Case Study chapter. The data came from two sources. The CB data, produced by the California Department of Education Research, Evaluation, and Technology Division, were obtained from the *College Bound Report, 1985 - 1992*. FTF data, consisting of first-time freshman numbers for Fall 1990, was provided by the Institutional Research Department at SJSU. The CB and FTF data did not require manipulation for analysis.

These two data elements produced not only the dependent variable for this study but also the known pool (KP) variable. The dependent variable, referred to as the percentage of first-time freshman (%FTF), is defined as the percentage of college bound seniors from a particular high school that enrolled at SJSU as a first-time freshman. The equation for this variable is %FTF = (FTF/CB) * 100. The other variable, KP, is a historic indicator of the number of students that were potential FTF but did not enroll at SJSU. The equation for this variable is 100-%FTF. A data table for the %FTF and KP variables is located in Appendix C, titled, "Table of Percent First-Time Freshmen and Known Pool."

Socioeconomic and Independent Variables

Other variables besides CB and FTF were used in the research to determine their effects on the spatial model. Two of the four independent variables were a subset of salient characteristics of college freshmen including two socioeconomic measurements. The two socioeconomic variables were median family income (I) and educational attainment level (ED) of adults over 25 years of age. These characteristics were also selected based on data availability. This study used the 1990 Census of Population and Housing by the United States Bureau of the Census. Most libraries have this information on CD-ROM for census tracts and ZIP Codes.

Another independent variable was the distance (D) from SJSU to the centroid of each spatial unit. Distance was selected as a variable because although we know that the majority of students that enroll at SJSU come from Santa Clara County, SJSU wanted to determine if there was a correlation between distance and numbers of FTF. MapInfo Professional Version 4.1, a geographic information system (GIS) from MapInfo Corp., Troy, N.Y., was used to measure the distance. The remaining independent variable was CB. See Table 4.1 for a summary of variables and data sources.

TABLE 4.1

VARIABLES AND DATA SOURCES

Variable / Agg. Level	% First Time Freshmen	College Bound	Income	Educational Attainment	Distance
High Sch.	CB Report/SJSU	CB Report	Census 1990	Census 1990	GIS
ZIP Code	aggregation	aggregation	Census 1990	Census 1990	GIS
County	mean	mean	n/a	n/a	n/a

n/a = not applicable

Digital Boundaries

Digital boundaries played an important role in this study. They were used to create thematic maps to analyze information loss among the three levels of data aggregation. For this study information loss is defined as the loss of spatial information when comparing the results to the base level's results. Information loss was determined by mapping the residual values for the levels of data aggregation and then comparing these thematic maps. The thematic maps were produced using MapInfo software. The digital boundaries for the three levels were obtained in two ways. Two of the three levels chosen for this study, ZIP Code and county, were available in digital format and were accessible through MapInfo.

The remaining level of data aggregation, High School Service Area, was not available in digital format and had to be constructed. Four steps were required for creating these boundaries: 1) locating a hard copy of digital census tract boundaries; 2) acquiring hard copy outlines for the high school service areas from the six school districts in Santa Clara County; 3) tracing the census tract and high school service area boundaries onto a county map; and 4) constructing the three sub cases for the High School Service Area level. MapInfo was used to manipulate the digital census tract boundaries to create

these three sub cases. The section "Levels of Data Aggregation" in the Case Study chapter described how the census tracts were aggregated for these sub cases.

Datasets

A dataset was created for each level of aggregated data. The datasets were formatted with the 36 schools listed vertically and the dependent and independent variables listed horizontally. This format of having a record for each of the 36 schools allowed for different levels to be correlated with the known pool variable. If the format was by enumeration unit, then the results would not be conducive for correlation because each level had varying number of units.

The datasets were compiled with the data collected for the variables. They were then screened to uncover any inconsistencies or errors that might jeopardize or bias the outcome of the study. This procedure revealed two outliers for the dependent variable of %FTF. Two of the schools sent more than 100% of their college bound students to SJSU. School 33 and school 34 respectively sent 188% and 272% of their college bound seniors meaning other seniors not classified as "college bound" also enrolled at SJSU. These two schools were removed from the dataset because these sent more than 100% of their college bound students to SJSU and because this study is searching for untapped sources of students. See Appendix C for the datasets used in spatial analysis.

Regression Residuals

This research was interested in the residuals produced by regression analysis. The regression models provided the estimated number of FTF who should have enrolled at SJSU. The estimates (residuals) describe the recruiting potential. For example, if 15

students enrolled at SJSU from a specific high school and the predicted value was 21, the residual is recorded as –6 (15 – 21). The negative residuals indicate where the potential pools of students are located and the size of these pools.

The residuals for the first two levels were obtained by running several regression models. A regression model assumes normal distribution of the dependent variable for every combination of the values of the independent variables in the model. This assumption was checked and the results showed normal distribution for the combinations of dependent and independent variables. Verifying other parametric measures from the regression model was not applicable to the scope of this study. However, further investigation in this area could concentrate on optimizing the model before using the residuals in rank correlation.

Four regression models were generated using combinations of the four independent variables along with the dependent variable. This approach allowed for checking the sensitivity of the spatial distribution due to the variables. If the results were similar for the four models, then there is no sensitivity among the variables, else the models would require further investigation. The output from the models indicated minimal sensitivity among the variables.

The first model included the variable CB and used simple linear regression. The remaining three models included the variables CB&D, CB&D&I, and CB&D&I&Ed, respectively. The residual output from each model was sorted in ascending order to list the potential pools first. Along with the residual output, the corresponding school numbers were also reported to help identify the spatial distribution by schools.

Regression was conducted using the statistical package available with Microsoft Excel 7.0 (Microsoft Corp,. Redmond, WA).

The residuals for the County level were obtained using another method. At this level, the variables were aggregated to constants, making it inappropriate for regression. Instead, the residuals were calculated by subtracting the county mean of %FTF from %FTF for each school.

After the residuals were sorted in ascending order they were ranked by size from 1 to N. Schools having large negative values were ranked the highest starting with 1. This procedure was repeated for the known pool variable. See Appendix C for the residuals and ranking results.

Spearman's Rank Correlation Coefficient

To determine if the results from the levels of data aggregation were statistically acceptable a correlation coefficient procedure was necessary to measure the correlation between the two variables of the potential pool and the known pool. When selecting a valid correlation coefficient procedure certain assumptions need to be met. For example, a commonly used procedure, the Pearson's Product Moment Coefficient is used to measure the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height in inches and weight in pounds. However, this procedure makes the implicit assumptions that the two variables are jointly normally distributed. This assumption is not justified based on the non-linear relationship between the two variables in question for this study, therefore, using a non-

parametric measure, such as the Spearman's Rank Correlation Coefficient, was more suitable.

As correlation is a ratio, it is a unitless number. Values for a correlation fall between −1.0 and +1.0. A value of 0.0 indicates no correlation exists between the two variables. A value close to +1.0 suggests a near perfect positive correlation; as one variable increases, the other will increase in near unison. A value near −1.0 indicates a near perfect negative correlation; as one variable increases, the other decreases.

This study anticipated positive correlation for the three levels of data aggregation. A positive correlation was expected because the known pool and the predicted pool are describing the same situation. The known pool is the historic indicator while the predicted pool is anticipating what the known pool will be. Therefore, as the known pool variable increases or decreases in size for a specific school, the values of the potential pool should also increase or decrease in unison.

The results from running Spearman's Rank Correlation Coefficient, r, were compiled in a data table. The format of the table includes listing the levels of data aggregation vertically and the (type of) spatial models horizontally. This setup provided a means to quickly glean information from the table. In particular, how r varied between the levels of aggregated data and among the spatial models. The next chapter analyzes the results produced from these methods.

CHAPTER V

RESULTS AND DISCUSSION

This study was designed to compare several appropriate levels of data aggregation using spatial analysis and other criteria as outlined in the Methods chapter to determine the optimum level of data aggregation for recruiting objectives. This study was also designed to ascertain the effect aggregation and zone design had on the results. In particular, was our ability to identify high schools inhibited as the data were aggregated, and was there information loss from the base level when compared to the other two levels?

This chapter analyzes the results keeping it focused on the problem statement: can an optimum level of data aggregation be identified among several appropriate levels for selecting high schools to recruit students, and also describes what was revealed in the results. Issues that were addressed in the previous chapters are assessed in this chapter. An overview of the assessment includes 1) comparing the residual results for the levels of aggregation; 2) evaluating zone design effects; 3) evaluating information loss; 4) determining if the High School Service Area level provided more useful information when compared to the other two levels; and lastly, 5) comparing the Spearman's Rank Correlation Coefficient output to determine whether there is statistical evidence that

supports identifying a "better" level of geography. The last section explores potential directions for future research.

Residual Results Comparison

The residual results from regression analysis for the three levels of data aggregation were scrutinized for two issues that were of interest to this research. First, did the residual output vary for the three levels of data and for zone design, and second, did the schools identified as having potential pools of students fluctuate in the output? For example, were the same schools identified for all three levels of geography or were some dropped and others introduced. Observations and discussion follow for these two issues.

Residual Output

The five tables containing the residual rankings for the three levels of data aggregation and for zone design are located in Appendix C. Table 5.1 condenses the residual output into this one table for ease of evaluation. Five columns were omitted from this table because these columns had the same results of other columns already present in the table. Models #1 and #2 for High School Service Area sub cases produced the same results because the same data were used in the models. Hence, the results for models #1 and #2 of sub case 1a are present while the same results for sub cases 1b and 1c have been omitted. This is also true for the ZIP Code's model #1. The residual output was then compared and three observations are noteworthy.

First, as the background research suggested, the results were different for the three selected levels and for the sub cases. Table 5.1 shows the output for the three levels are

different; the sub cases representing zone design also varied. While the output is different, there is not enough of a difference to deem one level is better than the others. From this observation, the author agrees with D. Sawicki in his finding that there isn't always a "best" unit of analysis to explain one specific kind of social behavior. Behavior may be explainable at a number of levels (Sawicki, 1973). The behavior being evaluated in this research is choosing a college to attend after high school. Each level produced similar behavioral results.

The second observation concerns residual size. A glance at Table 5.1 shows there are differences among the residual output; however, the magnitude of the residuals appear in a common range and pattern. The low end of the range starts at –1, meaning there is one potential student available for recruiting tactics, and the high end is –32 or 32 potential students. Of the residual output, two columns are very similar. The output for the first model (CB) at the High School Service Area and ZIP Code levels and the County output show the residuals being generated and are very close in magnitude. This could be due to the data being used. The CB model uses the dependent variable of %FTF and the independent variable of CB while the County level uses solely %FTF. As variables were added to the other models the magnitude of the residuals decreased in size. This behavior is explained in the next observation.

The third observation suggests a relationship is present between residual output and the number of variables introduced to a model. A possible explanation for this relationship involves the variance of the data. Adding variables to a model contributes to

TABLE 5.1

RESIDUAL OUTPUT FOR THE THREE LEVELS OF AGGREGATION AND ZONE DESIGN

HIGH SCHOOL SERVICE AREA								ZIP CODE			COUNTY
Case 1a				Case 1b		Case 1c		Case 2			Case 3
#1	#2	#3	#4	#3	#4	#3	#4	#2	#3	#4	
-29	-30	-30	-28	-27	-26	-29	-26	-32	-32	-23	-21
-22	-25	-27	-23	-23	-21	-28	-20	-20	-20	-17	-21
-21	-18	-16	-17	-15	-16	-22	-16	-18	-18	-14	-21
-17	-17	-16	-13	-15	-14	-15	-15	-17	-15	-11	-20
-17	-15	-13	-12	-13	-14	-14	-13	-15	-14	-10	-19
-17	-13	-13	-12	-12	-13	-12	-12	-15	-9	-10	-19
-13	-12	-11	-10	-10	-11	-11	-12	-13	-9	-9	-18
-13	-12	-11	-9	-9	-10	-9	-11	-11	-9	-8	-18
-12	-12	-10	-9	-8	-9	-9	-9	-9	-8	-7	-17
-12	-10	-8	-7	-7	-7	-8	-8	-6	-8	-5	-17
-11	-7	-8	-6	-7	-6	-6	-6	-6	-8	-5	-16
-11	-6	-5	-6	-6	-5	-5	-5	-6	-6	-5	-14
-8	-6	-5	-6	-5	-3	-4	-5	-5	-6	-4	-11
-8	-6	-4	-3	-4	-2	-4	-4	-5	-6	-3	-8
-7	-4	-2	-2	-4	-1	-3	-2	-5	-2	-3	-6
-3	-3	-2	.	-2	-1	-2	-1	-4	-1	-2	-2
-3	-2	-2	.	-2	.	.	.	-4	-1	-2	-1
-2	-1	.	.	.	.	.	.	-3	.	-2	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

#1 – CB Model
#2 – CB & D Model
#3 – CB & D & I Model
#4 – CB & D & I & Ed Model

explaining the data, so the measure of the spread of data is accounted for and this results in increasing correlation coefficient values and decreasing residual values. This observation was also noted in work performed by Fotheringham and Wong. They reasoned as variables are added to a model, the process involves a smoothing effect so that the variation of a variable tends to decrease as aggregation increases (1990).

Potential Schools

A primary objective for SJSU is to identify schools with potential college recruits. A part of this research was to investigate if there was a marked difference in the schools that were being identified for the three levels. The schools identified as having potential are shown in Table 5.2. The schools have been sorted in descending order with the school having the largest pool being first. Two observations were noted when comparing the residual output by school. First, the schools identified were different for each level. This is similar to the residual results and although the same schools were not identified among the levels of aggregation, the schools being identified were nearly the same. Thirteen out of the eighteen schools are common for the three levels. This observation reinforces the conclusion made in the residual section; while the schools are different, the majority of them are the same, therefore there is not enough of a difference to declare one level better than the other levels at this point in the analysis. Regardless of the level of data aggregation selected, SJSU would identify similar schools as having potential pools of FTF.

The second observation concerns the schools' rankings. The rank order of the schools was also different for the three levels. The rank order defines which schools have the greatest recruiting potential. They are ranked in descending order so the schools with the greatest potential are listed first. The school rankings were consistent for the first two levels while the remaining level shows no similarity or discernable pattern. Although the rankings may differ, the objective SJSU is seeking is still the same: identifying schools with recruiting potential. Referring back to Table 5.1, the residual pools are fairly

consistent among the levels. For example, school #23 has a potential pool of 21 students for one level and 13 in another level. SJSU needs to evaluate if this difference of 8 potential students is acceptable.

Effects of Zone Design Evaluation

Openshaw and Rao (1991) introduced the idea of zone design to achieve a suitable set of regions based on re-aggregating data. They noted that different zoning of the same data may or may not provide deviant results. This case study observed deviant results for the High School Service Area sub cases (Table 5.1 and 5.2). The residuals varied among the three sub cases but the differences are insignificant. The residuals differ by four at the most. For example, school #4 had 30 potential students for sub case 1a and 26 students for the other two zone designs.

For this case study, the effects of zone design were minimal. From this observation one might conclude that if the High School Service Area level was to be used, the zone design to implement would be the one that requires minimal time by the staff to create and then maintain. In this instance, it would be sub case 1a. This zone design required fewer census tracts to aggregate translating to less time and money to gather and compile the data. One drawback to this zone design would be the graphic representation. The areal units are not contiguous and they would appear to be islands floating in the boundary of Santa Clara County. The better zone design for mapping purposes would be sub case 1c. This zone design mimics the existing high school service area boundaries that are contiguous; however, this design requires the greatest amount of time for boundary definition and data collection. Depending upon the decision making

process for identifying where to recruit, either by tabular or spatial representation, sub cases 1a and 1c have distinct advantages.

TABLE 5.2

SCHOOLS IDENTIFIED AS HAVING POTENTIAL POOLS OF FTF

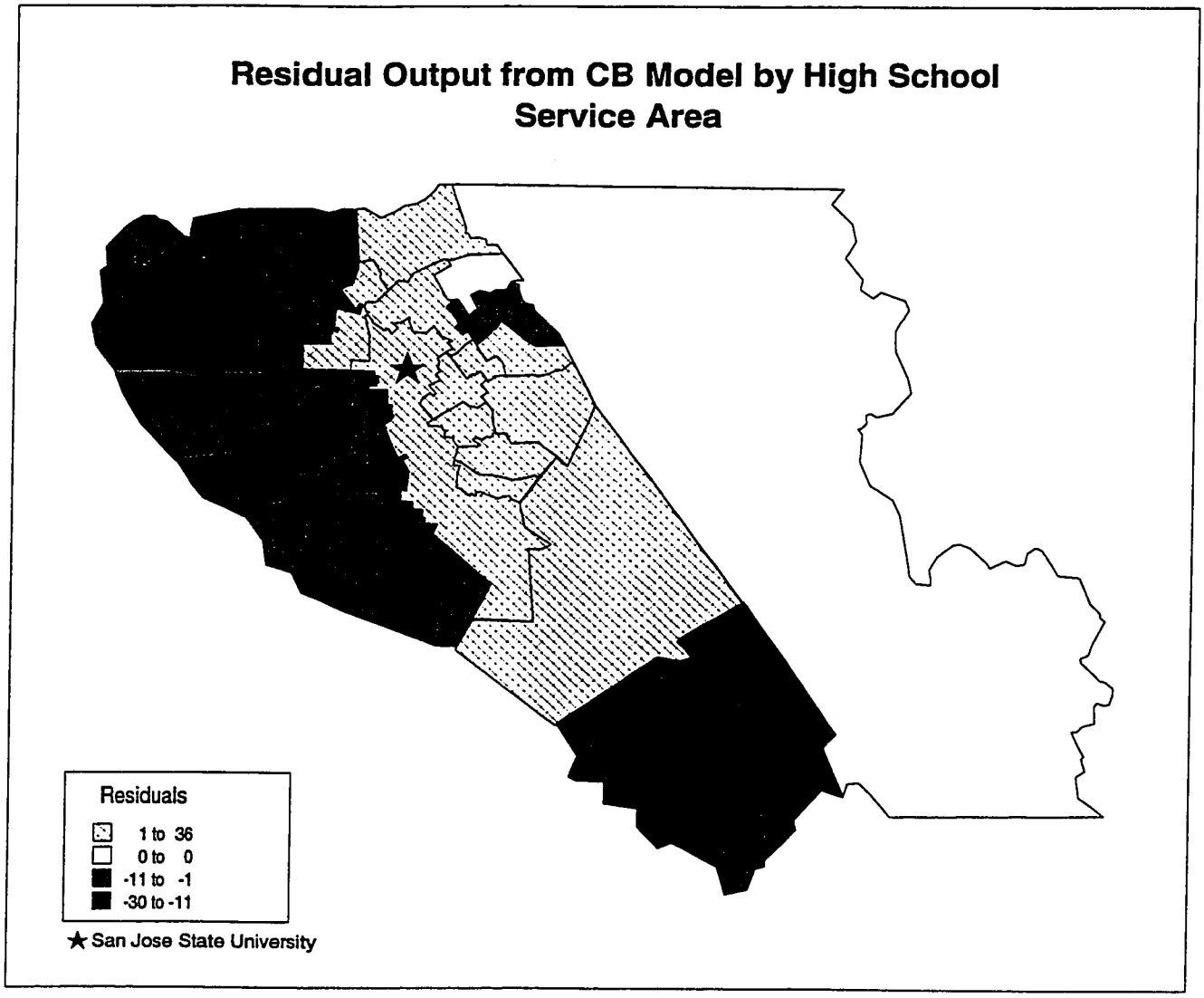
HIGH SCHOOL SERVICE AREA								ZIP CODE			COUNTY
Case 1a				Case 1b		Case 1c		Case 2			Case 3
#1	#2	#3	#4	#3	#4	#3	#4	#2	#3	#4	
4	4	4	4	4	4	4	4	4	4	4	27
3	3	3	3	3	3	3	3	2	23	23	28
23	2	2	23	2	16	2	7	3	2	7	4
2	23	23	2	16	2	16	23	23	1	18	21
21	21	1	22	23	23	23	16	18	18	2	17
24	18	16	16	1	7	15	2	21	27	17	23
27	15	15	24	17	1	1	1	1	31	27	19
22	1	18	1	15	24	17	22	16	17	19	18
17	17	7	7	7	17	18	24	17	16	8	22
18	16	22	15	36	22	36	17	36	32	36	3
15	22	17	18	30	36	7	21	27	36	24	24
16	19	27	17	27	15	22	18	31	19	1	20
28	7	36	36	24	27	20	36	15	7	16	2
19	27	30	21	35	18	30	15	19	24	21	16
1	36	35	35	22	35	27	35	22	15	14	15
36	11	19	.	19	6	19	27	7	20	3	36
7	30	20	.	18	.	.	.	20	.3	.31	30
20	20	.	.	.	.	.	.	11	.	.28	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	

#1 – CB Model
#2 – CB & D Model
#3 – CB & D & I Model
#4 – CB & D & I & Ed Model

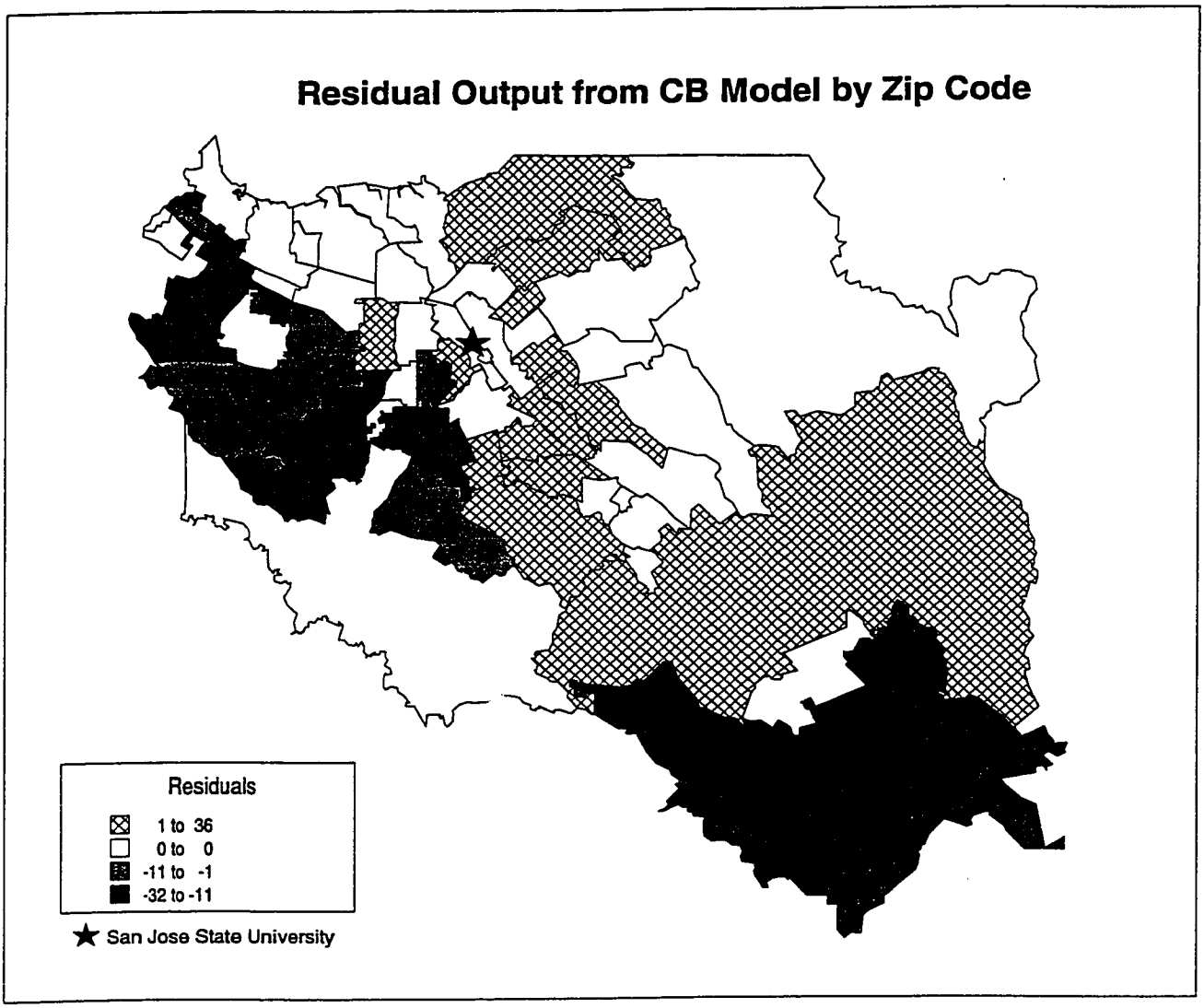
Information Loss Evaluation

Information loss was evaluated by comparing the three thematic maps that displayed the residual values. In particular, the maps were compared to note the areas of high concentration for potential pools of FTF. These maps are shown in Maps 5.1 – 5.3. If the same areas of the county were identified as having recruiting potential for all three

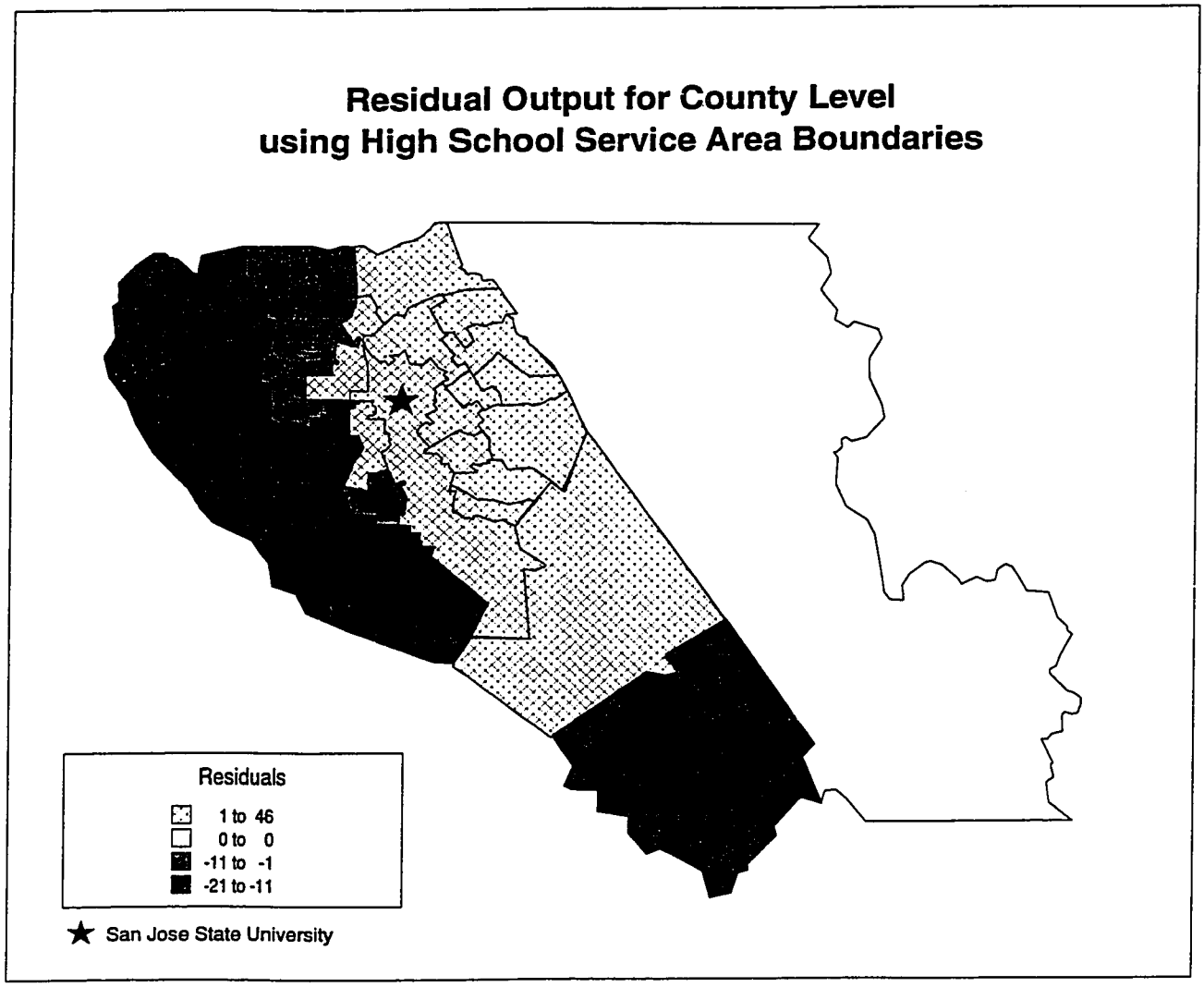
MAP 5.1


Residual Output from CB Model by High School Service Area
Residuals
1 to 36
0 to 0
-11 to -1
-30 to -11
San Jose State University

MAP 5.2


Residual Output from CB Model by Zip Code
Residuals
1 to 36
0 to 0
-11 to -1
-32 to -11
San Jose State University

MAP 5.3


Residual Output for County Level
using High School Service Area Boundaries
Residuals
1 to 46
0 to 0
-11 to -1
-21 to -11
San Jose State University

maps, then there is no information loss. This would indicate that from a spatial perspective, any one of the three levels would be adequate for SJSU to use.

After viewing the maps it is evident that all three maps indicate the same area in Santa Clara as having recruiting potential. The western side of the county shows the largest concentration of potential pools of FTF. This phenomenon might be attributed to the demographic make-up of this area. The 1990 census data shows that the high school students living in this area come from higher income backgrounds than the mean income for Santa Clara County, which is approximately $45,000. Also, this area shows a higher percentage of individuals with educational attainment beyond a four-year college education. In addition, two other renowned educational institutes, Santa Clara University and Stanford University, reside on this side of the county. One might conclude that the high school students in this area have a greater influence to further their education beyond high school and they have the means to attend colleges that are more expensive than SJSU. Of the three maps, the County level indicates more area with potential than the other two.

While information loss may be apparent on a school basis, viewing the results at the county level shows similar results. This spatial representation would help SJSU to recruit more effectively because they can identify the area in which to focus their attention. Instead of visiting schools in a random pattern, they could focus their energies in one area and target several schools to reduce costs.

Base Level Evaluation

The definition of base level was introduced in the Case Study chapter. The base level for this study was the High School Service Area level. The results from this level were compared to the other two levels to determine if this level provided more useful information to outweigh the factors of time, cost, and data analysis. The output for residual values, schools, and the maps were useful for the evaluation. As stated in the earlier sections on Residual Results Comparison and Information Loss Evaluation, there is no evidence that supports the theory that the base level provided additional information with respect to the other two.

The comparison of the results showed several similarities. The magnitudes of the residuals were similar, the majority of the schools were the same, and the same area of Santa Clara County was identified as having the highest recruiting potential. The only advantage the author sees in the base level is the spatial representation. The ZIP Code level appears scattered when compared to the base level boundaries. SJSU needs to determine which type of spatial representation would suit their recruiting objectives.

Spearman's Rank Correlation Coefficient Evaluation

The Spearman's Rank Correlation Coefficient was used in this study because it is an easy and quick way to determine what type of relationship exists between two variables, in this case the known pool and the potential pool. The Spearman's Rank Correlation Coefficient was calculated for each level of aggregation and for the three zone designs. The results are presented in Table 5.3.

The results in Table 5.3 agree with the background research that predicted the correlation coefficient would be inflated as units of observation were grouped together. This effect was noticed in a wide variety of applications such as correlation of rental values and delinquency rates for city sub areas and correlating race and literacy in the United States (Hannan and Burstein, 1974). The County level, the largest grouping in this study, had a near perfect positive correlation of .99 when correlated with the known pool. This means that the County level acts very similar to the historic event.

TABLE 5.3

RESULTS OF RANK CORRELATION

Level	CB	CB,D	CB,D,I	CB,D,I,Ed	%FTF
County					.99
Zip Code	.88	.75	.64	.62	
Case 1b	.88	.77	.62	.64	
Case 1c	.88	.77	.72	.67	
Case 1a	.88	.77	.69	.65	

Rank Correlation by model for each level of aggregation and zone design.

Based on criteria established for this study and the empirical evidence derived from this study, a conclusion can now be drawn to answer the problem statement. This study set out to determine whether or not an optimum level existed among three appropriate levels selected for this case study. Given that SJSU is looking for a way to determine the potential pools of FTF by high school in addition to keeping costs down, and using data that would be statistically acceptable, the level that meets all the criteria is the County level. The County level satisfied all the criteria outlined in the Methods chapter:

1. The County level met the objective of identifying the largest pools of potential students. Table 5.1 clearly shows the County column having a distinct advantage over the other two levels in this area.
2. The County level provided potential pools that correlated with the known pool as well as the other levels.
3. The County level required only the FTF and CB data, minimizing the cost of data collection and manipulation as compared to the other levels.

Essentially, this study has determined that the County level does as well as the other levels in terms of identifying the potential pool, even accounting for the problems of aggregation, and, in addition, the County level meets the third objective with lower costs.

Future Research

Recommendations for further study in this area would be to investigate the historic aspect of the study. A comparison among several years using CB and FTF data at the County level should reveal trends of where potential pools of students have been identified over time. This information could help SJSU to evaluate the accuracy of the predicted pools. If the predicted pools follow the trend of known pools, then the methods defined for predicting the pools are valid.

Another avenue for continued research is in improving the residual values. This effort would be to fine-tune the regression model. For instance, this could be achieved by examining the explanatory parameters of the model before the residual values are used in the Spearman's Rank Correlation Coefficient. A comparison could then be made on the correlation coefficient values in this study to assess the value of tuning the model.

A final suggestion would be to repeat this study at another major metropolitan university to see if similar results repeat. This would answer the question if the results were site specific. Many similar sites exist nationwide, such as the Denver metropolitan area in Colorado and the St. Louis area in Missouri. If the results proved similar, then this could convince universities to implement a recruitment methodology that was introduced in this study to help them maintain their enrollment figures.

CHAPTER VI

CONCLUSION

The primary objective of this study is to determine if an optimum level of data aggregation can be derived among several appropriate levels using spatial analysis and a case study. Determining an optimum level benefits SJSU in the recruitment process and helps SJSU to maintain enrollment numbers and to secure future funding. Results from regression analysis and Spearman's Rank Correlation Coefficient show that the County level is the better level based on the criteria established for this study. The County level is the most cost-effective in terms of data collection time and manipulation; this level identifies the largest potential pools by high school, and is statistically acceptable with a rank correlation coefficient of .99.

Research results also reveal that effects from both zone design and aggregation are minimal. Residual sizes and schools being identified are fairly consistent among the three levels of aggregated data. Information loss is not an issue when selecting the optimum level because the spatial distributions of the potential pools of students indicate the same area in all three cases. Lastly, the base level of data does not provide additional information that warrants the extra time and cost required for this level.

Based on the conclusions from this study, SJSU and other similar institutions can benefit from the methodology introduced in this research. This methodology suggests

that universities can effectively target high schools that provide a high return of enrolled freshmen students based on using CB and FTF data at the County level.

REFERENCE LIST

Batty, M, and P. K. Sikdar. "Spatial Aggregation in Gravity Models: 1. An Information-Theoretic Framework." *Environment and Planning A*, 14 (1982a): 377-405.

Batty, M, and P. K. Sikdar. "Spatial Aggregation in Gravity Models: 2. One-Dimensional Population Density Models." *Environment and Planning A*, 14 (1982b): 525-553.

Batty, M, and P. K. Sikdar. "Spatial Aggregation in Gravity Models: 3. Two-Dimensional Trip Distribution and Location Models." *Environment and Planning A*, 14 (1982c): 629-658.

Blalock, H. M. "Aggregation and Measurement Error." *Social Forces*, 50 (1971): 151-165.

Clark, W. A. V., and K. Avery. "The Effects of Data Aggregation in Statistical Analysis." *Geographical Analysis*, 8 (1976): 428-438.

Cliff, A., and J. K. Ord. *Spatial Autocorrelation*. London: Pion, 1973.

Curry, L. "A Note on Spatial Association." *Professional Geographer*, 18 (1966): 97-99.

Fotheringham, A. S., and D. W. S. Wong. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A*, 23(7) (1990): 1025-1044.

Gehlke, C. E., and K. Biehl. "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association Supplement*, 29 (1934): 169-170.

Hannan, M. T. and L. Burstein. "Estimation from Grouped Observations." *American Sociological Review*, 39 (1974): 374-392.

Kendall, M. G. and G. U. Yule. *An Introduction to the Theory of Statistics*. London: Griffin, 1950.

Openshaw, S. "A General Method for Identifying Scale and Aggregation Effects on any Statistical Model of Pattern and Process in a Spatial Domain." *Advanced Applications in Probability*, 8 (1976): 656-657.

Openshaw, S. "A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modeling." *Transactions Institution of British Geographers*, 2 (1977): 459-471.

Openshaw, S. and L. Rao. "*Re-engineering 1991 Census Geography: Serial and Parallel Algorithms for Unconstrained Zone Design.*" School of Geography, Leeds University, Leeds LS2 9JT, 1991.

Robinson, W. S. "Ecological Correlation and the Behavior of Individuals." *American Sociological Review,* 15 (1950): 351-357.

Sawicki, David, S. "Studies of Aggregated Areal Data: Problem of Statistical Inference." *Land Economics*, 49 (1973): 109-114.

Spartan Daily. 9/1/95 Vol. 105 No. 3, "Recruiters Hired to Raise Enrollment."

Spartan Daily. 11/2/95 Vol. 105 No. 45, "Minority Enrollment on the Rise."

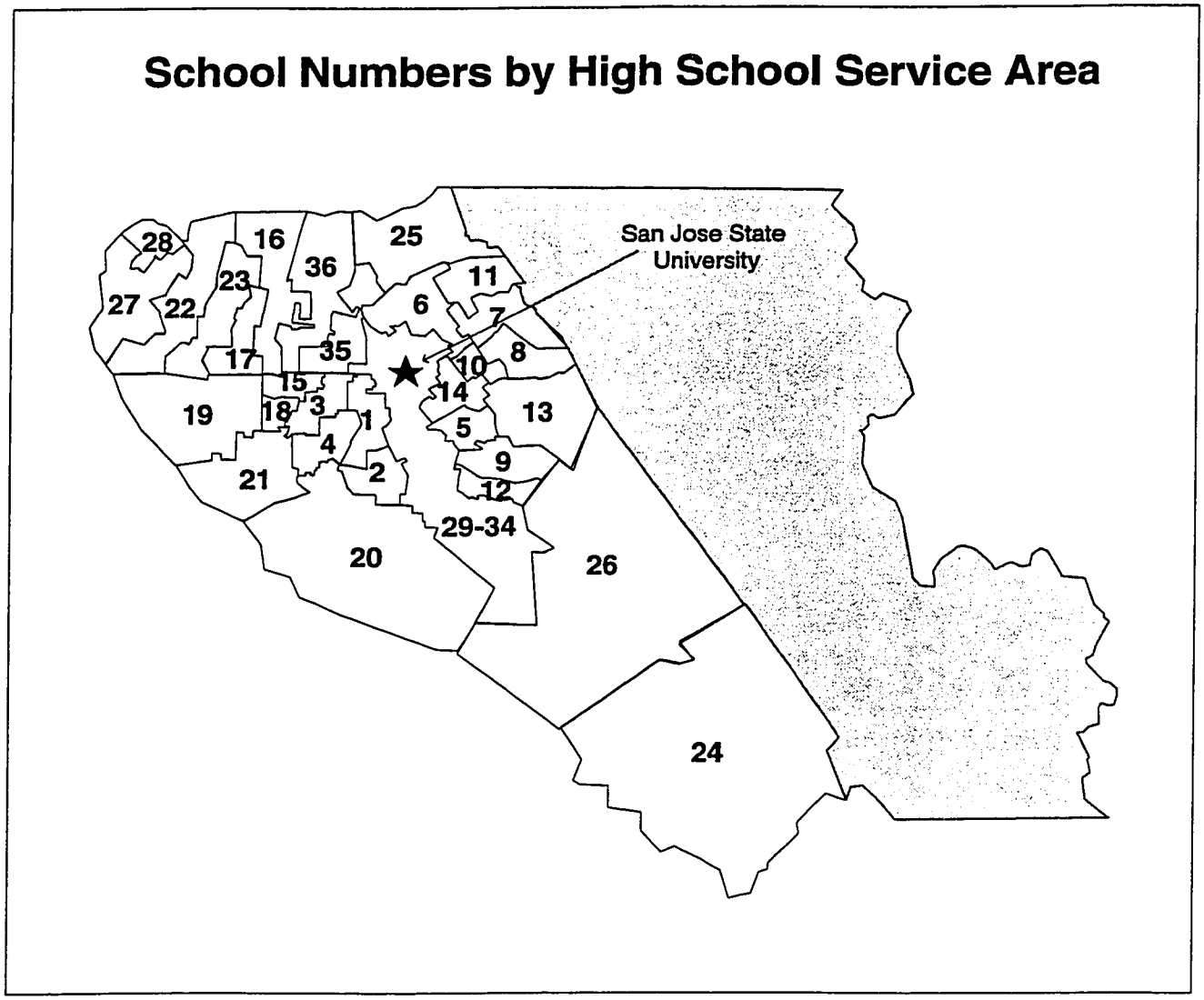
Spartan Daily. 11/6/95 Vol. 105 No. 47, "Campus Tuneup Needed."

Steel, D. G., D. Holt, and M. Tranmer. "*Modeling and Adjusting Aggregation Effects.*" Technical Report No. 94-3, March 1994.

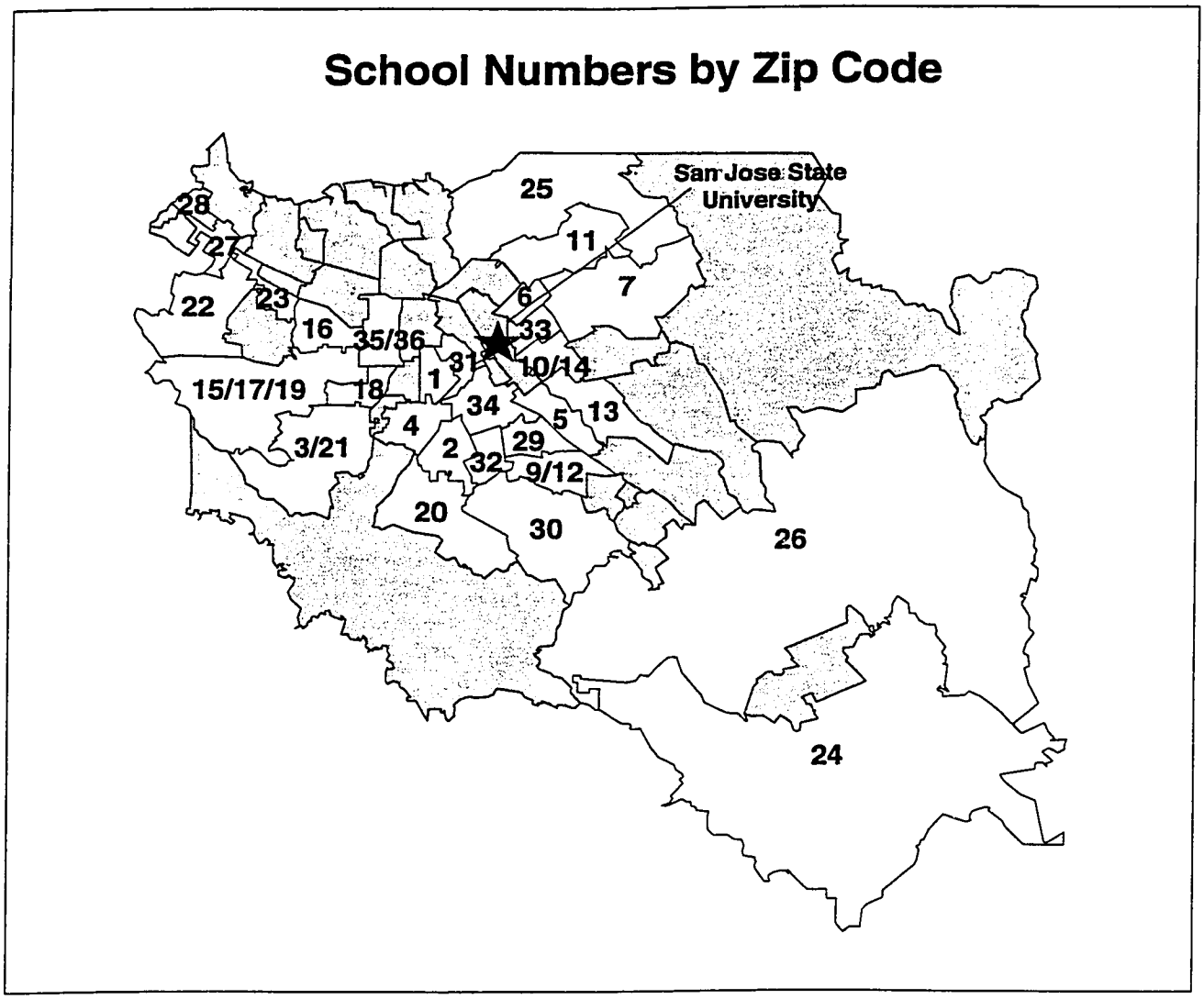
Thrall, Grant Ian, and J. C. deValle. "Calibrating an Applebaum Analog Market Area Model with Regression Analysis." *Geo Info Systems*, November. 1996, 52-55.

Wong, David. "Evaluating Spatially Aggregated Data for Socioeconomic/Policy Analysis in GIS." *GIS/LIS 1995.*

APPENDIX A


School Numbers by High School Service Area
San Jose State University
28
16
25
36
23
11
27
22
6
7
17
35
8
10
15
14
19
18
3
1
13
5
4
9
21
2
12
29-34
20
26
24

APPENDIX B


School Numbers by Zip Code
San Jose State University
25
28
27
11
7
22
23
6
16
35/36
33
31
10/14
15/17/19
18
1
34
5
13
4
3/21
2
29
32
9/12
20
30
26
24

APPENDIX C

Data Used in the Study

TABLE OF HIGH SCHOOL DATA BY DISTRICT

1989-1990

Number	High School	College Bound	SJSU First Time Freshmen
1	Del Mar	81	20
2	Leigh	95	12
3	Prospect	104	7
4	Westmont	84	2
5	Hill	106	46
6	Independence	308	98
7	Lick	71	22
8	Mt Pleasant	123	44
9	Oak Grove	144	57
10	Overfelt	53	37
11	Piedmont	115	31
12	Santa Theresa	136	41
13	Silver Creek	132	63
14	Yerba Buena	75	47
15	Cupertino	104	18
16	Fremont	118	18
17	Homestead	181	7
18	Lynbrook	173	10
19	Monta Vista	201	10
20	Los Gatos	209	19
21	Saratoga	155	5
22	Los Altos	162	10
23	Mt View	123	5
24	Gilroy	128	10
25	Milpitas	124	45
26	Live Oak	71	37
27	Gunn	185	4
28	Palo Alto	217	5
29	Gunderson	87	32
30	Leland	150	33
31	Lincoln	97	31
32	Pioneer	74	27
33	San Jose	11	13
34	Willow Glen	11	30
35	Santa Clara	146	43
36	Wilcox	129	28

Source: College Bound Report 1985-86 to 1991-92 by California Department of Education Research, Evaluation and Technology Division First Time Freshmen data for the 1990-1991 school year provided by Institutional Research, San Jose State University, San Jose, California

TABLE OF % FIRST TIME FRESHMEN AND KNOWN POTENTAIL POOL

1989-1990

Number	High School	College Bound	SJSU First Time Freshmen	% First Time Freshmen	Known Potential Pool (%)
1	Del Mar	81	20	25	75
2	Leigh	95	12	13	87
3	Prospect	104	7	7	93
4	Westmont	84	2	2	98
5	Hill	106	46	43	57
6	Independence	308	98	32	68
7	Lick	71	22	31	69
8	Mt Pleasant	123	44	36	64
9	Oak Grove	144	57	40	60
10	Overfelt	53	37	70	30
11	Piedmont	115	31	27	73
12	Santa Theresa	136	41	30	70
13	Silver Creek	132	63	48	52
14	Yerba Buena	75	47	63	37
15	Cupertino	104	18	17	83
16	Fremont	118	18	15	85
17	Homestead	181	7	4	96
18	Lynbrook	173	10	6	94
19	Monta Vista	201	10	5	95
20	Los Gatos	209	19	9	91
21	Saratoga	155	5	3	97
22	Los Altos	162	10	6	94
23	Mt View	123	5	4	96
24	Gilroy	128	10	8	92
25	Milpitas	124	45	36	64
26	Live Oak	71	37	52	48
27	Gunn	185	4	2	98
28	Palo Alto	217	5	2	98
29	Gunderson	87	32	37	63
30	Leland	150	33	22	78
31	Lincoln	97	31	32	68
32	Pioneer	74	27	36	64
33	San Jose	11	13	118	-218
34	Willow Glen	11	30	273	-173
35	Santa Clara	146	43	29	71
36	Wilcox	129	28	22	78

% First Time Freshmen = First Time Freshmen / College Bound
Known Potential Pool = 100 - % First Time Freshmen

RANK ORDER OF KNOWN POTENTIAL POOL

Number	High School	Known Potential Pool (%)	Rank Order
4	Westmont	98	2
27	Gunn	98	2
28	Palo Alto	98	2
21	Saratoga	97	4
17	Homestead	96	5.5
23	Mt View	96	5.5
19	Monta Vista	95	7
18	Lynbrook	94	8.5
22	Los Altos	94	8.5
3	Prospect	93	10
24	Gilroy	92	11
20	Los Gatos	91	12
2	Leigh	87	13
16	Fremont	85	14
15	Cupertino	83	15
30	Leland	78	16.5
36	Wilcox	78	16.5
1	Del Mar	75	18
11	Piedmont	73	19
35	Santa Clara	71	20
12	Santa Theresa	70	21
7	Lick	69	22
6	Independence	68	23.5
31	Lincoln	68	23.5
8	Mt Pleas	64	26
25	Milpitas	64	26
32	Pioneer	64	26
29	Gunderson	63	28
9	Oak Grove	60	29
5	Hill	57	30
13	Silver Creek	52	31
26	Live Oak	48	32
14	Yerba Buena	37	33
10	Overfelt	30	34
33	San Jose	-18	35
34	Willow Glen	-173	36

CASE 1

HIGH SCHOOL SERVICE LEVEL OF AGGREGATION

Case 1a – Aggregated Data for Median Family Income and % Less than High School Education from Census Tracts that are 100% within a High School Service Area

High School	% First Time Freshmen	College Bound	Distance To SJSU (miles)	Median Family Income ($)	% Less than High School
1	25	81	4	46,103	816
2	13	95	7	55,304	403
3	7	104	6	47,534	892
4	2	84	6	50,582	823
5	43	106	4	47,948	2178
6	32	308	3	50,319	4139
7	31	71	5	39,593	1466
8	36	123	5	58,348	2165
9	40	144	7	50,724	978
10	70	53	3	35,041	2885
11	27	115	6	59,139	915
12	30	136	8	60,580	338
13	48	132	7	58,349	1460
14	63	75	2	33,945	3083
15	17	104	7	54,093	390
16	15	118	9	48,988	1466
17	4	181	10	70,235	544
18	6	173	8	65,220	288
19	5	201	11	73,500	395
20	9	209	11	63,622	146
21	3	155	10	111,151	199
22	6	162	13	59,406	2023
23	4	123	11	57,768	1220
24	8	128	27	42,683	3963
25	36	124	7	51,881	1861
26	52	71	16	56,004	678
27	2	185	15	66,592	294
28	2	217	16	71,964	244
29	37	87	7	52,604	520
30	22	150	7	52,604	244
31	32	97	7	52,604	1463
32	36	74	7	52,604	139
35	29	146	5	47,394	1780
36	22	129	7	53,823	1468

Case 1b – Aggregated Data for Median Family Income and % Less than High School education from Census Tracts that are within and overlap with the High School Service Area

High School	% First Time Freshmen	College Bound	Distance to SJSU (miles)	Median Family Income ($)	% Less than High School
1	25	81	4	49,052	2870
2	13	95	7	57,593	1589
3	7	104	6	55,170	1714
4	2	84	6	56,430	1563
5	43	106	4	47,133	3819
6	32	308	3	47,285	9663
7	31	71	5	45,150	5407
8	36	123	5	60,348	2923
9	40	144	7	52,913	2023
10	70	53	3	40,826	6948
11	27	115	6	58,319	1475
12	30	136	8	58,140	1177
13	48	132	7	53,280	2927
14	63	75	2	37,572	10073
15	17	104	7	58,143	553
16	15	118	9	48,500	3613
17	4	181	10	64,794	1517
18	6	173	8	78,311	405
19	5	201	11	71,082	504
20	9	209	11	78,400	556
21	3	155	10	103,100	318
22	6	162	13	67,239	2287
23	4	123	11	62,885	1424
24	8	128	27	44,802	3963
25	36	124	7	49,000	1861
26	52	71	16	55,500	1334
27	2	185	15	66,408	332
28	2	217	16	73,689	348
29	37	87	7	50,000	778
30	22	150	7	50,000	414
31	32	97	7	50,000	3045
32	36	74	7	50,000	798
35	29	146	5	47,130	2958
36	22	129	7	51,680	2600

Case 1c – Aggregated Data for Median Family Income and % Less than High School Education from Census Tracts that are 100% within and that are n% contained within the High School Service Area

High School	% First Time Freshmen	College Bound	Distance to SJSU (miles)	Median Family Income ($)	% Less than High School
1	25	81	4	37,183	2164
2	13	95	7	31,897	781
3	7	104	6	32,385	1135
4	2	84	6	41,347	1243
5	43	106	4	38,372	2922
6	32	308	3	30,995	5518
7	31	71	5	35,486	3896
8	36	123	5	48,640	2474
9	40	144	7	40,207	1555
10	70	53	3	28,195	4838
11	27	115	6	46,970	1058
12	30	136	8	39,256	705
13	48	132	7	50,430	2528
14	63	75	2	23,766	5916
15	17	104	7	41,116	454
16	15	118	9	30,900	2229
17	4	181	10	53,947	1039
18	6	173	8	57,282	361
19	5	201	11	59,242	482
20	9	209	11	43,300	335
21	3	155	10	82,500	301
22	6	162	13	54,328	2201
23	4	123	11	53,454	1288
24	8	128	27	44,802	3963
25	36	124	7	49,000	1861
26	52	71	16	49,900	1006
27	2	185	15	60,951	313
28	2	217	16	58,135	263
29	37	87	7	37,600	631
30	22	150	7	37,600	311
31	32	97	7	37,600	1843
32	36	74	7	37,600	495
35	29	146	5	40,200	2695
36	22	129	7	34,700	1910

CASE 2

ZIP CODE LEVEL OF AGGREGATION

High School	% First Time Freshmen	College Bound	Distance from SJSU (miles)	Median Family Income ($)	% Less than High School
1	25	81	4	43,360	1270
2	13	95	6	56,053	949
3	5	129	11	95,287	337
4	2	84	6	50,570	943
5	43	106	5	43,381	3797
6	32	308	3	47,165	1728
7	33	97	6	46,615	4195
8	33	97	6	46,615	4195
9	35	140	7	54,840	1192
10	66	64	3	37,069	7266
11	27	115	6	60,015	1237
12	35	140	7	54,840	1192
13	48	132	6	52,412	2013
14	66	64	3	37,069	7266
15	9	162	12	66,826	808
16	15	118	9	62,650	832
17	9	162	12	66,826	808
18	6	173	7	61,644	659
19	9	162	12	66,828	808
20	9	209	9	1,699	296
21	5	129	11	95,287	337
22	6	162	15	96,107	150
23	4	123	12	53,458	1140
24	8	128	27	45,908	362
25	36	124	8	57,024	2407
26	52	71	19	58,452	1428
27	2	185	15	63,221	458
28	2	217	17	78,609	118
29	37	87	5	55,102	768
30	22	150	10	80,160	398
31	32	97	2	39,368	1900
32	36	74	6	30,401	882
35	26	137	6	51,852	1711
36	26	137	6	51,852	1711

CASE 3

COUNTY LEVEL OF AGGREGATION

Number	% First Time Freshmen
1	25
2	13
3	7
4	2
5	43
6	32
7	31
8	36
9	40
10	70
11	27
12	30
13	48
14	63
15	17
16	15
17	4
18	6
19	5
20	9
21	3
22	6
23	4
24	8
25	36
26	52
27	2
28	2
29	37
30	22
31	32
32	36
35	29
36	22

CASE 1

HIGH SCHOOL SERVICE LEVEL OF AGGREGATION

Case 1a – Residual Ranking and Rank Order Results

CB			CB,D			CB,D,I			CB,D,I,Ed		
Sch	RR	RO	Sch	RR	RO	Sch	RR	RO	Sch	RR	RO
4	-29	1	4	-30	1	4	-30	1	4	-28	1
3	-22	2	3	-25	2	3	-27	2	3	-23	2
23	-21	3	2	-18	3	2	-16	3.5	23	-17	3
2	-17	5	23	-17	4	23	-16	3.5	2	-13	4
21	-17	5	21	-15	5	1	-13	5.5	22	-12	5.5
24	-17	5	18	-13	6	16	-13	5.5	16	-12	5.5
27	-13	7.5	15	-12	8	15	-11	7.5	24	-10	7
22	-13	7.5	1	-12	8	18	-11	7.5	1	-9	8.5
17	-12	9.5	17	-12	8	7	-10	9	7	-9	8.5
18	-12	9.5	16	-10	10	22	-8	10.5	15	-7	10
15	-11	11.5	22	-7	11	17	-8	10.5	18	-6	12
16	-11	11.5	19	-6	13	27	-5	12.5	17	-6	12
28	-8	13.5	7	-6	13	36	-5	12.5	36	-6	12
19	-8	13.5	27	-6	13	30	-4	14	21	-3	14
1	-7	15	36	-4	15	35	-2	16	35	-2	15
36	-3	16.5	11	-3	16	19	-2	16	.	.	.
7	-3	16.5	30	-2	17	20	-2	16	.	.	.
20	-2	18	20	-1	18	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Sch = School Number
RR = Residual Ranking
RO = Rank Order

Case 1b – Residual Ranking and Rank Order Results

CB			CB,D			CB,D,I			CB,D,I,Ed		
Sch	RR	RO	Sch	RR	RO	Sch	RR	RO	Sch	RR	RO
4	-29	1	4	-30	1	4	-27	1	4	-26	1
3	-22	2	3	-25	2	3	-23	2	3	-21	2
23	-21	3	2	-18	3	2	-15	3.5	16	-16	3
2	-17	5	23	-17	4	16	-15	3.5	2	-14	4.5
21	-17	5	21	-15	5	23	-13	5	23	-14	4.5
24	-17	5	18	-13	6	1	-12	6	7	-13	6
27	-13	7.5	15	-12	8	17	-10	7	1	-11	7
22	-13	7.5	1	-12	8	15	-9	8	24	-10	8
17	-12	9.5	17	-12	8	7	-8	9	17	-9	9
18	-12	9.5	16	-10	10	36	-7	10.5	22	-7	10
15	-11	11.5	22	-7	11	30	-7	10.5	36	-6	11
16	-11	11.5	19	-6	13	27	-6	12	15	-5	12
28	-8	13.5	7	-6	13	24	-5	13	27	-3	13
19	-8	13.5	27	-6	13	35	-4	14.5	18	-2	14
1	-7	15	36	-4	15	22	-4	14.5	35	-1	15.5
36	-3	16.5	11	-3	16	19	-2	16.5	6	-1	15.5
7	-3	16.5	30	-2	17	18	-2	16.5	.	.	.
20	-2	18	20	-1	18	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Sch = School Number
RR = Residual Ranking
RO = Rank Order

Case 1c – Residual Ranking and Rank Order Results

CB			CB,D			CB,D,I			CB,D,I,Ed		
Sch	RR	RO	Sch	RR	RO	Sch	RR	RO	Sch	RR	RO
4	-29	1	4	-30	1	4	-29	1	4	-26	1
3	-22	2	3	-25	2	3	-28	2	3	-20	2
23	-21	3	2	-18	3	2	-22	3	7	-16	3
2	-17	5	23	-17	4	16	-15	4	23	-15	4
21	-17	5	21	-15	5	23	-14	5	16	-13	5
24	-17	5	18	-13	6	15	-12	6	2	-12	6.5
27	-13	7.5	15	-12	8	1	-11	7	1	-12	6.5
22	-13	7.5	1	-12	8	17	-9	8.5	22	-11	8
17	-12	9.5	17	-12	8	18	-9	85	24	-9	9
18	-12	9.5	16	-10	10	36	-8	10	17	-8	10
15	-11	11.5	22	-7	11	7	-6	11	21	-6	11
16	-11	11.5	19	-6	13	22	-5	12	18	-5	12.5
28	-8	13.5	7	-6	13	20	-4	13.5	36	-5	12.5
19	-8	13.5	27	-6	13	30	-4	13.5	15	-4	14
1	-7	15	36	-4	15	27	-3	15	35	-2	15
36	-3	16.5	11	-3	16	19	-2	16	27	-1	16
7	-3	16.5	30	-2	17	.	.	.	.	.	.
20	-2	18	20	-1	18	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Sch = School Number
RR = Residual Ranking
RO = Rank Order

CASE 2

ZIP CODE LEVEL OF AGGREGATION

Residual Ranking and Rank Order Results

CB			CB,D			CB,D,I			CB,D,I,Ed		
Sch	RR	RO	Sch	RR	RO	Sch	RR	RO	Sch	RR	RO
4	-29	1	4	-32	1	4	-32	1	4	-23	1
3	-22	2	2	-20	2	23	-20	2	23	-17	2
23	-21	3	3	-18	3	2	-18	3	7	-14	3
2	-17	5	23	-17	4	1	-15	4	18	-11	4
21	-17	5	18	-15	5.5	18	-14	5	2	-10	5.5
24	-17	5	21	-15	5.5	27	-9	7	17	-10	5.5
27	-13	7.5	1	-13	7	31	-9	7	27	-9	7
22	-13	7.5	16	-11	8	17	-9	7	19	-8	8
17	-12	9.5	17	-9	9	16	-8	10	8	-7	9
18	-12	9.5	36	-6	11	32	-8	10	36	-5	11
15	-11	11.5	27	-6	11	36	-8	10	24	-5	11
16	-11	11.5	31	-6	11	19	-6	13	1	-5	11
28	-8	13.5	15	-5	14	7	-6	13	16	-4	13
19	-8	13.5	19	-5	14	24	-6	13	21	-3	14.5
1	-7	15	22	-5	14	15	-2	15	14	-3	14.5
36	-3	16.5	7	-4	16.5	20	-1	16.5	3	-2	17
7	-3	16.5	20	-4	16.5	.3	.-1	16.5	.31	.-2	17
20	-2	18	11	-3	18	.	.	.	.28	.-2	17
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Sch = School Number
RR = Residual Ranking
RO = Rank Order

CASE 3

COUNTY LEVEL OF AGGREGATION

Residual Ranking and Rank Order Results

School No.	Residual Rank	Rank Order
27	-21	2
28	-21	2
4	-21	2
21	-20	4
17	-19	5.5
23	-19	5.5
19	-18	7.5
18	-18	7.5
22	-17	9.5
3	-17	9.5
24	-16	11
20	-14	12
2	-11	13
16	-8	14
15	-6	15
36	-2	16
30	-1	17
.	.	.
.	.	.
.	.	.