SPATIAL INTERPRETATION AND STATISTICAL MODELING OF BOSTON HIGH SCHOOL DROPOUTS

by

JENNIFER A. MILLER

B.A.S. Mathematical Sciences and Anthropology
Stanford University
(1986)

Submitted to the Department of Urban Studies and Planning
in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE
IN OPERATIONS RESEARCH

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1991

Signature of Author .....
Department of Urban Studies and Planning
May, 1991

Certified by .....
Joseph Ferreira, Jr.
Associate Professor, Urban Studies and Operations Research
Thesis Supervisor

Accepted by .....
Donald A. Schon
Chairman, Department of Urban Studies and Planning

Accepted by .....
Amedeo R. Odoni
Co-Director, Operations Research Center

SPATIAL INTERPRETATION AND STATISTICAL MODELING OF BOSTON HIGH SCHOOL DROPOUTS

by

JENNIFER A. MILLER

Submitted to the Department of Urban Studies and Planning
on May 16, 1991 in partial fulfillment of the
requirements for the Degree of Master of Science in
Operations Research

ABSTRACT

Education is one of our most cherished values, hence we offer a free public education to all, regardless of race, gender, wealth, or ability. Furthermore, the children we educate today constitute the labor force of tomorrow, and only a well-educated, skilled, and trained workforce will give the United States an edge in the increasingly competitive global marketplace. Yet, many of our urban school districts have dropout rates exceeding forty percent. In the Boston public school district, greater than one in three high school students eventually drop out. Many of these high school dropouts look forward to a life of little opportunity, high unemployment, and low wages. Although the dropout problem has received national attention, there is little agreement on how to prevent dropouts, and as yet, no comprehensive solutions have been implemented.

This study examines the high school dropout problem in a particular urban setting, Boston, Massachusetts, in an effort to determine the characteristics of dropouts versus non-dropouts, to identify groups of students who are most "at-risk", and to analyze residential patterns of dropouts in order to find the characteristics of neighborhoods in Boston with the most severe dropout problem. By identifying at-risk neighborhoods, community-specific dropout prevention programs can be targeted to those communities with the greatest need.

To analyze the high school dropout problem in Boston, this study uses a three-tiered approach by investigating: 1) the student, 2) the neighborhood, and 3) the school. Data from the Boston public school district, on the cohort group which entered ninth grade in 1982, is used to formulate and interpret a multidimensional log-linear model which describes the relationships among several explanatory variables correlated with dropping-out. Lastly, maps of student residences are analyzed to detect spatial patterns of dropouts in Boston, and to confirm visually the results of the statistical model.

Thesis Supervisor: Dr. Joseph Ferreira, Jr.

Title: Associate Professor of Urban Studies and Operations Research

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I
Introduction

A. Introduction

A public education is the right of every child, regardless of race, gender, wealth, or ability. Furthermore, the children we educate today constitute the labor force of tomorrow and only a well-educated, skilled, and trained work force will give the United States an edge in the increasingly competitive global marketplace.

However, our schools are riddled with problems: insufficient funding, deteriorating facilities, frustrated and complacent teachers, readily available drugs, and widespread violence in the schools. Recently the public outcry over the deteriorating quality of public education has gained national attention, and an appallingly high dropout rate is perhaps the most oft-quoted evidence of the decline in our public school systems.

Many of our teenagers consider high school a punishment, not a right, and they seek voluntary parole as soon as possible. To them, education is a nuisance, not a value. The dropout problem is worst in our urban school systems such as Los Angeles, Chicago, Detroit, and New York which all have dropout rates exceeding forty percent (Grossnickle 1986).

Numerous researchers have studied the dropout problem, yet there is little agreement on a solution. Even the Network of Urban Superintendents cannot agree on the best way to combat the dropout problem (Paulu 1987). Some superintendents believe that a back-to-basics education with higher academic expectations and disciplinary standards will reduce the number of dropouts. However, others argue that this will push borderline students out-the-door, and that a more caring, flexible, individualized school atmosphere is needed to encourage these students to stay in school.

Currently most dropout prevention programs are local in scope and offer a variety of piecemeal solutions. The effectiveness of these programs has yet to be determined, and

no programs have been implemented on a national scale. Thus although the dropout problem has been studied extensively, we still do not understand it well enough to fix it.

B. Research Questions

This thesis will examine the high school dropout problem in a particular urban setting, namely Boston, Massachusetts. Boston is no exception to the dropout problem in urban school districts. In Boston, the annual dropout rate peaked in 1984-85 at 17.2 percent, and has since decreased, but remains above ten percent (Horst 1990). The cohort dropout rate ranges between 33 and 40 percent which means that less than 60 percent of students who start ninth grade in a given year will finish high school four years later (Horst 1990). Thus one out of every three high school students in Boston will drop out.

Often the phenomenon of high school dropouts is characterized solely as an issue of race. But is race the only explanatory factor, or does it mask other important variables such as socio-economic status? This thesis will investigate the determinants of dropping-out in Boston in order to answer the following questions.

1) What are the characteristics of dropouts versus those who finish school normally in the Boston public school district?
2) Are there groups of students who drop out in mass (i.e., have an extremely high dropout rate) and if so, what are their characteristics?
3) Are high school dropouts concentrated in particular areas of Boston, and if so, what are the characteristics of those neighborhoods with higher dropout rates?

C. Research Design

To study these questions, we have data from the Boston public school district on the cohort group which entered ninth grade in September, 1982. These students normally would have graduated in June 1986, however some students dropped out, others

transferred out of the district, and some required a fifth year of high school to finish. The data set includes one record for every student in the cohort group for each year the student was enrolled in the Boston public high schools. Each record contains assorted demographic data on the student such as race, gender, age, and address, and administrative data such as whether or not the student is enrolled in the bilingual or the vocational programs. Lastly, each record contains a progress code indicating the student's status at year's end which may be: normal, dropout, probable dropout, transfer, or other loss such as death or expulsion.

This thesis approaches the dropout problem from two perspectives: 1) the statistician, and 2) the policy analyst. From the perspective of the statistician, this thesis represents an application of formal statistical techniques to an interesting data set on high school dropouts. In contrast, from the perspective of the policy analyst I hope to understand the mechanisms underlying a real-world problem, the high school dropout phenomenon, in order to propose better solutions and thereby inform practice. Fortunately, these two different approaches are more symbiotic than mutually exclusive, and hopefully this application of statistics to the dropout data will improve our understanding of the high school dropout problem in Boston.

This research has two parts: 1) statistical analysis and modeling of the dropout data, and 2) spatial analysis of the dropout problem in Boston. In the first part of the research, I formulate a statistical model, specifically a log-linear model, which describes the relationships among several qualitative explanatory variables for dropping-out. The model takes a three-tiered approach to analyzing the dropout problem by looking at the individual student, the neighborhood, and the school. Thus the model includes individual characteristics such as race, gender, and age. By linking the individual students to neighborhoods through student addresses, we can incorporate census data into the model and thus use type of neighborhood as the best approximation for individual socio-economic status. Finally, the type of school is included in the model to differentiate among public

schools in Boston. After formulating the model, I interpret the results and identify the characteristics of students with especially high dropout rates.

The second part of this research is the spatial analysis of the dropout problem in Boston. This consists of interpreting maps of selected student groups in Boston and identifying spatial patterns in the city. I also map dropout rates by census tract to identify neighborhoods in the city with the most severe dropout problem.

In the future, these "at-risk" neighborhoods of Boston could be targeted for community-specific dropout prevention programs. In the era of a recessionary economy and reduced government spending, it is crucial to make social programs efficient as well as effective. Hopefully this research will provide better identification of at-risk neighborhoods and thereby allow more precise targeting of social programs to those areas with the greatest need.

D. Research Findings

This thesis will show that dropping-out is significantly correlated with all five explanatory variables examined in detail in this study: the student's race, gender, age, neighborhood, and school. We will find that males drop out more often than females. Over-age students, defined as at least two years older than normal upon entrance to ninth grade, drop out significantly more than normal-age students; in fact, more than seventy percent of over-age students drop out. The examination school students rarely drop out, whereas the magnet and regular high schools have higher dropout rates. The special-needs schools, which cater to students with behavioral problems, have the highest dropout rates.

In terms of socioeconomic status, this thesis will show that the students from the poorest neighborhoods drop out the most. Students from the well-off homeowner and yuppie-renter communities drop out the least, with students from the average working-class neighborhoods somewhere in between. Furthermore, we will find a strong interaction

between race and neighborhood. In particular, the statistical model indicates that the well-off and average neighborhoods are primarily White and Asian, while the poorest communities are Black and Hispanic.

In analyzing the bivariate relationship between dropping-out and race, we will find that Hispanics drop out the most. Whites drop out slightly more than Blacks, and Asians drop out the least. However, when we control for other variables, and in particular study the experience of the prototypical public high school student in Boston, we will find that the White students are the most likely to drop out. In this case, the Hispanic students have the next highest dropout rate, followed by Blacks, and finally Asians.

After formulating a statistical model which describes the dropout phenomenon, we will then examine maps of student residences in Boston and interpret any apparent spatial patterns. By and large, the maps will confirm the results of the statistical model. Lastly, we will map the residences of a few groups of students with high dropout rates indicated by the model. These maps will indicate that there is some evidence of dropout clustering in Boston, but more investigation is needed.

E. Chapter Summary

Chapter Two reviews the literature on high school dropouts, in particular recent research on who drops out and why, the implications of dropping-out of high school, and the current practice in dropout prevention programs.

Chapter Three contains a description of the Boston dropout data and a discussion of the cohort and dropout definitions used in this research. This chapter also discusses the explanatory variables used in the statistical model. Finally, the chapter concludes with an explanation of address-matching and its use in determining the census tract for each

student, and in calculating the longitude and latitude coordinates of student residences for maps.

Chapter Four describes the clustering of census tracts based on four socioeconomic variables which measure wealth, family structure, type of housing, and educational attainment.

Chapter Five is an introduction to the statistical methodology of log-linear models which are used to describe categorical, as opposed to continuous, data. This chapter discusses log-linear model specification, parameter estimation, and goodness-of-fit.

Chapter Six includes frequencies which illustrate the structure of the underlying student population, and two-way models which test for independence between dropping-out and race, gender, age, neighborhood-type, and school-type.

Chapter Seven describes two higher-order log-linear models: a four dimensional model which does not include the age variable, and a five-dimensional model which does include age. The models are compared based on the goodness-of-fit and the number of higher-order terms, and the model including age is selected for further analysis.

Chapter Eight interprets the unsaturated five-dimensional log-linear model which best fits the dropout data by examining the direction and magnitude of the effects, and the differential in dropout rates across student groups.

Chapter Nine contains the spatial analysis of the dropout data. In particular , we examine the maps of tract clusters, dropout rates by tract, students by race and by school-type, and selected groups of students with high dropout rates.

Chapter Ten summarizes the results and concludes with suggestions for further research.

Chapter II
Literature Review:
High School Dropouts

Education is often considered a vital stepping-stone to a successful career and therefore a happy, productive life. Furthermore, an education is the right of every child regardless of race, gender, or wealth. Yet in today's urban schools, the educational reality is a far stretch of the imagination away from this utopia.

> "We have not realized the American ideal of a free public education for all. Our nation is clearly 'at risk' when large numbers of students leave before taking advantage of the opportunities schools have to offer." (Grossnickle 1986)

Many of our children choose to leave high school without graduating which raises several questions: how many students leave, why do they leave, what are the consequences of leaving high school without a diploma, and what can we do to prevent this national tragedy? This chapter will review the literature on high school dropouts, specifically:

A. the magnitude of the dropout problem,
B. the consequences of dropping out,
C. who drops out and why,
D. dropout prevention programs,
E. dropout definition and data issues.

A.The Magnitude of the Dropout Problem

The phenomenon of high school dropouts is not new to the United States, although the intense government and media attention on dropouts in the last few years suggests that the dropout problem has worsened. In fact, the dropout rate has steadily decreased over the last hundred years (Grossnickle 1986).

Year	Nationwide Dropout Rate
1900	90%
1930s	66%
1950	41%
1970s	28%
1980s	27%

Even though the dropout rate has declined over the years, it remains significantly high (almost thirty percent). There are two interpretations of this historical trend in the dropout rate. First, the decrease in the dropout rate to twenty-eight percent in 1970 and its subsequent levelling off suggests that we, as a nation, have reached the natural limit for the number of people capable of pursuing formal education (Catterall 1985). Thus we have done what we can and should not worry about those people who simply cannot succeed in our educational system due to some inherent personal defect. The second, and perhaps more optimistic, interpretation is that the current dropout rate, although lower than in times past, is unacceptable and represents a great under-utilization of our human resources which is detrimental to society as well as to the individual dropouts (Catterall 1985). According to this interpretation, attrition from high school is definitely a national concern.

Given this historical perspective, what are the dimensions of the high school dropout phenomenon today? The numbers are actually quite staggering. Nationally in 1986 there were approximately three million graduates and one million potential graduates who dropped out (Grossnickle 1986). In 1985-1986, 682,000 students dropped out of high school nationwide. Therefore we can estimate that 3,800 students drop out of high school every school day (Paulu 1987). About forty-five percent of Massachusetts' Hispanic students will drop out before finishing high school, and these students are concentrated in the most impoverished urban centers which, in general, have lower quality public school systems as compared to the more affluent and less diverse suburban public schools (Ribadeneira 1990).

In fact, the dropout problem is worst in urban school systems and future demographic changes will only exacerbate this problem as the urban minority populations, which are the most prone to dropping out, grow. For instance, Chicago had a dropout rate greater than forty percent in 1986, and in New York City, forty-five percent of students in the 1980 cohort were no longer in school by their senior year (Grossnickle 1986).

In Massachusetts, the state defines "at-risk" as any student who fails one or more of the State Basic Skills Tests on reading, writing, and math. In 1987, forty-five percent of all Boston Public School ninth graders were at-risk whereas the statewide average was twenty-three percent (Citywide Educational Coalition 1988). Moreover, twenty-five percent of ninth graders in the Boston Public Schools were not promoted to tenth grade in 1988 while the national non-promote rate is around five percent (Citywide Educational Coalition 1988). Perhaps most appalling, in 1987 Boston public school students took the Metropolitan Achievement Test (MAT), and "six out of ten students in non-exam high schools scored below the 40th percentile, meaning that they cannot read their textbooks with understanding" (Citywide Educational Coalition 1988). Thus the dropout problem is definitely significant, and the task of educating our youngsters is most worrisome in urban centers.

Although these statistics are quite shocking, critics argue that the high school diploma is an arbitrary cutoff, and as such, should not receive such critical examination. Is the high school diploma an arbitrary cutoff for defining someone as undereducated? Should the requirement instead be eighth grade completion, or possibly college graduation? In the 1960s several national campaigns arose to combat the dropout problem under the assumption that a high school diploma is a minimum requirement for an adequate education (Pallas 1984). In general, high school graduation is considered a reasonable measure of satisfactory education because it is the culmination of publicly funded education and there are certain public expectations of a minimum set of skills required to enter the workplace or to go on to higher education (Catterall 1985). "Every school district attends at least rhetorically to the goal of guiding all of its children to the diploma. This distance between this expression and reality suggests and identifiable gap in our public performance. (Catterall 1985)" The high school diploma symbolizes that a student has acquired this minimum set of skills needed for future success, hence high school graduation is a useful measure of satisfactory educational attainment.

The magnitude of the dropout problem is well illustrated by a seed analogy from Grossnickle (Grossnickle 1986). Students can be thought of as seeds which need to be planted in a safe, protected, nourishing environment. They need to be nurtured and fertilized to grow. Lastly, there is great sadness in the lost potential of a seed (student) which does not sprout (succeed).

B. The Consequences of Dropping Out

Clearly if there were no negative consequences associated with dropping out of high school, then it would not be considered a problem. Indeed, there are substantial costs for both the individual and society.

On the individual level, dropouts suffer from a higher unemployment rate, periodic loss of employment, and lower-paying occupations (Catterall 1985). Dropouts often have problems with the law. For example, seventy to ninety percent of the prison populations in some states are high school dropouts (Grossnickle 1986). High school graduates, on the other hand, tend to earn more money, have jobs with greater prestige, and are less likely to be unemployed (Pallas 1984). In a follow-up study of high school graduates and dropouts in Illinois in the early 1960s, Gallington found that only eight percent of graduates had poor pay compared to seventy-two percent of dropouts (Gallington 1966). Moreover, the contribution of education to earnings is not linear because graduation from high school gives far greater advantage in the workplace than the completion of any other single year in high school (King 1978). The picture is grim even for those dropouts who eventually complete a GED (general education documentation) because recent research shows that employers prefer job applicants with high school diplomas (Paulu 1987).

Furthermore recent changes in the job market are detrimental to dropouts. Dropouts usually work in low-skill jobs, but the number of low skill jobs is decreasing as our economy shifts from manufacturing to service-based. According to former-Governor

Dukakis at a 1990 symposium on Hispanic education, eighty-five percent of the jobs created in the next decade will require at least two years of education beyond high school (Ribadeneira 1990). Clearly we have a problem when only two-thirds of the student population graduates from high school. More education is needed now than ever before to compete in today's technologically advanced job market.

In addition to individual consequences, the societal costs of dropping out are staggering. The United States needs a well educated, trained workforce to compete in the increasingly competitive global markets. But dropouts generate less income over their working lifetimes, and there is a shortage of skills for production and job-training. As a result, our nation's productivity and output is smaller than it could be, and lower national income reduces tax revenues. Ironically, dropouts often require welfare/unemployment subsidies and health services which are funded through these reduced tax revenues. Another requirement for a successful society is a responsible citizenry, but dropouts participate less in elections and other civic activities. Grossnickle sums it up well: "a high school education should be considered the minimum survival kit for coping with today's world" (Grossnickle 1986).

It is actually possible, albeit difficult, to quantify the societal costs of youth dropping out of high school. In 1972, Levin studied the lost economic activity due to non-completion of high school on a national scale (Catterall 1985). Based on 1968 Census income data for 25 to 34 year old males, he estimated that the total loss of lifetime earnings due to dropping out is $237.6 billion (in 1968 dollars). The corresponding loss in tax revenues was $71 billion. Catterall performed a similar study in 1981 to update these figures taking into account changes in earnings patterns, price and income levels, and dropout rates (Catterall 1985). He estimates that in 1981, male dropouts earned $266,000 less and female dropouts earned $199,000 less than their counterparts over their working lifetimes. This translates into a 228 billion dollar loss in earnings nationwide and 68 billion

dollars in lost tax revenues for every high school class. Clearly our nation's productivity and economic health is compromised by the lost potential of high school dropouts.

Although the dropout problem has significant economic consequences, the amount of resources devoted to solving the problem is relatively quite small. For example, the Los Angeles School District's response to a grim study on dropouts was to begin a $1 million per year pilot program in 21 of 600 schools (Catterall 1985). The school district allocated only one-half of one percent (0.005) of its financial resources to a problem which affects almost half of its students (since the dropout rate in Los Angeles is between forty and fifty percent). Hence there is a puzzling discrepancy between the magnitude of the dropout problem and the amount of resources allocated to address it.

What are the reasons for this discrepancy between the economic loss of dropouts and the resources devoted to solving the problem? According to Catterall, the costs of dropouts are either underestimated or unappreciated by the necessary institutions (Catterall 1985). Also, addressing the dropout problem requires resource allocation now, but the benefits will not be realized for several years into the future. In this era of tight budgets, expenditures without immediate results are often questioned. Lastly there is a lack of consensus on the best strategy to prevent dropouts, and any large-scale program requires a strong belief that it will work.

C. Who Drops Out and Why

There have been several studies attempting to identify the characteristics of dropouts so that schools can intervene and help at-risk children as early as possible. In 1978, Lloyd found factors which can discriminate between dropouts and graduates with seventy-five percent accuracy as early as the third grade (Capuzzi and Gross 1989). The four factors he identified were reading achievement scores, IQ tests, socioeconomic background and family characteristics, and retention in the first three years of school. Pallas analyzed data

from the 1980 High School and Beyond national survey of high school sophomores and 1982 follow-up study of seniors and dropouts (Pallas 1984). He developed a model which expresses the probability of dropping out of high school as a function of an individual's social background characteristics, academic performance, social disability, and accelerated role transitions. Research indicates, however, that not all dropouts are the same. From extensive interviews with dropouts, teachers, counselors, and school administrators, Byrne identified four different types of dropouts: structural, contagious, immigrant, and capable (Byrne 1988).

Although not all dropouts are the same, there is a common set of characteristics which is consistent across many of these studies (Byrne 1988, Capuzzi 1989, Frase 1989, Gallington 1966, Horst 1990, Pallas 1984). In general, minority and disadvantaged students have the highest dropout rates. Hispanics drop out most often, followed by Blacks, Whites, and finally Asians. Males are slightly more prone to dropping out than females. Low socioeconomic status is a good indicator of future dropout rates, as well as a non-English speaking home environment, full-time employment in high school, and early marriage or pregnancy. Often dropouts have behavioral problems such as misbehavior in school, truancy, and delinquency. But according to Paulu, "poor academic performance is the single best predictor of who drops out" (Paulu 1987). Thus low grades, low test scores, and grade retention are excellent indicators of at-risk youth.

Even though dropouts usually have these characteristics, the question remains: why do they drop out of school? Capuzzi remarks that a list of "red-flag" characteristics of dropouts has limited usefulness. "This type of identification, however, touches only the surface of the problem. The causal factors that lead to these behaviors must be addressed if intervention programs are to be effective" (Capuzzi and Gross 1989).

The prevailing theory is that students drop out of school in a last ditch attempt to escape failure (Grossnickle 1986). They face constant negative feedback in terms of low grades and test scores, they feel that nobody cares either in school or at home, and finally

they feel overwhelmed and give up school to pursue success in other walks of life. A complementary theory is the social disability perspective (Pallas 1984). This theory postulates that dropouts are essentially misfits who fail to adjust personally or socially as expected in the transition to high school. Thus delinquency is both a cause and a consequence of dropping out. A third theory is the accelerated role transitions hypothesis (Pallas 1984). It states that youth who assume adult roles while still in high school have a greater propensity for dropping out because of the difficulty in managing a complex set of roles. Therefore students who work full-time, get married, or have children, will leave school early in order to resolve the conflict between the adolescent and adult roles in their lives. Furthermore, cultural groups have different time-tables for these role transitions which explains, according to this theory, the observed variability in racial/ethnic dropout rates.

D. Dropout Prevention Programs

Dropout Prevention is a field in its early infancy. As stated previously, there have been many studies examining the characteristics of dropouts, a few studies analyzing why students drop out, and even fewer studies researching effective dropout prevention techniques. There has been little objective research on whether dropout prevention programs are truly effective, and as yet, no national solutions have been proposed (Gainer 1987). In 1987, the General Accounting Office surveyed 1,015 high school dropout programs nationwide and found that "most of the surveyed programs have not been independently evaluated and, therefore, there is a lack of definitive evidence to prove what works" (Gainer 1987). This does not mean, however, that nothing is being done about the problem. There are perhaps as many local dropout prevention programs throughout the

United States as there are different types of dropouts, and "the survey's results provide information about programs that almost all local administrators perceive as effective" (Gainer 1987).

Currently there is a debate over the effects of back-to-basics education on high school dropouts. In the past few years, there has been a push for back-to-basics education with more stringent requirements, longer school days, and higher academic and disciplinary standards, in an effort to improve the quality of education in our public school systems. Some researchers fear that a more rigid school system will actually push borderline students out the door and exacerbate the dropout problem (Grossnickle 1986). They believe that at-risk students on the edge, who already have serious academic difficulties, will be unable to cope with stricter requirements, and will leave school as a result. The alternative view is that students perform according to the standards we set (Paulu 1987). If we expect little, we get little; whereas if we have high expectations for our students, they will perform well in school. It remains to be seen which theory will prevail.

Conventional wisdom suggests that there is no one solution to the dropout problem because different students have different needs: "what works for a bored but gifted youngster from the Bronx may be inappropriate for a chronically truant adolescent from Portland, Oregon" (Paulu 1987). Rather, any successful dropout program must be cooperative in nature with help from parents, teachers, school administrators, community agencies, industry, and government. Thus most dropout prevention programs provide a myriad of services to the student, and perhaps most importantly, try to create a flexible, supportive, enthusiastic environment for learning. The programs generally have multiple objectives (Gainer 1987):

--improve academic performance
--create positive attitude toward school
--reduce absenteeism
--provide job training/placement
--provide prenatal care/parenting support services.

To satisfy these objectives, dropout programs typically provide a wide range of services including (Gainer 1987):

--basic education
--career counseling
--encourage parental involvement
--job search assistance
--job skills training
--part-time employment/placement
--pregnancy counseling
--GED preparation
--day care
--instruction in English as a second language.

Many dropouts complain about the lack of attention and apparently uncaring teachers and administrators in the public schools (Byrne 1988). Recent evidence supports this criticism of the public schools. For instance in New York City, "researchers estimated that the average students received less than ten minutes a day of individual attention from instructors" (Gainer 1987). Thus the General Accounting Office survey found that a caring and committed staff, a secure environment, small classes, and personalized instruction were critical to the success of any dropout program (Gainer 1987).

To be successful, dropout prevention programs must overcome some serious obstacles. First, many students are far behind academically before they even get to high school, and it may be very difficult, if not impossible, to catch up and graduate on time. It is often too late to treat the dropout problem successfully in high school, hence the importance of early identification and intervention (perhaps even in pre-school). Second, programs may require overall school improvements to deal with problems such as bad facilities, drugs, and violence in the schools. Third, program implementation is not trivial. Programs may be ineffective, for example, until they are fully operational and it may take longer to provide services than originally planned. It is not easy to coordinate the activities of several agencies, and it may be hard to find a staff committed to helping difficult students. Finally, the image of the program is extremely important. No dropout program is successful if it stigmatizes youngsters as dumb or no good. Many potential dropouts

already have low self-esteem and labelling them as inferior only causes more problems. For these reasons, solving the dropout problem is not easy and should not be taken lightly.

The major thrust in dropout prevention is currently local in scope. The general accounting office survey describes numerous local programs which are tailored to individual communities and situations (Gainer 1987). Many urban superintendents support the local approach of developing strategies which are unique, and most appropriate, for a given community (Paulu 1987). Yet, some other urban superintendents argue that a major restructuring of public education is required to eliminate the dropout problem, as opposed to a band-aid here or a q-tip there (Paulu 1987). Since there is currently little objective evidence regarding the effectiveness of local programs, this question remains open. In any case, all involved agree that something must be done to help the dropouts, hence the "urban superintendents' call to action" (Paulu 1987).

E. Dropout Definition and Data Issues

Throughout this discussion, I have used the apparently innocuous term "dropout", yet this term is much more complicated and ambiguous than at first glance. Currently there is no standard definition of dropout, and as a result, federal, state, and local officials often use quite different definitions in their analyses of the dropout problem. Moreover,

> "estimates of the dropout rate can vary to a surprising degree depending on how 'dropout' is defined and measured....As yet, there is no agreed-upon method for estimating the problem. The reader is cautioned against comparing dropout rates from different sources." (Capuzzi and Gross 1989)

In calculating the dropout rate, many states differ on whether they include transfers to private schools, military enlistees, students completing GEDs, students educated at-home, or expelled students. The definition of "dropout" ranges from a list of typical characteristics to "broad projections of what students may be at risk for" (Capuzzi and Gross 1989). There is no national standardization of the definition of a dropout, the time

period during which data is collected, the method of data collection, the calculation of dropout rate, or the tracking of dropouts to see if they later return to finish high school or get a GED. In Massachusetts for example, "annual dropout rates reported by the state differ from those reported by BPS [Boston Public Schools] in part because October 1 enrollment is used instead of cumulative enrollment" (Horst 1990). Thus, comparing data across local, state, or national jurisdictions is problematic.

Some researchers argue that the term "dropout" is too strict because there are many different reasons for leaving high school. "For many students, dropping out is not so much an event that occurs at a specific point in time, but a process representing a gradual disengagement from school over time" (Frase 1989). Therefore the term "dropout" may be too restrictive and should be further disaggregated. For instance, instead of "dropout" we can analyze students who "fade out", are "pushed out", or are "pulled out" to understand their differences (Capuzzi and Gross 1989).

The data collection methodology may also affect dropout rates. If surveys are used, the survey respondents may overexaggerate the educational attainment of household members. Does high school completion mean receipt of a diploma only, or should a GED also be included? Clearly the answer will affect the dropout rate calculation. Also, if the measurement method changes, how can new figures be compared to the old figures? Another concern is systematic bias introduced by the experimental design of a study. For example, many national studies of dropouts use data from the High School and Beyond survey of 1980 sophomores and 1982 seniors/dropouts. Yet it has been shown that ninth graders have a significantly higher dropout rate than any other high school class (Byrne 1988). Thus the High School and Beyond survey ignores a large segment of the dropout population, specifically those students who dropped out freshman year in 1979 and therefore are not available to be counted as sophomores in 1980. Hence the analysis of dropouts involves some sticky methodological issues which may affect the results.

After the term dropout is suitably defined, there are two primary methods for computing dropout rates. First, the annual rate is the number of students who drop out in a given year divided by the total number of students. This gives a snapshot look at what is happening during the school year (Horst 1990). Second, the cohort rate is the percent of students entering ninth grade in a specific year who graduate at the end of a given time period (usually four years). The Boston Public Schools use a five-year period to count, as graduates, those students who were held back a year in high school but did go on to finish by the end of the fifth year (Horst 1990). In contrast to the annual rate, the cohort rate gives a long term view of the "holding power" of schools (Horst 1990).

Byrne demonstrated that the denominator (total number of students) used in either calculation may seriously affect the results (Byrne 1988). First, if students who transfer out of the school district are included, the dropout rate is artificially decreased. Second, if students held back in the ninth grade are included in the following year's cohort, this inflates the cohort rate because these students are much more likely to dropout than their counterparts. The following chapter discusses the specific dropout definition and the methodology used to compute dropout rates in this analysis of Boston high school dropouts.

Chapter III
Setting The Stage

This chapter sets the stage for the statistical analysis of the high school dropout problem in Boston. In particular, this chapter includes a description of the database, the cohort group, and the definition of dropout used in the analysis. Then the three levels of explanatory variables (individual, neighborhood, and school) are described, and the chapter concludes with a discussion of the use of address-matching in generating census tract numbers and coordinates for mapping student residences.

A. Description of the Database

The data used in this research was acquired from the Boston public school district in 1988. It consists of records for all students enrolled in the Boston public high schools during the years 1982-83 through 1986-87, plus an additional, more limited set of records for students enrolled during 1981-82. The data set contains one record for every student for each year that he or she is enrolled in the Boston public high schools. There are approximately 5,000 students in each high school class, hence in any given year there are 20,000 students enrolled in the Boston public high schools (5,000 students per class * 4 classes (9th-12th grade)). Since this data covers a five-year period, it includes 120,000 records (20,000 students per year * 5 years). This five-year time span is used because, in contrast to examining annual dropout rates, the purpose of this analysis is to examine the progress of one particular cohort group, students first enrolled in ninth grade during 1982-83, through the Boston public schools. The data from 1981-82 was used to determine which students entered ninth grade for the first time in 1982-83.

The records provided by the school district contain as much information about the individual student as possible without compromising confidentiality. Each record contains demographic, administrative, and progress data. The demographic data includes the

student's gender, race, and age. Some of the information is administrative (and could also be considered demographic) such as the student's address, which school he attends, and whether or not he is enrolled in a bilingual, special education, or vocational program. Finally, the records contain progress data including a code for the student's progress during the year (normal, transfer, dropout, probable dropout, and other loss due to death or expulsion), and the date of withdrawal if applicable. In order to preserve student confidentiality, the data does not include student names, grade point averages, or test scores. The unique identifier for each record is the student's number.

Processing a data set of this size (120,000 records) without computers would be an administrative nightmare, if not impossible, and even with computers it is not a trivial task. To process the data, we used a relational database management system, specifically Ingres, which implements SQL (structured query language) to perform database manipulations such as queries and updates. The original data tables were created by Aurelio Menendez, a research assistant for Professor Joseph Ferreira, in association with Gregory Byrne's longitudinal study of Boston high school dropouts (Byrne 1988).

The database was originally composed of four main tables created by Aurelio: SCHOOL, SCHOOL3, PERSONAL, and RECORD. SCHOOL is the main table which contains all 120,000 records supplied by the Boston school district. Thus it contains one record for every student for each year that he/she is enrolled in the 1982-83 through 1986-87 time period, and each record contains the demographic, administrative, and progress data described earlier.

Since we are specifically interested in the cohort group which entered ninth grade for the first time in 1982-83, we need to narrow down the data set. The SCHOOL3 table contains only the students in the cohort group (5,393 students). SCHOOL3 is a summary table which has the student's number, the grade (9th, 10th,..), and the student's status at the end of each year. For instance, a student who completed high school normally in four years would have the following entries:

1982-83/Status	1983-84/Status	1984-85/Status	1985-86/Status	1986-87/Status
9th / Normal	10th/Normal	11th/Normal	12th/Normal	----------

On the other hand, a student who was held back in the ninth grade and subsequently dropped out the second time through, would have the following record in SCHOOL3:

1982-83/Status	1983-84/Status	1984-85/Status	1985-86/Status	1986-87/Status
9th/Normal	9th/Dropout	----------	----------	----------

The codes for a student's status are: normal, transfer, dropout, probable dropout, and other loss (death, expulsion). Hence in the case above, 'normal' progress is somewhat misleading because the student was held back to repeat the grade; however the student was still enrolled in school and thus is considered 'normal' at the end of 1982-83. (This methodology is used by the Boston Public School District.) From this table, we can determine a final outcome for every student in the cohort.

The remaining two tables: PERSONAL and RECORD contain a subset of the demographic and administrative information for each student in the cohort group. The PERSONAL table has the student's number, year of birth, sex, and race, and the RECORD table has the student's number, school, and address.

For this research, I used the tables previously created by Aurelio and generated two new tables: COHORTADD and SCHOOLDATA. COHORTADD contains the addresses and latitude - longitude coordinates for all students in the cohort group whose addresses could be successfully matched. These coordinates are then used to produce maps of student residences (see chapter IX). The SCHOOLDATA table contains the individual observations of the explanatory variables which are described later in this chapter. This includes data from the original tables (race, gender, age, school), and census data used to cluster tracts.

Finally, the BOSDIME table, which contains a street network file for Boston in DIME format (provided to MIT by Geographical Data Technology), is used for address-

matching student residences. This table consists of records which identify all valid street segments in Boston, and thus is used to match student addresses against valid street addresses to compute latitude and longitude coordinates for subsequent mapping.

B. Cohort Definition

In the following analysis, the definition of cohort was chosen to be as consistent as possible with the definition used by the Boston public school district (Horst 1990). The 'cohort' consists of all students who entered the ninth grade for the first time in September 1982. These students would normally graduate in June 1986, however we have data for the 1986-87 school year to track any students who were held back to repeat a grade during high school. The students who are in ninth grade for the second (or third, fourth,...) time in 1982-83 should be included in the previous year's cohort group, hence they are excluded here. Anyone who transfers into the district during the five-year period is not considered part of the cohort group under the Boston school district's definition.

Two groups of students have been omitted from the cohort group used in the statistical analysis. First, Native American students are not included in the model because there are too few of them to draw generalizable conclusions (22 out of 5,393 students are Native American), and they would be too sparse in the multi-dimensional contingency table. Second, students whose residences could not be address-matched have been excluded (453 out of 5,393 students) because it is impossible to determine in which census tracts these students live. (See the section in this chapter on address-matching for more detailed information.) Therefore the cohort group analyzed in the statistical analysis is somewhat reduced.

C. Dropout Definition

The definition of dropout is also similar to that used by the Boston public school district (Horst 1990). Each student in the cohort group is assigned one of five final outcomes depending on his or her last known status. In other words, the last non-blank entry in the SCHOOL3 summary table is the student's final status (a blank entry in the SCHOOL3 table means that the student is no longer enrolled in the school district).

The five possible outcomes are:

Final Outcome	Number of Students
Normal	2,585
Transfer Out	894
Dropout	1,661
Probable Dropout	173
Other Loss (death, expelled)	80
Total (Entire Cohort):	5,393

A dropout is defined as any student whose final status is 'dropout' or 'probable dropout', whereas a non-dropout is any student whose final status is 'normal'. (Note that 'normal' does not necessarily imply that a student graduated, but that his last known progress was normal. This terminology is taken directly from the Boston public school district (Horst 1990).)

The treatment of cohort students who transfer out of the district is an interesting issue. Usually the cohort dropout rate is calculated as the number of dropouts divided by the number of students in the cohort. Byrne demonstrated that the inclusion of transfers as non-dropouts artificially decreases the dropout rate because the denominator is larger than it should be (Byrne 1988). The argument is that the final outcome of transfer students is unknown, hence we should not assume that they finish high school normally. Thus Byrne advocated excluding transfers from the denominator in dropout rate calculations, and the Boston public school district has since changed its practices to follow his suggestion (Byrne 1988, Horst 1990).

Therefore, in order to be consistent with current Boston school district practices, I have excluded transfers from the statistical model. I have also excluded students whose final outcome is 'other loss' by the same reasoning. As a result, there are two student groupings in the statistical model: 1) Normal, and 2) Dropouts. Normal is anyone whose final status is 'normal', and dropout is anyone whose final status is 'dropout' or 'probable dropout'.

Label	Number of Students *	Final Status
Normal	2,372	'normal'
Dropout	1,658	'dropout', 'probable dropout'
Total:	4,030	

* These totals do not include Native Americans, or students whose addresses could not be matched.

Thus the students who transferred out of the district, died, or were expelled, have not been included in the analysis even though there are a significant number of them (almost 900 students, or nearly twenty percent of the cohort group). In future research, it would be interesting to incorporate into the model the fate of the students who transfer out of the district as well as that of the students who transfer into the district during the course of the five year period. Excluding the transfers does give a more accurate measure of the dropout rate. However, I believe that transfers into and out of the district should not be totally ignored in assessing the overall health of the district.

D. Explanatory Variables

There are numerous possibilities for explanatory variables in studying the high school dropout problem, and several studies (Pallas 1984, Paulu 1987, Lloyd 1967, Gallington 1966) have attempted to enumerate the characteristics of high school dropouts. For instance, it has been shown that academic achievement, socio-economic status, family

structure, and self-esteem are important factors in describing the characteristics of dropouts versus non-dropouts.

However, there are many different ways of measuring these factors. Academic achievement, for instance, can be measured using grade point averages, standardized test scores, intelligence quotient tests, or number of years held back in school. Socioeconomic status is usually measured as a composite variable taking into account the family's income, father's occupation, mother's occupation, father's educational attainment, and mother's educational attainment. Self-esteem can be measured through a variety of variables including degree of belief in self-worth, attainment or lack of social skills, and belief in one's ability to influence or control what happens in life. As these examples indicate, there is a myriad of potential explanatory variables to consider in describing the high school dropout phenomenon.

Furthermore, one would ideally like student-specific information to determine the characteristics of dropouts, yet this is not always possible due to privacy, time, and budget constraints. For example, it would be interesting to know for each student: hours spent at home alone, family income, family structure, amount of parental involvement and encouragement with school, reason for leaving school such as pregnancy or full-time employment, and so on. Yet, it takes a lot of time and money to collect student-specific information of this type, especially for a large sample. Also student confidentiality must be protected, so school districts are often reluctant to release academic achievement information such as test scores and grade point averages.

Thus for information which is unavailable at the student-specific level, I have resorted to using an approximation, namely neighborhood census data. Since the census delineates census tracts so that residents within a given tract have similar characteristics, I hope that the use of census tract data serves as a reasonably good approximation of individual characteristics.

In this research, there are three basic categories of explanatory variables to be considered: student-level, neighborhood-level, and school-level. First, the student-level variables provide information specific to the individual students, such as gender, age, race, and address. The second level of variables contains information about the neighborhoods of the students such as type of housing, family structure, and socio-economic status. Lastly, school-level variables comprise the third category of explanatory variables which provide some measure of the differentiation among public high schools in Boston. Thus this research takes a three-tiered approach to analyzing the dropout problem by examining: 1) the student, 2) the neighborhood, and 3) the school.

It should be noted that the term 'explanatory' variable is used somewhat loosely here. The log-linear model identifies correlation between and among variables (i.e., whether or not two variables are independent). However the log-linear model does not necessarily imply causation. It may identify a significant interaction between variables, but this does not mean that the relationship between the variables is causal in nature. For instance, the log-linear model may indicate that there is a dependency between race and dropping out, but this does not necessarily mean that the fact that a student is Hispanic causes him to drop out of high school.

The following is a description of the explanatory variables used in the statistical model.

Student-Level Variables

o *Race* - There are four categories for race: Black, White, Hispanic, Asian. Actually there is a fifth category, Native American, but as mentioned previously, there are so few students in this category that it has been omitted.

o *Gender* - Obviously there are two categories for gender: Male and Female.

o *Age* - Although the school district did not release individual test scores or grade point averages to protect privacy, I wanted to include some measure of academic performance in the model. Thus age serves as a proxy for previous academic achievement.

The age variable has two categories: Normal and Overage. Over-age students are at least two years older than normal for their class. So students who are one year older than usual due to circumstances such as being held back a year, entering school late, or prolonged illness, are still considered 'normal'. More specifically, any student who was at least sixteen years old upon entering ninth grade on September 1, 1982 is categorized as overage, while any student less than sixteen years old is labeled normal. (The Boston public school district uses the same definition of over-age (Horst 1990).) Over-age students have typically been retained more than once to repeat a grade, making age a good indicator of serious academic difficulties prior to high school.

Neighborhood-Level Variable

o *Census Tract of Residence* - This variable represents the effects of the neighborhood in which the student lives. The tracts are separated into four groupings according to the residents' socioeconomic characteristics: well-off homeowners, yuppie renters, average working-class families, and poor broken-home households (see Chapter IV on tract clustering for more detail on how these groupings were selected). Ideally the model would include socioeconomic information specific to individual students such as parents' income, occupation, and educational attainment. Unfortunately this information was unavailable at the student level, so I have used the neighborhood's status as the best approximation of individual socio-economic status.

Furthermore, there may be neighborhood effects other than socio-economic status which are incorporated into the model through the use of this variable. For instance, some neighborhoods may exert neighborhood-pressure, similar to peer-pressure, on students to drop out, while other neighborhoods may be more supportive and encourage students to finish high school.

School-Level Variable

o *Type of High School* - At the time (1982-1986), there were four different types of high schools in Boston: examination schools, magnet schools, regular schools, and miscellaneous other schools.

The examination schools require students to pass an examination prior to entrance, and generally provide the highest quality public education in Boston. Thus these schools attract the highest academic achievers in the public school system and have excellent reputations. The three examination schools are: Boston Latin, Latin Academy, and Boston Technical.

During the period 1982-86, there were five magnet schools which offered special programs to attract a particular set of students. Boston High School offered separate morning and afternoon sessions to cater to working students. English High School had a school within a school organization. Copley Square High School (which is now named Snowden International) focused on international studies. Madison Park High School, conveniently located near an occupational resource center, offered a half-day school and half-day vocational program. Lastly, Umana High School in East Boston featured science programs but did not require any entrance examination (Umana no longer exists). It is believed that attendance at a magnet as opposed to a regular high school reflected a greater interest in education on the part of the student's family.

The majority of Boston public high school students attended the regular high schools which consequently represent the kernel of public secondary education in Boston. The regular high schools were: Brighton, Burke, Charlestown, Dorchester, East Boston, Hyde Park, Jamaica Plain, South Boston, and West Roxbury High School.

Finally there are several other schools which offer extremely specialized programs for students with unusual needs. For instance, two such schools are the Horace Mann School for the hearing-impaired and the McKinley School for students with behavioral problems. I have grouped these schools into a single category called: miscellaneous other.

E. Address-Matching

Address-matching is a technique used to reference geographical information. In this case, I used address-matching for two purposes:

1) to determine the census tract in which each student resides
2) to find the longitude and latitude coordinates of student residences for later location on a map of Boston.

Basically address-matching is used to match a particular address (street number, street name, street type, and zip code) against a street network file which contains records for every street segment in the city. To do the address-matching, I used two database tables: one with the student addresses (COHORTADD) and another with the street network records (BOSDIME).

Address-matching may sound simple but there are some underlying subtleties which can cause problems. For instance, all of the student address data was originally entered by hand and hence is not error-free. Thus some street names have typos such as 'Masachusetts Ave' instead of 'Massachusetts Ave', and as a result did not match the Boston Dime file even though it really is a valid street. Furthermore some mismatches are procedural as opposed to typographical. For example, the street type designation 'Avenue'

may be abbreviated as 'Ave' instead of 'Av', and therefore does not match the Dime file. Thus the first step was to clean up the data as much as possible to avoid needless mismatches.

Another issue in address-matching is which address to use for students who moved during the course of the five-year period (1982-1987). In fact, about twenty percent of the cohort group moved during their high school years, and many students moved multiple times. Since we are interested in the characteristics of the neighborhood prior to the student's dropping out, I decided to use the first address for each student who moved. Also this provides a comparative geographical reference for each student at the same point in time (upon entrance to ninth grade in 1982).

Unfortunately, some student addresses could not be matched against the Boston Dime file even after all of the identified typos and abbreviations were fixed. There are a variety of reasons for these mismatches. First, the Boston Dime file is not infallible. It does contain a few, though not many, mistakes such as typos and bad street numbers which can cause mismatches. Second, some of the student addresses are invalid because the students provided erroneous information to the school district, or perhaps even lied intentionally. Since the statistical model uses the type of census tract in which a student lives as a proxy for individual socioeconomic status, the students whose addresses could not be matched to determine the census tract were eliminated from the analysis.

Now if large numbers of students are systematically excluded from the analysis because their addresses do not match the Dime file, there is a legitimate concern that this introduces bias into the analysis, particularly if these students have a higher than average or lower than average dropout rate. Usually address-matching achieves about a seventy-percent success rate. Fortunately, in this case I was able to address-match 4,940 out of 5,393 students in the cohort which is about ninety-two percent. Thus bias is not a great concern in this instance because relatively few students were excluded due to address-matching problems.

The first application of address-matching was to determine in which census tract each student lived. Then census data could be used as the best approximation for information unavailable at the individual student level. The Boston public school district keeps information on student addresses but does not include the census tract number. However, the Boston Dime file does include census tract numbers. Therefore I matched each student's address against the Dime file and pulled off the census tract number from the corresponding street segment in the Dime file.

The second address-matching application was to compute the longitude and latitude coordinates of each student's address. Using SQL queries, it is possible to find the street segment record in the Dime file which matches the student's address, and to compute longitude and latitude coordinates for the student's residence by using the street number and interpolating between the longitude and latitude coordinates for the street segment. Then I used a coordinate conversion routine to convert from longitude/latitude to X,Y coordinates which represent inches on the digitizer. Finally I mapped the residences of specific groups of students, such as all overage male Hispanic dropouts in Boston. (See Chapter IX on spatial analysis for a more detailed description of the mapping.)

Summary

This chapter described the database, the definition of dropout and cohort, the explanatory variables to be used in the statistical model, and the use of address-matching to find census tract numbers and longitude/latitude coordinates of student residences. The next chapter will discuss the clustering of Boston census tracts based on socioeconomic characteristics.

Chapter IV
Tract Clustering

A. Cluster Analysis

In order to make the number of different census tracts more manageable for the log-linear model formulation, I decided to use a statistical technique called clustering. In 1980 there were 145 census tracts in Boston. If I did not group the census tracts in any way, then the multidimensional contingency table would have:

(dropout * race * gender * age * school * tract) = (2 * 4 * 2 * 2 * 4 * 145)
= (128 * 145)
= 18,560 cells.

Clearly this is an intractable number of cells since there are less than 5,000 students available to distribute across the contingency table. Hence the need for a natural grouping of census tracts is apparent for inclusion in the statistical model.

Cluster analysis is a multivariate statistical technique used to find natural groupings of items, in this case census tracts. It makes no prior assumptions about the number of groups or the structure of groups, rather the delineation of groups is based upon similarities or distances computed from a set of observations on the tracts. In most cluster analyses, this case included, there are too many possible groupings to examine every one and to choose the best grouping. However, there are algorithms which search through a set of groupings to find a good one (Johnson and Wichern 1988).

For the purposes of this study, I am interested in the clustering of census tracts based on socioeconomic characteristics since I am using the neighborhood socioeconomic status as a proxy for the individual student's situation. Thus the clustering is based on the values of four socioeconomic variables taken from the 1980 census data for each tract:

-- Median Income,
-- Percent Single Parent Households,
-- Percent of Occupied Housing which is Rental,
-- Percent High School Graduates.

Presumably these variables measure the composite level of wealth, family structure, type of housing, and educational background of the tract residents.

Two comments are in order here. First, the use of 1980 census data, as opposed to 1985 updates or 1990 census data, is not an arbitrary choice. The log-linear model examines students who entered high school in 1982, and we would like to use data which approximates the flavor of the neighborhood when the students began high school. As a rule of thumb, more recent data is usually preferable, but in this case we are not modelling the most recent students. The 1985 updates would reflect the neighborhood towards the end of these students' high school careers, and by then, many students have already dropped out. Thus I believe the 1980 census data is the best choice given these circumstances.

Second, critics may argue that the percent of high school graduates should not be included as a clustering variable because it is too highly correlated with the issue at hand, namely whether or not students drop out. In fact, it has been shown that low educational attainment of parents is a good indicator of at-risk students, because dropping-out is a generational problem. Parents who drop out tend to have children who drop out. Hence, ideally I would like to include the parents' educational level in the model, but once again, this information is unavailable at the individual level. As a result, I have included it here to take the neighborhood educational level into account as a proxy for parental educational attainment.

In this cluster analysis, we have observations on four continuous variables for each census tract: median income, percent single parent households, percent renters, and percent high school graduates.

Let X1 = income
X2 = percent single parent
X3 = percent renters
X4 = percent graduates

The values of these variables for each census tract can be compiled into a data table:

Tract ID	X1	X2	X3	X4
1	15,437.00	4.568	71.885	73.4
2	14,889.00	5.919	64.770	75.7
3	20,059.00	3.476	60.624	78.2
...	...	...	...	...
1404	16,045.00	5.934	39.819	68.2

Then we can think of each census tract as a point in four-dimensional space defined by the values of X1, X2, X3, and X4. Thus **X** = [X1, X2, X3, X4]' in vector form.

The goal of clustering is to divide the tracts into clusters by minimizing the distance among tracts within a cluster, and maximizing the distance between tracts in different clusters. Therefore, to measure the 'closeness' of tracts, we need some measure of distance. The most commonly used measure is Euclidean (straight-line) distance.

Suppose we want to determine the 'closeness' of two census tracts, **X** and **Y**.

$$\mathbf{X} = [\ X1, X2, X3, X4\ ]' \quad \text{and} \quad \mathbf{Y} = [\ Y1, Y2, Y3, Y4\ ]'$$

Then the Euclidean distance between **X** and **Y** is:

$$d(\mathbf{X}, \mathbf{Y}) = [\ (\mathbf{X} - \mathbf{Y})'\ (\mathbf{X} - \mathbf{Y})\ ]\ \text{^}1/2$$

$$= [(X1 - Y1)\text{^}2 + (X2 - Y2)\text{^}2 + (X3 - Y3)\text{^}2 + (X4\ Y4)\text{^}2]\text{^}1/2$$

Prior to clustering the tracts, I standardized the observations so that the variables with high values would not dominate the clustering. For example, the value of income is in thousands of dollars while the other three variables are expressed in percentages (percent single parent households, percent renters, and percent high school graduates) which are all less than or equal to one by definition. If we compared these values directly, the clustering

would focus on income since its values are four orders of magnitude larger than the other variables. In other words, the income variable would dominate the distance calculations because the other values are negligible in comparison. Therefore in order to avoid this problem, I standardized the observations as follows:

standardized observation $Z_i = (X_i - \text{mean}) / \text{standard deviation}$

and performed the cluster analysis on the standardized, rather than the original, observations.

To do the actual clustering, I used the non-hierarchical K-means method implemented by Systat (a PC-based statistical software package). This method groups tracts into a collection of K clusters where K is specified as part of the procedure. Basically, the algorithm assigns each tract to the cluster with the nearest centroid (mean vector) in terms of Euclidean distance. The algorithm is (Johnson and Wichern 1988):

K-means Algorithm

Step 1)	Specify K initial seed points (tracts).
Step 2)	For each tract, assign the tract to the cluster with the nearest centroid (mean vector) where 'nearest' means smallest Euclidean distance using the standardized observations. Then recalculate the centroid for the cluster receiving the new tract and for the cluster losing the tract (if applicable).
Step 3)	Repeat step 2 until no more tracts are reassigned to new clusters.

The choice of K, the number of clusters, is subjective and depends upon subject matter knowledge as well as data considerations. In general, one should choose K so that the between-cluster variability is maximized relative to the within-cluster variability. In this case, I wanted only a small number of clusters in order to minimize the number of cells in the multidimensional contingency table and also to make the interpretation easier. I tried clustering with K = 3, 4, and 6 which produced the following results (where SS = sum of squared deviations from the mean, a measure of variability).

Summary Statistics for 3 Clusters:

Variable	Between SS	DF	Within SS	DF
Income	76.936	2	67.064	142
PercSingleParent	64.379	2	79.621	142
PercRenter	80.813	2	63.187	142
PercGrad	88.097	2	55.903	142

Summary Statistics for 4 Clusters:

Variable	Between SS	DF	Within SS	DF
Income	88.852	3	55.148	141
PercSingleParent	111.120	3	32.880	141
PercRenter	93.574	3	50.426	141
PercGrad	88.184	3	55.816	141

Summary Statistics for 6 Clusters:

Variable	Between SS	DF	Within SS	DF
Income	99.319	5	44.681	139
PercSingleParent	110.708	5	33.292	139
PercRenter	100.916	5	43.084	139
PercGrad	95.656	5	48.344	139

When K = 3 clusters, there is not a very big difference in the 'between cluster' variability and the 'within cluster' variability as compared to the K = 4 and K = 6 options. Therefore the 3 cluster grouping is not as effective. Now the 6-cluster option has the greatest difference between 'within cluster' variability and 'between cluster' variability, however this option has another problem. Specifically the fifth cluster has only one tract in it (tract 807), and the sixth cluster has only five tracts. This grouping would produce very sparse counts in the cells of the contingency table since there are few students, relatively speaking, which live in any one tract. Hence I ruled out the 6-cluster grouping since the whole point of the clustering is to produce tract categories which make the log-linear model workable. This leaves the K = 4 cluster option, which appears to be reasonable. It has a

large difference between 'within cluster' and 'between cluster' variability, and it has a sufficient number of tracts in each cluster. Therefore the optimal grouping in this case is the four-cluster set.

Assuming that we use the four clusters of tracts, it is possible to rank the variables in terms of importance in the clustering using the univariate F-ratios. The F-ratio is the ratio of between-cluster variability to within-cluster variability. For example, the income variable has F-ratio:

$$\text{F-ratio} = \frac{\text{Between SS / DF}}{\text{Within SS / DF}} = \frac{(88.852 / 3)}{55.148 / 141)} = 72.725$$

For all four variables we have:

Variable	F-Ratio
PercentSingleParent	158.840
PercentRenter	87.216
Income	75.725
PercentGraduates	74.255

A large F-ratio indicates that the clusters are widely separated with respect to that variable, whereas there is little within cluster variation for that variable. Therefore, the percent single parent households appears to be most important in defining tract clusters, followed by percent renters, income, and percent high school graduates.

After separating the tracts into four clusters, the next step is to interpret the values of the four variables for each cluster. These are displayed in the following table.

Cluster	Mean Income	Mean Percent Single Parent	Mean Percent Renters	Mean Percent Graduates
1	-0.41	0.05	-0.01	-0.59
2	0.29	-1.16	0.85	1.26
3	1.15	-0.39	-1.19	0.60
4	-1.10	1.77	1.01	-0.80

Since the observations were standardized prior to clustering, these means should be interpreted relative to a mean of zero (in standard deviation units). For example, cluster 1 has negative mean income (-0.41). This indicates that the mean income of the tracts in cluster 1 is smaller than the overall mean income for all tracts. This sort of analysis leads to the following interpretations of the clusters. Such interpretation and "naming" of clusters is common in the practice of cluster analysis.

B. Cluster Interpretation

1) Cluster 1 has lower than average income, about average percent single parent households and percent renters, and fewer than average high school graduates. This suggests that the residents of the tracts in cluster 1 are average working-class families.

2) Cluster 2 has slightly higher than usual income, very few single parent households, a high percentage rental housing, and a much larger than usual percentage of high school graduates. Therefore it appears that the residents of the tracts in cluster 2 are young professional renters (yuppies) who are well educated, but do not have extremely high incomes because they are early in their careers.

3) Cluster 3 has very high income, low percent single parent households, very low percentage rental housing, and higher than usual percentage of high school graduates. This indicates that the residents of cluster 3 tracts are well-off homeowner families.

4) Cluster 4 has very low income, extremely high percent single parent households, very high percent rental housing, and very low percent high school graduates. Therefore the residents of the tracts in cluster 4 are the poorest in Boston, many of whom did not

finish high school and will never be able to purchase a home. Moreover, students from these tracts come from typically broken homes, with only a single parent to rely on for guidance, support, and encouragement. It is difficult to produce a single label for tracts of this type, however 'poorest households' seems fitting.

Hence I have given the four clusters of tracts the following labels, based upon the values of the socioeconomic variables, for use in the log-linear model.

Cluster	Category Label	Socioeconomic Ranking (highest to lowest)
1	Average Working-Class Families	(3)
2	Yuppie Renters	(2)
3	Well-Off Homeowner Families	(1)
4	Poorest Households	(4)

Chapter IX includes a map of the tract clusters in Boston. As expected, the yuppies populate the Back Bay and the well-off homeowners live in West Roxbury and other neighborhoods bordering Brookline. The average working-class families live in South Boston, East Boston, and Charlestown. And finally, the poorest neighborhoods are Mattapan, Roxbury, and Dorchester.

Summary

This chapter described the use of clustering to identify four groups of census tracts in Boston with similar socioeconomic characteristics. These tract clusters will later be used in the statistical model as the best approximation we have for the socioeconomic status of individual students. The next chapter provides an introduction to the type of statistical model used in this research, specifically the log-linear model which is useful for describing interactions among categorical variables.

Chapter V
Log-Linear Models

A. Introduction

In this research, I will formulate a statistical model, more specifically a log-linear model, which describes the high school dropout phenomenon in Boston. In general, a statistical model of this sort serves two purposes: first, it aids in understanding a large and complex data set, and second, it can be used to assess the relationships between explanatory variables. The following is a brief description of the statistical methodology underlying log-linear models.

There are basically two types of variables used in statistical analyses: discrete, also called categorical, and continuous, also called metric. In this case each student in the cohort has a set of identifying characteristics such as race, gender, age, type of neighborhood, and type of school. These variables are called 'categorical' because they take on discrete values which form a finite number of categories, whereas continuous variables can take on infinitely many values. For example, the race variable has four categories: Asian, Black, Hispanic, White, hence it is 'categorical', while a variable such as height is measured on a continuous scale. 'Dropout' is also a categorical variable.

The two types of variables are not mutually exclusive, however, because it is possible to categorize a continuous variable. For instance, the age variable could be measured on a continuous scale (0 years to infinity) but I chose to divide it into two discrete categories: normal-age and over-age upon entrance to ninth grade. There are no hard rules guiding the choice of categories for a continuous variable, and in fact the choice of categories may affect the results. In this case, I chose the categories for age based on substantive knowledge of current Boston public school district practices (Horst 1990).

Contingency tables are often used to organize categorical, as opposed to continuous, data such as that provided by the Boston public school district. Students are

cross-classified according to several categorical variables (dropout, race, gender, age, neighborhood, and school) to form a multidimensional contingency table where each variable represents one dimension.

Prior to 1970, the standard statistical technique for analyzing a multidimensional contingency table was to separate the problem and examine each two-way table individually (i.e., examine only two variables at a time). Unfortunately, this approach has several shortcomings (Fienberg 1977). For instance, it only analyzes marginal relationships between variables and does not consider their relationship when other variables are also present. Furthermore there is no way to compare these pairwise relationships simultaneously. Lastly this approach does not consider higher order interactions among variables.

In the 1970s the development of modern high-speed computers allowed researchers to tackle large-scale data-analysis problems. This fueled the most recent revolution in categorical data analysis when Cox, Bishop, Fienberg, and Holland developed the methodology of the log-linear model used in this thesis (Fienberg 1977). The log-linear model solves the problems inherent in analyzing multidimensional contingency tables because it does consider the relationships between variables when others are present, and it does take into account higher-order interactions among variables.

Categorical data analysis is actually a special case of multivariate analysis in which all the variables are discrete. Standard multivariate techniques, including ANOVA (analysis of variance) and regression, have a continuous dependent variable known as the response variable, and a set of explanatory variables which may be either continuous or discrete. On the other hand, the log-linear model has categorical response and explanatory variables. The log-linear model is analogous to the ANOVA model, however there is a subtle difference in interpretation. ANOVA assesses the effects of independent variables on dependent variables by partitioning the overall variability, while log-linear models describe the structural relationships among variables.

In this study, the log-linear model is particularly appropriate because many of the variables describing high school dropouts are categorical. For instance, race, gender, and school-type are all categorical variables. The 'dropout' variable has two categories indicating the student's final status: dropout or non-dropout. Furthermore, the continuous variables can be separated into logical groupings. For example, the age variable can be divided into two groups: normal-age and over-age. The tract clusters are separated according to different socioeconomic characteristics. Thus, the log-linear model is especially useful in modeling the characteristics of high school dropouts.

B. Model Specification

The following is a more formal, mathematical description of the log-linear model illustrated by a hypothetical two-dimensional example.

Suppose we are interested in whether there is a relationship between a student's gender and propensity for dropping out of high school. First, we cross-classify students according to gender and whether or not they dropped out of high school, to get the (2 x 2) contingency table.

		Gender		
		Male	Female	total:
Dropout	Yes	X_{11}	X_{12}	X_{1+}
	No	X_{21}	X_{22}	X_{2+}
	total:	X_{+1}	X_{+2}	1

where X_{ij} = the observed number of students in cell (i, j)

ith row total: $X_{i+} = \sum X_{ij}$ (over j) = $X_{i1} + X_{i2}$

jth column total : $X_{+j} = \sum X_{ij}$ (over i) = $X_{1j} + X_{2j}$

If N is fixed, then Xij is an observation from a Multinomial Distribution with total sample size N and cell probabilities Pij. (This assumes that we are using the Multinomial sampling model where we take a fixed sample of size N and then cross-classify each item of the sample according to its characteristics (i.e., its values for the underlying variables). Other sampling schemes include Poisson and Product-Multinomial sampling which are not discussed here.) And we can produce the following table with probabilities, instead of counts, for cell entries.

		Gender		
		Male	Female	total:
Dropout	Yes	P_{11}	P_{12}	P_{1+}
	No	P_{21}	P_{22}	P_{2+}
	total:	P_{+1}	P_{+2}	1

where $P_{11} + P_{12} + P_{21} + P_{22} = 1.0$
$(P_{1+}) + (P_{2+}) = 1.0$
$(P_{+1}) + (P_{+2}) = 1.0$

Thus
N = total sample size
Pij = probability of a student being in cell (i, j)
Mij = expected count (number of students) in cell (i, j)

$E(X_{ij}) = M_{ij}$ = expected count in cell (i, j)

==> $M_{ij} = N * P_{ij}$

Now suppose that dropping-out of high school and a student's gender are independent. If gender and dropping-out are independent, then

P_{ij} = Prob{ dropout = i and gender = j }
= Prob{ dropout = i } * Prob{gender = j}
= $(P_{i+}) * (P_{+j})$

Therefore under the model of independence, the cell probability is just the product of the marginal probabilities.

$$\Longrightarrow \quad M_{ij} = N * P_{ij}$$
$$= N * (P_{i+}) * (P_{+j})$$
$$= N * (X_{i+} / N) * (X_{+j} / N)$$
$$= (X_{i+} * X_{+j}) / N$$
$$= \text{expected count in cell (i, j) if dropout and gender are independent}$$

$$\Longrightarrow \quad \log M_{ij} = \log(X_{i+}) + \log(X_{+j}) - \log N$$

This formula can be expressed in a form which is analogous to ANOVA (analysis of variance):

$$\Longrightarrow \quad \log M_{ij} = \mu + \mu 1(i) + \mu 2(j)$$

Since the model is linear in the logarithmic scale, we call it a 'log-linear' model. For log-linear models, the dependent variable is the logarithm of the expected counts in the cells of the contingency table. The μ-terms represent deviations from the grand mean in this two-dimensional case, and in higher dimensional models they represent deviations from the lower-order terms.

μ = 'grand mean' of the logarithms of the expected cell counts

$$= 1/IJ \sum\sum \log M_{ij}$$

$$= 1/2*2 \sum\sum \log M_{ij}$$

$$= 1/4 (\log M11 + \log M12 + \log M21 + \log M22)$$

$\mu 1(i)$ = deviation of (the mean of the logs of the expected counts of J cells at level i of variable 1) from the grand mean

$$= [1/J \sum \log M_{ij} \text{ (over j)}] - \mu$$

$$= 1/2 (\log Mi1 + \log Mi2) - \mu$$

$\approx$ effect for being in row i

$\mu 2(j)$ = deviation of (the mean of the logs of the expected counts of I cells at level j of variable 2) from the grand mean

$$= [1/I \sum \log M_{ij} \text{ (over i)}] - \mu$$

$$= 1/2 (\log M1j + \log M2j) - \mu$$

$\approx$ effect for being in column j

If the independence model does not hold, then dropping-out and gender are correlated and we must include an interaction term to account for this dependence. (If variable 1 and 2 are independent, then there is no additional effect for being in row i and column j, so $\mu12(ij) = 0$, and this interaction term drops out of the model.)

Without independence, the model becomes:

$$\Longrightarrow \quad \log Mij = \mu + \mu1(i) + \mu2(j) + \mu12(ij)$$

where

$\mu12(ij)$ = deviation of (the logarithm of the expected count in cell (i,j)) from the lower-order mean effect ($\mu + \mu1(i) + \mu2(j)$)

$= \log Mij - [\mu + \mu1(i) + \mu2(j)]$

$= \log Mij - \mu - [1/2 \sum \log Mij \text{ (over j)} - \mu] - [1/2 \sum \log Mij \text{ (over i)} - \mu]$

$= \log Mij - 1/2(\sum \log Mij \text{ (over j)}) - 1/2(\sum \log Mij \text{ (over i)}) - \mu + 2\mu$

$= \log Mij - 1/2 \sum \log Mij \text{ (over j)} - 1/2 \sum \log Mij \text{ (over i)} + \mu$

$= \log Mij - 1/2(\log Mi1 + \log Mi2) - 1/2(\log M1j + \log M2j) + \mu$

$\approx$ additional effect for being in row i and column j

Since these u-terms represent deviations from the grand mean μ we need the following restrictions:

	$\sum \mu1(i) = 0$	$\Longrightarrow$ $\mu1(1) + \mu1(2) = 0$	$\Longrightarrow$	$\mu1(1) = -\mu1(2)$
	$\sum \mu2(j) = 0$	$\Longrightarrow$ $\mu2(1) + \mu2(2) = 0$	$\Longrightarrow$	$\mu2(1) = -\mu2(2)$
over i:	$\sum \mu12(ij) = 0$	$\Longrightarrow$ $\mu12(11) + \mu12(21) = 0$	$\Longrightarrow$	$\mu12(11) = -\mu12(21)$
		$\mu12(12) + \mu12(22) = 0$	$\Longrightarrow$	$\mu12(12) = -\mu12(22)$
over j:	$\sum \mu12(ij) = 0$	$\Longrightarrow$ $\mu12(11) + \mu12(12) = 0$	$\Longrightarrow$	$\mu12(11) = -\mu12(12)$
		$\mu12(21) + \mu12(22) = 0$	$\Longrightarrow$	$\mu12(21) = -\mu12(22)$

Actually there is only one interaction parameter which must be specified since the others can be derived using: $+ \mu12(11) = -\mu12(21) = + \mu12(22) = -\mu12(12)$.

Alternatively, the μ-terms in the log-linear model can be expressed in terms of cross-product ratios, also called odds-ratios. The cross-product ratio measures association between two variables. In terms of the expected cell counts, the cross-product ratio is:

$$\alpha = M11 * M22 / M12 * M21$$

or in terms of the cell probabilities:

$$\alpha = P11 * P22 / P12 * P21.$$

(Note: if $\alpha = 1$, then the two variables are independent.) The parameters of the log-linear model are functions of cross-product type terms, hence we can write the μ-terms as functions of the cross-product ratio as follows (for example, let i=1 and j=1):

$$\mu12(11) = 1/4 \log \alpha = 1/4 \log [M11*M22 / M12 * M21]$$

$$\mu1(1) = 1/4 \log [M11*M12 / M21*M22]$$

$$\mu2(1) = 1/4 \log [M11*M21 / M12*M22]$$

To check this, we have:

$$\log Mij = \mu + \mu1(i) + \mu2(j) + \mu12(ij)$$

$$\Longrightarrow \log M11 = \mu + \mu1(1) + \mu2(1) + \mu12(11)$$

$$\log M11 = 1/4 (\log M11 + \log M12 + \log M21 + \log M22) + 1/4 (\log M11 + \log M12 - \log M21 - \log M22) + 1/4 (\log M11 + \log M21 - \log M12 - \log M22) + 1/4 (\log M11 + \log M22 - \log M12 - \log M21)$$

$$\log M11 = \log M11 + 1/2 \log M12 - 1/2 \log M12 + 1/2 \log M21 - 1/2 \log M21 + 1/2 \log M22 - 1/2 \log M22$$

$$\log M11 = \log M11.$$

If we are interested in formally testing whether or not dropping-out and gender are independent, we can specify two log-linear models, one model with the two-factor interaction term and another model without the two-factor interaction, and see which model provides the best fit to the data. This suggests a useful hierarchy for model building.

C. Hierarchical Models

Clearly, as the number of dimensions increases, the number of potential log-linear models grows exponentially depending on the number of μ-terms and which combinations of μ-terms are included. Thus we restrict ourselves to the consideration of only hierarchical models. The saturated model includes all possible μ-terms, and hence fits the data perfectly because the number of independent parameters equals the number of cells in the contingency table. However, more interesting models usually do not include every possible μ-term in order to model independence as well as dependence among variables. The hierarchy principle states that no higher-order terms can be included in the model unless all related lower-order terms are also included. We need this restriction to maintain the interpretation of higher-order μ-terms as deviations from lower-order μ-terms.

Perhaps an example will best illustrate the hierarchy principle. Suppose we have a three-dimensional contingency table and are interested in specifying a log-linear model to describe the data. The saturated model is:

$$\log Mijk = \mu + \mu1(i) + \mu2(j) + \mu3(k) + \mu12(ij) + \mu13(ik) + \mu23(jk) + \mu123(ijk)$$

which includes all main effects, all three two-factor effects, and the three-factor effect. The following model, which contains a subset of these μ-terms, is <u>not</u> hierarchical

$$\log Mijk = \mu + \mu1(i) + \mu2(j) + \mu3(k) + \mu123(ijk)$$

because it does include the higher-order μ123(ijk) term, but it does not include the related lower order two-factor terms μ12(ij), μ13(ik), μ23(jk). Since the related lower-order terms have been excluded, this model is non-hierarchical and would not be considered in this analysis. On the other hand, a model such as

$$\log Mijk = \mu + \mu 1(i) + \mu 2(j) + \mu 3(k) + \mu 12(ij)$$

is hierarchical since the higher-order term μ12(ij) and both related lower-order terms, μ1(i) and μ2(j), are present. Thus the hierarchy principle is useful in model specification as well as interpretation.

D. Parameter Estimation

After specifying a particular log-linear model by choosing appropriate μ-terms, the next step is to estimate the expected cell counts under this model. In general, we use maximum likelihood estimation to derive estimates of the expected cell counts. Maximum likelihood estimation uses the likelihood function, which is actually the probability density function, interpreted so that the observed data is given and the density function parameters are random variables (as opposed to the parameters being given, and the data being random variables). In this case, we use the multinomial density function given our sampling scheme, however it has been shown that all three sampling schemes yield the same maximum likelihood estimates (Bishop, Fienberg, Holland 1975). Maximum likelihood estimation makes intuitive sense. Briefly, the idea is to select values of the parameters which maximize the likelihood (probability) of observing the data that we actually did observe.

It is possible to derive the MLEs (maximum likelihood estimates) of cell counts using an iterative proportional fitting algorithm which successively fits the sufficient configurations of a given log-linear model (Fienberg 1977). This algorithm has several desirable properties. First, it always converges to the expected cell values. Second, the convergence threshold, and hence the desired accuracy of the estimates, can be selected by the user. Third, the algorithm can use any set of starting values for the initial cell estimates, although most computer programs start with all one's. Fourth, there is another method for computing cell estimates directly, however this algorithm converges to these direct estimates in one cycle, if they exist. Lastly, this method does not require the costly calculation of μ-terms to derive the cell estimates, rather it uses a set of marginal totals specified by the model.

Before starting the iterative proportional fitting algorithm, it is necessary to determine the sufficient configurations (marginal totals) for the given log-linear model. Fienberg suggests the following general procedure (Fienberg 1977):

0) Select a log-linear model.

1) Look at the highest-order effect in the model for each variable.

2) Compute the observed marginal total which corresponds to the highest-order effect for each variable.

3) Compute the estimates of expected cell values using the set of observed marginal totals found in step 2. (Use the iterative proportional fitting algorithm described below.)

Thus the MLE estimates, M_{ijk}, are functions of a set of marginal totals dictated by the form of the log-linear model, specifically by which μ-terms are included in the model.

The following is a general description of the iterative proportional fitting algorithm for deriving cell estimates which is implemented by most commercial software packages with log-linear modeling capabilities.

Iterative Proportional Fitting Algorithm

(Step 1) Choose a set of starting values for all cells (usually use 1's).

(Step 2) Calculate the expected cell values which satisfy the first set of marginal constraints (found in the general procedure described above).

(Step 3) Recalculate the expected cell values so that they satisfy the second set of marginal constraints.

...

(Step p) Recalculate the expected cell values so that the pth set of marginal constraints is satisfied.

(Step p+1) Return to step 2 and repeat the cycle until the change in the cell estimates from one cycle to the next is below the convergence threshold.

In the course of successively fitting marginal constraints, a previous set of marginal constraints may no longer fit temporarily, however this procedure will eventually converge yielding the estimated expected cell values (Mijk). The Systat statistical package allows up to twenty iterations with a convergence threshold of 0.1, but the user can easily change the default convergence threshold and maximum number of iterations.

E. Goodness-of-Fit

Now that we have selected a log-linear model and obtained estimates of the cell counts expected under this model, the next step is to check how well the model fits the observed data. There are two statistics used to check the summary goodness-of-fit of a model: the Pearson Chi-Square statistic (X2) and the Likelihood-Ratio statistic (G2).

$$X2 = \sum [(\text{observed count} - \text{expected count})^2 / \text{expected count}] \quad \text{(over all cells)}$$

$$G2 = \sum [(\text{observed count}) * \log (\text{observed count} / \text{expected count})] \quad \text{(over all cells)}$$

Both of these statistics measure the discrepancy between the observed cell counts and what we would expect under the log-linear model. If there is a big difference between the observed data and that expected under the model, the X2 and G2 will both be large and we conclude that the model does not fit the data well. On the other hand, if the difference between the expected and observed counts is small, then X2 and G2 will both be small and we conclude that the model does fit the data.

In fact, X2 and G2 are each asymptotically distributed chi-square with the degrees of freedom depending on the log-linear model. The degrees of freedom are the number of cells in the contingency table minus the number of parameters (μ-terms) fitted. For example, if we have a two-by-two table and we fit the independence model:

log Mij = μ + μ1(i) + μ2(j), we have

# of cells = I * J = 2 * 2 = 4

# of parameters fitted = 3.

Therefore we have 4 - 3 = 1 degree of freedom in this case. The following table gives the degrees of freedom for all three-dimensional (2 x 2 x 2) log-linear models:

Model	Shorthand	# cells	# pars fit	df
μ + μ1 + μ2 + μ3 + μ12 + μ13 + μ23 + μ123	[123]	8	8	0
μ + μ1 + μ2 + μ3 + μ12 + μ13 + μ23	[12][13][23]	8	7	1
μ + μ1 + μ2 + μ3 + μ12 + μ13	[12][13]	8	6	2
μ + μ1 + μ2 + μ3 + μ12 + μ23	[12][23]	8	6	2
μ + μ1 + μ2 + μ3 + μ13 + μ23	[13][23]	8	6	2
μ + μ1 + μ2 + μ3 + μ12	[12][3]	8	5	3
μ + μ1 + μ2 + μ3 + μ13	[13][2]	8	5	3
μ + μ1 + μ2 + μ3 + μ23	[23][1]	8	5	3
μ + μ1 + μ2 + μ3	[1][2][3]	8	4	4

When the sample size is large (approximately ten times the number of cells in the table) and the null hypothesis is true (the model describes the structure in the data), then X2 and G2 are asymptotically equivalent, and their values will be very close. However, if the observed data is sparse and several cells in the multidimensional contingency table contain zeros, then the chi-square distribution approximation for X2 and G2 breaks down. In this case, any hypothesis tests based on these statistics are suspect. A general rule-of-thumb is that no more than one-fifth of the cells should have sparse entries (counts less than five) for the chi-square approximation to hold (Bishop, Fienberg, and Holland 1975).

Summary

This chapter gave a brief introduction to log-linear models which are used to analyze categorical data. We examined the structure of the log-linear model, parameter estimation, and goodness-of-fit statistics. The following chapter will use log-linear models in a specific application area, namely to examine the data on Boston high school dropouts, in order to understand the interactions between dropping-out and the explanatory variables: race, gender, age, tract-type, and school-type.

Chapter VI
Frequencies and Two-Way Models

This chapter examines the frequencies of students in the cohort group according to race, gender, age, tract-type, and school-type, in order to understand the characteristics of the underlying student population. The chapter also investigates the correlation between dropping-out and these explanatory variables by formulating two-way log-linear models and testing the fit of the models without the two-factor interaction term. In every case, the two-way independence model is rejected indicating that dropping-out is significantly correlated with each of the explanatory variables. Lastly, the pattern of residuals is analyzed to determine which students are more likely to drop out than others.

A. Frequencies

Before examining who drops out of high school, it is important to understand the structure of the underlying student population. As mentioned previously, the cohort group used in the following analysis is somewhat smaller than the original 1982 ninth grade cohort group. Specifically, the cohort group used here does not include:

1) students whose addresses could not be matched to determine census tract of residence
2) students who transferred out of (or into) the Boston public school district
3) Native American students (because there are too few of them).

Hence this cohort group has only 4,030 students whereas the original group contained 5,393 students.

1. Racial Composition

The racial composition of the 1982-86 Boston public high school cohort group is indicated in the table below.

ASIAN	BLACK	HISPANIC	WHITE	TOTAL:
279	2,000	509	1,242	4,030

This table indicates that the largest racial group is Black, followed by Whites, Hispanics, and finally Asians. The minority students (Blacks, Hispanics, and Asians) actually constitute the majority (69%), since less than fifty percent of the cohort group is white. There are fewer White students in the cohort group because many Whites attend private or parochial schools, by choice, instead of the public schools. Many Blacks, however, do not have the financial means to send their children to private schools. For example, only forty percent of the White students in Boston attend the public schools, whereas seventy-six percent of Boston's Black students attend the public schools (Byrne 1988).

2. Gender Composition

The next table gives the breakdown of the cohort group by gender.

FEMALE	MALE	TOTAL:
1,941	2,089	4,030

The cohort group is approximately forty-eight percent female and fifty-two percent male, so there are slightly more male students than females.

3. Age Distribution

The following table gives the distribution of students by age, specifically how many students are normal-age (less than sixteen years old upon entrance to ninth grade in 1982) and how many students are over-age which is defined as two years older than normal (at least sixteen years old upon entrance to ninth grade in 1982).

NORMAL	OVERAGE	TOTAL:
3,356	674	4,030

Clearly the vast majority of students in the cohort group are normal-age. Although the percentage of over-age students is fairly small (seventeen percent), there are still a significant number of students (almost seven hundred) who are at least two years over-age, indicating that most likely they were held back at least two years in school and have experienced serious academic difficulties prior to high school. However, another possibility is that they emigrated to Boston while in primary school and were placed in younger grades in order to alleviate bilingual difficulties or weakness of previous schooling.

4. Tract Distribution

As described previously, the Boston census tracts were clustered into four groups based on a set of four socioeconomic variables (median income, percent single parent families, percent rental housing, and percent high school graduates). The next table shows the number of students who live in each type of census tract.

POOREST	AVERAGE	YUPPIE	WELL-OFF	TOTAL:
1,012	1,686	281	1,051	4,030

The Yuppie-renter census tracts have very few students which makes sense because young professionals rarely have high-school aged children. The distribution of students across the poor broken-home, average working-class family, and well-off homeowner tract groupings is fairly even with the largest number of students in the average group as one

might expect. All three groups have a significant number of students, and no one type of neighborhood appears to dominate. However, the large number of students from the poorest broken-home neighborhoods is a definite concern.

5. School Distribution

This table disaggregates the cohort group by type of high school.

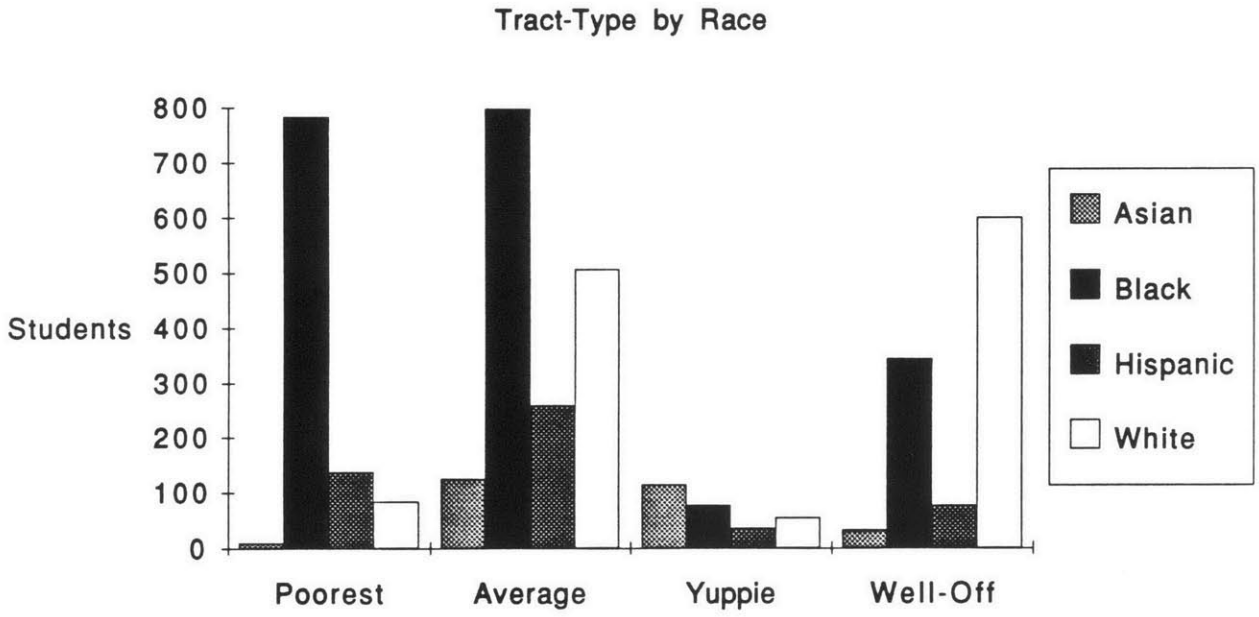
EXAM	MAGNET	REGULAR	OTHER	TOTAL:
708	1,015	1,933	374	4,030

Almost half of the students attended the regular public high schools, while a quarter of the students attended magnet schools which offered special programs such as half-day work/study schedules. Nearly one-fifth of the students attended the examination schools which have the highest academic reputation of any public schools, and the remaining students attended other schools which cater to students with specialized needs such as behavioral problems. Thus the majority (73%) of Boston public school students attend the regular and magnet high schools, and these schools constitute the kernel of public secondary education in Boston.

6. Racial Composition by Tract

The distribution of students by race across the four types of census tracts is definitely not uniform. Figure 1 on the following page indicates that the Black students are highly concentrated in the poor census tracts, while the White students inhabit the tracts at the higher end of the socioeconomic spectrum (i.e., the average working-class and the well-off tracts). The Asian students come from mostly average working-class neighborhoods with only a few students from the very poor or the very well-off tracts.

Figure 1


Tract-Type by Race
800
700
600
500
Students 400
300
200
100
0
Poorest
Average
Yuppie
Well-Off
Asian
Black
Hispanic
White

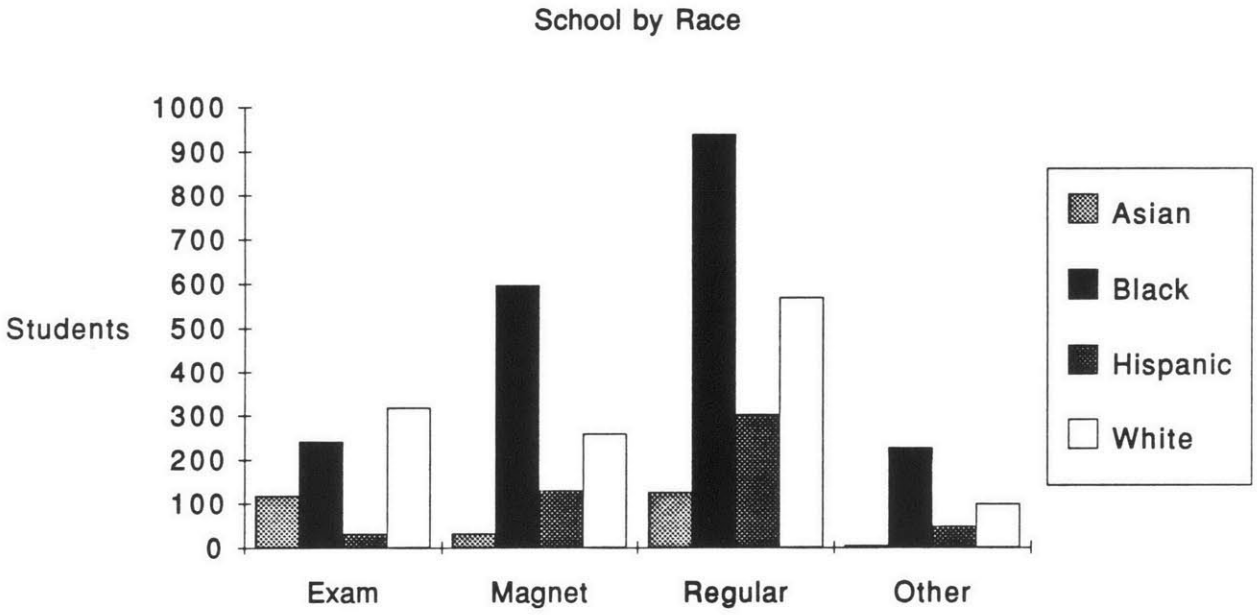
Interestingly, there are a large number of Asian students from the yuppie tracts, in fact more than any other racial group. Similarly to the Blacks, the Hispanic students come from lower socioeconomic neighborhoods. But, by far the largest number of students from the poorest neighborhoods are Black (seventy-seven percent).

7. Racial Composition by School

Perhaps analogously to the racial composition across tracts, the distribution of students by race across different types of schools is also not uniform, as can be seen in Figure 2 on the following page. The biggest non-uniformity is in the examination schools. Only six percent of Hispanics and twelve percent of Blacks attend the exam schools, whereas twenty-six percent of Whites and forty-two percent of Asians attend the exam schools. On the other hand, almost the same percentages of Asians (45%), Blacks (47%), and Whites (46%) in the public school system go to the regular high schools. Also a majority of the Hispanics (59%) attend the regular schools. Relatively more Blacks and Hispanics go to magnet schools, and there are very few Asians in either the magnet schools or the 'other' schools for students with special needs. About the same proportion of Blacks, Hispanics, and Whites (8-11%) attend the special-needs schools.

In terms of racial composition within each type of school, the exam schools are mostly White (45%), with fewer Blacks (34%), Asians (17%), and very few Hispanics (4%). In contrast, the regular public schools are mostly Black (49%), followed by Whites (29%), Hispanics (16%), and lastly Asians (6%). The magnet and the 'other' schools are similar to the regular schools except that there are more Blacks and fewer of the other races. Thus the examination school student-body is primarily White, whereas the regular schools are mainly Black.

Figure 2


School by Race
Students
1000
900
800
700
600
500
400
300
200
100
0
Exam
Magnet
Regular
Other
Asian
Black
Hispanic
White

8. Number of Dropouts

This last frequency table gives the number of dropouts in the cohort group and the number of students who finished high school normally.

NORMAL	DROPOUT	TOTAL:
2,372	1,658	4,030

Thus the dropout rate for this cohort group, excluding transfers from the denominator, is approximately forty-one percent. Clearly high school dropouts are a significant problem in Boston since greater than one in three students drop out.

B. Two-Way Models

Now that we understand the composition of the cohort group, the next step is to examine the bivariate relationships between dropping-out of high school and the five explanatory variables: race, gender, age, tract, and school. In particular, we want to know whether dropping-out of high school is independent of any of these variables. To test for independence, the first step is to crosstabulate the dropout variable by an explanatory variable such as race. Then we formulate the two-dimensional log-linear model, excluding the interaction term. After estimating the expected cell frequencies and computing the goodness-of-fit statistics, if the model fits then we conclude that the two variables are independent; otherwise we conclude that there is a statistically significant interaction. The significance level is $\alpha = .05$.

This procedure examines the explanatory variables individually and does not consider the relationship when other explanatory variables are taken into account. Thus

this bivariate evaluation is only a first step, and in later sections we will consider higher-dimensional models.

(1) Dropping-out vs. Race

Is dropping-out independent of race, or are students of one racial group more likely to drop out than others? To answer this, we can test the appropriate log-linear model. In this case, the two-dimensional saturated model is:

$$\log Mij = \mu + \mu d(i) + \mu r(j) + \mu dr(ij)$$ where d = dropout variable and r = race variable.

or in shorthand: [D][R][DR]. To check for independence, we formulate the log-linear model without the two-factor interaction term ($\mu dr(ij)$) which is [D][R] in shorthand or in long form:

$$\log Mij = \mu + \mu d(i) + \mu r(i)$$

and check how well this model fits the observed data.

I used Systat, a commercial statistical software package, to calculate the expected cell frequencies and the goodness-of-fit statistics. The table below gives the expected cell frequencies under the independence model, and the observed cell frequencies in parentheses.

	ASIAN	BLACK	HISPANIC	WHITE	TOTAL:
DROPOUT	114.78 (74)	822.83 (791)	209.41 (259)	510.98 (534)	1658
NORMAL	164.22 (205)	1177.17 (1209)	299.59 (250)	731.02 (708)	2372
TOTAL:	279	2000	509	1242	4030

Thus for the independence model, the goodness-of-fit statistics are:

$$X2 = \Sigma [(obs - exp)^2 / exp] = 48.43 \quad ==> \text{Probability} = 0.000$$

and

$$G2 = 2 \Sigma [(obs) \log (obs / exp)] = 49.45 \quad ==> \text{Probability} = 0.000$$

with degrees of freedom = (I - 1) * (J - 1) = 1 * 3 = 3.

By definition, the probability or the P-Value is the smallest significance level α for which the null hypothesis could be rejected. Since the probability = 0.000 is less than the commonly used significance level $\alpha = 0.05$, we reject the null hypothesis of independence. In other words, these two values for X2 and G2 are too far out into the right tail of the chi-square distribution, indicating that the expected counts, assuming that dropout and race are independent, are significantly different from the observed cell counts, and it is unlikely (less than a five percent chance) that this difference is a result of random chance. Therefore we conclude that dropping-out and race are correlated.

Next we can examine the residuals to see which races are more or less likely to drop out. The residuals are standardized in order to compare the differences between observed and expected cell counts across cells. (If two cells have the same residual, but one cell has a large count and the other cell has a small count, the residual represents a relatively larger deviation for the cell with the small count. By standardizing, we take into account the size of the cell counts and thus can compare residuals across cells with small and large counts.) Systat computes the standardized residuals with the following formula:

Standardized Residual = (Observed - Expected) / SQRT (Expected).

In this case, the residuals are:

	ASIAN	BLACK	HISPANIC	WHITE
DROPOUT	-3.81	-1.11	3.43	1.02
NORMAL	3.18	0.93	-2.87	-0.85

The residuals suggest some rather surprising results given the literature on high school dropouts described in Chapter II. First, as expected, Asians are much less likely to dropout than their counterparts, and Hispanics are much more likely to dropout. But surprisingly, Blacks actually drop out less frequently than Whites. However the difference between Blacks and Whites is not great percentage-wise. Approximately forty percent of Blacks in the cohort group dropped out, whereas forty-three percent of whites dropped out. Thus the two-way model indicates that Hispanics drop out the most, followed by Whites, then Blacks, and finally Asians.

(2) Dropping-Out vs. Gender

Next we can determine whether males are more likely to drop out than females, or vice versa, or whether there is no difference between the sexes. The saturated two-dimensional log-linear model is:

$$\log Mij = \mu + \mu d(i) + \mu g(j) + \mu dg(ij)$$ where d = dropout variable, g = gender variable

and the independence model which excludes the two-factor interaction term is:

$$\log Mij = \mu + \mu d(i) + \mu g(j)$$

The expected cell values under the independence model are contained in the next table.

	FEMALE	MALE	TOTAL:
DROPOUT	798.56 (689)	859.44 (969)	1658
NORMAL	1142.44 (1252)	1229.56 (1120)	2372
TOTAL:	1941	2089	4030

The goodness-of-fit statistics for the independence model, with 1 degree of freedom, are:

X2 = 49.26 ==> Probability = 0.000

G2 = 49.43 ==> Probability = 0.000

Since the probability is less than .05, we reject the null hypothesis that dropping-out and gender are independent. Furthermore the residuals indicate that males are slightly more likely to drop out than females, as seen in the table below.

	FEMALE	MALE
DROPOUT	-3.88	3.74
NORMAL	3.24	-3.12

(3) Dropping-Out vs. Age

The next explanatory variable is age, and we can test whether dropping out of high school is independent of a student's age upon entering ninth grade. The fully saturated model which includes the two-factor interaction term is:

$\log Mij = \mu + \mu d(i) + \mu a(j) + \mu da(ij)$ where d = dropout variable
a = age variable.

The independence model to be tested is:

$$\log Mij = \mu + \mu d(i) + \mu a(j).$$

The fitted values under the independence model are:

	NORMAL	OVERAGE	TOTAL:
DROPOUT	1380.71 (1171)	277.29 (487)	1658
NORMAL	1975.29 (2185)	396.71 (187)	2372
TOTAL:	3356	674	4030

and the summary statistics, with 1 degree of freedom, are:

X2 = 323.56 ==> Probability = 0.000

G2 = 322.38 ==> Probability = 0.000

The size of the goodness-of-fit statistics indicate that there is a large discrepancy between the observed cell counts and those expected under the independence model. As expected, the residuals in the table below indicate that the students who are at least two years older than normal upon entrance to high school are much more likely to drop out.

	NORMAL	OVERAGE
DROPOUT	-5.64	12.59
NORMAL	4.72	-10.53

Moreover, the large magnitude of the residuals indicates that the difference between normal-age and over-age students is pronounced. This makes sense since over-age students often have had serious academic problems prior to high school, and eventually they give up and drop out.

(4) Dropping-Out vs. Tract

Is the type of neighborhood in which a student lives an indicator of future risk of dropping out? The following test shows that tract-type is indeed correlated with propensity for dropping-out. The saturated model is:

$$\log Mij = \mu + \mu d(i) + \mu t(j) + \mu dt(ij) \qquad \text{where } d = \text{dropout variable}, \; t = \text{tract variable}$$

and the model with no two-factor interaction is:

$$\log Mij = \mu + \mu d(i) + \mu t(j).$$

The table of expected cell values under the independence model is given below.

	POOREST	AVERAGE	YUPPIE	WELL-OFF	TOTAL:
DROPOUT	415.35 (480)	693.64 (734)	115.61 (100)	432.40 (344)	1658
NORMAL	595.65 (532)	992.36 (952)	165.39 (181)	618.60 (707)	2372
TOTAL:	1012	1686	281	1051	4030

These observed and expected cell values give the following summary statistics, with 3 degrees of freedom:

X2 = 54.80 ==> Probability = 0.000

G2 = 55.45 ==> Probability = 0.000

Hence, at the five percent significance level, it appears that dropping-out and tract-type are not independent. The residuals are:

	POOREST	AVERAGE	YUPPIE	WELL-OFF
DROPOUT	3.12	1.53	-1.45	-4.25
NORMAL	-2.61	-1.28	1.21	3.55

This pattern of residuals suggests that, as one might expect, students from the poorest neighborhoods are the most likely to drop out, followed by those from the average working-class neighborhoods. The students from yuppie neighborhoods are less likely to drop out, and the students from well-off homeowner family neighborhoods are the least likely to drop out of high school. Thus it appears that neighborhood socioeconomic status is inversely correlated with dropping-out. The students from higher socioeconomic status communities drop out the least, and the students from the poorest neighborhoods drop out the most.

(5) Dropping-Out vs. School

One would expect the exam-school students to be less likely to drop out than the regular high school students because the exam students must pass an entrance examination prior to enrollment and presumably are more motivated academically. In fact, the data below do indicate a differential in dropout patterns across school types. First, the saturated model is:

$$\log Mij = \mu + \mu d(i) + \mu s(j) + \mu ds(ij) \quad \text{where } d = \text{dropout variable}, \; s = \text{school variable}$$

and the independence model is: $\log Mij = \mu + \mu d(i) + \mu s(j)$.

The fitted cell values, assuming that school-type and dropping-out are independent, are:

	EXAM	MAGNET	REGULAR	OTHER	TOTAL:
DROPOUT	291.28 (85)	417.59 (435)	795.26 (895)	153.87 (243)	1658
NORMAL	416.72 (623)	597.41 (580)	1137.74 (1038)	220.13 (131)	2372
TOTAL:	708	1015	1933	374	4030

Hence the summary statistics, with 3 degrees of freedom, are:

X2 = 358.40 ==> Probability = 0.000

G2 = 400.04 ==> Probability = 0.000

Since X2 and G2 are so large, the school-type and dropping-out are clearly not independent. The residuals in the table below indicate which schools have higher dropout rates.

	EXAM	MAGNET	REGULAR	OTHER
DROPOUT	-12.09	0.85	3.54	7.19
NORMAL	10.11	-0.71	-2.96	-6.01

As expected, the residuals suggest that exam students are the least likely to drop out. The regular-school students drop out more often than the magnet-school students, which makes sense since families who send their children to magnet schools are generally thought to show greater interest in their children's education. Lastly, students from the other specialized schools are most likely to drop out, which is also in agreement with the view that many of these students have academic difficulties associated with behavioral problems.

Summary

The frequency tables indicated that the cohort group has slightly more males than females, and that the majority of students are normal-age as opposed to over-age. The largest racial group is Black, followed by Whites, Hispanics, and then a small number of Asians. More students come from average working-class neighborhoods, and there are about as many poor students as well-off students. Lastly, the vast majority of the cohort group attends the regular and the magnet high schools.

The five hypothesis tests described above indicate that dropping-out is not independent of race, gender, age, tract-type, or school-type, when analyzed separately. In every case, the null hypothesis of independence was rejected. By analyzing the pattern of residuals, we found that Hispanics are the most likely to drop out, followed by Whites, Blacks, and finally Asians. Males dropped out more often than females, and over-age students are far more likely to leave high school without graduating. As expected, students from the poorest neighborhoods are most likely to drop out, while students from well-off neighborhoods tend to complete their high school educations. Finally, the examination school students are the least likely to leave school, and regular school students drop out more often than those attending magnet schools.

Given these conclusions, there are two statistical issues which should be addressed: 1) repeated testing on the same data, and 2) the bivariate approach to analyzing multidimensional data.

First, there is a problem with performing several hypothesis tests on the same set of data. For instance, if we choose a significance level of .05 and do twenty hypothesis tests on the same data, then we would expect at least one out of the twenty hypothesis tests to yield a significant result by chance alone. However in the analysis performed above, the test statistics were not even close to the critical values in every case. Every test rejected the null hypothesis resoundingly, and given the size of the test statistics, it is extremely doubtful that this is a result of chance alone.

The second concern is that the bivariate approach to analyzing multidimensional data is misleading. The tests performed above examined the relationship between dropping-out and the explanatory variables one variable at a time. Thus this approach does not consider the relationship when more variables are introduced into the analysis. For example, it is possible that dropping-out is independent of race if tract-type is controlled or school-type is controlled. The bivariate approach suffers from the same missing variables problem which often plagues regression analysis. The examination of the bivariate relationships is useful, but the analysis should not end here. Once the bivariate relationships are understood, the next step is to examine higher dimension log-linear models which will be the focus of the next chapter.

Chapter VII
Higher-Order Models

A. The Multi-Dimensional Contingency Table

Now that we have examined the two-way models, the next step is to consider even higher dimensional models, specifically models involving all five explanatory variables: race, gender, age, tract, and school. If we crosstabulate the students by all five of these characteristics plus the dropout variable, we get a contingency table with dimensions:

dropout * race * gender * age * tract * school = (2) * (4) * (2) * (2) * (4) * (4)
= 512 cells.

This is a very large contingency table and if the students were uniformly spread throughout the table, there would be (4030 students / 512 cells) = 7.9 or approximately eight students per cell. However, the students are not evenly distributed across the table. (If they were, this analysis would have no point.) Therefore some cells in the table have more than their share of students while other cells have fewer.

This presents a problem because there are several cells containing zero to five students. In fact, more than one-fifth of the cells in this contingency table have sparse entries, and as mentioned in the chapter on log-linear models, the chi-square approximation breaks down in this case. Thus I had to find a way to condense the size of the table without losing too much of the dimensionality. To do this, I examined two variables: tract-type and school-type.

1) Tract Variable

In clustering the census tracts based on socioeconomic data, there were four tract or neighborhood groupings: well-off homeowners, yuppie renters, average working-class families, and poor broken-homes. The yuppie and the well-off tracts were similar in that they both had higher than average income and higher than average educational attainment, but the yuppie tracts had much more rental housing and lower income than the well-off,

presumably because the yuppies are early in their careers. Thus the yuppies and the well-off are alike in that they are both above average socioeconomically, even though their stage in life differs (yuppies are young while the well-off are more established). This suggests that the yuppie and the well-off tracts could be combined into one category, namely the well-off.

To test this more formally, I checked whether dropping-out is independent of tract-type for the yuppie and the well-off tracts by testing the log-linear model with the two-factor interaction term set equal to zero:

$\log Mij = \mu + \mu d(i) + \mu t^*(j)$ where d = dropout variable
t* = tract variable (yuppie, well-off).

The expected cell values assuming independence are:

	YUPPIE	WELL-OFF	TOTAL:
DROPOUT	93.67 (100)	350.33 (344)	444
NORMAL	187.33 (181)	700.67 (707)	888
TOTAL:	281	1051	1332

and the summary statistics, with 1 degree of freedom, are:

X2 = 0.81 ==> Probability = 0.367

G2 = 0.81 ==> Probability = 0.369 .

Clearly, the expected cell counts, assuming independence, are very close to the observed cell counts as indicated by the small X2 and G2 statistics and the high probability values. Therefore we cannot reject the null hypothesis of independence at the .05 significance level, and we conclude that dropping out and tract-type <u>are independent</u> if we

consider only the well-off and the yuppie tracts. Thus it seems reasonable to collapse the yuppie and well-off tracts into one category designated 'well-off' for the higher-order model analysis.

(2) School Variable

Similarly to the tract variable, we want to collapse the school variable in order to reduce the size of the contingency table and eliminate sparse cells. The school variable has four categories: examination, magnet, regular, and other special-needs schools. I decided to condense the school variable into a single category, which effectively removes it from the model as a separate dimension, for the following reasons.

The examination and the special-needs schools are really special cases which are not representative of the general public school experience in Boston. Students must be highly motivated academically to gain admittance to the examination schools, and these schools historically have a very low dropout rate because of the student selection process. The special-needs schools, on the other hand, generally cater to students at the opposite end of the spectrum who exhibit behavioral or disciplinary problems. (There are some exceptions, for instance the Horace Mann school for the hearing-impaired.) These students are more likely than not to dropout, and these schools represent a last ditch attempt to help them.

Thus we know that the exam students probably will not drop out, and the special-needs students most likely will. Therefore I decided to eliminate the exam students and the special-needs students from the higher-order model analysis since we already understand their situations. The more interesting question is what happens to the typical student in the regular and the magnet public high schools. In the typical high schools, which students are more likely to encounter difficulties and drop out?

After eliminating the exam and the special-needs students, the next question is whether to use only regular students or whether to combine the regular and magnet students

into one category. If magnet and regular schools do not have significantly different dropout rates, then it makes sense to combine them. We can test whether dropping-out is independent of school-type for the regular and magnet schools as follows.

The independence model is:

$\log Mij = \mu + \mu d(i) + \mu s^*(j)$ where d = dropout variable
s* = school variable (regular, magnet).

Assuming independence, the fitted cell values are:

	MAGNET	REGULAR	TOTAL:
DROPOUT	457.92 (435)	872.08 (895)	1330
NORMAL	557.08 (580)	1060.92 (1038)	1618
TOTAL:	1015	1933	2948

and the goodness-of-fit statistics, with 1 degree of freedom, are:

X2 = 3.19 ==> Probability = 0.074

G2 = 3.19 ==> Probability = 0.074

Since X2 and G2 have probability greater than .05, we accept the null hypothesis that dropping-out and school-type are independent, if only the magnet and the regular schools are included. Although the magnet schools have a slightly smaller dropout rate than the regular schools (43% vs. 46%), the difference is not statistically significant at the .05 level. Therefore it makes sense to consolidate the magnet and the regular school students into a single category.

Reduced Table

Therefore, to reduce the number of sparse cells in the full multidimensional contingency table, we have collapsed two variables:

1) Tract now has three categories: poorest, average, and well-off (where well-off is the combination of yuppie and well-off).

2) School now has a single category combining regular and magnet schools. (Exam and special-needs students are excluded from the higher-order model analysis.)

The reduced multidimensional contingency table has:

dropout * race * gender * age * tract * school = (2) * (4) * (2) * (2) * (3) * (1)
= 96 cells

which is much more manageable.

B. Fitting Higher-Dimensional Models

To describe the structure of the Boston high school dropout data, I formulated two higher-dimensional log-linear models, one with and one without the age variable, using the reduced contingency table described above. The first model is four-dimensional with the dropout, gender, race, and tract variables, whereas the second model is five-dimensional including dropout, gender, race, tract, and age.

The previous use of log-linear models in Chapter VI was confirmatory in nature, however in this case, the focus is exploratory. In the chapter on two-way models, we performed specific hypothesis tests to determine, for example, if dropping-out is independent of race. Now we want to find the log-linear model which best fits the full data array in order to describe the relationships and interactions among all of the explanatory variables. To find the best-fitting model, I will use two techniques: uniform-order fitting and then backward elimination.

Uniform-Order Fitting

Uniform-order fitting is a technique used to narrow down the class of models under consideration. It assumes that the log-linear models are hierarchical, so no higher-order terms are included unless the related lower-order terms are also included. As applied to a four-dimensional contingency table, the technique is:

(1) fit the model with only main effects: [1][2][3][4] (in shorthand notation)
(2) fit the model with all two-factor interactions: [12][13][14][23][24][34]
(3) fit the model with all three-factor interactions: [123][124][134][234]
(4) fit the saturated model with the four-factor interaction: [1234] (perfect fit)

where 1 = first variable, 2 = second variable, 3 = third variable, and 4 = fourth variable.

The next step is to compare the goodness-of-fit of each model. (The saturated model will fit the data perfectly since the number of parameters equals the number of cells.) Because of the hierarchy principle, the μ-terms in model (1) are a subset of those in model (2) which are a subset of those in model (3) and so on. Thus there are four possible outcomes given the goodness-of-fit statistics:

a) model (4) fits, but model (3) doesn't
b) models (4) and (3) fit, but model (2) doesn't
c) model s (4), (3), and (2) fit, but model (1) doesn't
d) models (4), (3), (2), and (1) fit.

In each case, a higher-order model fits, but the model one order lower does not fit. This suggests that some model in between the two will fit the data with a minimum number of μ-terms, and we can use backward elimination to find the intermediary model.

Backward Elimination

Backward elimination is a stepwise procedure used to find the best-fitting intermediary model. First we start with the highest uniform-order model which fits the data. Then the idea is to eliminate μ-terms one at a time until we cannot eliminate any more terms without destroying the fit. At each step, the procedure eliminates the term which causes the smallest increase in the goodness-of-fit statistic. (A model with fewer terms will not fit the data as well, hence the difference between the observed and expected cell counts

increases, and the X2 and G2 statistics increase as we delete μ-terms.) Eventually the goodness-of-fit statistic will get so big that we cannot eliminate any terms without rejecting the model. When this happens, we conclude that the current model is the best-fitting model.

C. The Four-Dimensional Model

This model includes the four variables: dropout, race, gender, and tract. The following table contains the observed cell counts when students are cross-classified according to these four variables.

TRACT	GENDER	RACE	DROPOUT	NORMAL
POOREST	MALE	ASIAN	2	2
		BLACK	141	158
		HISPANIC	34	23
		WHITE	24	12
POOREST	FEMALE	ASIAN	0	1
		BLACK	130	191
		HISPANIC	32	32
		WHITE	26	5
AVERAGE	MALE	ASIAN	29	19
		BLACK	127	158
		HISPANIC	64	49
		WHITE	118	90
AVERAGE	FEMALE	ASIAN	2	20
		BLACK	120	194
		HISPANIC	42	55
		WHITE	79	106

WELL-OFF	MALE	ASIAN	22	30
		BLACK	59	99
		HISPANIC	22	17
		WHITE	107	102
WELL-OFF	FEMALE	ASIAN	8	22
		BLACK	45	112
		HISPANIC	25	36
		WHITE	72	85

This table should be read from left to right across rows. For example from the second row, there are 141 Black males from the poorest tracts who dropped out of high school, and 158 who finished high school normally.

After compiling the four dimensional contingency table, the next step is to fit the models of uniform order in order to determine where to start the backward elimination. Thus I fit the models with only main effects, with all two-factor interactions, and with all three-factor interactions. (The saturated model is the only model with the four-factor interaction term and it fits the data perfectly.) The goodness-of-fit statistics are given in the table below with the notation:

{ D = dropout variable, T = tract variable, R = race variable, G = gender variable}.

MODEL	DF	X2	PROB	G2	PROB
[DTRG]		0.00		0.00	
[DTR][DTG][DRG][TRG]	6	10.61	0.101	11.06	0.087
[DT][DR][DG][TR][TG][RG]	23	38.83	0.021	42.75	0.007
[D][T][R][G]	40	513.82	0.000	579.04	0.000

As expected, the X2 and G2 statistics decrease while the probability increases when we fit higher-order models. Thus the model with all three-factor terms fits better than the model with all two-factor terms, which in turn fits better than the model with only main effects. Furthermore, at the .05 level of significance, the model with all three-factor effects does fit the data (since the probabilities are greater than .05), but the model with all two-factor effects does not fit (since the probabilities are less than .05). Therefore we need a model with at least some of the three-factor effects to fit the data, and we will use the model with all three-factor effects to begin backward elimination of terms.

To perform backward elimination, we try eliminating each three-factor term one at a time and select the model with the least decrease in probability at each step. The next table summarizes the backward elimination by giving the term eliminated at each step and the fit of the subsequent model. (Obviously several models were fit at each step, but the details have been omitted for clarity.)

Step	Term Eliminated	Model	DF	G2	Prob
0	---	[DTR][DTG][DRG][TRG]	6	11.06	0.087
1	[DTG]	[DTR][DRG][TRG]	8	12.70	0.123
2	[TRG]	[DTR][DRG][TG]	14	21.50	0.090
3	[DTR]	[DRG][DT][TR][TG]	20	32.46	0.039 *

At Step 3, the [DTR] term is eliminated producing a model with probability .039 which is less than the significance level .05. Therefore, the model without the [DTR] term does not fit the data, and as a result, we must keep the [DTR] term in the model. Thus the model found in step 2 of the backward elimination fits the data with the least number of μ-terms.

Final Four-Dimensional Model

The final four-dimensional model is [DTR][DRG][TG] in shorthand, or in long form:

$$\log M_{ijkl} = \mu + \mu_{d(i)} + \mu_{r(j)} + \mu_{g(k)} + \mu_{t(l)} + \mu_{dt(il)} + \mu_{dr(ij)} + \mu_{rt(jl)} + \mu_{dg(ik)} + \mu_{rg(jk)} + \mu_{gt(kl)} + \mu_{drt(ijl)} + \mu_{drg(ijk)}.$$

This model includes all four main effects: dropout, race, gender, and tract, and all possible two-way interactions:

dropout * race
dropout * gender
dropout * tract
race * gender
race * tract.

Therefore none of the variables are independent of any other variable. Furthermore, the model includes two three-factor effects: dropout * race * tract, and dropout * race * gender. These three-factor terms suggest that dropping-out and race are not independent even if we control for tract and gender.

In conclusion, this model portrays a complex relationship between dropping-out and the other three explanatory variables (race, gender, tract). All the variables are correlated, and in particular, dropout and race are correlated even when other factors such as type of neighborhood and student's gender are taken into consideration. Given this model, clearly there is no one indicator, but rather a set of factors, which describe who drops out and who does not.

The next section examines a five-dimensional model with an additional explanatory variable, the student's age upon entrance to ninth grade, to see how the inclusion of age enhances the model's descriptive power.

D. The Five-Dimensional Model

The next model includes the age variable, hence it has five dimensions: dropout, race, gender, tract, and age. Similarly to the four-dimensional case, the goal of this analysis is to find the best-fitting log-linear model with a minimum number of μ-terms which describe the interactions among the explanatory variables for the high school dropout data.

First we crosstabulate students according to the five characteristics, giving the table below. This table should be read from left to right across rows. For instance, the second row indicates that 95 normal-age Black males from the poorest tracts dropped out, and 140 finished high school. In contrast, 46 over-age Black males from the poorest tracts dropped out, and only 18 finished high school normally.

			NORMAL-AGE		OVER-AGE	
			DROPOUT	NONDROP	DROPOUT	NONDROP
POOREST	MALE	ASIAN	1	1	1	1
		BLACK	95	140	46	18
		HISPANIC	19	20	15	3
		WHITE	14	12	10	0
POOREST	FEMALE	ASIAN	0	1	0	0
		BLACK	91	179	39	12
		HISPANIC	23	26	9	6
		WHITE	19	3	7	2
AVERAGE	MALE	ASIAN	8	13	21	6
		BLACK	85	140	42	18
		HISPANIC	43	44	21	5
		WHITE	92	84	26	6

AVERAGE	FEMALE	ASIAN	1	15	1	5
		BLACK	92	179	28	15
		HISPANIC	30	52	12	3
		WHITE	63	101	16	5
WELL-OFF	MALE	ASIAN	5	18	17	12
		BLACK	37	89	22	10
		HISPANIC	13	14	9	3
		WHITE	79	96	28	6
WELL-OFF	FEMALE	ASIAN	3	15	5	7
		BLACK	34	105	11	7
		HISPANIC	19	33	6	3
		WHITE	58	82	14	3

Now that we have tabulated the five-dimensional contingency table, the next step is to determine the best fitting uniform-order model. In the five-dimensional case, we fit the models with only main effects, with all two-factor interactions, with all three-factor interactions, and with all four-factor interactions. Presumably, one of the higher-order models will fit, while the next lower model does not. The table below gives the results of the uniform order fitting using the shorthand model notation: {D = dropout variable, T = tract variable, R = race variable, G = gender variable, and A = age variable}.

MODEL	DF	X2	PROB	G2	PROB
[DTRGA]		0.00		0.00	
[DTRG][DTRA][DTGA] [DRGA][TRGA]	6	7.91	0.245	7.89	0.246
[DTR][DTG][DTA][DRG] [DRA][DGA][TRG][TRA] [TGA][RGA]	29	21.75	0.830	21.94	0.823
[DT][DR][DG][DA][TR] [TG][TA][RG][RA][GA]	63	55.13	0.749	57.85	0.660
[D][R][G][T][A]	87	980.01	0.000 *	900.42	0.000 *

All the uniform-order models fit the data well except for the model with only main effects ([D][R][G][T][A]) which has probability less than the significance level of .05. The model with only main effects does not fit the data, but the model with all two-factor interactions does. Therefore we will use the model with all two-factor interactions as the starting point for backward elimination.

In backward elimination, we start with the model with all two-factor interactions and eliminate terms which have the least detrimental effect on the goodness-of-fit statistic. At each step, we fit every model with an additional two-factor term missing, and select the model with the smallest increase in G2 and subsequent decrease in probability. When we can no longer eliminate two-factor terms without rejecting the model, we are done. The next table summarizes the results of the backward elimination at each step.

Step	Term Eliminated	Model	DF	G2	Prob
0	---	[DT][DR][DG][DA] [TR][TG][TA][RG] [RA][GA]	63	57.85	0.660
1	[TG]	[DT][DR][DG][DA] [TR][TA][RG] [RA][GA]	65	58.46	0.704
2	[TA]	[DT][DR][DG][DA] [TR][RG] [RA][GA]	67	61.50	0.667
3	[GA]	[DT][DR][DG][DA] [TR][RG] [RA]	68	75.79	0.242
4	[DT]	[DR][DG][DA] [TR][RG] [RA]	70	93.28	0.033 *

In the last step, the two-factor term [DT] is eliminated, and as a result, the model no longer fits the data since the probability is less than .05 that the discrepancy between the observed and the expected cell values is a result of chance alone. Therefore [DT] must be included in the model, and no other terms can be eliminated without rejecting the model. Thus the model from step 3 is the final model.

Final Five-Dimensional Model

In shorthand, the final model is: [DR][DG][DT][DA][TR][RG][RA]

and in long form, the final five-dimensional model is:

$$\log M_{ijklm} = \mu + \mu_{d(i)} + \mu_{r(j)} + \mu_{g(k)} + \mu_{t(l)} + \mu_{a(m)} + \mu_{dr(ij)} + \mu_{dg(ik)} + \mu_{dt(il)} + \mu_{da(im)} + \mu_{rt(jl)} + \mu_{rg(jk)} + \mu_{ra(jm)}.$$

In the five-dimensional case, the final model includes all five main effects: dropout, race, gender, tract, and age, and some, but not all, of the two-factor interactions:

dropout * race
dropout * gender
dropout * tract
dropout * age
race * tract
race * gender
race * age.

This indicates that dropping-out depends on race, gender, tract, and age. In addition, a student's tract, gender, and age are also related to his race. However, three of the two-factor effects are not included in the model. In particular, the μ-terms for:

gender * tract
gender * age
tract * age

are set equal to zero. Therefore a student's age is independent of gender and tract, and gender is also independent of tract. It makes sense that males and females are uniformly distributed across neighborhoods, however the age-tract independence is more interesting and somewhat surprising. This suggests that over-age students, who typically suffer serious academic problems prior to high school, are not heavily concentrated in some parts of the city and lightly in others. Also since gender and age are independent, females are as likely to be over-age as males.

E. Comparison of the Four and Five-Dimensional Models

At this point, we have two higher-dimensional log-linear models which describe the structure in the high school dropout data well. The four-dimensional model does not include the age variable, whereas the five-dimensional model does. The table below compares the two models.

Model	Number of Variables	Number of μ-terms	G2	DF	Probability
Four-Dim	4	13	21.50	14	0.090
Five-Dim	5	13	75.79	68	0.242

Although the five-dimensional model has more variables (5 instead of 4), both models have the same number of μ-terms (13). The four-dimensional model appears to fit the data better since it has a smaller G2 statistic (21.5 versus 75.79), however the sacrifice in degrees of freedom takes its toll on the probability value. Even though the five-dimensional model has a larger G2 statistic, it has many more degrees of freedom, and hence the corresponding probability value is much higher than that for the four-dimensional model. The probability that the discrepancy between the expected and the observed values is a result of chance alone is 24.2 percent for the five-dimensional model, but it is only 9 percent for the four-dimensional model. Furthermore, the four-dimensional model includes some three-factor terms whereas the five-dimensional model has only main effects and two-factor terms.

Since three-factor terms are more difficult to interpret and the five-dimensional model provides a better fit in terms of the probability value, I believe the five-dimensional model including the age variable is more useful in understanding the structure of the dropout data. Therefore, in the next chapter we will examine the five-dimensional model more closely in order to interpret what the model tells us about the determinants of dropping-out.

Chapter VIII
Interpretation of the Five-Dimensional Model

In the previous chapter, we formulated two higher-dimensional models, one with and one without the age variable, which fit the high school dropout data with a minimum number of μ-terms. In comparing the models, the five-dimensional model provided a better fit for the number of degrees of freedom, and it included only two-factor interactions which are easier to interpret. Hence this chapter will focus on the interpretation of the five-dimensional model in describing who drops out versus who stays in school.

A. The Model

The final five-dimensional model, derived in the previous chapter, has five variables: dropout, race, gender, tract, and age, and the following form:

$$==> \quad [DR][DG][DT][DA][TR][RG][RA]$$

$$==> \quad \log M_{ijklm} = \mu + \mu_{d(i)} + \mu_{r(j)} + \mu_{g(k)} + \mu_{t(l)} + \mu_{a(m)} + \mu_{dr(ij)} + \mu_{dg(ik)} + \mu_{dt(il)} + \mu_{da(im)} + \mu_{rt(jl)} + \mu_{rg(jk)} + \mu_{ra(jm)}.$$

Table 1 on the following page contains the expected cell counts given the five-dimensional unsaturated log-linear model.

According to this model, dropping-out depends on all four explanatory variables: race, gender, tract, and age. Moreover, there is correlation between race and tract-type, gender, and age. Not all of the two factor effects are present however. Specifically, age is independent of gender, and both are independent of tract-type.

Table 1. Predicted Cell Counts Under The Five-Dimensional Model

			Normal-Age		Over-Age	
			Dropout	Non-Drop	Dropout	Non-Drop
Poorest	Male	Asian	0.94	1.87	2.13	0.97
		Black	105.15	139.49	45.80	13.95
		Hispanic	24.38	21.49	12.81	2.59
		White	18.93	13.61	5.86	0.97
Poorest	Female	Asian	0.40	1.16	0.91	0.60
		Black	90.28	172.73	39.33	17.27
		Hispanic	21.60	27.46	11.35	3.31
		White	13.07	13.55	4.04	0.96
Average	Male	Asian	6.84	17.12	15.46	8.88
		Black	88.96	148.50	38.75	14.85
		Hispanic	37.48	41.58	19.70	5.02
		White	95.73	86.60	29.62	6.15
Average	Female	Asian	2.94	10.61	6.64	5.50
		Black	76.38	183.90	33.27	18.39
		Hispanic	33.21	53.14	17.46	6.41
		White	66.09	86.23	20.45	6.12
Well-Off	Male	Asian	7.02	21.65	15.87	11.23
		Black	41.25	84.85	17.97	8.48
		Hispanic	16.33	22.33	8.59	2.69
		White	80.51	89.74	24.91	6.37
Well-Off	Female	Asian	3.02	13.42	6.82	6.96
		Black	35.42	105.08	15.43	10.51
		Hispanic	14.47	28.53	7.61	3.44
		White	55.58	89.35	17.20	6.35

Although it is important to identify correlations among variables, we must go beyond simply stating that dropping-out is dependent on race. The remainder of this chapter will examine the magnitude and direction of these dependencies for specific groups of students to answer questions such as: how does dropping-out depend on race when we take into consideration the student's gender age, and tract-type? In this case, are Blacks or Whites more inclined to drop out?

B. The μ-Terms

The log-linear model computes expected cell counts as linear combinations of variable "effects" where there is an overall effect (μ), five main effects (μd, μr, μg, μt, μa), and seven two-factor effects (μdr, μdg, μdt, μda, μrg, μrt, μra). By summing the logarithms of the predicted cell counts under the log-linear model, we can compute the magnitude and the direction of these μ-terms using formulas such as:

grand mean: $\mu = 1 / IJKLM \sum \log M_{ijklm}$ (over i,j,k,l,m)

main effect [D]: $\mu_{d(i)} = [\ 1 / JKLM \sum \log M_{ijklm}$ (over j,k,l,m)] - μ

two-factor effect [DR]:

$$\mu_{dr(ij)} = [1 / KLM \sum \log M_{ijklm} \text{ (over k,l,m) }] - [\ \mu + \mu_{d(i)} + \mu_{r(j)}\ ].$$

Then we can interpret the magnitude and the direction of the μ-terms for particular groups of students.

Since the μ-terms represent deviations from lower-order effects, we can add up all the μ-terms for a particular (i,j,k,l,m) to get the logarithm of the expected count for that cell in the multidimensional contingency table. In other words, the sum of the μ-terms gives the logarithm of the number of students which have the five characteristics (i,j,k,l,m). A positive μ-term means that we add students to the count for that cell because there are more

students than average who have that characteristic. Similarly, a negative μ-term means that we subtract students from the count for that cell since there are fewer students with that characteristic. The absolute value of the μ-term indicates the magnitude of the effect. Hence a large μ-term implies a large deviation from the mean, whereas a small μ-term indicates only a minor deviation for that effect. Thus large μ-terms add or subtract relatively large numbers of students to the cell count, whereas small μ-terms modify the cell count only slightly.

The following tables contain the μ-terms for the unsaturated five-dimensional log-linear model.

Grand Mean		
μ-term	Estimated Value based on model	Students
μ	+2.666	All

This is the overall mean effect which contributes equally to the logarithm of the expected count in every cell of the five-dimensional table.

Main Effects

The main effects reflect the underlying structure of the student population, and in general, they confirm the results of the frequency analysis in Chapter VI.

1. Dropout Effect

Dropout		
μd(1)	+0.100	Dropout
μd(2)	-0.100	Non-dropout

The sign of this main effect is bewildering. Since there are fewer dropouts than non-dropouts in the cohort group, we would expect this μ-term to be negative for dropout and positive for non-dropout. Nevertheless I checked it several times and confirmed that the sign is correct. However, the magnitude of this μ-term indicates that the dropout effect is relatively small. In fact, the overall dropout rate is close to fifty percent (45%), and this may account for the discrepancy in the sign. Moreover, the dropout variable is highly correlated with the other explanatory variables, and the sign of the main dropout effect may change as a result of controlling for the other variables in the five-dimensional model.

2. Race Effect

Race		
μr(1)	-1.232	Asian
μr(2)	+1.107	Black
μr(3)	-0.073	Hispanic
μr(4)	+0.198	White

The main effect for race indicates that the students are not uniformly distributed across racial groups, as we found in the Chapter VI discussion of frequencies. In particular, there are more Blacks in the cohort group than Whites, more Whites than Hispanics, and lastly more Hispanics than Asians. Thus the Blacks are the largest group and the Asians are the smallest.

3. Gender Effect

Gender		
μg(1)	+0.095	Male
μg(2)	-0.095	Female

There are slightly more males than females in the cohort group, but the difference is not great since the magnitude of the gender-effect is relatively small.

4. Tract Effect

Tract		
μt(1)	-0.611	Poorest
μt(2)	+0.471	Average
μt(3)	+0.140	Well-Off

As expected, the tract effect indicates that more students live in the average working-class neighborhoods than in the poorest or the most well-off communities.

5. Age Effect

Age		
μa(1)	+0.596	Normal-Age
μa(2)	-0.596	Over-Age

The last main effect, age, is large in magnitude and shows that there are many more normal-age students in the cohort group than over-age students.

Two-Factor Effects

The signs and magnitudes of the two-factor μ-terms in the tables below indicate which student groups are more or less likely to drop out. The three other two-factor effects (race * tract, race * gender, and race * age) suggest additional interactions among the explanatory variables in the data.

1. Dropout and Race Effect

Dropout * Race	Dropout	Non-Dropout
Asian	μdr(11) = -0.279	μdr(21) = +0.279
Black	μdr(12) = -0.076	μdr(22) = +0.076
Hispanic	μdr(13) = +0.127	μdr(23) = -0.127
White	μdr(14) = +0.228	μdr(24) = -0.228

Since the model includes the two-factor interaction term for dropout and race, we know that dropping-out of high school is not independent of race, instead some racial groups are more prone to dropping-out than others. In Chapter VI, we analyzed the bivariate relationship between dropping-out and race, but did not take into account the other explanatory variables. In the bivariate case, we found that Hispanics drop out the most, followed by Whites, then Blacks, and Asians dropped out the least. In contrast, the five-dimensional model evaluates the relationship between race and dropping-out in the presence of the other factors (age, gender, and tract) and the results are different.

Similarly to the bivariate case, the Asians are the least likely to drop out since the corresponding μ-term is negative, and the Hispanics are more likely than average to drop out. However, when the five-dimensional model controls for other factors, the Black students are less likely than average to drop out (since the μ-term is negative, -0.076), whereas the White students drop out more than any other racial group. In terms of percentages, 40% of Asians, 41% of Blacks, 51% of Hispanics, and 52% of Whites dropped out.

Thus if we consider the experience of the 'average' Boston public school student, the Whites are most likely to drop out. The average Black high school student is, in fact, less likely to drop out. Moreover, recently there has been great concern over the high dropout rate among Hispanics (Ribadeneira 1990), but according to this analysis, the average White student is worse off than the average Hispanic in the regular public high schools.

This result is unexpected for three reasons. First, the general consensus in the literature is that Hispanic and Black students are more "at-risk" than Whites and Asians (Byrne 1988, Capuzzi 1989, Frase 1989, Gallington 1966, Horst 1990, Pallas 1984). Second, the overall dropout rate for Whites is relatively low compared to Blacks and Hispanics in Boston (Byrne 1988, Horst 1990). Third, in Chapter VI the dropout vs. race two-way model indicated that the Hispanic students have the highest dropout rate. Thus, it is unexpected that the average White students have the highest dropout rates.

This surprising finding is probably a by-product of who attends the regular Boston public high schools. First, this model does not include the examination school students, many of whom are White and rarely drop out. Hence the overall dropout rate for Whites is lower because the examination school students, who have extremely low dropout rates, pull down the average. When the examination students are excluded, the White dropout rate increases substantially.

Second, the White students who attend the regular public high schools are not a representative sample of the White population in Boston. In general, the Whites are better off socioeconomically and hence choose to send their children to private or parochial schools. Only forty-percent of the White students in Boston actually attend the public schools, and still fewer attend the non-examination schools (Byrne 1988). On the other hand, the Blacks are generally less well-off and do not have the financial means to send their children to private schools. Thus the average Black student has no choice but to attend the regular public schools, whereas only the worst-off Whites attend these schools. Therefore the high dropout rate for Whites is indicative of the differential in socioeconomic status across racial groups and the subsequent imbalance in the public school student-body.

2. Dropout and Gender Effect

Dropout * Gender	Dropout	Non-Dropout
Male	$\mu dg(11) = +0.091$	$\mu dg(21) = -0.091$
Female	$\mu dg(12) = -0.091$	$\mu dg(22) = +0.091$

This table indicates that males drop out more than females since the (dropout * male) μ-term is positive (49% of males and 41% of females dropped out). This result is not unusual as it has been shown that males usually have a higher dropout rate than females (see Chapter II on high school dropouts). However, this μ-term is quite small in magnitude (less than 0.1 in absolute value). This suggests that, although there is a significant difference between males and females in dropping out, the difference is not large in comparison with the other two-factor effects.

3. Dropout and Tract Effect

Dropout * Tract	Dropout	Non-Dropout
Poorest Broken-Home	$\mu dt(11) = +0.111$	$\mu dt(21) = -0.111$
Average Working-Class	$\mu dt(12) = -0.003$	$\mu dt(22) = +0.003$
Well-Off	$\mu dt(13) = -0.108$	$\mu dt(23) = +0.108$

The inclusion of this term in the model suggests that neighborhood socioeconomic status is a factor in describing the differential in dropout rates. In fact these μ-terms show that socioeconomic status is inversely correlated with dropout rates. Students from the well-off homeowner and yuppie renter census tracts in Boston drop out the least (41.7 percent). At the other end of the socioeconomic spectrum, students from the poorest broken-home neighborhoods drop out the most (47.8 percent). The students from average working-class neighborhoods lie somewhere in between (45.7 percent dropped out). The magnitudes of these μ-terms are relatively small suggesting that this two-factor effect is not extremely large although it is significant.

4. Dropout and Age Effect

Dropout * Age	Dropout	Non-Dropout
Normal-Age	$\mu da11) = -0.367$	$\mu da(21) = +0.367$
Over-Age	$\mu da(12) = +0.367$	$\mu da(22) = -0.367$

According to these μ-terms, over-age students drop out much more than normal-age students which is expected. Less than forty percent of normal-age students dropped out, but over seventy percent of over-age students did not finish high school. Again, we

defined over-age to mean that the student entered high school at least two years older than normal. Most of these students were held back repeatedly in elementary school, thus over-age is indicative of past academic problems and most likely future difficulties as well. Often these students suffer constant failure and dislike school as a result (Capuzzi 1989, Horst 1990). Eventually they just give up and drop out. The magnitude of this μ-term is high (almost 0.4), so the dropout * age effect is large.

5. Race and Tract Effect

Race * Tract	Asian	Black	Hispanic	White
Poorest	μrt(11) = -0.832	μrt(21) = +0.867	μrt(31)=+0.489	μrt(41) = -0.524
Average	μrt(12) = +0.184	μrt(22) = -0.268	μrt(32) = -0.048	μrt(42) = +0.129
Well-Off	μrt(13) = +0.648	μrt(23) = -0.599	μrt(33) = -0.441	μrt(43) = +0.395

The magnitudes of some of these μ-terms are very high (from 0.5 to 0.8 in absolute value) suggesting that the race * tract effect is not only significant, but very large and important in describing the dropout data. The signs of the μ-terms indicate that most Asians and Whites live in the well-off neighborhoods (since the μ-terms are positive 0.648 and 0.395 respectively) while very few live in the poorest broken-home areas. On the other hand, most of the Blacks live in the poorest areas of the city. Few Black students live in the average working-class neighborhoods, and still fewer live in the well-off communities. For example, forty-percent of the Black students live in the poorest census tracts as compared to only eight percent of the Whites. The Hispanics are similar to the Blacks (more live in the poorer neighborhoods), but the magnitude of the effect is not as great for Hispanics. Interestingly, more Asians than Whites live in the well-off tracts. In

summary, the Black and Hispanic students come from poor communities, and the Asian and White students live in the average and well-off neighborhoods.

Typically we would expect students from the poorer communities to drop out more often because of financial pressures to work full-time, lack of parental support, and so on. However this data presents a paradox: the White students from the well-off communities drop out more, and the Black students from the poorer neighborhoods drop out less.

Perhaps this paradox stems from the residential segregation in Boston. Boston has White neighborhoods and Black neighborhoods, and there is very little integration within the city (see Chapter IX for maps). One explanation of the apparent paradox is that these White students come from the generally well-off "White" section of town, but these particular families are in fact relatively poor. Thus the average socioeconomic status for the neighborhood is relatively high, but the only kids who go to the non-private, non-parochial, non-examination schools are from the few poor families in the neighborhood. In this case, the use of census data as a general approximation for individual characteristics breaks down.

Another possibility is that programs in the minority communities have had success in reaching some youngsters, and providing them with the means or the desire to stay in school, whereas the average White public school student sees little value in completing a high school education. Clearly both of these explanations are speculative, and more information on the individual students is needed. Nevertheless, the paradox is intriguing

6. Race and Gender Effect

Perhaps surprisingly, there are also significant differences across races by gender. The μ-terms in the table below show that there are more Asian males than females, whereas there are more female Blacks and Hispanics than males in this student group. Apparently there are about the same number of White males and females since the μ-term is approximately zero.

Race * Gender	Male	Female
Asian	$\mu rg(11) = +0.237$	$\mu rg(12) = -0.237$
Black	$\mu rg(21) = -0.110$	$\mu rg(22) = +0.110$
Hispanic	$\mu rg(31) = -0.126$	$\mu rg(32) = +0.126$
White	$\mu rg(41) = -0.002$	$\mu rg(42) = +0.002$

The μ-terms are not large except for the Asians. In this particular cohort group, sixty-six percent of the Asian students are male and thirty-four percent are female. This could be because there are fewer Asian students overall and excluding the examination students further decreases the number of students thereby creating an imbalance between males and females due to the small sample size. This gender imbalance for Asians could also be due to migration patterns, especially in light of the next table which indicates that Asian students are most likely to be over-age.

7. Race and Age Effect

Race * Age	Normal-Age	Over-Age
Asian	$\mu ra(11) = -0.635$	$\mu ra(12) = +0.635$
Black	$\mu ra(21) = +0.187$	$\mu ra(22) = -0.187$
Hispanic	$\mu ra(31) = +0.093$	$\mu ra(32) = -0.093$
White	$\mu ra(41) = +0.357$	$\mu ra(42) = -0.357$

These μ-terms suggest that the White students are most likely to be normal-age, followed by Blacks, and then Hispanics. When compared to the Whites and Blacks, there are relatively more over-age Hispanics indicating that more Hispanic students have

encountered serious academic difficulties prior to high school. (A "more negative" μ-term means that fewer than average students are in this category.) However, the race * age effect is much smaller for Whites, Blacks, and Hispanics than for the Asians. More Asian students are over-age than any other racial group, even the Hispanics, which is unexpected given the current literature on dropouts (Ribadeneira 1990).

Furthermore, the magnitude of these μ-terms is high, so the race * age effect is large, especially for Asians, in comparison to the other two-factor effects. In particular, almost fifty percent of the Asian students are over-age, but only fifteen percent of whites are over-age. Thus the "average" Asian public high school student in Boston is much more likely to be over-age than his counterparts. Again this presents a paradox since more of the Asian students are over-age but fewer of them drop out.

Perhaps in this case, over-age indicates a language problem, specifically a lack of English proficiency, as opposed to an academic dysfunction. If these Asian students are recent immigrants, they may be older simply because they have been out of school for a couple of years in the midst of migrating to this country. Or, they may be over-age because it has taken them longer to learn English. This would explain why the Asians are most likely to be over-age, but have the lowest dropout rate. In this case, age is a measure of English language proficiency instead of academic deficiency. The same may be true for the Hispanics, although to a lesser extent, since fewer Hispanics than Asians are over-age.

Check of μ-Terms

Now that we have calculated all of the μ-terms, we can combine them to get the expected counts for each group of students with a particular set of characteristics. For example, consider the poorest normal-age Asian male dropouts. Cell (i,j,k,l,m) = (1,1,1,1,1) contains the model's predicted count for these students.. Using the μ-terms, we can verify the count as follows.

$$\log M_{ijklm} = \mu + \mu d(i) + \mu r(j) + \mu g(k) + \mu t(l) + \mu a(m) + \mu dr(ij) + \mu dg(ik) + \mu dt(il) + \mu da(im) + \mu rt(jl) + \mu rg(jk) + \mu ra(jm)$$

$$\Longrightarrow \quad \log M_{11111} = \mu + \mu d(1) + \mu r(1) + \mu g(1) + \mu t(1) + \mu a(1) + \mu dr(11) + \mu dg(11) + \mu dt(11) + \mu da(11) + \mu rt(11) + \mu rg(11) + \mu ra(11)$$

$$\Longrightarrow \quad \log M_{11111} = 2.666 + 0.100 - 1.232 + 0.095 - 0.611 + 0.596 - 0.279 + 0.091 + 0.111 - 0.367 - 0.832 + 0.237 - 0.635$$

$$\Longrightarrow \quad \log M_{11111} = (\text{add for mean effect}) + (\text{add because dropout}) + (\text{subtract because Asian}) + (\text{add because male}) - (\text{subtract because from poor tract}) + (\text{add because normal-age}) - (\text{subtract because Asian dropout}) + (\text{add because male dropout}) + (\text{add because poor dropout}) - (\text{subtract because normal-age dropout}) - (\text{subtract because poor Asian}) + (\text{add because Asian male}) - (\text{subtract because normal-age Asian})$$

$$\Longrightarrow \quad \log M_{11111} = -0.06$$

$$\Longrightarrow \quad M_{11111} = 0.94$$

Thus, this cell count is the same as that in the table of expected cell counts given at the beginning of the chapter. This also verifies that the sign of the main dropout effect is correct even though it is counter-intuitive.

C. Dropout Rates

Now that we understand the direction and the magnitudes of the variable effects on different groups of students, the next step is to examine dropout rates and identify groups of students, if any, experiencing a mass dropout phenomena. Table 2 on the next page gives the dropout rates predicted by the five-dimensional log-linear model for student groupings based on race, gender, age, and tract.

Table 2. Predicted Dropout Rates Under The Five-Dimensional Model

Dropout Rates				
Tract	Gender	Race	Normal-Age	Over-Age
Poorest	Male	Asian	0.33	0.69
		Black	0.43	0.77
		Hispanic	0.53	0.83
		White	0.58	0.86
Poorest	Female	Asian	0.26	0.60
		Black	0.34	0.69
		Hispanic	0.44	0.77
		White	0.49	0.81
Average	Male	Asian	0.29	0.64
		Black	0.37	0.72
		Hispanic	0.47	0.80
		White	0.53	0.83
Average	Female	Asian	0.22	0.55
		Black	0.29	0.64
		Hispanic	0.38	0.73
		White	0.43	0.77
Well-Off	Male	Asian	0.24	0.59
		Black	0.33	0.68
		Hispanic	0.42	0.76
		White	0.47	0.80
Well-Off	Female	Asian	0.18	0.49
		Black	0.25	0.59
		Hispanic	0.34	0.69
		White	0.38	0.73

These dropout rates confirm the previous analysis of the μ-terms. According to the model, the White students are consistently the worst off since they have the highest dropout rate in every category of tract, gender, and age. The Hispanics have the next highest dropout rates, while the Asians and the Blacks are comparatively less likely to drop out. The dropout rates range from a low of 18% for normal-age well-off Asian females to a high of 86% for the over-age poorest White males. Thus the model indicates that there is significant variation in dropout rates across student groups.

As expected, the over-age students have substantially higher dropout rates than the normal-age students. Several of the over-age groups have dropout rates exceeding eighty percent which means that only one in five of these students finishes high school.

For the normal-age students, the differential in dropout rates across races decreases slightly as socioeconomic status increases. For example, there is a smaller difference in dropout rates between Asians (lowest dropout rate) and Whites (highest dropout rate) in the well-off neighborhoods than in the poor neighborhoods.

In every category of race, gender, tract, and even age, the dropout rates for females are lower than those for males. Also, the dropout rates are consistently higher in the poorer tracts than in the average and well-off tracts. For the normal-age students, the difference in dropout rates across tract-type is approximately the same for every racial group. However, for the over-age students, there is more variation in dropout rates across tract-type for the Asians and the Blacks. This means that tract-type has less effect on the dropout rates of over-age Hispanics and Whites than on the Asians and Blacks, but the difference is not large.

Summary

In this chapter, we interpreted the unsaturated five-dimensional log-linear model which describes the structure in the Boston high school dropout data. First we examined

the μ-terms to determine the direction and magnitude of the main effects and the two-factor effects included in the model, and then we looked at the dropout rates for different groups of students.

The signs of the individual μ-terms indicated the direction and magnitude of the effects for different groups of students. We found that males drop out more than females, normal-age students drop out much less often than over-age students, and the students from the poorest census tracts drop out more than those from the average working-class or well-off neighborhoods. More of the Black and Hispanic students live in the poorest communities, while the Whites and the Asians inhabit the average and well-off areas of the city. We also found that there are more Asian males than females, but fewer males than females in the other racial groups.

When the five-dimensional model controlled for the other explanatory variables, we found that the Whites are more likely to drop out than any other racial group. Hispanics have the next highest dropout rate, followed by Blacks, and lastly Asians. This is unexpected because the overall dropout rate is lower for Whites, and the literature on high school dropouts suggests that the minority students (excluding the Asians) are most at-risk. This result is also counter to expectations because the Whites live in the neighborhoods with higher socioeconomic status, and usually we associate high socio-economic status with lower dropout rates (Pallas 1984).

The direction of the race * age interaction was also unexpected. We found that the Asian students were the most likely to be over-age, especially male Asians. Whites, on the other hand, were least likely to be over-age. Since Asians actually have the lowest dropout rate, it appears that age may measure language proficiency as opposed to academic difficulties.

In the next chapter, we will generate maps of student residences in Boston for different groups of students. Then we can visually interpret any spatial patterns in the dropout data.

Chapter IX
Spatial Analysis

A. Introduction

In the previous chapters, we crosstabulated students according to six characteristics: their race, gender, age, type of neighborhood, type of school, and whether or not they dropped out. Then we formulated a multi-dimensional log-linear model which describes the structure in the dropout data, and interpreted the model to identify groups of students which are more or less likely to drop out of high school. This chapter will examine the geographical aspect of the dropout data, in particular where different groups of students live in Boston and whether there are any spatial trends which can be more readily identified visually.

If there are residential pockets in Boston where students dropout in mass, then the schools, the parents, the community, and business leaders can use this information to target specific neighborhoods for dropout prevention programs. However, we must be careful not to label or stigmatize a set of neighborhoods negatively as dropout-prone for this may only exacerbate the problem.

B. Description of Maps

There are two types of maps included in this chapter: thematic maps and point maps. The thematic maps show the Boston census tracts shaded according to the level of some attribute such as socioeconomic status or dropout rate. In contrast, the point maps use points to indicate the location of student residences within Boston. Thus we can select a subset of students, for instance all overage Hispanic males who dropped out, and map their residences to see if they live throughout the city or if they tend to live nearby one another.

To generate the thematic and point maps, I used the XMAP[1] mapping program on an IBM RT workstation running UNIX (Ferreira and Wiggins 1990). XMAP requires a boundary file, an attribute file for thematic maps, and a point file for point maps. The boundary file contains the map boundaries. For the thematic maps, I used a boundary file which outlines each census tract in Boston, and for the point maps, I used a less detailed boundary file which shows the neighborhood statistical areas in Boston. I used the less detailed boundary file for the point maps in order to avoid too much clutter in the image. Thus the two types of maps appear slightly different because they do not have the exact same boundaries.

The point coordinates for each student's residence were calculated as described previously in the section on address-matching. Basically the students' addresses were matched against the Boston Dime (street network) file in order to compute longitude and latitude coordinates. Then the longitude, latitude coordinates were converted into X,Y coordinates which represent inches on the digitizer. The point coverages which represent student residences can be combined with the neighborhood statistical area boundary map of Boston to form a single image. Unfortunately, the boundary map is not well registered so the two coordinate systems are slightly off, and as a result, some of the points do not fall exactly within the map boundaries. However, they are close enough to convey the general idea of where the students live.

For the thematic maps, the boundary map outlines all the cities in Suffolk County. Hence, Winthrop and Chelsea are also included in the map of "Boston". Since this study examines only the city of Boston, there is no data for these surrounding communities, and the corresponding census tracts are left blank on the thematic maps.

---

[1] written at the MIT Computer Resource Laboratory, in the School of Architecture and Planning, by Phil Thompson, et al, under the supervision of Professor Joseph Ferreira. It is a thematic mapping tool for use in research and provides a high-degree of interaction and tools to juxtapose data obtained from many sources.

The following maps are described below and attached at the end of this chapter:

1. all students
2. students by race (4 maps)
3. students by school-type (3 maps)
4. tract clusters
5. dropout rate by tract
6. selected groups with high dropout rates (4 maps).

Map 1 - All Students (Figure 3)

This is a point-map which shows the residences of all students in the 1982-86 cohort group who attended the Boston public high schools (except those students whose addresses could not be matched against the Boston street network file.) Since there are many students (4,940 to be exact), portions of the map are blackened with points indicating that the student population is quite dense in those areas of the city.

Clearly the students are not uniformly spread throughout the city. For instance, the airport and Franklin Park have no students which makes sense because these areas are not residential. The most dense student population forms a crescent shaped moon around the Northeast end of Franklin Park in Roxbury, Mattapan, and Dorchester. On the other hand, the West Roxbury and Brighton areas are less dense with students, and the Back Bay has even fewer students.

Maps of Students by Race

The next four maps depict the residences of students by race. The degree of segregation in the residential communities of Boston is astounding. The Black students live in one part of the city and the White students live in another, with very little overlap. In fact, even the Asian and Hispanic families live in their own isolated communities, although the Blacks and the Hispanics appear to be the most integrated of any of the racial groups.

Map 2A - Black Students (Figure 4)

This map shows the residences of the Black students in the cohort group. The Black students are concentrated in the area surrounding the Northern and Eastern sides of Franklin Park in communities such as Dorchester, Roxbury, and Mattapan. The points are very close together indicating that the Black students inhabit the most dense areas of Boston.

Map 2B - White Students (Figure 5)

In stark contrast to the Blacks, the White students live on the peripheries of Boston with concentrations in West Roxbury, Brighton, South Boston, Charlestown, and East Boston. Moreover, the residential areas of White students appear to be less dense when compared to the neighborhoods of the Black students.

Map 2C - Asian Students (Figure 6)

The Asian student population is the smallest of any racial group so there are fewer points on this map. The Asians appear to live primarily in two clusters. There is one group in Chinatown and a second cluster on the western border of Brookline.

Map 2D - Hispanic Students (Figure 7)

Similarly to the Asians, there are fewer Hispanic students and hence fewer points on the map of their residences. The Hispanics overlap somewhat with the Blacks, and there appear to be three of four clusters of Hispanic residents in the central part of the city just north and on either side of Franklin Park.

Maps of Students by School-Type

The next three maps show the location of students attending the three main types of schools in Boston: regular, magnet, and examination public high schools.

Map 3A - A Regular High School (Figure 8)

This map shows the cohort students who attended Burke High School (mapped by residence). Burke is a regular public high school, and as such most of its students come from the local surrounding area. There are only one or two students who attend Burke but live elsewhere in the city. Most of the regular schools are similar to Burke, however some of the schools draw students from elsewhere in the city as a result of busing (for desegregation).

Map 3B - A Magnet High School (Figure 9)

This is a map of the residences of cohort students from Madison Park High School. At the time (1982-86), Madison Park was a magnet high school located near an occupational resource center. It attracted students by offering a half-day school and half-day vocational program. As expected, Madison Park had a wider draw of students than the regular schools. The points on the map are more dispersed indicating that students came from elsewhere in the city to attend Madison Park. It appears that most of Madison Park's students came from the Black neighborhoods.

Map 3C - An Examination School (Figure 10)

Boston Latin High School is one of the three examination schools in the Boston public school district, and this map shows the residences of the students in the cohort group who attended Boston Latin. Clearly Boston Latin has the widest draw of the three schools mapped since it attracts students from all over Boston. The largest residential concentration appears to be students from Chinatown, but in general the students live throughout Boston including the minority communities. Thus the examination school's reputation for quality is indeed powerful in attracting students from throughout the city.

Map 4 - Tract Clusters (Figure 11)

In order to reduce the multidimensional contingency table to a manageable size for the statistical model, the census tracts were clustered according to four socioeconomic variables: income, percent renters, percent single parent households, and percent high school graduates. The cluster analysis yielded four groupings of tracts: well-off homeowners, yuppie renters, average working-class families, and poor broken-home households.

This thematic map shows the partitioning of Boston by tract-cluster and thus indicates the socioeconomic status of different areas of the city. The shading is inversely correlated with socioeconomic status, so darker shades imply lower socioeconomic status. Thus the well-off homeowners inhabit the outskirts of Boston, while the average working class-families and poor households are located in the central core of the city. The yuppie renters live in the Back Bay area, on the Shawmut Peninsula, and along the western border of Brookline. The average working-class families live in South Boston and East Boston, while the well-off live in the West Roxbury region near Brookline and part of Charlestown. The very poor live in Roxbury, Dorchester, Mattapan and Jamaica Plain.

The geographical location of tract clusters confirms the race-by-tract results from the statistical model. In particular, the model indicated that whites live in the well-off tracts whereas more Blacks live in the poor neighborhoods. The maps of student residences showed that the Whites live on the outskirts of the city especially in the lower end, and according to the tract-cluster map, these regions are mostly well-off. On the other hand, the Black students live in the core of the city, and this map suggests that those communities are either very poor or average working-class. Thus the map of census tract-clusters and the maps of student residences by race verify the results of the statistical model.

Map 5 - Dropout Rate by Tract (Figure 12)

This thematic map shows the dropout rate by census tract in Boston. For each tract, the dropout rate was calculated as the number of dropouts divided by the sum of the number of dropouts and the number of students who finished normally. (The denominator did not include transfers.) Also these calculations were done for the reduced cohort group used in the five-dimensional log-linear model. Hence it does not include Native American students, transfers, exam school students, and special-needs school students. Therefore this map depicts the likelihood that the 'average' public high school student who lives in a particular neighborhood in Boston will drop out.

The darker shades on the map indicate a higher dropout rate. Because the dropout rates were calculated for the reduced cohort group, some of the census tracts did not have any student residents, particularly in the Back Bay. These tracts have been left blank to indicate that more information is needed. The blank tracts should not be interpreted as having extremely low dropout rates.

The map indicates that there is extensive variation in dropout rates across neighboring census tracts. For instance, the well-off tracts range from very low dropout rates (0 to 37%) to high dropout rates (46% - 57%). However many of the average and poor neighborhoods have relatively low dropout rates which confirms the results of the statistical model. Nevertheless, there are several tracts with dropout rates exceeding sixty-percent. By identifying at-risk neighborhoods where dropping-out is a greater problem, we can then target dropout prevention programs to help these particular communities.

In fact, there are three tracts with dropout rates greater than eighty-five percent which means that only fifteen out of every hundred students from those neighborhoods graduate from high school. However, the number of cohort students living in each census tract is relatively small (less than thirty). Thus the large differences in dropout rates across tracts are not necessarily statistically significant, but may be a function of small sample size.

Maps of Selected Groups of Students with High Dropout Rates

The five-dimensional log-linear model suggested that there were several groups of students with higher than average dropout rates. The final four maps display the residences of four such groups of students: poor normal-age White females, well-off Whites, poor Black students, and overage male Hispanics, in an effort to identify clusters of dropouts. There is some evidence that the poor White female dropouts live nearby each other and that the poor Black dropouts cluster together, however there is little evidence that the well-off Whites or the overage Male Hispanics drop out in mass in some neighborhoods and not others.

Each map contrasts the location of the dropouts and the non-dropouts. The dropouts are indicated by circles, and the non-dropouts are indicated by stars.

Map 6A - Poor White Normal-Age Female Dropouts (Figure 13)

According to the model, almost fifty percent of this group of students, the normal-age White females from the poorest tracts, drop out. Thus one in two of these students do not finish high school. The map of these dropouts' residences suggests that they are clustered in one section of the city, namely South Boston. There appear to be three groups of poor normal-age White female dropouts who live nearby one another. It would be interesting to study whether peer pressure discourages these students from finishing high school.

Map 6B - Well-Off White Students (Figure 14)

This point-map shows where the well-off White students live. Apparently these dropouts come from all of the well-off white neighborhoods. In other words, there are no clusters of well-off Whites where they all dropped out, and clusters where none dropped out. Rather, each well-off White neighborhood has its share of dropouts and non-dropouts. One possible exception is a group of well-off whites just below South Boston

which is very dense and looks like mostly dropouts. This tract also borders average working-class and poor tracts.

Map 6C - Poorest Black Dropouts (Figure 15)

The poorest Blacks have the highest dropout rate of the Black students (forty-five percent). According to this map of their residences, the poorest Black dropouts are highly concentrated on the northern and eastern edges of Franklin Park. Moreover the dropouts appear to live in small, dense groups; however, it is very difficult to tell whether these groupings are meaningful or whether they are a by-product of the underlying population density.

Map 6D - Over-Age Male Hispanics (Figure 16)

This map shows the residences of the over-age male Hispanics in the cohort group. The dropout rate for over-age male Hispanics ranges from seventy-six percent for the well-off to eighty-three percent for the poor. Thus the dropout rate for these students is alarmingly high. However these students do not appear to be concentrated in only one or two neighborhoods. At first glance, it looks as if they are clustered in the poorer areas of the city, but this is because the majority of the Hispanic population lives in these areas.

Summary

This chapter analyzed some of the spatial patterns in the high school dropout data. The maps show a tremendous amount of racial segregation in Boston, and the map of tract clusters indicates that race is highly correlated with socioeconomic status. The maps of students by school-type show that the exam schools attract students from all over the cities, whereas the regular schools serve primarily local students. The map of dropout rates by tract indicates that there is substantial variation in dropout rates across neighborhoods,

including those with the same socioeconomic status. Lastly, the maps of selected groups of dropouts suggests that there may be a clustering phenomenon, but the evidence is not overwhelming because it is difficult to distinguish clusters of dropouts from the underlying clusters of students.

FIGURE 3

MAP 1

ALL STUDENTS

(N = 4,940)

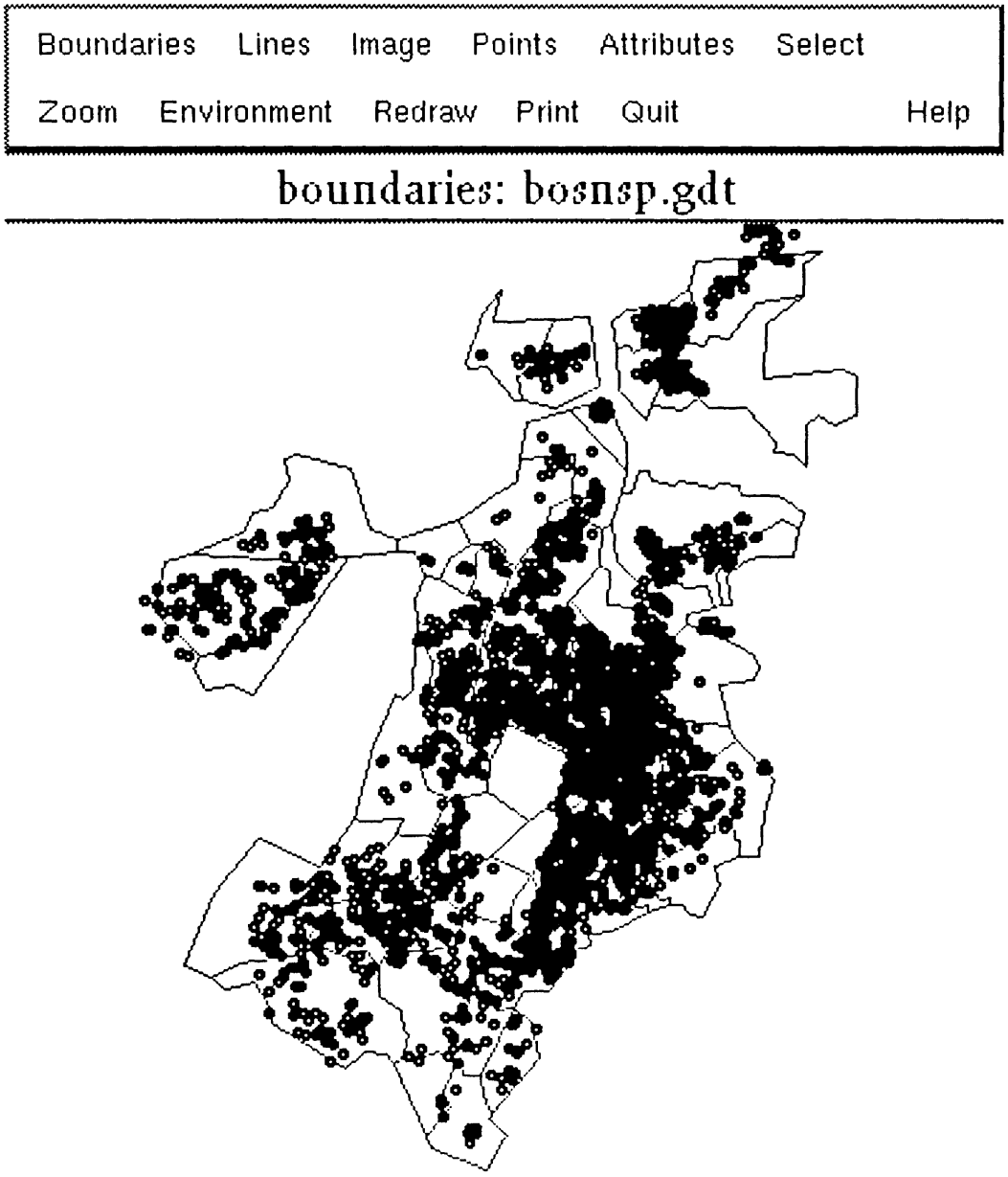


FIGURE 4

MAP 2A

BLACK STUDENTS

(N = 1,713)

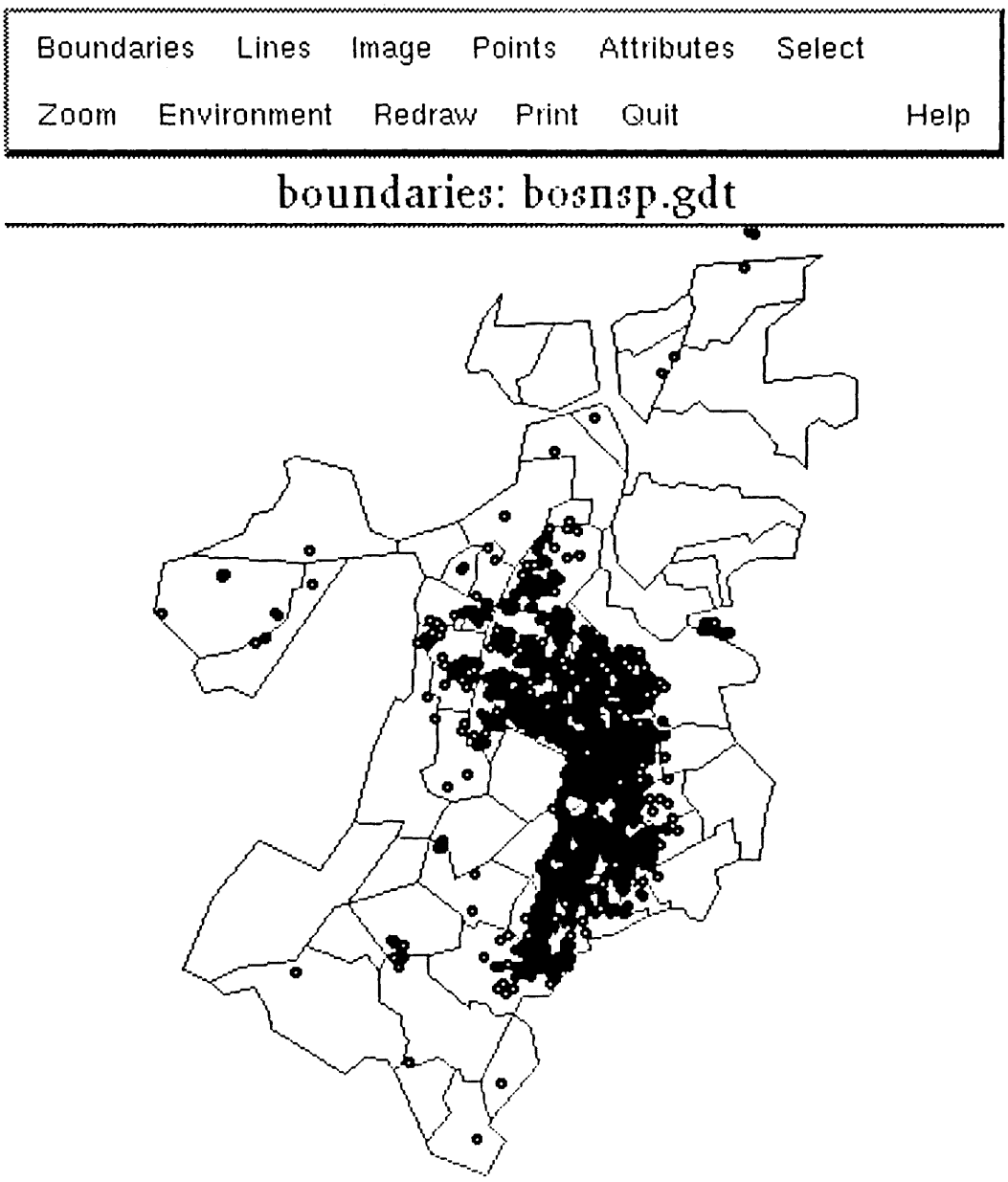


FIGURE 5

MAP 2B

WHITE STUDENTS
(N = 1,325)

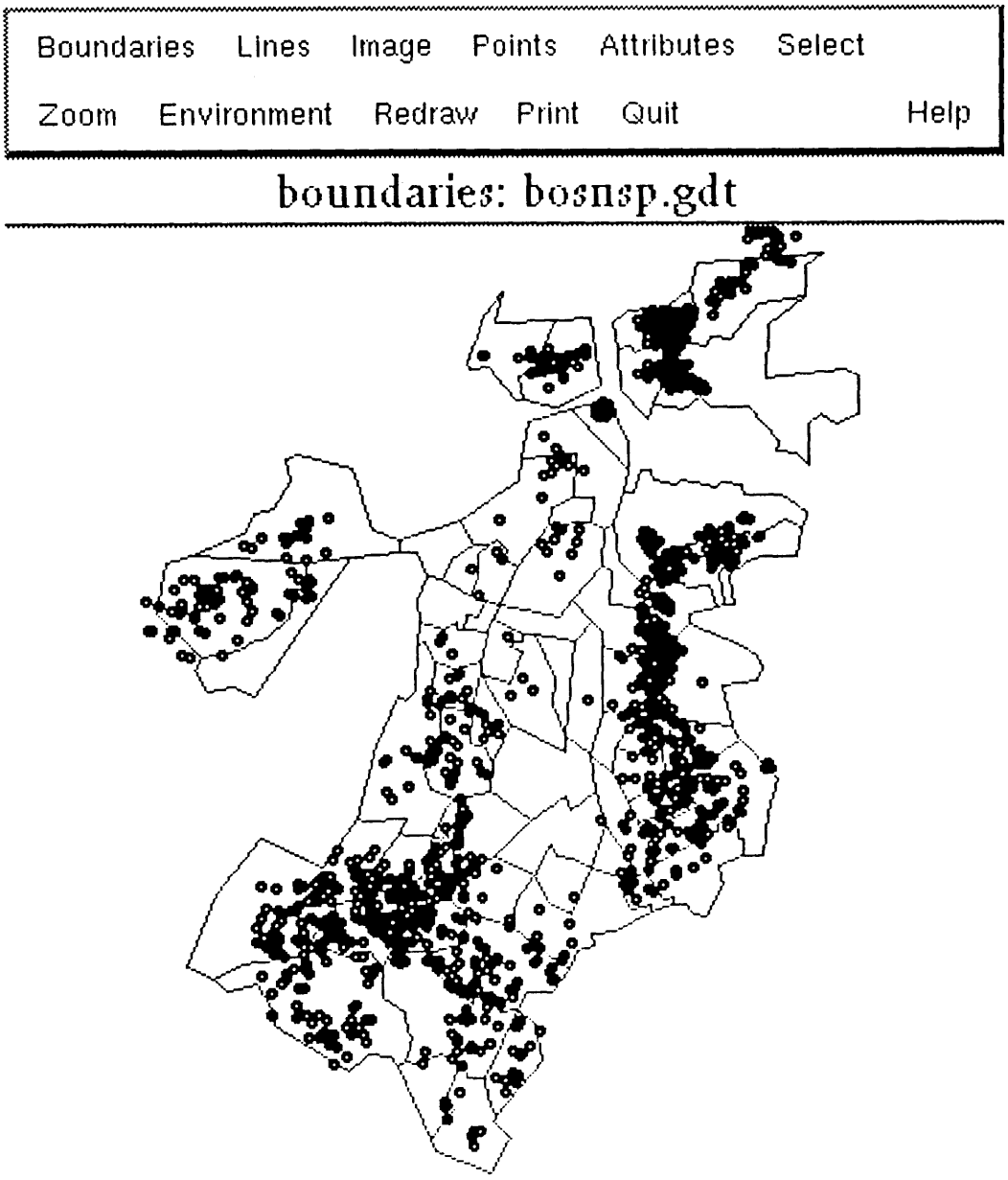


FIGURE 6

MAP 2C

ASIAN STUDENTS

(N = 254)

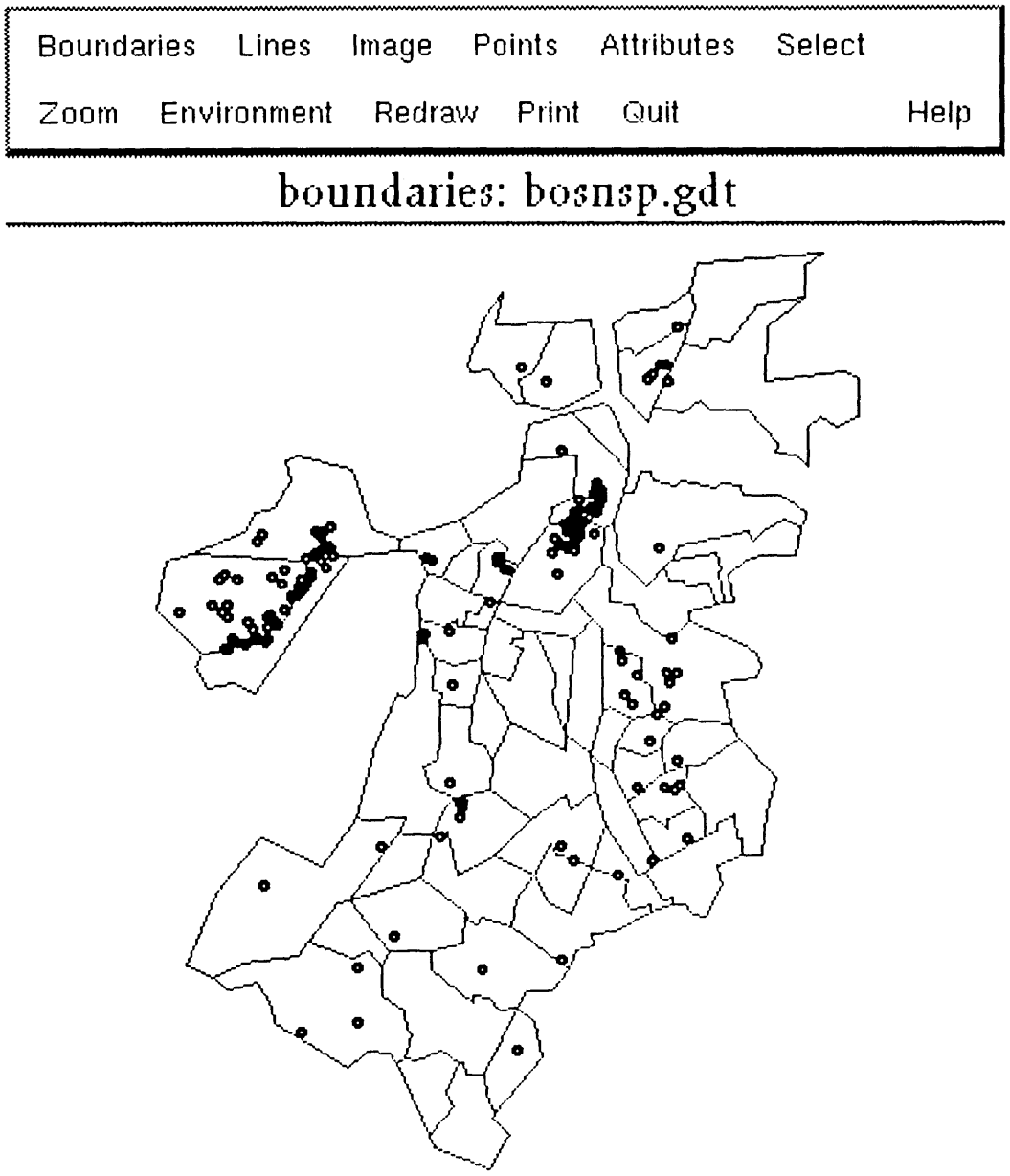


FIGURE 7

MAP 2D

HISPANIC STUDENTS

(N = 462)

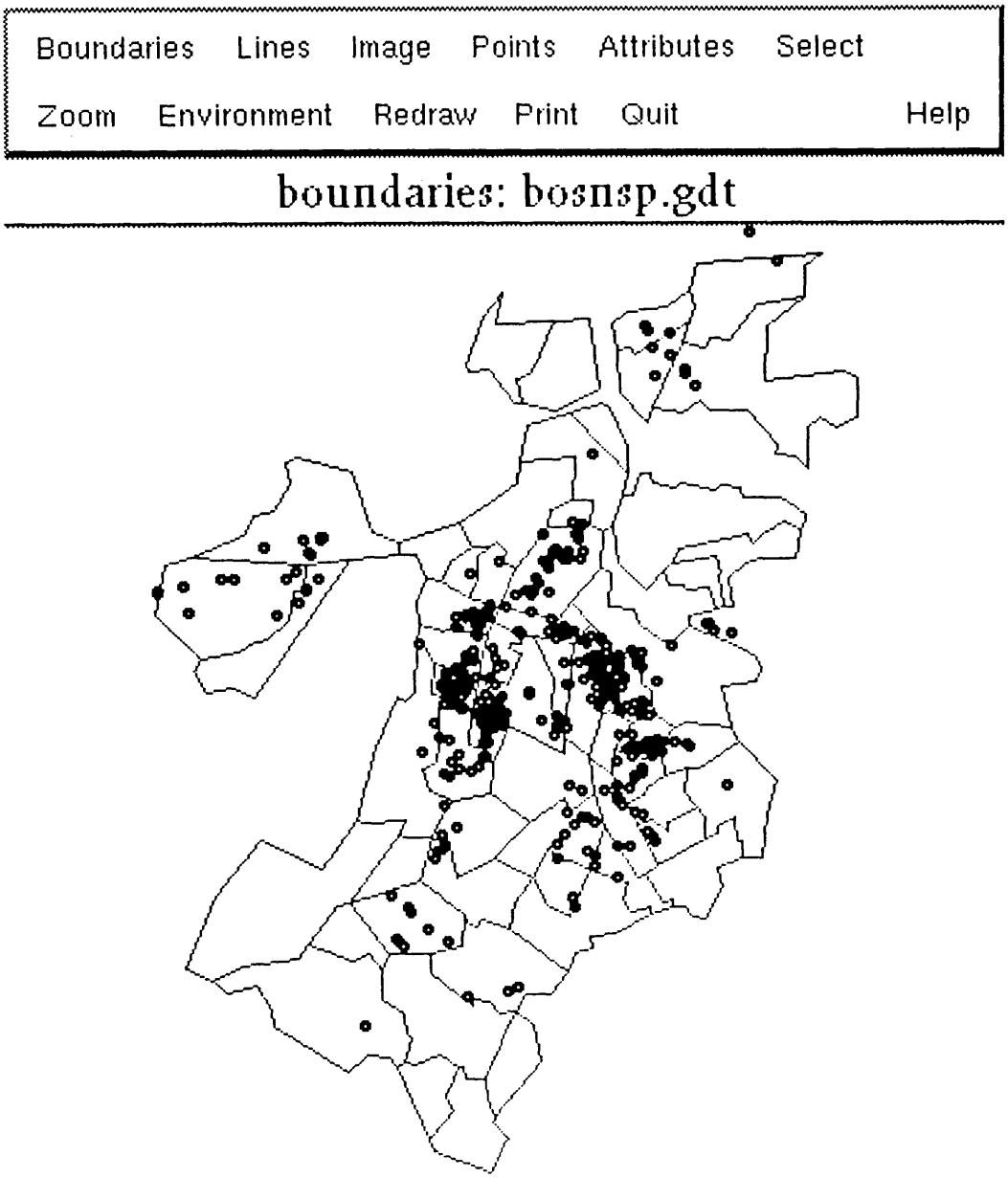


FIGURE 8

MAP 3A

A REGULAR HIGH SCHOOL

(BURKE HIGH SCHOOL)

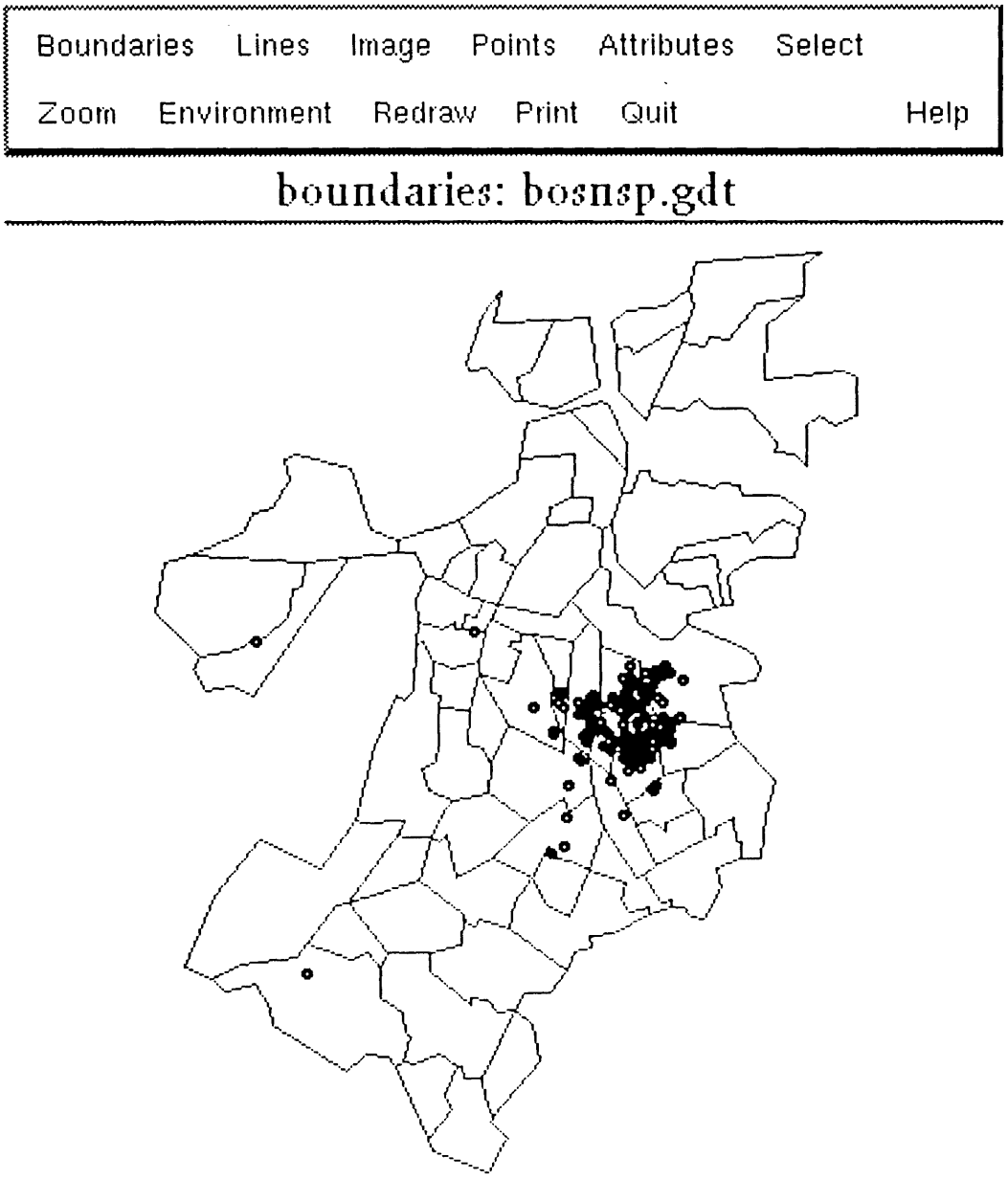


FIGURE 9

MAP 3B

A MAGNET HIGH SCHOOL

(MADISON PARK H.S.)

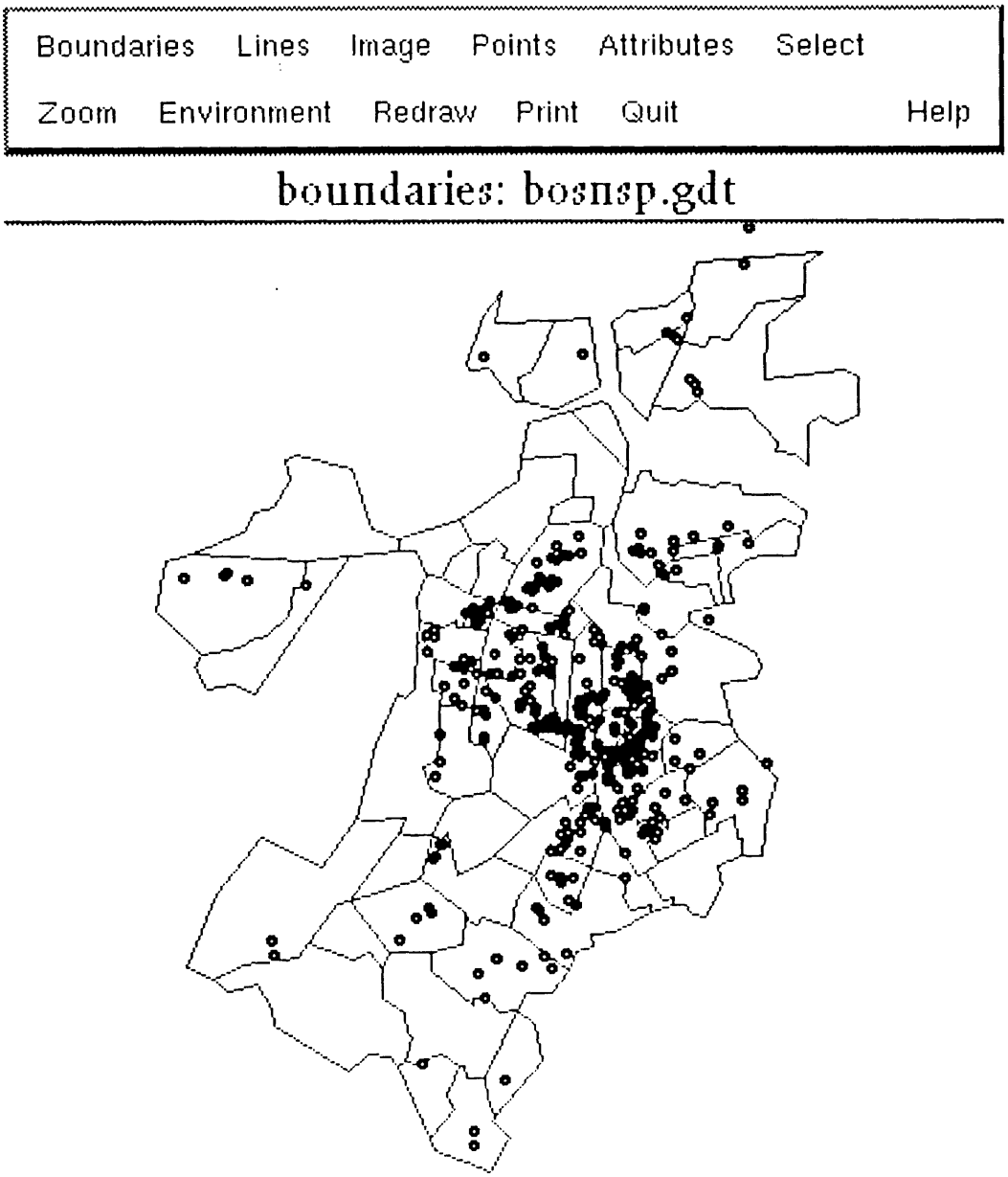


FIGURE 10

MAP 3C

AN EXAMINATION HIGH SCHOOL
(BOSTON LATIN)

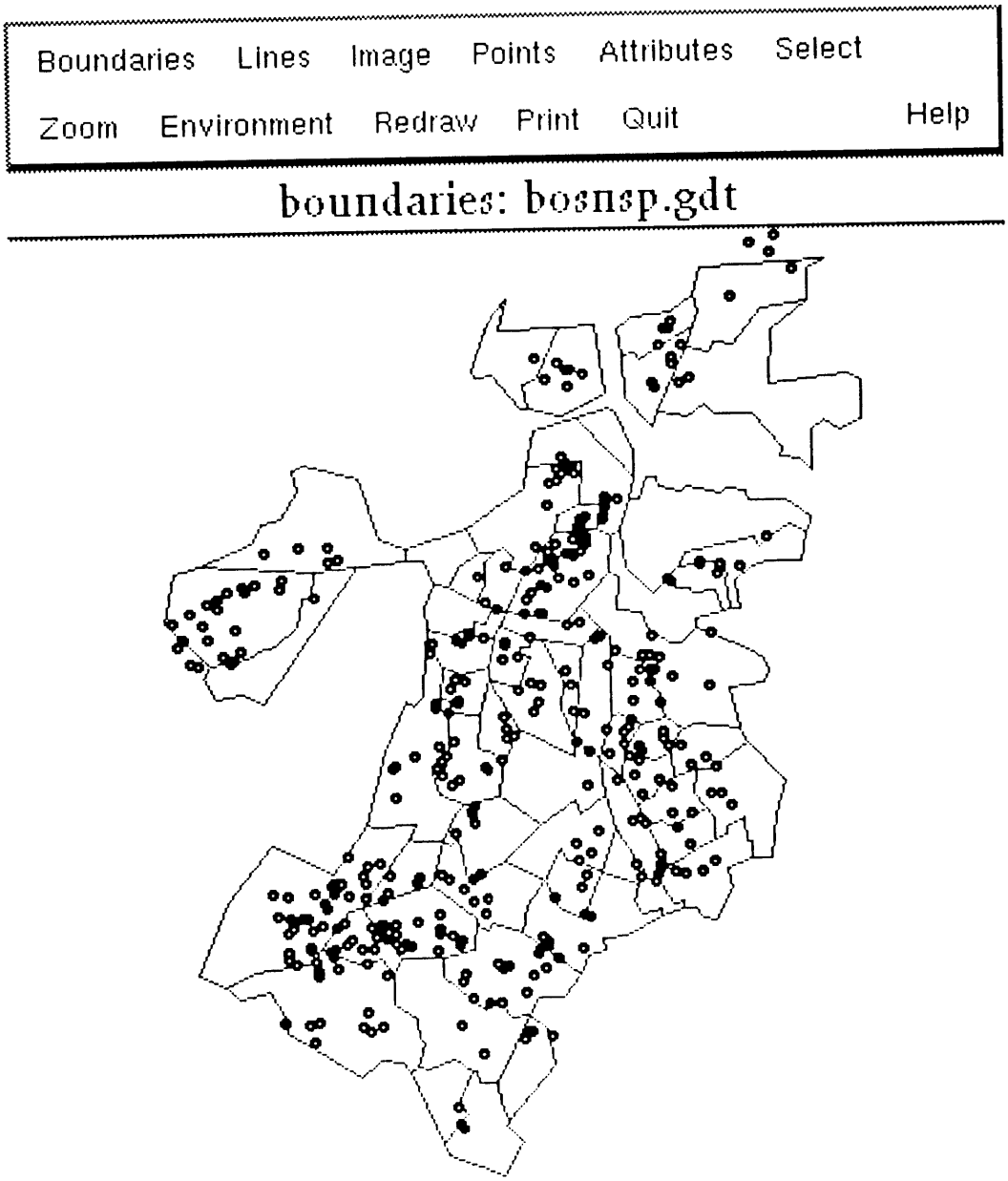


FIGURE 11

MAP 4

TRACT CLUSTERS

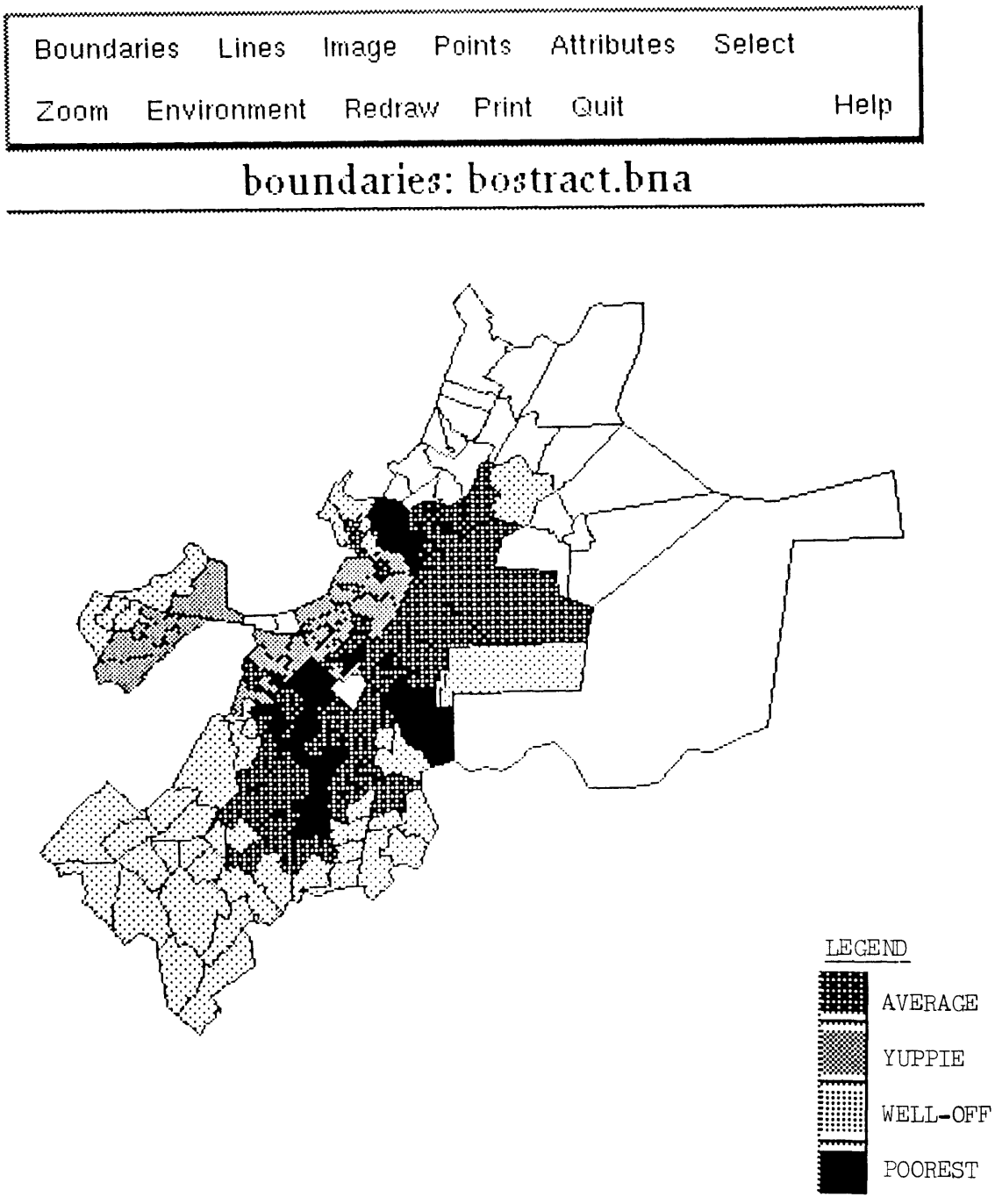


FIGURE 12

MAP 5

DROPOUT RATE BY TRACT

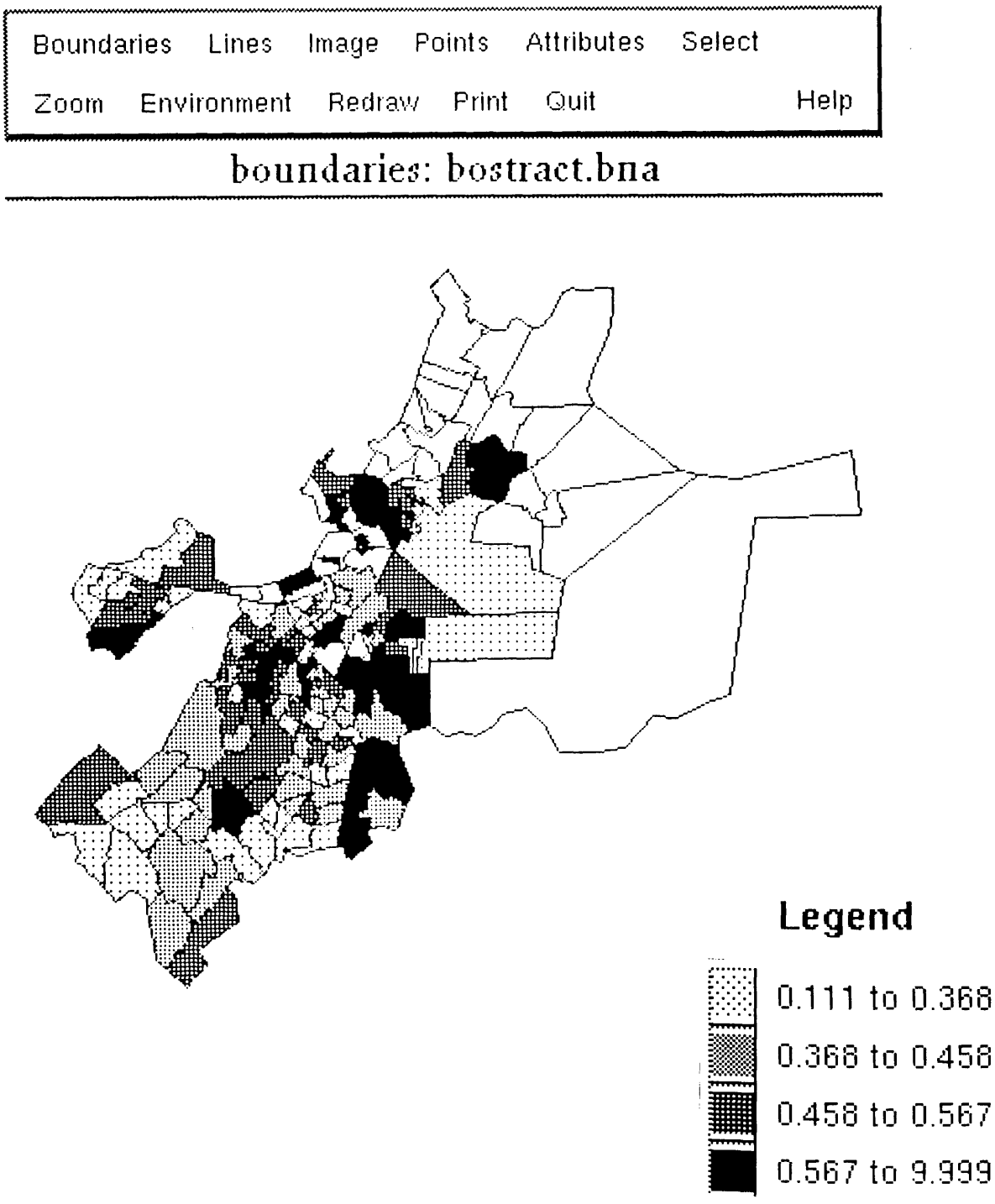


FIGURE 13

MAP 6A

POOREST WHITE NORMAL-AGE FEMALES

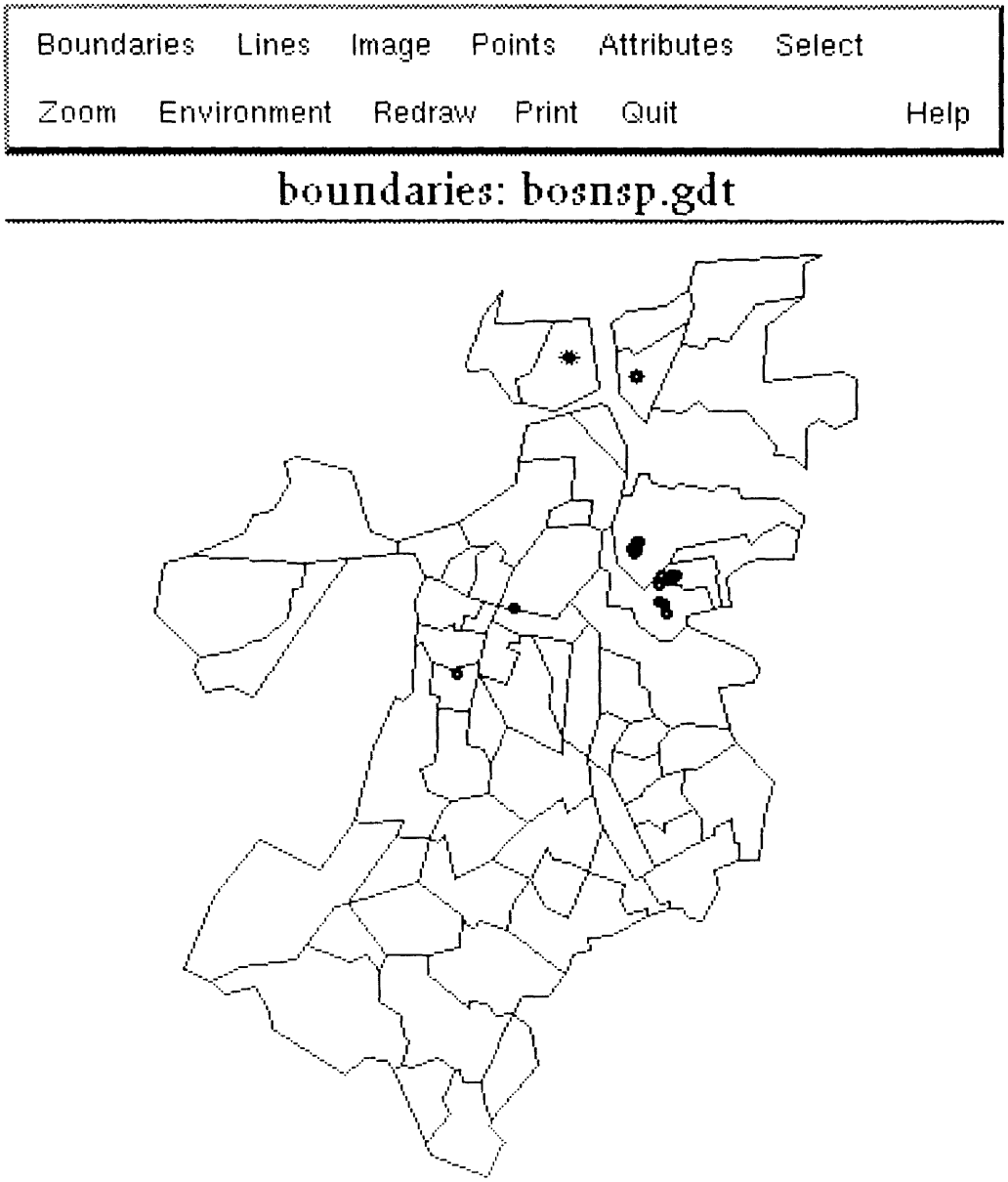


FIGURE 14

MAP 6B

WELL-OFF WHITE STUDENTS

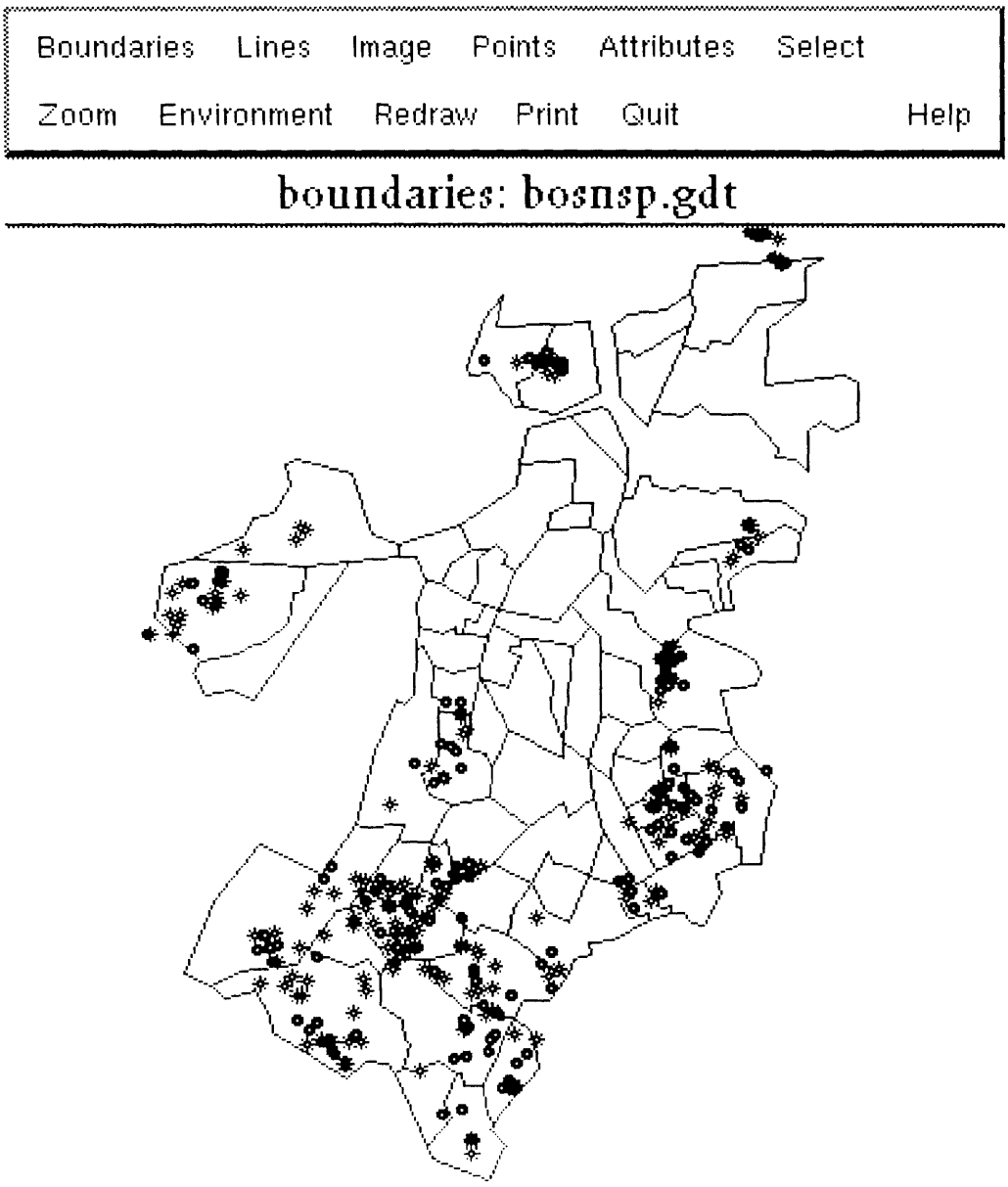


FIGURE 15.

MAP 6C

POOREST BLACK STUDENTS

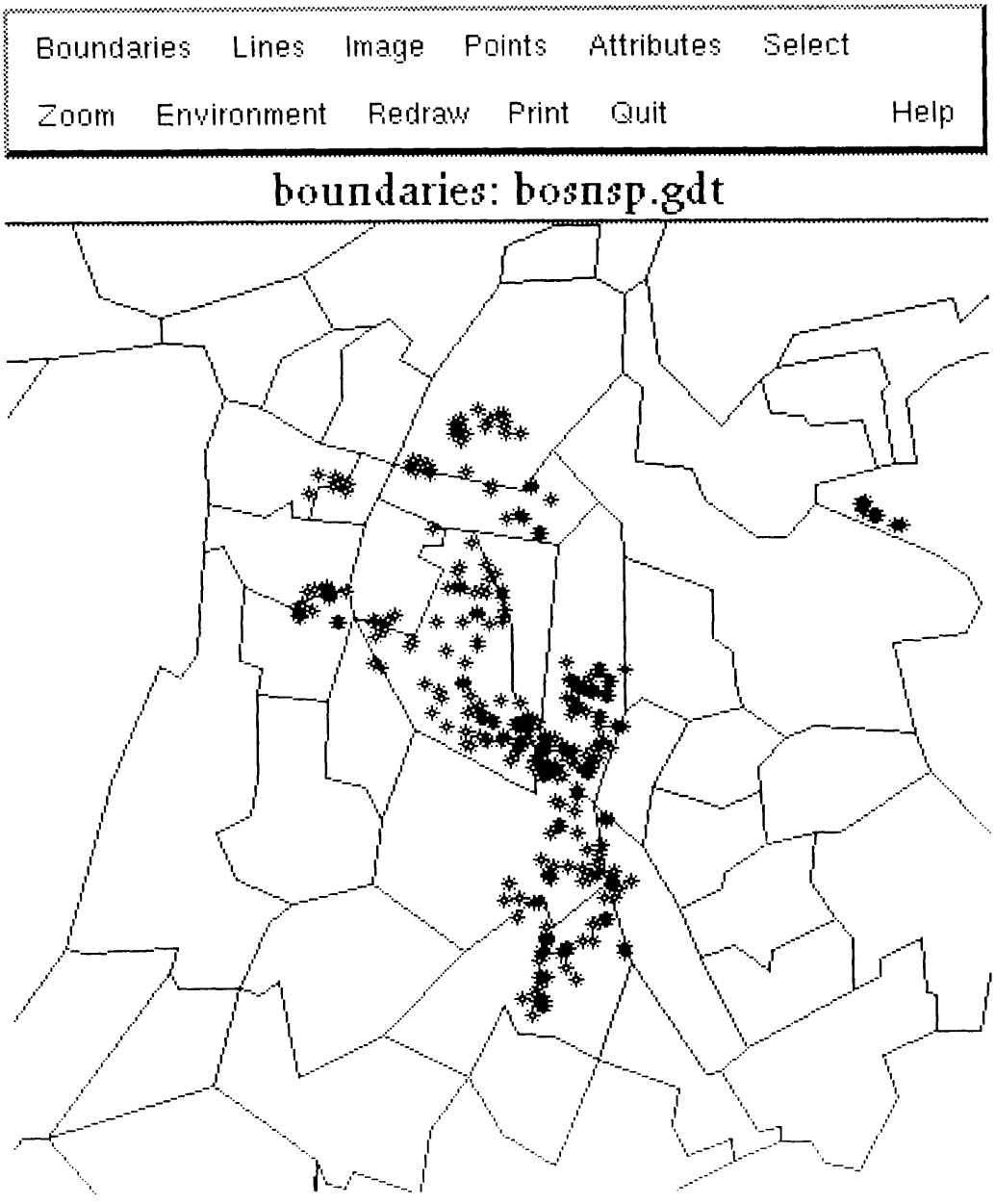
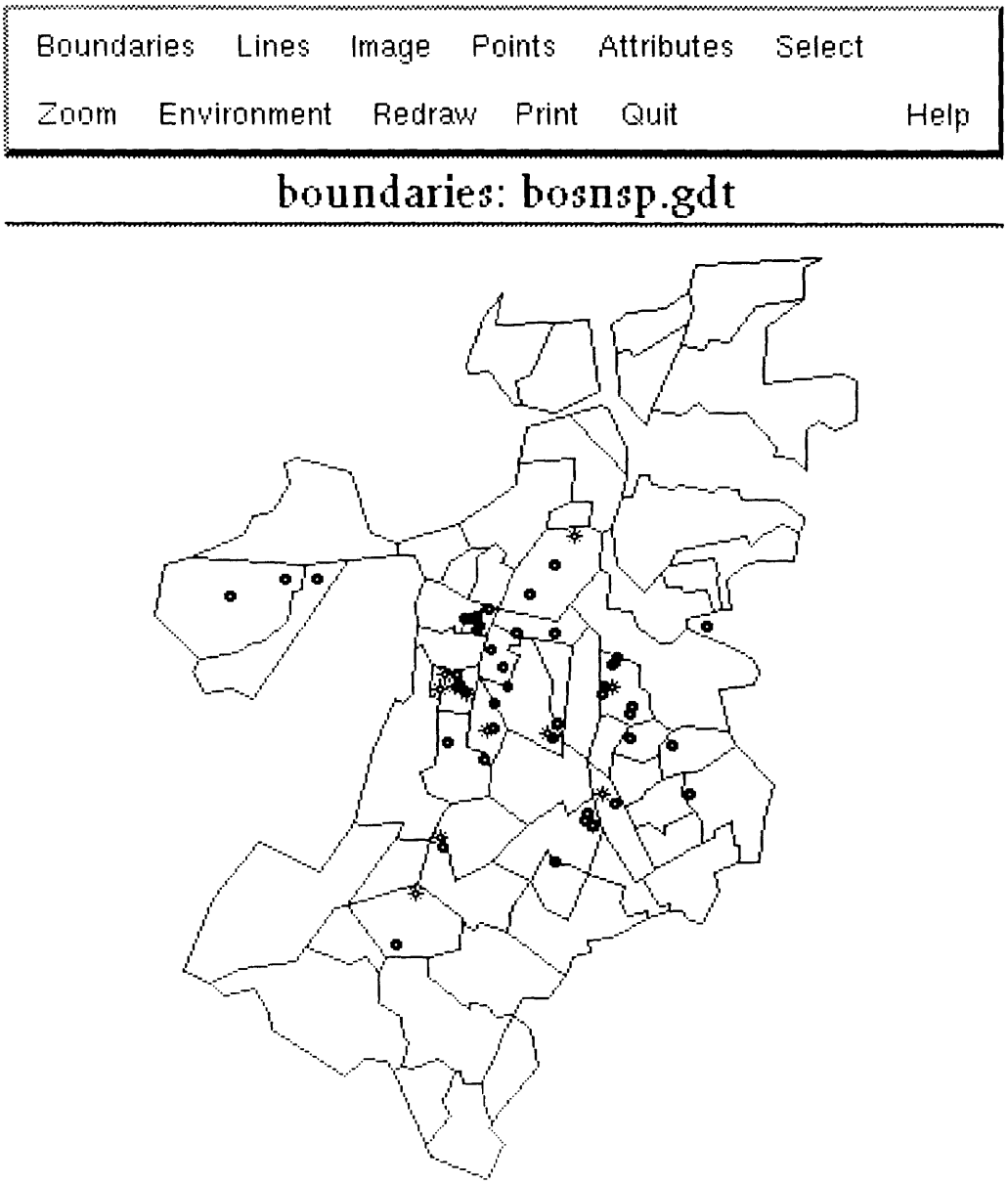


FIGURE 16

MAP 6D

OVER-AGE MALE HISPANICS



Chapter X
Conclusion

A. Summary of Research Findings

The literature on high school dropouts, described in Chapter II, indicates that dropping-out is a complicated problem, and there are no simple solutions. Although the national dropout rate has decreased over the past fifty years, it remains intolerably high (almost thirty percent), especially in our urban centers where the dropout rate often exceeds forty percent. Dropping-out has negative consequences for both the individual and society, including lower wages, higher unemployment rates, and lower national productivity. Furthermore, no comprehensive dropout prevention programs have been implemented, primarily because there is little agreement on the best method to prevent dropouts.

This research examined the high school dropout problem in a particular urban setting, specifically the Boston public school district, and went beyond the simple calculation of dropout percentages to try to understand the complex interactions which cause some students to drop out and others to stay in school. The research consisted of two parts: 1) we formulated and interpreted a formal statistical model which describes the complicated interactions among variables correlated with dropping-out, and 2) we generated maps in order to perform spatial analysis of the dropout problem in Boston.

Research Summary

Before examining who drops out of high school, we looked at who attends the Boston public high schools in Chapter VI. We found that the cohort group is primarily Black (49%). The White students are the next largest racial group (31%), then the Hispanics (13%), and the Asians (7%). There are slightly more males than females in the cohort group, and the majority of students are normal-age, as opposed to over-age. More students live in the average working-class neighborhoods, and there are approximately as many poor students as well-off students. Lastly, the majority of the Boston public school

students attend the regular and the magnet schools, in contrast to the examination and the special-needs schools.

After examining the structure of the cohort group, we analyzed the bivariate log-linear models to determine the characteristics of dropouts versus non-dropouts. In Chapter VI, the two-way models and tests of independence indicated that dropping-out is correlated with the student's race, gender, age, type of neighborhood, and type of school. In other words, dropout rates vary significantly across each of these five explanatory variables. Males are more likely to drop out than females, normal-age students drop out less often than over-age students, and students from the poor neighborhoods drop out more often than those from the average and the well-off neighborhoods. The two-way model of race versus dropping-out indicated that the Hispanics drop out the most, and Asians the least. Whites left school slightly more than Blacks, but the difference was not great in percentage terms.

We found that the examination school students rarely drop out, presumably because they are highly skilled and motivated academically. On the other hand, the special-needs schools, which serve students with severe behavioral problems, had extremely high dropout rates. Since we already understood the dynamics of the exam and the special-needs schools, we decided to concentrate on the experience of the average student in the Boston public school district who attends either the regular or the magnet high schools. Thus the five-dimensional model, which does not include the examination and special-needs students, describes the characteristics of the prototypical Boston public high school student.

Although the results of the five-dimensional log-linear model are similar to those of the two-way models, they are not identical because the five-dimensional model describes the relationships between variables while controlling for other variables. According to the five-dimensional log-linear model, dropping-out is simultaneously correlated with race, gender, tract, and age. In addition, there are three other two-factor effects: tract-type is

correlated with race, race is correlated with age, and race is correlated with gender. The model also includes main effects which describe the underlying structure of the cohort group.

Similarly to the two-way models, the five-dimensional model indicates that males drop out more than females. Also the students from poor neighborhoods are more likely to drop out than students from average working-class neighborhoods, who in turn are more likely to drop out than those from the well-off homeowner or yuppie renter communities.

Moreover, the five-dimensional model shows that over-age students drop out much more often than normal-age students. Usually over-age is an indicator of serious academic difficulties prior to high school. However, the Asians were most likely to be over-age, and at the same time dropped out the least often. Perhaps these students are recent immigrants, and their age indicates a lack of familiarity with English rather than severe academic problems. In this case, age measures English, as opposed to academic, proficiency.

As indicated by the table of expected cell counts under the five-dimensional model (Table 1), the distribution of racial groups across tract-types was far from uniform. The Black and Hispanic students live mostly in the poor communities, whereas the Whites and Asians populate the average and well-off neighborhoods. The maps of student residences confirm that there is little integration within the city of Boston. The Blacks live in the core of the city in the poorer areas such as Roxbury, Dorchester, Mattapan, and Jamaica Plain. The Whites, on the other hand, live in the relatively well-off surrounding communities of West Roxbury, Brighton, Charlestown, and South Boston.

According to the five-dimensional model which controls for other variables, the Whites are in fact the most likely to drop out of any racial group if we consider only the regular and magnet school students. As can be seen in the table of predicted dropout rates (Table 2), the White students have the highest dropout rate (52%), followed closely by the Hispanics (51%), and finally the Blacks (41%) and the Asians (40%). This result is unexpected given that the overall dropout rate is lower for Whites in Boston (Byrne 1988).

Furthermore, the general consensus in the dropout literature is that Hispanic and Black students are most "at-risk" of dropping-out (see Chapter II). Thus focusing on the average Boston public high school student changes the analysis. In this scenario, the plight of the White students is the most worrisome, although the dropout rate for all races is alarmingly high.

In general, lower socioeconomic status is associated with higher dropout rates (Pallas 1984). However, in this study we found that the Whites have the highest dropout rate even though they come from higher socioeconomic status neighborhoods. Perhaps this apparent paradox stems from the racial and socioeconomic segregation in Boston communities.

The maps of students by racial group (Maps 2A - 2D) show that there is very little integration in Boston. In comparing these maps with the map of tract clusters, we found that the White neighborhoods in Boston are generally better off socioeconomically. Since they are relatively well-off financially, many White families send their children to private or parochial schools (Byrne 1988). According to Byrne, only forty percent of the White students in Boston attend the public schools (Byrne 1988). Moreover, only thirty percent of White families with incomes above $40,000 / year send their children to the public schools, as compared to over seventy percent of White families with incomes below $10,000 / year (Byrne 1988). Thus the only White students remaining in the regular high schools are those who cannot afford private school and cannot get into the examination schools.

In contrast, the maps indicate that the Black neighborhoods are less well-off socioeconomically in comparison to the White communities. Fewer Black families have the choice of sending their children to non-public schools out of financial necessity. According to Byrne, seventy-six percent of Black students in Boston attend the public schools (Byrne 1988). Therefore the majority of the Black student population attends the regular public schools, while only the most disadvantaged Whites attend these schools. On average, the

Black students perform better than the Whites even though in general they are worse off socioeconomically.

In the second part of the research, the maps of student residences were used to interpret the spatial patterns in the dropout data. The maps showing the location of student residences for different racial groups dramatically illustrate the extent of racial segregation in Boston. In addition, the map of tract clusters confirms the link between race and neighborhood socioeconomic status.

The map of exam school students shows that the exam schools and their reputation for quality attracts students from all over Boston. On the other hand, the regular public high schools serve mostly students from the local neighborhood and a few students bused in from other parts of the city. The magnet schools did succeed in attracting students from elsewhere in the city, but not nearly as well as the exam schools.

The maps of selected groups of students with high dropout rates, suggested by the log-linear model, were less informative. There is some evidence of clusters of dropouts, particularly for the poorest White females. However, it is difficult to distinguish dropout clusters because the overall student population is not uniformly spread throughout the city. Thus the clusters of dropouts often overlap clusters of students, and without more information it is impossible to determine whether there is a dropout cluster effect or whether the grouping is simply a function of the population density in that area of the city. Hence more research is needed in this area.

Finally, the map of dropout rates by tract and Table 2 show that there is substantial variation in dropout rates across neighborhoods. The dropout phenomenon cannot be described as simply a minority problem or a poor families' problem. The map shows that even the well-off communities have areas with above average dropout rates, and some of the poor neighborhoods have below average dropout rates. Three tracts had dropout rates exceeding eighty-five percent which means that only one in seven students finishes high

school in these neighborhoods. Clearly these neighborhoods are excellent candidates for community-specific dropout prevention programs.

However, a note of caution is warranted on drawing conclusions from these census tract dropout rates. There are relatively few cohort students living in each census tract (less than thirty), hence these high dropout rates may be a product of the small sample size rather than any statistically significant difference. More investigation is needed to determine whether these neighborhoods truly experience a mass dropout phenomenon.

Research Questions Revisited

I began this study with three research questions and the following is a summary of the results.

(1) What are the characteristics of dropouts versus those who finish school normally in the Boston public school district?

As described previously, we found that the dropouts are more likely to be males who attend the regular or the magnet public high schools. They are at least two years older than normal for their class, and they come from the poorest neighborhoods in Boston. The average Boston high school dropout is typically White or Hispanic, and less likely to be Black or Asian. On the other hand, the non-dropouts are more likely to be female, and attend the examination schools. Most non-dropouts are normal-age, and live in the more well-off communities of Boston. The Asian and Black students are more likely to be non-dropouts than the Hispanic or White students in the regular and magnet high schools.

(2) Are there groups of students who drop out in mass (i.e., have an extremely high dropout rate), and if so, what are their characteristics?

The log-linear model indicated that there are statistically significant variations in dropout rates across groups of students separated according to race, gender, age, tract-type, and school-type. As can be seen in Table 2, the dropout rates range from a low of eighteen percent, for the normal-age well-off female Asians, to a high of eighty-six percent, for the over-age poorest male Whites.

The dropout rates for over-age students are all extremely high (from fifty to ninety percent). For over-age male Whites, the dropout rate exceeds eighty percent in every socioeconomic category. This suggests that these students dropout almost in mass, regardless of whether they are poor, average, or well-off. According to the log-linear model's expected cell counts in Table 1, the White students have the highest dropout rates in every category, followed by the Hispanics.

Although there is considerable variation in dropout rates across student groups, nearly all of these groups of students have dropout rates exceeding twenty percent which is alarming, perhaps intolerable, and suggestive of a cry for help.

(3) Are high school dropouts concentrated in particular areas of Boston, and if so, what are the characteristics of those neighborhoods with higher dropout rates?

In Chapter IX on spatial analysis, we found clear evidence of racial and economic segregation in Boston. The map of tract clusters (Map 4) showed that the yuppies live in the Back Bay and Beacon Hill, the well-off live in the peripheral areas of Boston such as West Roxbury, and the average working-class live in South Boston, East Boston, and

Charlestown. The poorest neighborhoods, including Roxbury, Dorchester, and Mattapan, surround Franklin Park in the center of the city.

Since the poorest neighborhoods generally have higher dropout rates, it appears that the dropouts are concentrated in the center of the city in Roxbury, Dorchester, Mattapan, and Jamaica Plain. However, this apparent clustering of dropouts is partly due to the high population density in these areas of Boston. In addition, we found a strong correlation between socioeconomic status and race. The Whites tend to live in the yuppie and well-off neighborhoods, while the majority of the Blacks live in the poorer communities. Thus, it appears that the high school dropouts are concentrated in the poor, Black neighborhoods even though the model indicates that the Whites have higher dropout rates.

In looking at maps of selected groups of students suggested by the log-linear model, we found some evidence of clustering, but again it is difficult to distinguish clusters of dropouts from clusters of students. Perhaps, a more refined unit of analysis, such as a block-level instead of a tract-level model, would more readily reveal residential clusters of dropouts.

B. Suggestions for Further Research

Additional research in the following areas could further improve our understanding of the high school dropout problem in Boston.

1) *Path Analysis* - The log-linear model can identify significant interactions among variables, but it does not specify the direction of causality. For example, the five-dimensional log-linear model indicated that dropping-out is correlated with neighborhood-type, but we should not conclude that growing-up in a poor neighborhood causes one to drop out. On the other hand, a path analysis could be used to determine the links and causal relationships between explanatory variables.

2) *Predictive Model* - Now that the log-linear model has identified the important determinants of dropout in Boston, the next step is to develop a predictive model which classifies students as potential dropouts or non-dropouts. However, a predictive model should be used with caution for fear of labeling specific youngsters as probable failures.

3) *Block-Level Analysis* - This thesis used the census tract as the basic neighborhood unit of analysis, but there are other possible levels of neighborhood disaggregation. A block-level analysis might be particularly revealing of clusters of dropouts in the city. In addition, a block-level analysis could incorporate housing-type into the model. For instance, one could disaggregate students into public housing and non-public housing groups, to see if there are differences in dropout rates across housing-type.

4) *Dropout Definition* - In this study of high school dropouts in Boston, I adopted a single definition of "cohort" and "dropout" (described in Chapter III). As mentioned in the literature review on dropouts, there are many different ways of defining dropouts and measuring the dropout problem. For example, I did not include transfer students in the formal statistical model because their final outcome was unknown. However, it would be useful to follow up the transfer students (both into and out of the district) and incorporate their fate into the model, in order to get a better sense of the overall health of the district. Future research could examine the dropout problem in Boston using multiple definitions of dropout and cohort, and compare the results.

C. Conclusion

Education is one of our most cherished values, hence we offer a free public education to all. Yet many of our children choose not to exercise their right to a public

education. The phenomenon of high school dropouts is a complicated problem, and there are no simple solutions. Hopefully this research has clarified our understanding of the specific dropout problem in Boston, in particular the fate of the average public high school student. Today we teach spelling and arithmetic. Hopefully tomorrow we can teach all of our children that education is indeed valuable.

BIBLIOGRAPHY

Bishop, Y., Fienberg, S., and Holland, P. 1975. Discrete Multivariate Analysis. Cambridge, Massachusetts: The MIT Press.

Byrne, Gregory. 1988. "High School Dropouts In Boston." M.C.P. diss., Cambridge, Massachusetts: Massachusetts Institute of Technology.

Capuzzi, Dave and Douglas R. Gross. 1989. Youth At Risk: A Resource for Counselors, Teachers, and Parents. Alexandria, Virginia: American Association for Counseling and Development.

Catterall, James S. 1985. "On The Social Costs of Dropping Out of School." Palo Alto, California: Stanford Education Policy Institute, School of Education, Stanford University.

Citywide Educational Coalition. 1988. "BPS Students and Student Achievement." Boston, Massachusetts: Boston Chamber of Commerce.

Ferreira, Joseph and Lyna L. Wiggins. 1990. "The Density Dial: A Visualization Tool for Thematic Mapping." Eugene, Oregon: GeoInfo Systems, Aster Publishing Company.

Fienberg, Stephen E. 1977. The Analysis of Cross-Classified Categorical Data. Cambridge, Massachusetts: The MIT Press.

Frase, Mary J. 1988. Dropout Rates in the United States. Washington D.C.: National Center for Education Statistics, U.S. Department of Education.

Gainer, William J. 1987. School Dropouts: Survey of Local Programs. Washington D.C.: United States General Accounting Office.

Gallington, Ralph O. 1966. "The Fate and Probable Future of High School Dropouts and the Identification of Potential High School Dropouts." Carbondale, Illinois: Southern Illinois University.

Goodman, Leo A. 1978. Analyzing Qualitative / Categorical Data. Cambridge, Massachusetts: Abt Books.

Grossnickle, Donald R. 1986. High School Dropouts: Causes, Consequences, and Cure. Bloomington, Indiana: Phi Delta Kappa Educational Foundation.

Horst, Leslie. 1990. "Annual and Cohort Dropout Rates in Boston Public Schools: Focus on Programmatic and Demographic Characteristics." Boston, Massachusetts: Office of Research and Development, Boston Public Schools.

Johnson, Richard A. and Dean W. Wichern. 1988. Applied Multivariate Statistical Analysis. Englewood Cliffs, New Jersey: Prentice Hall.

King, Randall Howard. 1978. The Labor Market Consequences of Dropping Out of High School. Columbus, Ohio: Center for Human Resource Research, The Ohio State University.

Larsen, Richard and Morris Marx. 1981. Mathematical Statistics and its Applications. Englewood Cliffs, New Jersey: Prentice Hall.

Lloyd, Dee Norman. 1967. Multiple Correlation Analysis of Antecedent Relationships to High School Dropout or Graduation. Washington D.C.: National Institute of Mental Health.

Menendez, Aurelio. 1988. "The Boston Public School System." Cambridge, Massachusetts: Massachusetts Institute of Technology, Computer Resource Laboratory Working Papers.

Pallas, Aaron M. 1984. "The Determinants of High School Dropout." Ph.D. diss., Baltimore, Maryland: The Johns Hopkins University.

Paulu, Nancy. 1987. Dealing with Dropouts: The Urban Superintendents' Call to Action. Washington D.C.: Office of Educational Research and Improvement, U.S. Department of Education.

Ribadeneira, Diego. 1990. "Schools Are Seen Failing Hispanics." Boston, Massachusetts: The Boston Globe, October 26, 1990.

Ribadeneira, Diego. 1990. "Symposium Outlines Problems, Solutions in Hispanic Education." Boston, Massachusetts: The Boston Globe, October 27, 1990.

U.S. Department of Education. 1990. "National Education Longitudinal Study of 1988: A Profile of the American Eighth Grader." Washington D.C.: Office of Educational Research and Improvement.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Joseph Ferreira, for all of his technical guidance, encouragement, enthusiasm, and patience. Also I would like to thank Lyna Wiggins for her technical advice and continuing support. I am greatly indebted to Gregory Byrne and Aurelio Menendez, for their previous work in organizing the Boston high school dropout database. Phil Thompson helped me considerably with the XMAP thematic mapping program, and Leslie Horst was helpful in providing information on current and past practices in the Boston public school district. Finally, I would like to thank my family for their constant support and encouragement.