The Dissertation Committee for Gladys Helena Krause
certifies that this is the approved version of the following dissertation:

An Exploratory Study of Teacher Retention Using Data Mining

Committee:

Jill Marshall, Supervisor

Guadalupe Carmona, Co-Supervisor

James Barufaldi

Susan Empson

Pradeep Ravikumar

An Exploratory Study of Teacher Retention Using Data Mining

by

Gladys Helena Krause, B.A., M.A

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2014

Este documento está dedicado a mi Nono, quien siempre me decía: "...ojalá que Dios me dé licencia de verla convertida en una doctorcita ...". A Juan Sebastián y María Isabel, por soportar cada minuto de este doctorado sin tener fines de semana divertidos e interminables horas en la universidad esperando por mí. A mi papá y mi mamá, por siempre creer en mí a pesar de mi terquedad. A mi tía, quien siempre me ayudó y acompañó con sus oraciones. Para tí papo, quien hizo este sueño una realidad.

Acknowledgments

My deepest thanks are extended to Dr. Guadalupe Carmona and Dr. Jill Marshall, who served as my advisors. Their advice and feedback guided me through the completion of my doctoral studies. I thank Dr. Marshall for her providing insight and instruction, and for keeping me sane on the path to completion. I thank Dr. Carmona for supporting me on many levels: for accepting me into the graduate program and for introducing me to the academic work of education. I owe my committee members Dr. Susan Empson, Dr. James Barufaldi and Dr. Pradeep Ravikumar my sincerest gratitude for their support. I thank Dr. Empson for opening new doors: for allowing me to teach the classes that I love and for forcing me always to consider the question "what have I learned form this?" I thank Dr. Barufaldi for asking the right questions. I thank Dr. Ravikumar for walking me through the world of Data Mining. I greatly appreciate the committee's expert guidance, time, and effort throughout my graduate journey. I thank Lynn Kirby and Jason Ermer for not only providing me with a job, but more importantly for teaching me what teaching is really about. I thank all my friends for keeping me laughing and for their earnest support. I must thank my children, Juan Sebastián and María Isabel. I know they are ready for me to finish. I thank Oma and Opa for helping me every minute and dropping by the house to take care of my children so I could go to school. I thank my parents, brother and sister for their support, especially when I felt defeated. I thank my loving husband for

helping me with literally everything. I could not have done this without you!

An Exploratory Study of Teacher Retention Using Data Mining

Publication No. _____

Gladys Helena Krause, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Jill Marshall
Co-Supervisor: Guadalupe Carmona

The object of this investigation is to report a study of mathematics teacher retention in the Texas Education System by generating a model that allows the identification of crucial factors that are associated with teacher retention in their profession. This study answers the research question: given a new mathematics teacher with little or no service in the Texas Education System, how long might one expect her to remain in the system? The basic categories, used in this study to describe teacher retention are: long term (10 and more years of service), medium term (5 to 9 years of service), and short term (1 to 4 years of service). The research question is addressed by generating a model through data mining techniques and using teacher data and variables from the Texas Public Education Information Management System (PEIMS) that allows a descriptive identification of those factors that are crucial in teacher retention. Research on mathematics teacher turnover in Texas has

not yet focused on teacher characteristics. The literature review presented in this investigation shows that teacher characteristics are important in studying factors that may influence teachers' decisions to stay or to leave the system. This study presents the field of education, and the state of Texas, with an opportunity to isolate those crucial factors that keep mathematics teachers from leaving the teaching profession, which has the potential to inform policy makers and other educators when making decisions that could have an impact on teacher retention. Also, the methodology applied, data mining, allows this study to take full advantage of a collection of valuable resources provided by the Texas Education Agency (TEA) through the Public Education Information Management System (PEIMS), which has not yet been used to study the phenomenon of teacher retention.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

My investigation restricts its focus to minority mathematics teachers in the state of Texas. Specifically, I am focusing on teachers of hispanic background, although other ethnicities that are also considered minorities, such as Asians, African Americans, and Native Americans, are studied and mentioned in my investigation in order to inform this research and assess its impact. Focusing on hispanic teachers has an impact on the development of teacher preparation programs, on professional development, and ultimately on student achievement due to the changing demographics of the state and the population of Texas schools. According to Stevens (n.d.) teachers do not show the same diversity found among students in Texas schools. In a recent report by Stutz (2010), during the 2008–2009 school year the population of hispanic teachers in the state of Texas was only 22%, while 48% of the students in the entire state were hispanic. In the same report Dr. Edward Fuller, at that time a senior research associate for the Center for Teaching Quality (CTQ) in Chapel Hill, North Carolina, commented that "[t]he research shows that if you can match the ethnicity and race of teachers and students, teachers tend to be more effective" (Stutz, 2010). Grant (1992), The Institute for Education in Transformation (n.d.), and Meier and Stewart (1991) have also presented concrete

data that confirms that the demographics in the teaching force is important to the learning outcomes of minority students. In Hemphill and Rahman (2011)'s latest report on the achievement gap it is clear that closing the hispanic-white achievement gap remains a challenge throughout the country. The report also confirms that hispanics are "the fastest-growing segment of the United States population" (Hemphill & Rahman, 2011, p.iii). The same study provides data from the U.S. Census Bureau showing that the hispanic population "increased by about 58 percent, from 22 million in 1990 to 35 million in 2000, compared with an increase of about 13 percent for the total U.S. population" (Hemphill & Rahman, 2011, p.iii).

This study focuses on Texas mathematics teachers in large part due to the importance and weight that mathematics courses have in the state curriculum. Teachers of mathematics at the secondary level teach 5 to 6 classes per day throughout the 180–day school year (Texas Education Agency Policy Planning and Evaluation Division, 1993). This implies that every mathematics teacher in the state of Texas teaches about 900 classes per school year. Thus the impact that teachers of mathematics have on the educational system is very significant. It is not a surprise, then, that we find in the literature that any crises in the economic, social and political aspects of the country are constantly related to failure in the educational system of the nation, particularly in the fields of science and mathematics (Cuban, 2010).

This study addresses important gaps in the current literature. According to Torres, Santos, Peck, and Cortés (2004) the studies that have developed

theories on occupational choice and career development have sampled mostly middle class males and do not address aspects related to ethnicity, race and culture when choosing teaching as a profession. An additional benefit of this investigation, in the area of professional development, is that it informs our understanding of the reasons that might motivate people from ethnic minorities to remain as teachers of mathematics. Consequently, this study has the potential to improve the shaping of incentives to promote the field of mathematics teaching as a desirable career path for a wider group of people.

1.1 Research Question

My proposed study can be stated this way: **What are the characteristics of mathematics teachers that are related to teacher retention in Texas schools?** More specifically,

- What are the characteristics that most relate to retention of hispanic mathematics teachers in Texas schools?
- Do these characteristics change over time?

Chapter 2

Literature Review

This chapter presents a review of the relevant literature concerning the proposed study. It is divided into two parts: the first part describes a theoretical framework in which both teacher and student strive for success in accordance with the incentives which present themselves. The focus of this part is on the factors in and results of the interaction between teachers and students. The second part of this literature review focuses on research related to teacher retention, with a brief description of the research theory on minority teacher retention in mathematics, followed by a summary of the research on teacher characteristics.

2.1 Theoretical Framework

This study seeks to describe the characteristics of mathematics teachers that remain in the Texas educational system and how these might change over-time. Identifying the characteristics of the teachers that stay in their current positions or leave the profession helps to identify strategies for professional development that motivate teachers to remain in the system for longer periods of time. Chapman (1984) has found that teachers' first teaching experience

has a significant impact on the number of years teachers tend to stay in the system. If teacher training programs, school administrators and professional development programs can identify the salient teacher characteristics, adjustments could be made in their programs and schools to facilitate better teaching experiences that consequently might diminish teacher turnover rates.

2.2 Teacher Incentives

Teacher compensation is a relatively controversial topic when talking about teacher retention. Darling-Hammond (1984) considers teacher compensation to be fundamental in retaining teachers. As she indicates, other professional fields offer more competitive salaries as well as economic incentives in order to attract talented people, while the teaching profession is always associated with lower salaries. However, Rivkin and Hanushek (2007) present an investigation contradicting Darling-Hammond (1984)'s argument for the importance of compensation. A study from Rivkin and Hanushek (2007), on teacher incentives in Texas, showed that 6 percent of teachers in Texas move to different schools within the same district and that 5 percent of teachers switch districts. Rivkin and Hanushek (2007) also stated that 82 percent of teachers in Texas remain in the same school and 7 percent leave the profession. In this article, Rivkin and Hanushek (2007) showed that teachers in Texas tend to switch districts and schools according to students' characteristics. Salaries and other compensations are not as important for teachers as the characteristics of students in the school or district in which teachers choose

to teach. The same study showed that the academic achievement of students in the school in which teachers prefer to teach actually increases compared to the academic achievement of those students in the schools teachers tend to leave. In a recent study on teacher salaries in Texas, G. Krause (2011) did not find a definitive conclusion on teacher retention based on teachers' salaries. The report did not consider aspects of mobility, nor the various other characteristics such as ethnicity, years of experience and level of preparation. A three-year pilot investigation by the National Center on Performance Incentives (2010) on teacher incentives also concluded that there is no significant effect on teacher incentives and student performance. The study did not find evidence of an effect in teachers' practices due to incentives. This study did not consider teacher characteristics either; only teacher practices were linked to students' standardized test scores.

2.3 Teacher Characteristics and Student Performance

There is an extensive literature on studies that focus on the effect teacher quality has on student achievement. It is logical to derive conclusions on the interdependency of these two aspects of education, when the average gain in learning among students within the same schools is so different. Hanushek (2011) found that students assigned to certain teachers produce much higher test scores than students assigned to other teachers, even though the material covered is the same when they are in the same school and students share similar characteristics. "[T]eachers are very important; no

other measured aspect of schools is nearly as important in determining student achievement" affirms Hanushek (2011, p.467). Rockoff, Jacob, Kane, and Staiger (2008) found that students assigned to teachers with higher cognitive level score higher in mathematics tests. Also, Dobbie (2011), in a study based on data from Teach for America, found that students assigned to teachers with higher rating in leadership experience and perseverance score higher in academic tests. This same study also found that students who are taught by teachers with these characteristics are less likely to misbehave.

Other authors consider teacher characteristics, such as ethnicity and race, to have as important an impact on student achievement as teacher quality. Dee (2005) highlights different studies that present evidence of teachers' ethnic background influencing students' academic outcomes. Dee (2005) describes what is called "stereotype threat", "active teacher effect" and "passive teacher effect". He describes the first one as the situation where, for example, a black student might experience certain apprehension with white teachers that can cause the student to not feel comfortable in the class and consequently, affect the student's academic outcome. Dee (2005) defines the second as teachers' unintended bias in the expectations and interactions of students of different ethnicity, and finally, he defines the third as the effect triggered by the teacher's racial, gender or ethnic identity that positively affect students' outcomes. In this case teachers are seen as role models which can increase motivation and expectations for academic achievement.

2.4 Teacher Retention

Ingersoll and May (2012) trace the origins of investigating teacher retention to two decades ago when a group of researchers noted significant changes in the characteristics of the people that decided on teaching as their professional path. One investigation from Darling-Hammond (1984), gained attention from policy makers. The report describes several factors that led the country to an "[e]merging crisis in teaching" (Darling-Hammond, 1984, p.1). The author talks about a crisis in the general field of teaching; however, she noted that the content areas most affected by the lack of teachers were mathematics, physics, biology, chemistry, computer programming, and bilingual education. She identified that "[d]emographic trends, expanded opportunities for minorities and women, low salaries and lack of prestige associated with teaching" (Darling-Hammond, 1984, p.7) were reasons why people preferred other careers over teaching. She explains that in the 1970s there was a surplus of teachers in the U.S. According to her report, 20% of the bachelor's degrees in the 1970s corresponded to education. By the 1980s the surplus decreased significantly, with the addition of a simultaneous increase in the student population. With a higher number of students in the public schools, there was a greater need for teachers. But during the 1980s, minorities in general, and women in particular, were in high demand in the larger workforce. Women who traditionally would have majored in education were now choosing career paths such as business, commerce, and professional healthcare, with higher salaries and status.

Darling-Hammond (1984) also identified teacher dissatisfaction as one of the causes for teacher attrition. She says that "[c]onditions that undermine teacher efficacy, i.e. the teacher's ability to do an effective job of teaching, are strongly related to teacher attrition" (Darling-Hammond, 1984, p.13). The author presented data in her report that indicated that those teachers who possessed a master's degree and teacher certification in the subject they were teaching were the same teachers that reported higher dissatisfaction. She links this dissatisfaction at least in part to a systemic failure to see teachers as autonomous decision makers, but rather as agents that must follow decisions established by others. Darling-Hammond (1984) explains that when teachers are required to follow a standard pattern for teaching, they cannot be effective. If the teacher needs to teach according to the standard established, she cannot adjust her instruction according to the needs of her students. This standardization therefore lowers the quality of teaching imparted in the schools and consequently produces poor academic achievement.

As for recommendations for improving teacher retention, twenty years ago Darling-Hammond (1984) suggested that teachers' salaries should increase. She also suggested a change in teachers' roles. Teachers' lack of autonomy on the content and methodology for teaching in their own classrooms is the primary factor that influences lack of motivation to continue in a teaching career for most teachers. Darling-Hammond (1984) justifies the change in teachers' roles by explaining that their new roles would lead to administrative positions, and lead to changes in the resource allocations in the schools.

Additionally, she suggested that the following features should characterize teaching: "[r]igorous entry requirements, supervised induction, autonomous performance, peer-defined standards of practice, increased responsibility with increased competence" (Darling-Hammond, 1984, p.17).

Twenty years after Darling-Hammond (1984)'s report, Ingersoll and May (2012) investigated teacher retention. The authors not only identified the problem of retaining teachers, mostly in the fields of mathematics and science, but also raised a new issue which had received little attention: the cost of teacher turnover. Their investigation cited studies within the field of economics that demonstrated that in industry, if corporations' employees change jobs frequently, it typically does not have a negative impact on the companies' efficiency. Moreover, the paper cites Kimmit (2007), Deputy of the Treasury, who explains that employee turnover is beneficial for individuals and organizations. Ingersoll and May (2012) contrast these findings with other studies explaining that, in some cases, employee turnover has produced problems in the organizations, and that high turnovers in industry can have high costs and might be a sign of a poor administration. While there is ample discussion of the benefits and costs of employee turnover in industry, very little research can be found on the costs or benefits of teacher turnover in education. The financial cost of teacher turnover is further described in section 2.5.

In the intervening 20 years, some studies did address issues surrounding teacher turnover. Guarino, Santibañez, and Daley (2006) present results which help identify some of the predictors and sources of teacher turnover; Grissmer

and Kirby (1991) highlights that the fields of special education, mathematics, and science have the highest rates of teacher turnover; Rumberger, R. (1987) clarifies that a strong incentive for mathematics and science teachers to leave their teaching careers is their ability find other jobs outside education. These investigations suggest the complexity of understanding the factors contributing to teacher turnover and retention. Ingersoll and May (2012) identify these studies and present a more current analysis that complements these prior investigations. Ingersoll and May (2012)'s overall conclusions on the investigation of teacher turnover in science and mathematics confirmed that, contrary to the general belief, there is a large enough pool of science and mathematics teachers to meet demand. However, this particular group of teachers shows a high turnover rate, tending to leave the schools earlier than teachers in other areas. Ingersoll and May (2012) have also found that mathematics and science teachers do not tend to leave the educational system to pursue a different career path. Rather, when mathematics and science teachers leave the schools for other jobs, these jobs tend to be related to education: for example they may find work as educational administrators or developing new curricula.

Another important finding by Ingersoll and May (2012) was that mathematics and science teachers tend to move frequently between schools. In particular mathematics and science teachers show a higher turnover rate in poorer schools with a large proportion of minority students than in mostly white, non-poor public schools. Ingersoll and May (2012)'s own research suggests that mathematics and science teachers who leave poor and high-proportion

minority schools are not dissatisfied with the students, but instead, with the administration and poor management skills of the administrative staff. Especially important is their finding that those mathematics and science teachers who leave the educational system altogether have two completely different reasons motivating them to leave: mathematics teachers typically state that they leave because of dissatisfaction with their classrooms and issues related to classroom management, while science teachers typically state that they leave because the salaries are low. Ten years before Ingersoll and May (2012)'s study, work on teacher retention published in Ingersoll (2001) confirmed that teacher characteristics, especially age and subject taught, are among the most influential factors for teachers leaving the field.

2.4.1 Age, Gender and Ethnicity in Teacher Retention

The preceding discussion describes research surrounding salary and teacher retention. Among all the variables studied in this investigation, salary is most commonly cited in the literature as the reason why teachers leave their professions. The literature has relatively less to say regarding other characteristics such as the age, gender, and ethnicity of teachers and how they might affect their decisions to stay or leave the teaching profession.

Ingersoll (2011) presents a useful description of how these three characteristics influence teacher retention in the United States. His study suggests that the age of the teacher might be an important predictor of teacher turnover. The analysis suggests that teachers who are less than 30 years old or who are

over 50 tend to leave the system in greater number than middle-aged teachers. The report does not appear to delve into the reasons for such findings. But one may view these results in terms of other research. Ingersoll (2001) suggested that younger teachers appear to still be involved in the search for a career path and so tend to enter and leave the profession in greater numbers; and older teachers are nearing the point of retirement, and so have an incentive to leave. It should be noted that Ingersoll (2011) divides age into a categorical variable: younger (less than 30), middle (40–50 years old), and older (over 50). So there is some loss of resolution within these various age groups. In point of fact, Ingersoll (2011) states that this correlation with age is only the case for white teachers, while for minority teachers this tendency was not observed. Though not a point made explicitly in the study, this suggests that the predictive value of age for teacher turnover can not be evaluated in isolation from ethnicity. This will be a theme picked up in the description and conclusions of the current investigation.

Ingersoll (2011) also found that for minority teachers male teachers are more likely to depart than female teachers, and for whites, Ingersoll (2011) found no difference in gender. He also found that minority teachers tend to stay longer than white teachers. It is worth noting that the same study finds that mathematics and science teachers show no greater tendency to depart than other teachers.

2.5 The Cost of Teacher Turnover

In general, teacher turnover necessarily implies a financial cost. Different studies have presented varied results and data that demonstrate these costs. For Texas, Texas Center for Educational Research (2000) has explicitly identified three areas that represent costs due to teacher attrition: separation cost, hiring cost, and training and support cost. The report defines *separation cost* as the expenses caused by "exit interviews, notification to insurance companies, notification to payroll, completion of service records, and other exit reports" (Texas Center for Educational Research, 2000, p.1). *Hiring cost* is defined by the expenses related to advertisement, background checks, interviews, and processing applications. *Training and support cost* corresponds to the expense of training beginning teachers and the stipend paid to mentors to help train the new teacher. Overall, Texas Center for Educational Research (2000) states that the total cost of teacher turnover per district is about $329 million, when roughly 15.5% of teachers leave a district. The report also points out that this estimate is conservative, noting that other studies have calculated the cost for the same percentage of teacher turnover to be about $2.1 billion. Either number is alarmingly large, considering that the same money could otherwise be used directly in the classrooms supporting students' academic achievement. Barnes, Crowe, and Schaefer (2007) completed a study of the costs of teacher turnover in five school districts: Chicago Public Schools (Chicago, Illinois), Milwaukee Public Schools (Milwaukee, Wisconsin), Granville County Schools (Granville, North Carolina), Jemez Valley Public Schools (New Mexico), and

Santa Rosa Public Schools (New Mexico). The study found that the costs of recruiting, hiring, and training a new teacher are substantial. "In Granville County, North Carolina, the cost of each teacher who left the district was just under $10,000. In a small rural district such as Jemez Valley, New Mexico, the cost per teacher leaver is $4,366. In Milwaukee, the average cost per teacher leaver was $15,325. In a very large district like Chicago, the average cost was $17,872 per leaver. "The total cost of turnover in the Chicago Public Schools is estimated to be over $86 million per year" (Barnes et al., 2007, p.3). The same study also found that low performance and high poverty in schools were correlated with high teacher turnover. The data presented in these reports confirm the alarmingly high rates of teacher turnover in the nation's school system and consequently the high cost that this entails. At the same time, this is an opportunity to search for solutions that would help ameliorate this problem. Ultimately, the reports presented highlight the relevance of the proposed study.

2.6 Conceptual Framework

In this section I provide a conceptual overview of the present study based on the theoretical framework described above. Figure 2.1 depicts a logic flow describing how the data regarding teacher characteristics enters the model, informs the research, and is used to classify a new data entry.

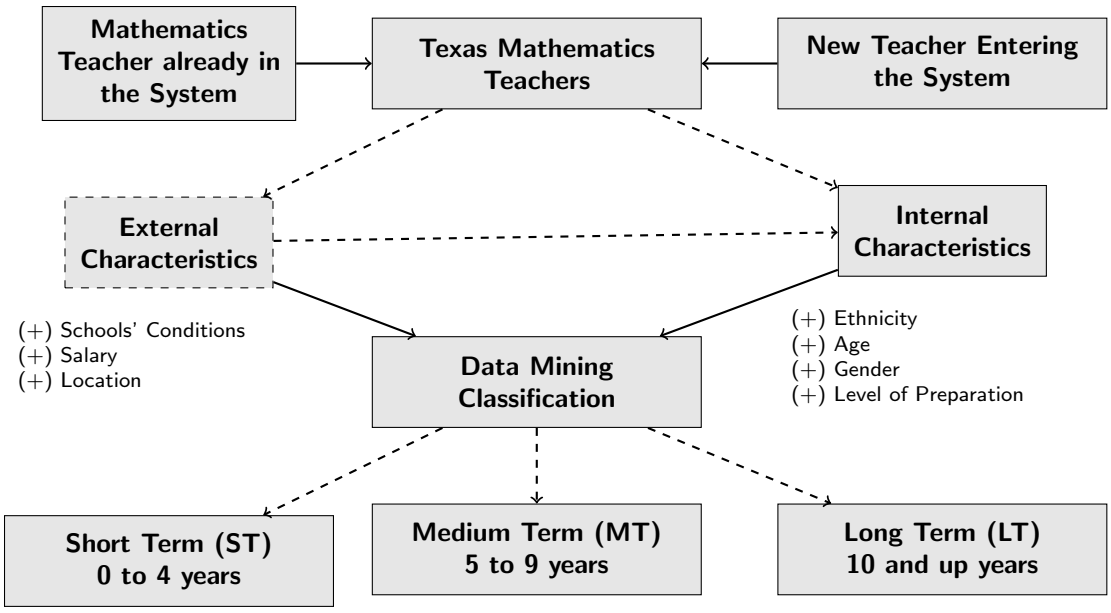


Figure 2.1: The framework shows how teacher characteristics help identify how long a new teacher might stay in the Texas educational system.

2.6.1 Teacher Characteristics: Internal and External

This study defines *teacher characteristics* as those properties which define or determine the state or situation of a teacher. Some of these properties may be inherent to the individual teacher, such as gender or year of birth; others may pertain to the overall situation in which the teacher finds him- or herself, such as school size and location. Moreover many characteristics may be shared by different teachers, e.g. gender; few characteristics, such as identification number, will be unique to a given teacher. Teachers will exert control over some of their characteristics, such as highest degree attained; while others will be under control of no individual teacher, such as school ranking, and that are either unchanging or cannot change unless the teacher decides to

take action. Those characteristics, which either define the physical makeup of the teacher (e.g. gender, ethnicity, age) or over which the individual teacher may exercise control (e.g. level of preparation), are termed by this study as *internal* characteristics. This study labels as *external* those characteristics which pertain to the environment in which the teacher is situated, but not properly under the control of any particular teacher (e.g. school condition).

As one might expect, many characteristics are interrelated. In particular, external characteristics such as school location can become linked with internal characteristics such as teacher salary.

2.6.2 Classification

The object of this study is to determine teacher characteristics according to the length of their terms of service in the Texas Education System. Fundamentally, the process this study seeks to implement is descriptive. The study seeks to answer the question: given a new teacher, defined by certain characteristics but with little or no service in the Texas Education System, how long might one expect the teacher to remain in the system? The basic categories, used in this study to describe teacher retention are: *long term, medium term,* and *short term*, whose numerical values are provided in Section 3.6.

The teacher classification scheme in its most basic form derives from similarities and dissimilarities among the teachers already in the system. For these teachers, their individual characteristics are known, along with their terms of service. When studying a new teacher, the study compares the char-

acteristics of the new teacher to those of teachers already in the system and arrives at a descriptive classification by looking at shared characteristics.

2.7 STEM Education and Data Mining

Data Mining has so far received little application in STEM education. This is unfortunate given the vastness of the field of education and the necessity of analyzing large amounts of data. There are some signs, however, that the situation is changing. Early in 2008 the first international data mining conference was held (International Educational Data Mining Society, 2011), and in July of 2011 the International Educational Data Mining Society was founded. Most of the research published during these years relates to students' gains, behavior, and use of technology in the classroom. Rather than techniques of data mining centering on classification, applications in education have focused more heavily on network analysis. McCornick, Carmichael, Fox, and Procter (2011) have noted that the majority of network research in education has been done in the area of teacher development. In particular, most of the studies up to now have focused on collaboration among teachers.

2.8 Current Issues

From the literature review described above, several aspects that are fundamental to the analysis of teacher retention in schools can be identified. One fact that is clear is the special attention that must be paid to teachers of mathematics and science. Mathematics and science teachers historically have

presented higher rates of teacher turnover. At the same time any decrease in the number of these teachers draws overwhelming attention from government and industry, since the nation's low academic achievement in science and mathematics is associated to slow economic growth and a lack of global competitiveness.

We have seen that certain factors are known to affect teacher attrition. Broadly speaking, these may be divided into two basic categories: internal and external factors. By internal factors we mean those factors which are particular to the teacher in question, such as age, subject taught, level of dissatisfaction with the job. External factors are those which are not properties of individual teachers. For example the school condition would be an external factor. School condition itself is a term that encompasses several different factors, such as the proportion of economically disadvantaged students and the proportion of minority students, among others. Unfortunately, studies show that the proportion of students from poor backgrounds (often measured by the number that qualify for lunch that is provided either free or at a reduced rate) is closely linked to the proportion of minority students within the school (Abbott & Joireman, 2001). Thus student poverty and ethnicity seem inextricably linked. Furthermore, we have seen that teacher and student ethnicity can interact in a way that can have a dramatic effect on student performance (Hemphill & Rahman, 2011). Student performance can in turn affect teacher dissatisfaction (Dee, 2005), which brings us full circle: teacher ethnicity can affect teacher dissatisfaction, and so can affect teacher attrition.

2.9 Importance of the Study

As presented in the literature review provided, the highest rates of teacher turnover are found in the fields of science and mathematics. Past studies have identified the problem and searched for methods to improve teacher attrition rates. Particularly in Texas, Mount (2012) did an extensive study on teacher migration focusing on science teachers. Mount (2012) found that 22 percent of science teachers move to different schools from year to year, and that this teacher migration occurred more from schools with a higher percentage of students eligible for free and reduced lunch.

However, research on mathematics teacher turnover in Texas has not yet focused on teacher characteristics. The literature review presented here has shown that teacher characteristics are important because they influence teachers' decision to stay or to leave the system. Moreover, the results of Mount (2012) have confirmed that, at least for science teachers, students' socioeconomic status, which is intimately related to their ethnicity, influences their decisions to stay or to leave a particular school.

This study presents the field of education, and the state of Texas, with an opportunity to isolate those crucial factors discouraging mathematics teachers from remaining in the teaching profession. Also, the methodology described in chapter 3 allows this study to take full avantage of a collection of valuable resources provided by TEA through the Public Education Information Management System (PEIMS) which, as explained in section 3.2, has not yet been used to study the growing phenomenon of teacher turnover.

Chapter 3

Methodology

3.1 Introduction

In this chapter I first describe the data set that is used for this investigation. I then describe the methodology for the current study and explain how the methodology is applied.

3.2 Data

The data set available for the proposed study comes from the Public Education Information Management System (PEIMS), compiled by the Texas Education Agency (TEA) from 2003 to 2007. PEIMS was established in 1986 by the State Board of Education with the purpose of improving education practices of local school districts (Texas Education Agency, 2011b). For each year, PEIMS "encompasses all data requested and received by TEA about public education, including student demographic and academic performance, personnel, financial, and organizational information" (Texas Education Agency, 2011b, ¶1). This data is requested by TEA from all school districts in Texas. PEIMS presents detailed information on teachers such as ethnicity, age, type of certification, subject taught, school and district where the teachers teach, salary,

years of experience, and other variables. See the appendix A for a complete list of the variables compiled in PEIMS. PEIMS is used by TEA to generate several reports such as:

- TEA Standard Reports. According to Texas Education Agency (2011b) the PEIMS Standard Reports are provided to fullfil requirements for information regarding public education in Texas. "Some of these reports comprise geographic information for each district and campus; while others contain listings of school districts or campuses by legislative districts. There are also reports that include information for high school graduates, or information for student enrollment. Other PEIMS Standard Reports include information concerning superintendents, staff or teachers employed by school districts."(Texas Education Agency, 2011b, ¶1)

- Academic Excellence Indicator System Accountability System (AEIS). The information collected here refers to the performance of students in each school and district in Texas. Texas Education Agency (2011b) reports that the indicators presented in AEIS are the following:

  - Results of the Texas Assessment of Knowledge and Skills (TAKS). For the 2011–2012 school year, TAKS data is only available for grades 10 and 11.
  - Exit-level TAKS Cumulative Passing Rates;
  - Progress of Prior Year TAKS Failers;

- Attendance Rates;
- Annual Dropout Rates (grades 7–8 and grades 9–12);
- Completion Rates (4-year and 5-year longitudinal);
- College Readiness Indicators;
- Completion of Advanced/Dual Enrollment Courses;
- Completion of the Recommended High School Program or Distinguished Achievement Program;
- Participation and Performance on Advanced Placement (AP) and International Baccalaureate (IB) Examinations;
- Texas Success Initiative (TSI) – Higher Education Readiness Component;
- Participation and Performance on the College Admissions Tests (SAT and ACT), and College-Ready Graduates.

Texas Education Agency (2011b) disaggregates these indicators by ethnicity, special education, and low income status. Starting in the 2002–2003 school year limited English proficient status is reported, in the 2003–2004 school year at-risk status is also reported by district, region and state, and in the 2008–2009 school year bilingual/ESL is also reported by district, region, and state. AEIS has its origins when the Texas Legislature saw the need for emphasizing the academic achievement of students as the foundation for accountability. Prior to this,

accountability was measured on the basis of schools following the rules and regulations. (Texas Education Agency, 2011b)

- Accountability System. The Texas Education Agency (TEA) is currently developing a new accountability system based on the STAAR (State of Texas Assessments of Academic Readiness) tests (Texas Education Agency, 2012). "In February of 2012, Commissioner Robert Scott amended the Texas Education Agency's House Bill 3 Transition Plan"(Texas Education Agency, 2012, ¶4). This amendment allowed changes in the public school accountability system of the state. Texas Education Agency, and Texas Higher Education Coordinating Board, and Texas Educators (2012) established the following indicators for determining accountability ratings:
  - Student performance on the STAAR grades 3-8 and End-of-Course (EOC) assessments.
  - Dropout Rates (including district completion rates) for grades 9 through 12.
  - High School Graduation Rates.

  Texas Education Agency, and Texas Higher Education Coordinating Board, and Texas Educators (2012) also specifies that the commissioner of education will determine how to assign ratings for accountability based on the recommendations from advisory groups and public input.

- Snapshot. According to Texas Education Agency (2011b) snapshot reports provide an overview of public education in Texas for a particular school year. The information provided in the snapshot report corresponds to state-level and a profile about the characteristics of each public school district.

- Pocket Edition. This report is published annually since the 1991–1992 school year. It reports state-level statistics on students; TAKS performance and participation; graduates and college admissions; attendance, completion, and dropouts; accountability ratings; personnel; and finances (Texas Education Agency, 2011b).

This study contributes to current research by taking advantage of the richness of data that has been systematically collected by TEA for many years. Despite having one of the largest education data bases in the world (Texas Education Agency, 2011b) and all the effort devoted by TEA to collect and analyze this incredibly valuable information, no study or report has yet appeared indicating how ethnicity and other teacher characteristics in Texas inform the increasing problem of mathematics teacher turnover.

3.3 Variables

As stated in Chapter 2 research on mathematics teacher turnover in Texas has not addressed teacher characteristics and the impact they might have on the teachers' decisions to stay or leave their profession. An investiga-

Table 3.1: The table provides a list and explanation of the variables encountered in the TEA data

Name in PEIMS	**Name in Study**	**Comments**
Scrambled_ID	Scrambled_ID	None
GENDER	Gender	None
ETHNICX	Ethnicity	Each ethnicity is identified by the following codes: White = 4; Hispanic = 3; African American = 2; Asian = 1; Native American = 0. This variable will be treated as a dichotomous variable. Cf. Section 3.6.
BIRTHDATE	Current Age	This variable comes from TEA using the following format: 14-Mar-49. The actual age will be used for the data analysis. It will be calculated based on the year of birth.
EXPER	Experience	This variable uses the following parameters: *short, medium* and *long term*. The justification for modifying this variable is presented in detail in Section 3.6.
BASEPAY	BasePay	None

tion by Allen (2005) has stated that in order to find real answers to the current problem of teacher turnover, investigators need to focus thoroughly on the "understanding of the characteristics of the teacher workforce and the impact those characteristics have on teachers' decisions to enter and remain in teaching and their success in the classroom" (Allen, 2005, p.iv). Another investigation by Ingersoll (2001) pointed out that research on where minority teachers tend to be employed has received little attention, as well as what happens with these teachers once they have been employed. He also stated that even less attention has been paid to the impact of minority teacher turnover. He stated, "As a recent review concluded, empirical research on minority teacher turnover has been limited, has had mixed findings, and, in general, has been inadequate ..." Ingersoll (2001, p.5).

The variables analyzed in this investigation were selected based on a combination of factors. Some, such as salary, were included because they have already received attention in the literature. Therefore they provide a benchmark to which to compare results of the study. Other variables, such as age and gender, were chosen because of a perceived paucity of adequate study within the literature. A final factor affecting the decision to include variables was the feasibility of their inclusion. Given the nature of the computational investigation, the inclusion of more than one factor was relatively easy to effect. However the inclusion of too many variables would have complicated and slowed the analysis (see Section 3.6). This led to the decision to include only a representative subset of characteristics, leaving others for subsequent

investigation. Now that the program is built, other variables can be added, the only cost being in program runtime. Variables such as school location, certification type, and number of classes taught by the teachers will be included in future work; their omission in the present study owes itself primarily to the need to streamline the data during initial phase of analysis as the methodology was developed.

To simplify how the information is displayed in this document some of the names of the variables have been modified. Table 3.1 describes the variables and modifications made to them.

3.4 Years

As noted in section 3.2, changes in how PEIMS disaggregates the data were introduced during the 2003–2004 school year. During this same school year other major changes were introduced. On the state level, a new accountability system was put in place. The new accountability system had to implement and report the results of the Texas Assessment of Knowledge and Skills (TAKS). The new TAKS was to include more subjects and grades and was supposed to be more rigorous than the previous system, the Texas Assessment of Academic Skills (TAAS) (Texas Education Agency, 2010, p.34). On the federal level, states were required to develop and submit Adequate Yearly Progress (AYP) criteria. "Each state was required to establish a timeline to ensure that not later than 12 years after the end of the 2001–2002 school year (2013–2014 school year), all students in each group will meet or

exceed the state's performance standards." (Texas Education Agency, 2010, p.34) In summary, the 2003–2004 school year was the starting point for a more rigorous accountability system, not only in the state of Texas, but also in the nation. The significance of this change made the 2003–2004 school year a natural starting point for the present investigation.

In addition, I analyzed the school year 2006–2007. Because of the classification scheme employed in the present study, the minimum term of service distinguished by the algorithm is 4 years or less. Thus in order to ensure the algorithm would retain the ability to identify, not only teachers who served medium– or long–term, but also short–term according to the classification scheme outlined above, it was deemed appropriate to look at a gap in school years where teachers new to the system in the first year could still have equal *a priori* chances of falling into any of the short–, medium–, or long–term categories. Given the initial reference point of 2003–2004, then in the school year 2006–2007 a teacher who was new in 2003–2004 could still be in the system yet maintain the possibility of being short–term.

3.5 Sample

As explained above, PEIMS collects data from all teachers in the Texas educational system in the state. Given that this study only focuses on the mathematics teachers within PEIMS, I have created a subset of PEIMS containing only the mathematics teachers for the 2003–2004 and 2006–2007 school years. Table 3.2 describes the total number of teachers and what courses those

teachers taught in each of the years analyzed in this study. For the 2003–2004 school year there is a total number of $32,253$ teachers and for the 2006–2007 school year there is a total of $32,933$ teachers. The first step to prepare the data set for this investigation shows an increment in the number of mathematics teachers in the state. 680 mathematics teachers joined the system during this 3 year period.

3.5.1 Extracting the Sample

I used the programming language Python (Rossum, 1995; Python Software Foundation, 1990–2013), an interactive shell for Python called IPython (IPython Development Team, 2008), and a Python library called Pandas (Lambda Foundry, Inc. & PyData Development Team, 2012) to extract and organize the data set used for this investigation. "Pandas provides rich data structures and functions designed to make working with structured data fast, easy, and expressive"(McKinney, 2013, p.4). Pandas allows Python to be a powerful data analysis tool, which is the main reason why this tool was chosen for this investigation.

3.6 Data Mining

The methodology to analyze this data set comes from the area of data mining and machine learning. The intent is to apply the following algorithm to the data: k–Nearest Neighbors.

The first step is the development of the *k–Nearest Neighbors* algorithm

Table 3.2: The table provides a list, from PEIMS, of the mathematics courses taught in middle and high school in the state of Texas and the total number of teachers teaching those courses during the 2003–2004 and 2006–2007 school years.

Mathematics Course	**2003–2004**	**2006–2007**
Algebra 1	7,515	7,581
Algebra 2	4,513	4,727
Algebra 1-4 Mathematics	101	0
Calculus AB	955	1,000
Calculus BC	190	240
Differential Equations	2	0
Discrete Mathematics	2	2
Foundations of Mathematics	2	2
Geometry	5,443	5,398
Mathematics Grade 7	5,664	5,699
Mathematics Grade 8	5,166	5,545
Mathematics Higher Level	14	0
Independent Study in Mathematics First Time	325	433
Independent Study in Mathematics Second Time	10	70
Mathematical Models with Applications	2,060	1,923
Problem Solving, Mathematical Models,and Computer Simulation	32	0
Statistics	245	312
Mathematical Methods Subsidiary	13	0
Number Theory	1	1

Table 3.3: The table provides a list, from PEIMS, of the mathematics courses taught in middle and high school in the state of Texas and the total number of teachers teaching those courses during the 2003–2004 that had from 0 to 4 years of teaching experience, and the list of teachers from that year that continued again in the 2006–2007 school years.

Mathematics Course	**2003–2004**	**2006–2007**
Algebra 1	2,379	1,669
Algebra 2	1,016	706
Algebra 1-4 Mathematics	32	25
Calculus AB	113	72
Calculus BC	10	5
Differential Equations	0	0
Discrete Mathematics	1	0
Foundations of Mathematics	0	0
Geometry	1,486	1,019
Mathematics Grade 7	1,983	1,387
Mathematics Grade 8	1,733	1,222
Mathematics Higher Level	0	0
Independent Study in Mathematics First Time	75	55
Independent Study in Mathematics Second Time	10	9
Mathematical Models with Applications	567	393
Problem Solving, Mathematical Models, and Computer Simulation	9	7
Statistics	44	28
Mathematical Methods Subsidiary	0	0
Number Theory	1	1
LDC Mathematics Grades 7–12	636	447

(Harrington, 2012). This algorithm has the virtue of being straightforward conceptually and computationally. In particular it converts each teacher's data to numeric values and views these as vectors in a state space. The data is then considered as a collection of some number n of vectors $\mathbf{v}$. The algorithm classifies a new data vector $\mathbf{w}$ by going through all n of the data vectors $\mathbf{v}$ in the data set, computing the distance $d(\mathbf{v}, \mathbf{w})$ between $\mathbf{w}$ and each vector, ordering the vectors $\mathbf{v}$ according to their distance from $\mathbf{w}$, and choosing the closest k vectors. The category that appears most often among those k vectors is the category to which the new data vector $\mathbf{w}$ is assigned.

An example of how *k–Nearest Neighbors* algorithm is applied to the data is the following:

Table 3.4 provides an excerpt of the information contained in the 2006 – 2007 PEIMS which is part of the data base used in this investigation. The names of the variables have been changed according to the description in Table 3.1. Also, the column identifying the vector has been added in order to explain how the calculations are done.

Table 3.4: Characteristics of Mathematics' Teachers. Data taken from 2006–2007 PEIMS. The *Vector* column has been added for purposes of explication. *Ethnicity* will be treated as a dichotomous variable.

Vector	Scrambled_ID	Current Age	Gen	Ethnicity	Exper
$\mathbf{v}_1$	3686	32	1	5	1 (short term)

continued on next page

Table 3.4: (continued)

Vector	Scrambled_ID	Current Age	Gen	Ethnicity	Exper
$\mathbf{v}_2$	4258457	64	0	5	14 (long term)
$\mathbf{v}_3$	4599759	53	0	4	22 (long term)
$\mathbf{v}_4$	14229703	44	0	3	1 (short term)
$\mathbf{v}_5$	14226919	39	1	2	5 (medium term)

For the example presented in this section only four variables are used: *Current Age, Gen, Ethnicity*, and *Exper*. This will help simplify the calculations for the purpose to exemplify the k-nearest neighbor methodology.

In this example there are 5 vectors, since there are 5 teachers. Teacher one is identified as $\mathbf{v}_1$, teacher two $\mathbf{v}_2$, and so on. If there is a new teacher entering the Texas educational system, it is unknown how long this teacher will stay in the system, but there is a way of making a prediction. The first step to make this prediction is calculating the Euclidean distance. Which for this example looks as follows:

$$\mathbf{v}_i = \begin{pmatrix} v_i^1 \\ v_i^2 \\ v_i^3 \\ v_i^4 \end{pmatrix} = \begin{pmatrix} \text{age of teacher } i \text{ [in years]} \\ \text{gender of teacher } i \\ \text{ethnicity of teacher } i \\ \text{years of experience of teacher } i \text{ [in years]} \end{pmatrix}$$

$$d(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\left(v_i^1 - v_j^1\right)^2 + \left(v_i^2 - v_j^2\right)^2 + \left(v_i^3 - v_j^3\right)^2 + \left(v_i^4 - v_j^4\right)^2}.$$

If we highlight the role of units in the above expression, this really means

$$d(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\left(v_i^1 - v_j^1\right)^2 [\text{years}^2] + \left(v_i^2 - v_j^2\right)^2 + \left(v_i^3 - v_j^3\right)^2 + \left(v_i^4 - v_j^4\right)^2 [\text{years}^2]}.$$

Since there are values that lie in different ranges, as shown in Table 3.4, it is common to normalize the data, otherwise Age would dominate, even though all variables are equally important. Also, ages and gender cannot be added together, this is an indicator that the distance function needs coefficients to handle the units. Specifically, the term

$$v_i^1 - v_j^1 \qquad \text{has units of} \qquad [\text{years}],$$

so that $\left(v_i^1 - v_j^1\right)^2$ has units of $[\text{years}]^2$. But the term $v_i^2 - v_j^2$ has no associated units (since one gender was assigned a one, the other gender a zero), and so neither does $\left(v_i^2 - v_j^2\right)^2$. It therefore makes no more sense to add these terms than it would to add two apples and two oranges. Similarly for the remaining terms in the expression for the Euclidean distance.

To handle such issues, it is necessary to slightly generalize the form of the Euclidean distance:

$$d(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{c^1 \left(v_i^1 - v_j^1\right)^2 + c^2 \left(v_i^2 - v_j^2\right)^2 + c^3 \left(v_i^3 - v_j^3\right)^2 + c^4 \left(v_i^4 - v_j^4\right)^2}.$$

If choosen, say, the number c^1 such that it has units of inverse–years-squared, then its units will cancel those of the square term which it multiplies:

$$c^1 \left[\frac{1}{\text{years}^2}\right] \cdot \left(v_i^1 - v_j^1\right)^2 [\text{years}^2] = c^1 \left(v_i^1 - v_j^1\right)^2 \left[\frac{\text{years}^2}{\text{years}^2}\right] = c^1 \left(v_i^1 - v_j^1\right)^2,$$

where the last expression is simply a numerical value, with no associated units. The units of c^4 are chosen in the same way to cancel the units in the last term in the distance; c^2 and c^3 need no associated units, since the associated terms in the respective products are themselves unitless.

Now, finally, the above procedure of resolving the units is combined with the issue of changing scales. Though the *units* of c^1 cancel those of $\left(v_i^1 - v_j^1\right)^2$, its *numerical value* is still undetermined. This numerical value is chosen in such a way as to make sure the product lies in the desired numerical range. For example, one might choose c^1 to be the inverse of the square of the maximum age in our data:

$$c^1 \left(v_i^1 - v_j^1\right)^2 = \frac{1}{(\text{maximum age})^2} \left(v_i^1 - v_j^1\right)^2.$$

Since the maximum age carries units of years, then c^1 chosen in this way has the units required above. Moreover, this term now always falls within the range $[0, 1]$. Thus, both problems are solved at once: unwanted units are removed, and the range of values obtained by this term is normalized. A similar procedure is applied to c^4, and to any of the other coefficients c^k in the distance function.

The variable Ethnicity also needs special consideration. There are five ethnicities represented in the data: White, African American, Hispanic, Asian, and Native American. TEA has coded them: 5, 4, 3, 2, and 1, respectively. The numbers identify the type of ethnicity, but the value of each number is not a metric, but rather a categorical variable. The fact that 5 is higher than

2, while 1 is lower, is irrelevant; the system merely notes that 5 and 2 are each different from 1 and therefore the ethnicities, respectively, are different. This means that a simple "distance" calculation, i.e. taking the magnitude of the difference between two values, is not quite appropriate in the case of the variable Ethnicity. Rather, we must specify a special distance function for this variable. If we let e_i represent the numerical value of the ethnicity of the ith teacher in the TEA scheme, then we may say that the *distance* d_{ij} between teachers i and j *in terms of ethnicity* is

$$d_{ij} := \begin{cases} 0, & \text{if } e_i = e_j, \text{ and} \\ 1, & \text{otherwise.} \end{cases}$$

That is, two teachers have an *ethnic distance* $d_{ij} = 0$ if their ethnicities are the same, and an ethnic distance $d_{ij} = 1$ if not. With this function the expression for the distance from the preceding discussion might take with the following form:

$$d(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{c^1 \left(v_i^1 - v_j^1\right)^2 + c^2 \left(v_i^2 - v_j^2\right)^2 + c^3 \left(v_i^3 - v_j^3\right)^2 + c^4 d_{ij}^2}.$$

One question the reader might have at this point is: between what are the distances being calculated? Between variables already in the data set? No. The interest is only in the distances a *new* data point is given, say $\mathbf{w}$, that comes from *outside* the data set. The data set is the list of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, for some integer n ($n = 5$ in Table 3.4), which can be written more simply as $\{\mathbf{v}_i\}_{i=1}^n$. This is a collection of vectors. The interest is in the collection of distances $\{d(\mathbf{w}, \mathbf{v}_1), d(\mathbf{w}, \mathbf{v}_2), \ldots, d(\mathbf{w}, \mathbf{v}_n)\}$, or simply $\{d(\mathbf{w}, \mathbf{v}_i)\}_{i=1}^n$, which is a collection of numbers.

After calculating the distances $\{d(\mathbf{w}, \mathbf{v}_i)\}_{i=1}^{n}$ the k nearest teachers are found by sorting the distances in increasing order. Assume, in this example, that $k = 2$. This value is much higher when using the complete data set. The parameter k is given by the programmer in a way that balances computational performance with accuracy of classification. As such, it can be adjusted after performing an error analysis based on the total error rate for the classifier. Harrington (2012) explains that the error rate is the number of misclassified pieces of data divided by the total number of data points tested. To simplify the classification procedure, three categories were created: *Short Term (st), Medium Term (mt),* and *Long Term (lt).* The first category covers a range from 0 to 4 years of experience, the second covers from 5 to 9 years, and the third covers 10 years of experience and above. These categories are based on Moreland, A. (2011)'s work, where she defines three categories: *Early, Senior,* and *Veterans* based on the teachers' years of experience. In Moreland, A. (2011) we find that the *Early* years are defined by the first five years of work experience. The *Senior* years are defined by the following four years of experience and the *Veteran* years are defined by nine or more years of experience. The classification used in the model of this investigation includes one more year of experience for the *Senior* and *Veteran* categories. The decision to include an extra year in each of the categories corresponding to *Senior* and *Veteran* in the work of Mount (2012) was made based on TEA's 2011–2012 Minimum Salary Schedule for Credited Years of Experienced Texas Education Agency (2011a).

3.7 Assumptions

In the exploration of data, the notion of "assumptions" comes into play as one decides what statistical tools to apply in order to better understand the model. Properly, this falls under the purview of statistical *inference*: that is, one may talk of *descriptive* statistics, where one strives to describe key measures of a data set, such as the mean, mode, and median, in order to simplify understanding of a vast amount of unwieldy information; or one may talk of *predictive* statistics, or statistical inference, where one strives to encounter a model, or mathematical framework, which not only serves to delineate certain of the descriptive parameters of the data on hand, but also supplies certain additional parameters or assumptions which, perhaps, in combination with the descriptive parameters, may serve as a guide to the prediction of future values of data collected under similar circumstances.

For example, having a set of student test grades, one may simply seek to describe the data, and may therefore calculate properties such as the average value. This average, however, tells nothing about what to expect when the test is re-administered in the future.[1]

However one may also suspect that a given student's test scores increase linearly with age. This marks the introduction of a *model*. By making

[1]This should not be confused with an inferential model with only one parameter. If one wishes to create a predictive model which has only one parameter, then one may go through the various calculations to find that, for a wide variety of situations, the average value is the single best "point–estimator". But this is a predictor, not a descriptor. It just happens that the former gives the value of the latter when applied to the same set of data.

this assertion, one can describe a way of predicting, given some information about a student (her age), a future data value (the score when the test is re-administered). Of course one may make all manner of assertions about sets of data, but ultimately one must see if the model fits the data on hand.

In this hypothetical example, a linear relationship is supposed, and in statistical terms, it is supposed that one is attempting to do linear regression on the data on hand. Should that analysis prove successful, then the trend line that results would provide with a way of incorporating new data: given a line relating age to score, in order to predict the score of a particular student on the next test, one looks at the age of the student, and then claims that the corresponding value on the line is the predicted score.

Of course, for the regression analysis to work, it is necessary to ensure that certain assumptions are valid. Which assumptions? The assumptions that occur in the theorems pertaining to the topic. For example, the typical case for linear regression requires that measurement errors be normally distributed and independent. (Certain theorems might weaken these restrictions.) Thus, before even considering the application of a particular model, one must investigate whether the data satisfies the requirements of the theorems that ensure the model's validity.

Data mining often applies statistical methods. It also finds application as a method of attacking data that is too large or too varied for a researcher to find an applicable statistical model, or to verify whether the data satisfies the specific requirements of the theorems pertaining to the desired model. Data

mining therefore provides in many cases a different, complementary mode of analysis. In particular, it often seeks to describe the data on hand, and even more importantly to predict aspects of new data, *in the absence* of a particular statistical model: data mining often makes no claims as to what statistical model best describes the data. And in that sense, it also frequently lacks any accompanying requirements that the data must satisfy for a given model to be applied (e.g. that the variables obey certain statistical properties required by theorems).

The proposed methodology, k–Nearest Neighbors, makes *no particular assumptions about the data* other than that it contains numeric values. This holds simply because the algorithm tries to calculate "distances", and this calculation requires numerical values.[2] This however is no different from any part of statistics: statistics is a mathematical discipline and therefore only applies to numbers, and data which involves text or other elements must first be converted to numerical values before statistical analysis can be applied. Thus for non-numeric values, it is necessary to devise a discrete numeric encoding corresponding to the categorical values, just as in any other setting.

[2]The term "distance" here is convention, nothing more. "Similarity" or "dissimilarity" are other common terms for much the same mathematical entity. There need not be any *a priori* interpretation as physical distance.

3.8 Distance

Many data mining techniques, and k–Nearest Neighbors in particular, require a so–called "distance" function. This is simply a function that, given any two data vectors $\mathbf{v}$ and $\mathbf{w}$, produces a number $d(\mathbf{v}, \mathbf{w})$ which we interpret as the distance between them in some generalized sense. That is, it provides a numerical measure of how "far" from one another those two points are. If the data points consist of coordinate locations, then most likely the distance function's value will have the interpretation of a true physical distance. But this need not be the case. In general it may provide the same notion of "far" as when we talk about two people as being "far apart on the political spectrum": it just means that, given a number of distinct data elements, $\mathbf{v}$ and $\mathbf{w}$ differ in many of those individual elements.

Mathematically a distance function, or technically a *metric*, satisfies some basic properties:

- It must be symmetric:

$$d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v}).$$

  That is, the distance from one point to another must be the same as from the other to the one.

- It must be positive definite:

$$d(\mathbf{v}, \mathbf{w}) \geq 0, \qquad \text{with equality only if } \mathbf{v} = \mathbf{w}.$$

That is, the distance should be a positive number, and it should be zero only if the two items are the same point.

- It should satisfy the triangle inequality:

$$d(\mathbf{v}, \mathbf{z}) + d(\mathbf{z}, \mathbf{w}) \geq d(\mathbf{v}, \mathbf{w}).$$

That is, as in a triangle, the distance between two points should be no greater than than the distance travelled going from the first point to an intermediate point, and then from the intermediate point to the last point.

To see this more clearly, we may imagine a function describing physical distance. In this case, when the distance between objects $\mathbf{w}$ and $\mathbf{v}$ is zero, they occupy the same physical position. If these are baseballs, say, then two baseballs cannot occupy the same physical location. For baseballs $\mathbf{w}$ and $\mathbf{v}$ to occupy the same position, they must be the same baseball: $\mathbf{w} = \mathbf{v}$. Thus, discussions of metrics typically assume that zero distance implies identity. In the present study, however, the situation is different. We can imagine that two teachers have the same ethnicity, have been teaching the same number of years, teach in the same school, etc., and yet are not the same actual person. Hence, in terms of the distance presented above, it may still be the case that $\mathbf{w} \neq \mathbf{v}$, even though $d(\mathbf{w}, \mathbf{v}) = 0$.

Similarly, the triangle inequality, as its name suggests, incorporates the intuition that the shortest distance between $\mathbf{w}$ and $\mathbf{v}$ is along the straight line

joining them; if one attempts to travel between the two points by traveling from the first to a third point, $\mathbf{z}$, and then from that third point to the other, then one will necessarily travel a longer distance, unless $\mathbf{z}$ lies on the straight line joining $\mathbf{w}$ and $\mathbf{v}$. Since the space under consideration is not a physical space to be traveled, but rather a space of points $\mathbf{w}$ consisting of teacher characteristics, the notion of straight-line travel might not carry all the intuition garnered from the study of situations in physics. So the triangle inequality might not need to be imposed as a condition for a viable "distance" function for the purposes of this study, even though the use of "distance" in this case would break with terminology standard in other subdisciplines of mathematics.

When values are categorical, this simply means that, in numeric terms, their possible values form a discrete set. In such circumstances one often tailors the distance function so that it has the desired interpretation. A common practice is to treat data points as having a distance of 0 when they share a categorical value, and as having a distance of 1 when they differ in the value of that categorical value.

This need not pose a problem when a data vector $\mathbf{v}$ contains components, some of which are continuous, others of which are discrete. The above prescription serves to compute the distance for the discrete components, while a normal difference $v_k - w_k$ would serve to compute a distance in continuous components. Providing all of the components were normalized (i.e. scaled to fall within a similar range of values), then combining components of different types should pose no *a priori* difficulties. The only difficulty arises when we

try to impose a particular interpretation, such as physical distance. But if we realize we are only concerned with "similarity" or "dissimilarity", then we need not be concerned with having differing measures of "likeness" for differing components of a data vector $\mathbf{v}$.

3.9 Utility

A typical question posed by a statistical investigation of data is, "What factors most contribute to the effect being studied?" For example, in the context of the present investigation, one might ask, "What factor most contributes to a teacher's longevity in the school system?" Why ask such a question? Presumably because one could think that understanding the relation between the factors and the effects will help discern in the future what characteristics are associated with those teachers who remain longest in the system based on the values they present for the relevant factors.

Investigation of such a question necessarily implies the use of a model. It was supposed right from the beginning that some factors are more important than others. Then it is necessary to go through the process of justifying how and why certain factors are discarded and not others, and it is necessary to ensure that the remaining factors are not interdependent, and so on. At each stage additional assumptions are brought in.

A data mining algorithm like k–Nearest Neighbors seeks to avoid this process. Though one can reduce the number of factors one wishes to process, this is not strictly necessary. The algorithm in principle is designed to

work with all factors present in a data set, limited only by the computational resources of the machine on which the algorithm is run.

How, then, one decides which factors really are important for determining the effect under consideration? In the present context, how can one determine which teacher characteristics contribute most greatly to teacher retention? There are two answers to this question.

The first is: the k–Nearest Neighbors algorithm takes as input *all* the data on hand. Once an algorithm is designed based on current data (and tested with a subset of that data to see that it yields accurate results), then as a new data vector $\mathbf{v}$ is fed, it will provide with an evaluation based on all information contained in that vector. One need not tell it beforehand to disregard certain factors and to pay attention to a particular few. The algorithm is fed with everything one has, and it comes back with an answer. In this sense it is truly holistic: it evaluates a teacher as a *whole*, using all available data.

The second answer is: Having a distance with a standard Euclidean form like

$$d(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{k=1}^{n} c_k (v_k - w_k)^2},$$

where the coefficients c_k take care of the units associated with each component of the data, as well as the scaling of the range, then one is free to adjust these parameters. In particular, one can make one coefficient so much bigger than the others such that their corresponding terms will outweigh other terms in the sum. In such a way we can provide more *weight* to some factors relative to

others. This amounts to asserting that some factors are more important than others.

How could one know which factors are more important? For this, one reserve a part of the data for testing the algorithm. As adjustments are made to the weights c_k, one can test to see whether the algorithm performs better or worse on test data (i.e. on data where one already knows what the answer should be). This might involve looking at false positives or other measures. This is a free–form process where the designer has free rein to change the weights, but at the same time she has a concrete method of evaluating those changes: she looks for an improvement of the algorithm's accuracy on the test data.[3] It is important to keep in mind: the adjustments are made not in the search for a particular, well–defined error rate. Rather the algorithm designer adjusts the weights to *improve* the error rate: given a test run with a particular error rate, the designer then adjusts the weights, runs the algorithm again, and compares the new error rate with the previous error rate to see if the adjustments improved performance or not. See Section 3.6 for further discussion of error rates.

The parameter k, on the other hand, is really an algorithm–intrinsic parameter. It bears little if any relation to the relative importance of the factors in the data. Rather, once a metric has been defined, k merely determines

[3]The weights of certain factors are increased, not based on whether they "should" be more important according to some theoretical consideration, but rather based on whether the algorithm *performs better* on the data under consideration.

how many of the "neighbors" of a new data point $\mathbf{v}$ get a "vote" as to what category $\mathbf{v}$ should belong to. Ultimately the k–Nearest Neighbors algorithm boils down to this: voting. Once the algorithm has been designed and tested, then we feed it new data points one at a time. For each new data point $\mathbf{v}$, the algorithm goes back through all the n original data vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$, calculates the distances $\{d(\mathbf{v}, \mathbf{w}_1), d(\mathbf{v}, \mathbf{w}_2), \ldots, d(\mathbf{v}, \mathbf{w}_n)\}$, orders them from decreasing to increasing, and figures out which collection of k data vectors ($k \leq n$) are "closest" to $\mathbf{v}$. Then, looking at these k vectors, it simply takes a vote: what category appears most? The answer to that question determines the category assigned to the new data vector $\mathbf{v}$.

By selecting different values for k, one selects how many vectors get to vote. For $k = 1$, we are saying that we will assign $\mathbf{v}$ the same category as that single vector which lies "closest" to it in terms of our metric. But perhaps they lie close to one another because of one single factor on which they agree, and that factor happens to dominate all others. So we should probably take more neighbors. Perhaps we should take $k = 3$: then whichever category appears two out of three times in the three nearest neighbors of $\mathbf{v}$ will be assigned as $\mathbf{v}$'s category. But what if all three neighbors have distinct categories? Then no single category appears more than any other, and so it is unclear as to which category $\mathbf{v}$ should have. So we should set k to a value high enough that we expect at least one category to occur with some frequency. But we must keep k low enough that its value does not serious slow the program's performance.

This manner of determining k is more art than science. It simply

balances concerns such as runtime with overall performance of the algorithm in its role of categorization. This is less a theoretical parameter than a pragmatic one. And its value does not so much affect the validity of the procedure as its effectiveness in completing its task.

3.10 Outliers

Any statistical analysis must take care to identify outliers, those points that lie "far" from the "typical" points of the data. But we must recognize that the sheer assertion of the existence of outliers presupposes some, perhaps hidden, theoretical assumptions.

If we simply look at descriptive statistics, outliers are simply the relatively few points that do not bunch with the relatively many. For example, if we simply describe data using box–and–whisker plots, then outliers fall beyond the whiskers (or beyond the box, depending on one's concept of what it means to fall outside of the central group). In such a context, an outlier is nothing of particular concern: it is a natural outcome of the non–uniformity of the data under consideration.

Where outliers become problematic is in statistical inference, when we try to construct a model based on the data on hand in order to predict likely outcomes of new measurements. The reason outliers become problematic is the following. The basic process of statistical inference is this: we collect a set of data, we describe a model with a relatively small number of parameters (much smaller than the size of the collection of data), we manipulate the data in such

a way as to determine those parameters, and then we use the resulting model (now with its parameters fixed) to predict outcomes of new measurements. In this process we use the data on hand to fix the parameters of the model.

The problem of outliers arises due to this use of the data to fix the parameters. Typically the process of fixing parameters requires the researcher to assume that the data on hand is actually a sample of a repeated set of identical experiments. This assumption generally factors crucially into how the parameters are determined from the data. But outliers are precisely those points of a sample which are "atypical": in repeated experimentation, those outliers should occur less frequently than the points in the "center" of the sample. But if we only have our one data set with which to determine the model parameters, then these outliers will contribute with equal weight to the central points, even though we know they are the less–typical values. In this sense, we say the outliers "distort" the fixing of model parameters: the parameter values will now include what should be rare and inconsequential data points on equal footing with more typical points.

But in an algorithm like k–Nearest Neighbors, we are not trying to fix parameters in this way. We are not trying to describe a robust set of data by a handful of parameter values. Instead, we are simply trying to find which values in our current data are "most like" (closer to) a new data point. Maybe our new data point is an "outlier", i.e. a point located far from a central group. Then hopefully it will be near other outliers, and our algorithm will classify it as such. If we try to remove points in our data that we think are outliers,

we will undercut the algorithm's ability to classify new outliers as what they really are. But we must keep in mind: removal of outliers is intimately tied to the notion that our data derives from a random sample of data encountered in repeated experiments. Data mining with algoritms like k–Nearest Neighbors does not necessarily make that assumption. From the point of view of the algorithm, the data we have is simply that: the data we have. From the algorithm's point of view, outliers are a mere curiosity: a special name given to points whose distance from other points is "large". But those points should remain in the data: if not, then how would the algorithm know to classify a new data point whose distance from the center is also "large" with those outliers?

3.11 The Model

3.11.1 File Manipulation

The first step in developing the algorithm that can run the data set for its analysis is to organize the files containing the data set. The original data set comes from TEA split in different files. Each file contains the information per year. I am only using the files for the 2003–2004 and 2006–2007, therefore the process described in this section needs to be repeated two times.

Each file is a comma separated file. Each entry in the files lists the following information per individual teacher:

- Scrambled_ID. This column identifies each teacher in the system using a

unique id per teacher.

- Year. This variable refers to the school year where the data is collected. The information in each raw is the same for all teachers in the file.
- District. This column lists the identification number of the district where the teacher teaches.
- Distname. This column lists the name of the school district where each teacher teaches.
- Campus. Lists the identification number of the school link to the previous variables.
- Campname. Lists the name of the school link to the previous variable.
- Fname. Lists the first name of the teacher.
- Lname. Lists the last name of the teacher.
- Ethnicx. Lists the ethnicity of each teacher using the TEA code for each ethnicity. See Table 3.1.
- Gender. Lists the gender of each teacher.
- Birthdate. Lists the date of birth of each teacher. See Table 3.1.
- Exper. Lists the years of experience of each teacher.
- Basepay. Lists the annual salary for each teacher.

- Service. Lists the number that identifies the course taught by each teacher.
- Servicex. Lists the name of the course link to the previous variable.
- Grade_levelx. Lists the grade level taught by each teacher.

Using Pandas, I first extracted the teachers that teach mathematics. Given the size of the original file, I decided to split the information in different files. I created a file per each mathematics course taught. As mentioned before, I needed to repeat this step using the file for the 2006–2007 school year. Then, I repeated the process one more time to extract only the mathematics teachers that had from 0 to 4 years of experience during the 2003–2004 school year. The next step in organizing the data was the identification of those teachers that remained in the system from the 2003–2004 school year to 2006–2007 school year. Once again with the help of Python and Pandas, I was able to extract the ids of those teachers that remained in the system. The logic of the program that extracted the data is the following:

- Step 1. Check the data for 2003: i.e. open the file, put it in a DataFrame
- Step 2. Go through each Scramble_ID in 2003. This is what the program does as it goes through it:
  - for each ID in 2003 data:
  - look for that ID in the 2007 data

- if it is there, keep that ID in a list
- otherwise, forget it
- output: the list of IDs which persist

3.11.2 Creating the Model

Having the data set organized in the files, the next step is to create the algorithm that defines the model to analyze the data set. The logic behind developing the algorithm is the following:

- I start by reading in the database assuming it is contained in a single CSV file.
  - The Input: 'infilename': the name of the CSV file containing the data base
  - The Output: 'dataBase': a Pandas DataFrame containing the data base
- The new point to be classified is read in assuming it is contained in a separate CSV file. In fact, I allowed for the possibility that the CSV file could contain multiple rows, each to be interpreted as a separate point that needs classification.
  - The Input: 'infilename': the name of the CSV file containing the points to be classified (one per line)

- The Output: 'newPoints': a Pandas DataFrame containing the points to be classified (one per row)

As explained in 3.3 some of the variables in the data set need to be modified in order to have the appropriate format for running the algorithm. These step was done as follow:

- *BIRTHDATE* was the first variable modified. The date of birth is given in the form 'dd-mm-yy' (e.g. 14-Mar-49), where the last two digits are the ones I am interested in. (The tacit assumption is that all teachers in the database will have been born before the year 2000). I added a new column to the data base with the current age. The name of the column is *Current Age*.

  - The Input:

    * 'dataBase': a Pandas DataFrame containing the data base
    * 'birth_col': the name of the column (as a string) containing the dates of birth of all teachers in the data base (default 'BIRTHDATE')
    * 'current_year': the numerical value of the current year (default '2013')

  - The Output: 'dB': the modified data base including columns for the 'Birth Year' and 'Current Age'

- The next variable modified was *GENDER*, which was handled as a categorical variable. I represented the opposing categories of 'male' and 'female' by '1' and '0', respectively. The following function allows me to make this kind of switch for an arbitrary list of categories, together with a corresponding list of numerical values by which to encode them. Then a new column, *GEN* was created.

  - The Input:

    * 'dataBase': the Pandas DataFrame containing the data base
    * 'col_orig': the string representing the name of the column whose categorical data we wish to represent numerically (default 'GENDER')
    * 'col_numb': the string representing the name of the column where the corresponding numerical data will be placed (default ''GEN'')
    * 'categories': a list of strings denoting the categories to be coded (default '['MALE', 'FEMALE']')
    * 'values': a list of numbers representing the numerical values corresponding to the categories (default '[1, 0]')
    * 'default': a default numerical value, used for creating the new column (default '0')

  - The Output: 'dB': a Pandas DataFrame containing the data base, with an added column containing the numerical representation of

the categorical data

The following function calculates the distance for quantities, such as ethnicities (*ETHNICX* third variable modified), where the only important information is whether the elements are identical or not. This goes by the name of the discrete distance in the topology of metric spaces. Suppose we have points x and y. Then the idea of the discrete distance is the following:

- The idea:
  - If 'x' and 'y' are the same, return 0;
  - If different, return 1.

  As such, the quantities do not even need to be coded as numerical values; they can remain strings if desired.

  - The Input:
    - 'newPoint': the new data vector you wish to classify
    - 'dataBase': the collection of data (as Pandas DataFrame) with which to compare new_point
    - 'cols_dscr': a list of columns requiring treatment with discrete distance, such as with ethnicity
  - The Output: 'spec_dists': a DataFrame with 0 in the rows where quantities were the same, 1 elsewhere, organized by column

The last step before creating the algorithm is the normalization of the data. It should take place before any data mining algorithm is applied. Ultimately this is part of the data preparation, and not part of the mining algorithm itself. In particular we need to make sure that the new point is included in any normalizations of date set; if not, then the data elements of the new point will be markedly off-scale and skew the results of the mining algorithm.

This function normalizes rows with numerical values:

- Input:
- 'newPoints': a list, Pandas Series, or Pandas DataFrame which we wish to classify (this should now be able to handle the case where 'newPoints' contains several points to be classified)
- 'dataBase': a Pandas DataFrame containing the (classified) vectors of the data base
- 'columns': a list of the names of the columns we wish to normalize
- 'constants': a list of numerical coefficients that can serve as weights for each column (defaults to 1.0 for each column)
- Output:
- 'nP_norm': a Pandas DataFrame with 'newPoint"s data in normalized form
- 'dB_norm': a Pandas DataFrame with 'dataBase"s data in normalized form

3.12 Future Use

What is the result of the implementation of the k–Nearest Neighbor algorithm applied to our teacher data? Ultimately it is a program (G. H. Krause, 2013–2014). It is a program we can run any time we are given a new piece of data, that is, the characteristics of a new teacher. The program will classify the new teacher based on similarities to characteristics of teachers in the existing data.

Now as we run the algorithm on a new data point $\mathbf{v}$, i.e. a new teacher introduced to the system, the algorithm will find which vectors $\mathbf{w}$ in the current data set have the closest characteristics and then categorize $\mathbf{v}$ as likely to remain for a short, medium or long term, based on the majority vote of those vectors $\mathbf{w}$ closest to $\mathbf{v}$. Such is the nature of the k–Nearest Neighbors algorithm.

Finally, one may analyze the program itself. Suppose the metric has the form stated below,

$$d(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{k=1}^{n} c_k (v_k - w_k)^2},$$

and it seems to do a reasonable job at classifying teachers according to longevity, then one can say that those factors with the largest weights c_k are the ones that most contribute to the process of classification. That is, they are the most important contributors to determining teacher retention.

The interpretation is slightly subtle, however. These factors themselves need not determine how long a teacher will remain in the system. Rather they will be the factors which most determine the *similarity* between teachers of a given length of time in the system. Why those teachers remain in the system might still remain unclear, but looking at a successful metric might give insight into what properties the teachers have in common.

Chapter 4

Results

4.1 Introduction

This chapter organizes and reports the main results of the study, including relevant quantitative (statistical and computational) data. The purpose of developing a model to examine the data was to identify the characteristics of minority mathematics teachers that remain in the Texas educational system for extended periods of time. This concurrent model explores all mathematics teachers from the 2003–2004 school year reported in PEIMS. It also explores all mathematics teachers from the 2006–2007 school year reported in PEIMS. The model used in this study is based on Data Mining techniques. Specifically, the intent is to apply the k–Nearest Neighbors algorithm to the data. The k–Nearest Neighbors algorithm allows for a novel look into the characteristics of hispanic mathematics teachers and provides the ability to test the notion of teacher retention against the backdrop of current literature on teacher turnover.

4.2 Descriptive Analysis

The overarching inquiry at the center of this research is, **What are the characteristics of mathematics teachers that are related to teacher retention in Texas schools?** Several salient characteristics of the mathematics teachers during the 2003–2004 school year are described in Tables 4.1–4.13.

There are no records of teachers with 0–4 years of experience in the following courses: Differential Equations, Foundations of Mathematics, Mathematics Higher Level and Mathematical Methods Subsidiary. The information provided by TEA contained no description of the content of the Mathematical Methods Subsidiary course. The TEKS did not contain any information regarding this course, nor is there any specification of the grade level corresponding to this course other than that it is a secondary mathematics course. For these reasons, the data on the characteristics of the mathematics teachers listed below does not contain data on those listed as teaching Mathematical Methods Subsidiary.

There was only one teacher teaching Discrete Mathematics with 0–4 years of experience (specifically, 3 years) in the 2003–2004 school year. The teacher was 33 years old at that time, male, Asian and his salary was $39,998 per year. He did not stay in the system for the 2006–2007 school year. This teacher is included in the data for this investigation.

There was also only one teacher teaching Number Theory during the 2003–2004 school year with 0–4 years of experience (specifically, 2). The

teacher was 38 years old, male, white and his salary was $38,491 per year. He did stay in the system through the 2006–2007 school year. His salary increased to $45,050 by that year. This teacher is also included in the data for this investigation.

The characteristics of the teachers teaching Algebra 1-4 Mathematics are not included in this analysis. As with the course Mathematical Methods Subsidiary, none of the records provided describe the content of the Algebra 1-4 Mathematics classes. Since there are only 28 teachers in the data set that appear to have from 0–4 years of experience in the 2003–2004 school year, they have not been included in the analysis.

Each table lists, by characteristic, the proportion of mathematics teachers present in the 2003–2004 school year who were also present in the 2006–2007 school year. As can be seen, throughout each of the tables there is a tendency to find female mathematics teachers staying in the system with a higher proportion than their male counterparts. Moreover, in all courses hispanic teachers tend to stay longer, as well as teachers in their 50s and those with a salary between $40,000 and $50,000.

At the same time, in each of the courses it is frequently seen that African American mathematics teachers leave the system at a much higher rate than any other ethnicity, as well as those teachers in their 60s and 30s, and those with a salary in the range of $30,000–$40,000 per year.

Table 4.1: Retention rates, broken down according to characteristic, of Algebra I teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Characteristics		**Algebra I Initial**	**Remaining**	**Retention**
Ethnicity	White	1,565	1,089	0.70
	Hispanic	450	345	0.77
	African American	271	173	0.64
	Asian	87	58	0.67
	Native American	6	4	0.67
Gender	Male	978	702	0.72
	Female	1,401	967	0.69
Age	20s	1,197	840	0.70
	30s	595	427	0.71
	40s	358	242	0.68
	50s	195	141	0.72
	60s	32	19	0.59
	70s	2	0	0.00
Salary	$\leq \$10,000$	31	21	0.67
	$\$10,001$ to $\$20,000$	94	61	0.65
	$\$20,001$ to $\$30,000$	492	338	0.69
	$\$30,001$ to $\$40,000$	1,704	1,213	0.71
	$\$40,001$ to $\$50,000$	52	32	0.62
	$\$50,001 \leq$	6	4	0.67
Total		2,379	1,669	0.70

Table 4.2: Retention rates, broken down according to characteristic, of Algebra II teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Algebra II				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	647	434	0.67
	Hispanic	223	180	0.80
	African American	96	64	0.67
	Asian	44	23	0.52
	Native American	6	5	0.83
Gender	Male	427	292	0.68
	Female	589	414	0.70
Age	20s	486	331	0.68
	30s	270	192	0.71
	40s	152	106	0.69
	50s	90	66	0.73
	60s	17	11	0.65
	70s	1	0	0.00
Salary	$\leq \$10,000$	6	4	0.67
	$\$10,001$ to $\$20,000$	32	21	0.66
	$\$20,001$ to $\$30,000$	250	168	0.67
	$\$30,001$ to $\$40,000$	695	491	0.71
	$\$40,001$ to $\$50,000$	29	19	0.66
	$\$50,001 \leq$	4	3	0.75
Total		1,016	706	0.69

Table 4.3: Retention rates, broken down according to characteristic, of Calculus AB (AP Calculus) teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Calculus AB (AP Calculus)				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	95	59	0.62
	Hispanic	7	5	0.72
	African American	5	3	0.60
	Asian	6	5	0.83
	Native American	0	0	0.00
Gender	Male	52	31	0.60
	Female	61	41	0.67
Age	20s	49	30	0.62
	30s	31	19	0.63
	40s	16	11	0.69
	50s	9	7	0.78
	60s	8	5	0.63
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $\$20,000$	3	2	0.67
	$\$20,001$ to $\$30,000$	42	27	0.64
	$\$30,001$ to $\$40,000$	61	38	0.62
	$\$40,001$ to $\$50,000$	7	5	0.72
	$\$50,001 \leq$	0	0	0.00
Total		113	72	0.64

Table 4.4: Retention rates, broken down according to characteristic, of BC (AP Calculus) teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

	Calculus BC (AP Calculus)			
	Characteristics	**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	7	4	0.57
	Hispanic	1	0	0.00
	African American	2	1	0.50
	Asian	0	0	0.00
	Native American	0	0	0.00
Gender	Male	8	3	0.37
	Female	2	2	1.00
Age	20s	3	2	0.67
	30s	4	2	0.50
	40s	1	1	1.00
	50s	1	0	0.00
	60s	1	0	0.00
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $\$20,000$	0	0	0.00
	$\$20,001$ to $\$30,000$	2	1	0.50
	$\$30,001$ to $\$40,000$	6	2	0.67
	$\$40,001$ to $\$50,000$	2	2	1.00
	$\$\$50,001 \leq$	0	0	0.00
Total		10	5	0.5

Table 4.5: Retention rates, broken down according to characteristic, of Geometry teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Geometry				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	957	632	0.67
	Hispanic	300	233	0.78
	African American	163	108	0.66
	Asian	59	39	0.66
	Native American	7	7	1.00
Gender	Male	663	452	0.68
	Female	823	567	0.69
Age	20s	679	458	0.67
	30s	394	273	0.69
	40s	247	173	0.70
	50s	135	94	0.70
	60s	29	20	0.69
	70s	2	1	0.50
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	71	42	0.59
	$\$20,001$ to $30,000$	335	222	0.66
	$\$30,001$ to $40,000$	1028	719	0.70
	$\$40,001$ to $50,000$	48	34	0.71
	$\$50,001 \leq$	4	2	0.50
Total		1486	1019	0.69

Table 4.6: Retention rates, broken down according to characteristic, of Independent Study First Time teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Independent Study First Time Characteristics		2003–2004	2006–2007	Retention
Ethnicity	White	40	26	0.65
	Hispanic	24	21	0.88
	African American	4	3	0.75
	Asian	7	5	0.71
	Native American	0	0	0.00
Gender	Male	40	29	0.73
	Female	35	26	0.74
Age	20s	40	29	0.73
	30s	18	13	0.72
	40s	11	8	0.73
	50s	0	0	0.00
	60s	0	0	0.00
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	2	2	1.00
	$\$20,001$ to $30,000$	14	10	0.72
	$\$30,001$ to $40,000$	56	40	0.72
	$\$40,001$ to $50,000$	3	3	1.00
	$\$50,001 \leq$	0	0	0.00
Total		75	55	0.73

Table 4.7: Retention rates, broken down according to characteristic, of Independent Study Second Time teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Independent Study Second Time Characteristics		2003–2004	2006–2007	Retention
Ethnicity	White	5	4	0.80
	Hispanic	5	5	1, 00
	African American	0	0	0.00
	Asian	0	0	0.00
	Native American	0	0	0.00
Gender	Male	7	6	0.86
	Female	3	3	1.00
Age	20s	4	4	1.00
	30s	2	1	0.50
	40s	2	2	1.00
	50s	2	2	1.00
	60s	0	0	0.00
	70s	0	0	0.00
	80s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	0	0	0.00
	$\$20,001$ to $30,000$	1	1	1.00
	$\$30,001$ to $40,000$	9	8	0.89
	$\$40,001$ to $50,000$	0	0	0.00
	$\$50,001 \leq$	0	0	0.00
Total		10	9	0.90

Table 4.8: Retention rates, broken down according to characteristic, of Mathematical Models with Applications teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Mathematical Models with Applications				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	406	275	0.68
	Hispanic	84	66	0.79
	African American	62	42	0.68
	Asian	14	9	0.64
	Native American	1	1	1.00
Gender	Male	266	186	0.70
	Female	301	207	0.69
Age	20s	233	165	0.71
	30s	154	111	0.72
	40s	95	65	0.68
	50s	68	45	0.66
	60s	15	7	0.47
	70s	2	0	0.00
Salary	$\leq \$10,000$	9	8	0.89
	$\$10,001$ to $20,000$	31	21	0.68
	$\$20,001$ to $30,000$	157	107	0.68
	$\$30,001$ to $40,000$	348	246	0.71
	$\$40,001$ to $50,000$	20	11	0.56
	$\$50,001 \leq$	2	0	0.00
Total		567	393	0.69

Table 4.9: Retention rates, broken down according to characteristic, of Grade 7 teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Mathematics, Grade 7				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	1,251	859	0.69
	Hispanic	389	299	0.77
	African American	628	186	0.69
	Asian	66	36	0.55
	Native American	9	7	0.78
Gender	Male	651	458	0.70
	Female	1,332	929	0.70
Age	20s	954	648	0.68
	30s	535	380	0.71
	40s	354	265	0.75
	50s	125	86	0.69
	60s	14	7	0.50
	70s	1	1	1.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	61	42	0.67
	$\$20,001$ to $30,000$	339	239	0.71
	$\$30,001$ to $40,000$	1,547	1,080	0.70
	$\$40,001$ to $50,000$	33	25	0.76
	$\$50,001 \leq$	3	1	0.33
Total		1,983	1,387	0.69

Table 4.10: Retention rates, broken down according to characteristic, of Grade 8 teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Mathematics, Grade 8				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	1,082	755	0.70
	Hispanic	328	268	0.82
	African American	267	164	0.62
	Asian	52	32	0.62
	Native American	4	3	0.75
Gender	Male	591	423	0.70
	Female	1,142	799	0.70
Age	20s	822	573	0.70
	30s	465	329	0.70
	40s	329	247	0.75
	50s	96	65	0.68
	60s	21	8	0.38
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	60	46	0.72
	$\$20,001$ to $30,000$	293	209	0.77
	$\$30,001$ to $40,000$	1,350	948	0.70
	$\$40,001$ to $50,000$	24	15	0.62
	$\$50,001 \leq$	6	4	0.67
Total		1,733	1,222	0.71

Table 4.11: Retention rates, broken down according to characteristic, of Problem Solving teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

Problem Solving				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	9	7	0.78
	Hispanic	0	0	0.00
	African American	0	0	0.00
	Asian	0	0	0.00
	Native American	0	0	0.00
Gender	Male	4	3	0.75
	Female	5	5	1.00
Age	20s	6	5	0.83
	30s	2	2	1.00
	40s	1	1	1.00
	50s	0	0	0.00
	60s	0	0	0.00
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	0	0	0.00
	$\$20,001$ to $30,000$	0	0	0.00
	$\$30,001$ to $40,000$	9	7	0.78
	$\$40,001$ to $50,000$	0	0	0.00
	$\$50,001 \leq$	0	0	0.00
Total		9	7	0.78

Table 4.12: Retention rates, broken down according to characteristic, of Ap Statistics teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

AP Statistics				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	33	23	0.70
	Hispanic	6	4	0.67
	African American	3	1	0.33
	Asian	2	0	0.00
	Native American	0	0	0.00
Gender	Male	26	15	0.58
	Female	18	13	0.72
Age	20s	24	16	0.67
	30s	11	6	0.55
	40s	8	5	0.63
	50s	1	1	1.00
	60s	0	0	0.00
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	0	0	0.00
	$\$20,001$ to $30,000$	4	3	0.75
	$\$30,001$ to $40,000$	35	20	0.57
	$\$40,001$ to $50,000$	5	5	1.00
	$\$50,001 \leq$	0	0	0.00
Total		44	28	0.67

Table 4.13: Retention rates, broken down according to characteristic, of LDC Mathematics teachers with 0–4 years of experience during the 2003–2004 school year, and who stayed through the 2006–2007 school year.

LDC Mathematics – Grade 7 to 12				
Characteristics		**2003–2004**	**2006–2007**	**Retention**
Ethnicity	White	477	325	0.68
	Hispanic	103	78	0.76
	African American	40	32	0.80
	Asian	14	11	0.79
	Native American	2	1	0.50
Gender	Male	238	172	0.72
	Female	398	275	0.69
Age	20s	330	224	0.68
	30s	152	109	0.72
	40s	97	72	0.74
	50s	50	39	0.78
	60s	6	3	0.50
	70s	0	0	0.00
Salary	$\leq \$10,000$	0	0	0.00
	$\$10,001$ to $20,000$	22	15	0.68
	$\$20,001$ to $30,000$	169	124	0.74
	$\$30,001$ to $40,000$	429	296	0.69
	$\$40,001$ to $50,000$	12	10	0.83
	$\$50,001 \leq$	4	2	0.50
Total		636	447	0.70

Table 4.14: Characteristics of hispanic Algebra I teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Algebra I – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	226	174
	Female	224	171
Age	20s	235	183
	30s	146	113
	40s	39	27
	50s	26	19
	60s	4	3
Salary	$\leq \$10,000$	7	4
	$\$10,001$ to $20,000$	17	11
	$\$20,001$ to $30,000$	74	54
	$\$30,001$ to $40,000$	356	276
	$\$40,001$ to $50,000$	4	4
	$\$50,001 \leq$	2	2

4.2.1 Hispanic Math Teachers: Descriptive

I have extracted the characteristics of hispanic mathematics teachers from the data set. The results are presented in Tables 4.14–4.24.

According to the characteristics describing hispanic teachers, the data shows that both hispanic men and women tend to stay in the system by the same proportion. Both, men and women in their 30s and 40s tend to remain longer in the system. For hispanic teachers it was not possible to establish what salary level tend to remain the longest.

Table 4.15: Characteristics of hispanic Algebra II teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Algebra II – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	113	92
	Female	110	88
Age	20s	122	98
	30s	67	54
	40s	15	12
	50s	16	14
	60s	3	2
Salary	$\leq \$10,000$	1	1
	$\$10,001$ to $20,000$	7	7
	$\$20,001$ to $30,000$	46	34
	$\$30,001$ to $40,000$	168	137
	$\$40,001$ to $50,000$	1	1
	$\$50,001 \leq$	1	1

Table 4.16: Characteristics of Calculus AB (AP Calculus) teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Calculus AB (AP Calculus) – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	3	1
	Female	4	4
Age	20s	3	3
	30s	2	2
	40s	2	0
	50s	0	0
	60s	0	0
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	0	0
	$\$20,001$ to $30,000$	2	2
	$\$30,001$ to $40,000$	5	3
	$\$40,001$ to $50,000$	0	0
	$\$50,001 \leq$	0	0

Table 4.17: Characteristics of Geometry teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Geometry – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	152	119
	Female	148	114
Age	20s	153	121
	30s	103	82
	40s	23	16
	50s	15	10
	60s	6	4
Salary	$\leq \$10,000$	9	4
	$\$10,001$ to $20,000$	20	13
	$\$20,001$ to $30,000$	53	38
	$\$30,001$ to $40,000$	229	182
	$\$40,001$ to $50,000$	2	2
	$\$50,001 \leq$	1	1

Table 4.18: Characteristics of Independent Study First Time teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Independent Study First Time – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	14	13
	Female	10	8
Age	20s	14	12
	30s	7	7
	40s	3	2
	50s	0	0
	60s	0	0
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	1	1
	$\$20,001$ to $30,000$	21	18
	$\$30,001$ to $40,000$	2	2
	$\$40,001$ to $50,000$	0	0
	$\$50,001 \leq$	0	0

Table 4.19: Characteristics of Independent Study Second Time teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Independent Study Second Time – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	4	4
	Female	1	1
Age	20s	4	4
	30s	1	1
	40s	0	0
	50s	0	0
	60s	0	0
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	0	0
	$\$20,001$ to $30,000$	0	0
	$\$30,001$ to $40,000$	0	0
	$\$40,001$ to $50,000$	5	5
	$\$50,001 \leq$	0	0

Table 4.20: Characteristics of Mathematical Models with Applications teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Mathematical Models with Applications – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	46	36
	Female	38	30
Age	20s	40	34
	30s	32	25
	40s	6	4
	50s	4	2
	60s	2	1
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	3	3
	$\$20,001$ to $30,000$	20	15
	$\$30,001$ to $40,000$	60	47
	$\$40,001$ to $50,000$	1	1
	$\$50,001 \leq$	0	0

Table 4.21: Characteristics of Mathematics, Grade 7 teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Mathematics, Grade 7 – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	175	132
	Female	214	167
Age	20s	188	148
	30s	126	97
	40s	48	38
	50s	23	14
	60s	4	2
Salary	$\leq \$10,000$	3	1
	$\$10,001$ to $20,000$	12	8
	$\$20,001$ to $30,000$	37	27
	$\$30,001$ to $40,000$	335	260
	$\$40,001$ to $50,000$	4	4
	$\$50,001 \leq$	1	0

Table 4.22: Characteristics of Mathematics, Grade 8 teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

Mathematics, Grade 8 – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	153	125
	Female	175	143
Age	20s	151	124
	30s	108	90
	40s	49	41
	50s	16	11
	60s	4	2
Salary	$\leq \$10,000$	1	0
	$\$10,001$ to $20,000$	14	11
	$\$20,001$ to $30,000$	32	27
	$\$30,001$ to $40,000$	277	226
	$\$40,001$ to $50,000$	3	3
	$\$50,001 \leq$	2	1

Table 4.23: Characteristics of Ap Statistics teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

AP Statistics – Hispanics			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	4	2
	Female	2	2
Age	20s	4	2
	30s	1	1
	40s	1	1
	50s	0	0
	60s	0	0
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	0	0
	$\$20,001$ to $30,000$	0	0
	$\$30,001$ to $40,000$	6	4
	$\$40,001$ to $50,000$	0	0
	$\$50,001 \leq$	0	0

Table 4.24: Characteristics of LDC Mathematics teachers who had from 0 to 4 years of experience during the 2003–2004 school year, and who stayed during the 2006–2007 school year.

LDC Mathematics – Grade 7 to 12 – Hispanic			
Characteristics		**2003–2004**	**2006–2007**
Gender	Male	52	41
	Female	51	37
Age	20s	56	39
	30s	31	27
	40s	12	8
	50s	4	4
	60s	0	0
Salary	$\leq \$10,000$	0	0
	$\$10,001$ to $20,000$	6	6
	$\$20,001$ to $30,000$	16	12
	$\$30,001$ to $40,000$	80	59
	$\$40,001$ to $50,000$	0	0
	$\$50,001 \leq$	1	1

4.2.2 Discussion

A report from Henke, Zahn, and Carroll (2001) shows a tendency in all occupations for employees to leave their jobs within their first 5 years on the job. Henke et al. (2001) also presents data that shows that, among all professions, teaching actually presents lower resignation rates than other professions for employees in the first five years on the job.

According to the data used in this investigation, $3,042$ mathematics teachers out of a total of $10,095$ left the system during their first four years in the job during the 2003–2004 school year. This indicates a retention rate of 0.70. Of the $10,095$ teachers in the system that year, $1,920$ were hispanic. Of these, 416 left, indicating a 0.78 retention rate. Among the minority teachers in the educational system, Native American mathematics teachers had the higher retention rate, 0.80, followed by hispanic teachers. African American teachers presented the lowest retention rate among all ethnicities, 0.5. Whites presented a 0.76 retention rate and Asians presented a 0.63 retention rate.

An investigation from Zumwalt and Craig (2005), using data from the National Center for Educational Statistics (NCES), showed that during the years 1999 and 2000, hispanic teachers were among the teachers that tended to stay longer in the educational system. This is consistent with the results found in this investigation. Zumwalt and Craig (2005)'s investigation also mentions that the number of hispanic teachers is increasing every year, which is also consistent with the results found in the present investigation. At the same time, Zumwalt and Craig (2005) explains that minority teachers, in gen-

eral, expressed that their motivation to become teachers started with their own experiences as students. They felt that their own education lacked quality and they felt obligated to go back and bring better opportunities to others. Contrary to Zumwalt and Craig (2005)'s results, indicating that African American teachers tend to stay longer as teachers, the data in this investigation suggests that in Texas, African American teachers tend to leave the system earlier than other ethnicities.

Zumwalt and Craig (2005) did not find differences between gender. For hispanic teachers in particular, the data presented here did not show substantial differences in gender either; however, looking at all the mathematics courses, there is a slight tendency to higher retention rates among female teachers than male teachers in Texas.

Consistent with Zumwalt and Craig (2005)'s investigation, teachers in their 50s, in Texas, tend to have a higher retention rate, and teachers in their 30s have the lower retention rate.

4.3 Inferential Analysis

The present section discusses the data generated by the application of the k–Nearest Neighbor algorithm to the data base. The ultimate goal of this portion of the investigation is to establish which values of the weights in the metric

$$d(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{k=1}^{n} c_k (v_k - w_k)^2},$$

yield the best performance of the algorithm. That is: which values c_k yield the best predictive ability of the algorithm? This predictive ability is measured in terms of the classification error rate as measured on a data base of test points, points whose classification is already known. Thus, I seek that collection of weights $\{c_1, c_2, c_3, c_4\}$ which minimizes the error rate. Once I find this combination of weights, I may then *infer* that the characteristic whose corresponding weight has a value higher than the others is the characteristic which has the most predictive ability for classifying new data. Phrased differently, that characteristic is more important or more relevant than the others in establishing the longevity of a given teacher in the system.

In this section, I discuss the data which bears on establishing the best–performing values of the weights, and which therefore establishes the relative ranking of characteristics in terms of predictive ability.

4.3.1 Math Teacher Characteristics: Static

Because computational resources for the current investigation were limited, proper reading of the data output by the program requires an understanding of the method in which it was generated. In the most general case, the algorithm sought to find the optimal values of four weights $\{c_1, c_2, c_3, c_4\}$, corresponding to *gender*, *ethnicity*, *current age*, and *salary*. The procedural goal would be to set specific values for each of the weights,

$$c_1 = \tilde{c}_1, \qquad c_2 = \tilde{c}_2, \qquad c_3 = \tilde{c}_3, \qquad c_4 = \tilde{c}_4,$$

subject to the constraint

$$\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3 + \tilde{c}_4 = 1,$$

then calculate the error rate for this particular combination of weights. The procedure would then assign new values

$$c_1 = \bar{c}_1, \qquad c_2 = \bar{c}_2, \qquad c_3 = \bar{c}_3, \qquad c_4 = \bar{c}_4,$$

subject to the same constraint, and calculate a new error rate. The computer would continue in this way, exhausting all possible combinations of weight values satisfying the constraint, and computing a new error rate each time. I would then look at the error rates, find the minimum, and say that the particular combination of weights corresponding to this minimum error rate is the desired combination.

Naturally there is an infinity of combinations of weights which satisfy the constraint, so that this procedure is not practical even provided ideal computational resources. But due to the severely limited resources available, I have followed a particularly restricted procedure. Specifically, the procedure is as follows:

1. The computer creates a list of all *distinct pairs* of weights (c_i, c_j).

    - The pairs (c_i, c_j) and (c_j, c_i) are not considered distinct (they will contain the same information), and so the computer only chooses one of these. This choice is made randomly (this simplifies the programming).

2. The computer assigns a value of 0.25 to each of the two *other* weights, and leaves them fixed.

    - Because of the overall constraint that all weights must sum to 1, this means we must have $c_i + c_j = 0.5$.

3. The computer initializes c_i with a value of 0, hence c_j receives a value of 0.5.

4. The computer calculates an error rate with this combination of values.

5. The computer then increments c_i by 0.125, necessarily decreasing c_j by the same amount to maintain the constraint $c_i + c_j = 0.5$.

6. The computer calculates a new error rate.

7. The computer repeats the procedure of incrementing c_i and calculating a new error rate until c_i reaches a maximum value of 0.5 (and c_j a minimum of 0).

8. The computer repeats the entire procedure for the next pair of weights (c_m, c_n) until all pairs are exhausted.

This procedure, while hardly optimal or even exhaustive, provides a small search of the space of possible values of the weights which nevertheless lends some insight into relative importance and at the same time performs well with the available computational resources. Given more robust computational resources, this procedure could be improved dramatically.

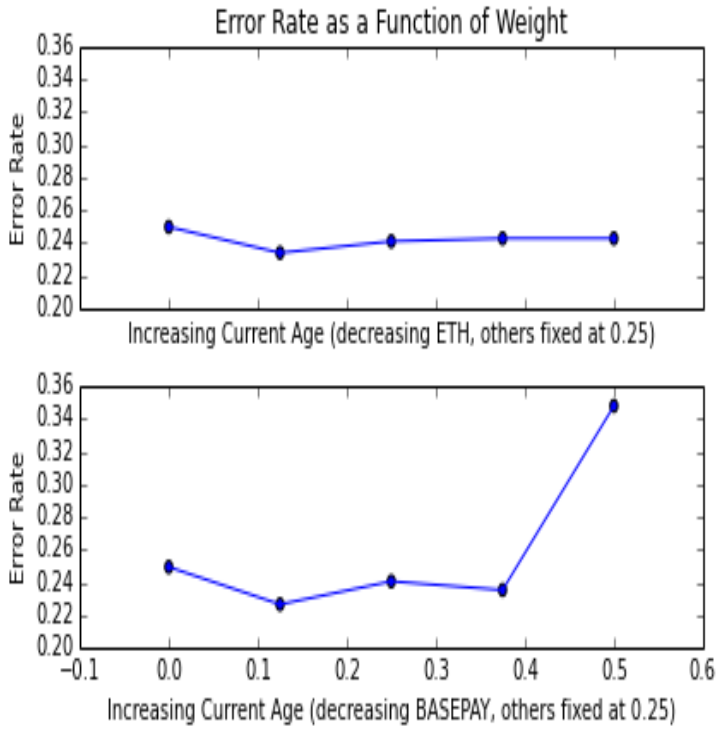
Given the above procedure, the output produced comes in the form of weight combinations together with the error rate calculated for each combination. But because of the particular procedure employed, two weights are always held fixed at 0.25 each, one weight c_i is incremented, and the final weight is always known implicitly from the relation

$$c_j = 0.5 - c_i.$$

Thus, I may graph the error rates as functions of the changing value of the weight c_i, as long as I stipulate which weight c_j will form its conjugate (i.e. will decrease in lock-step with c_i's increase). Figures 4.1–4.3 present the error rate for running k–Nearest Neighbors one time through the above procedure using the 2003–2004 data.

Let us focus on the first of the graphs in Figure 4.1. For each point in this graph, the weight assigned to Base Pay and Gender are held constant. Moving from left to right, the weight assigned to Current Age increases, while that assigned to Ethnicity decreases so as to keep their sum constant. One can see from the left–most point that removing Current Age from consideration (setting its weight to zero, and thus setting the weight of Ethnicity to 0.5) leads to slightly worse performance. In the remainder of the graph, one can see that the error rate remains roughly constant as Current Age is weighted more heavily, Ethnicity less heavily. This suggests that there is little reason to weight one of these characteristics more heavily than the other, though perhaps this is a slight improvement in performance if Current Age receives

Figure 4.1: 2003–2004. Error Rates for Current Age vs. Ethnicity and for Current Age vs. Base Pay



the lower weight of 0.125. In terms of the actual numerical output of the program, Table 4.25 lists the data plotted in the first graph in Figure 4.1.

The graph and accompanying table show that the classification procedure performs slightly better (has a lower error rate) when the weight of Current Age falls somewhere between 0.125 and 0.25. When the weight of Current Age takes on the value 0.125, the weight for Ethnicity takes on the corresponding value $1 - 0.125 = 0.375$; and so on. Other weights stay fixed

Table 4.25: Tabular display of data represented in visual form by the upper graph of Figure 4.1.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.25	0.0	0.5	0.249561
0.25	0.25	0.125	0.375	0.233743
0.25	0.25	0.25	0.25	0.240773
0.25	0.25	0.375	0.125	0.242531
0.25	0.25	0.5	0.0	0.242531

at values of 0.25. The change in the error rate over the full range of values employed for Current Age is, compared to other graphs, rather modest. From this graph what can be concluded with relative certainty is, given that Base Pay and Gender weights are set at 0.25 each, *excluding* Current Age as a factor (i.e. setting its weight to zero) in favor of including Ethnicity on a higher footing (i.e. setting its weight to 0.5) leads to slightly *worse* performance of the algorithm, i.e. to a modestly higher error rate. Thus Current Age should be incorporated as a determining characteristic, in comparison with Ethnicity.

Now let us move on to the lower graph in Figure 4.1. Here again the weight of Current Age increases from left to right, but since it is paired with the weight for Base Pay, it is the Base Pay weight that decreases over the same range, while the weights for Gender and now Ethnicity stay fixed at 0.25. The numeric data represented by the graph is listed for convenience in Table 4.26.

We see here that, once again, the algorithm seems to perform better for the Current Age weight in the range $[0.125, 0.375]$. Again, the performance is

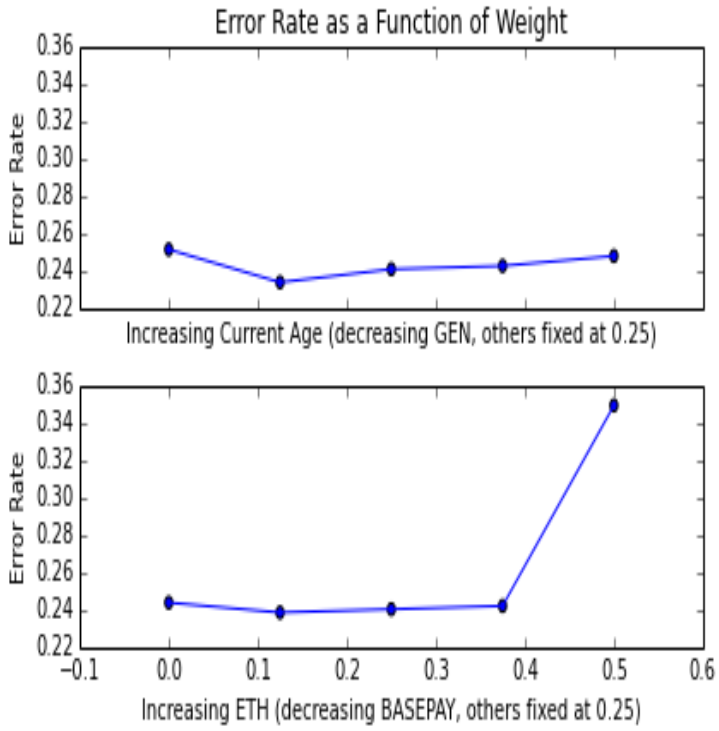
Table 4.26: Tabular display of data represented in visual form by the lower graph of Figure 4.1.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.5	0.25	0.0	0.25	0.249561
0.375	0.25	0.125	0.25	0.226714
0.25	0.25	0.25	0.25	0.240773
0.125	0.25	0.375	0.25	0.235501
0.0	0.25	0.5	0.25	0.347979

worse (i.e. leads to a higher error rate) at the extremes: poorer performance when Current Age is removed (has a weight of 0) in favor of Base Pay, all other weights being equal; and noticeably terrible performance, relatively speaking, when Base Pay is eliminated in favor of Current Age, all other weights being equal. Thus, I should eliminate neither Current Age nor Base Pay in favor of the other, and the error rate shows lower values for weights of these two characteristics in a middle range. Moreover, performance is slightly better for the lower weighting of Current Age relative to Base Pay, though it is hard to make too much of this, since the next lowest error rate corresponds to a higher weighting of Base Pay relative to Current Age.

We may consider the data outlined in Figure 4.2 using the same procedure. In the top graph, we find that, for constant Base Pay and Ethnicity weights, as the weight for Current Age increases and the weight for Gender decreases, the error rate increases after an initial dip. This initial dip suggests that Current Age should not be removed from consideration, since this leads

Figure 4.2: 2003–2004. Error Rates for Current Age vs. Gender and for Ethnicity vs. Base Pay



to a higher error rate. The next highest error rate in the graph, however, occurs when Gender is removed from consideration. Thus Gender too should remain as a factor in the distance calculation. The intermediate points of the graph show similar performance, except for the the point where $c_{\text{age}} = 0.125$, where the error rate shows a minimum in the graph. Thus, if this graph shows a preference, it might be towards a stronger weighting of Gender relative to Current Age. But again the difference in error rates is slight, and the safest

Table 4.27: Tabular display of data represented in visual form by the upper graph of Figure 4.2.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.5	0.0	0.25	0.251318
0.25	0.375	0.125	0.25	0.233743
0.25	0.25	0.25	0.25	0.240773
0.25	0.125	0.375	0.25	0.242531
0.25	0.0	0.5	0.25	0.247803

interpretation is that there is little preference for one over the other.

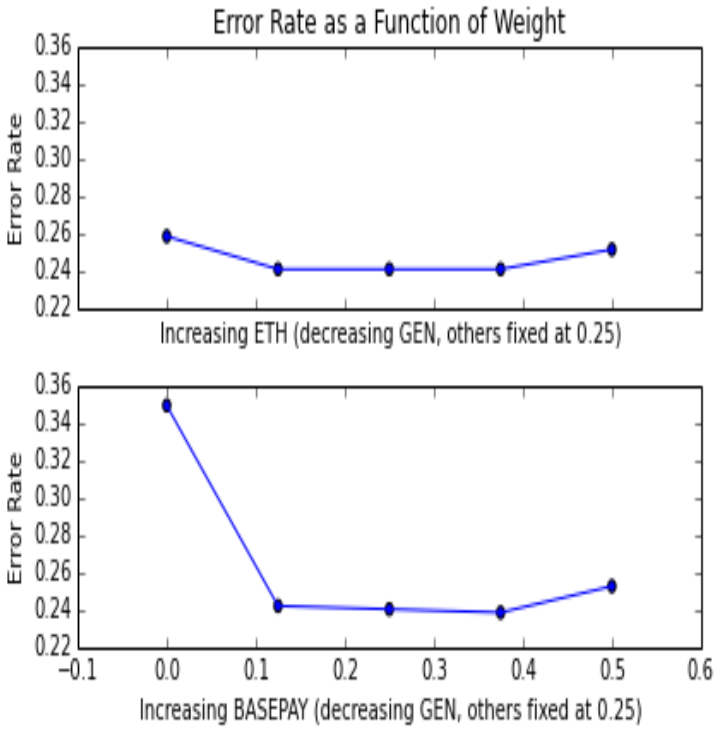
Table 4.28: Tabular display of data represented in visual form by the lower graph of Figure 4.2.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.5	0.25	0.25	0.0	0.244288
0.375	0.25	0.25	0.125	0.239016
0.25	0.25	0.25	0.25	0.240773
0.125	0.25	0.25	0.375	0.242531
0.0	0.25	0.25	0.5	0.349736

The second graph in Figure 4.2 provides little clear information in terms of what values of the Ethnicity weight optimize the performance of the algorithm, compared with the weight for Base Pay. For most weight values, the algorithm functions almost equally well regardless of whether more weight is given to Ethnicity or to Base Pay. But it does highlight that the classification algorithm shows distinctly worse performance when the weight for Ethnicity is increased to 0.5 and the weight for Base Pay set to 0. Thus I can conclude that

it is definitely less optimal to exclude Base Pay as a characteristic included in the distance metric if this is done in order to amplify the contribution of Ethnicity.

Figure 4.3: 2003–2004. Error Rates for Ethnicity vs. Gender and for Base Pay vs. Gender



The upper graph of Figure 4.3 does not show any clear tendency, but it suggests that performance is similar regardless of whether Ethnicity or Gender receives the higher weight. It does however make clear that neither Ethnicity nor Gender should be removed in favor of the other. The lower graph of

Table 4.29: Tabular display of data represented in visual form by the upper graph of Figure 4.3.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.5	0.25	0.0	0.258348
0.25	0.375	0.25	0.125	0.240773
0.25	0.25	0.25	0.25	0.240773
0.25	0.125	0.25	0.375	0.240773
0.25	0.0	0.25	0.5	0.251318

Table 4.30: Tabular display of data represented in visual form by the lower graph of Figure 4.3.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.0	0.5	0.25	0.25	0.349736
0.125	0.375	0.25	0.25	0.242531
0.25	0.25	0.25	0.25	0.240773
0.375	0.125	0.25	0.25	0.239016
0.5	0.0	0.25	0.25	0.253076

Figure 4.3 similarly shows little preference between weights, this time Base Pay and Gender. However it makes clear that Base Pay should not be eliminated in favor of Gender, since this leads to dramatically poorer performance.

4.3.2 Summary of Implications

We may now take a moment to understand the inferential evidence supplied by the output of the computational procedure applied. We are left with a series of binary comparisons. The following list summarizes the conclusions

outlined above.

- Incorporate Current Age, when compared to Ethnicity.
- $c_{\text{age}} \sim c_{\text{ethnicity}}$.
- Eliminate neither Base Pay nor Current Age in favor of the other.
- $c_{\text{age}} \sim c_{\text{basepay}}$.
- Do not eliminate Gender in favor of Current Age.
- $c_{\text{age}} \sim c_{\text{gender}}$.
- Do not eliminate Base Pay in favor of Ethnicity.
- $c_{\text{base pay}} \sim c_{\text{ethnicity}}$.
- Remove neither Ethnicity nor Gender in favor of the other.
- $c_{\text{ethnicity}} \gtrsim c_{\text{gender}}$.
- Do not eliminate Base Pay in favor of Gender.

From this list it is safe to say that there is no clear evidence that any of the four characteristics under consideration should be eliminated. They all contain some predictive value. Moreover, it is hard to establish any clear evidence that one particular characteristic should be ranked higher than the others:

$$c_{\text{age}} \sim c_{\text{ethnicity}} \sim c_{\text{base pay}} \sim c_{\text{gender}}.$$

Table 4.31: 2003–2004. Weights for smallest error rates achieved in Figures 4.1–4.3.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.25	0.125	0.375	0.233743
0.375	0.25	0.125	0.25	0.226714
0.125	0.25	0.375	0.25	0.235501
0.25	0.375	0.125	0.25	0.233743
0.375	0.25	0.25	0.125	0.239016
0.375	0.125	0.25	0.25	0.239016

Nevertheless one may still estimate which particular set of values for the weights provides the best performance by looking at the numerical data. Table 4.31 provides a list of the lowest error rates achieved in the graphs, along with the corresponding values of the weights. The absolute minimum is 0.226714, achieved for the values

$$c_{\text{basepay}} = 0.375, \qquad c_{\text{gender}} = 0.25, \qquad c_{\text{age}} = 0.125, \qquad c_{\text{ethnicity}} = 0.25.$$

This suggests that Base Pay should receive the highest weight, Current Age the lowest, and Gender and Ethnicity should receive equal intermediate weights. This point is depicted in the lower graph of Figure 4.1. This would suggest a ranking

$$c_{\text{base pay}} \gtrsim c_{\text{ethnicity}} \sim c_{\text{gender}} \sim c_{\text{age}}.$$

4.3.3 Discussion

How can we understand the relative ranking suggested by the preceding discussion?

One might have expected Current Age to show a noticeably higher importance for two main reasons. In the first instance, one may imagine that the older a teacher, the greater the possibility that the teacher has been in the system for a longer time. Since in our test data base a teacher's term of service was taken as a proxy for how long the teacher would eventually serve, it stands to reason that the procedure would select age as the more useful parameter with which to influence the distance calculation. The fact that this did not happen is somewhat heartening: it suggests that use of the term of service as a proxy does not automatically bias the results. The proxy serves its purpose adequately.

On the other hand, one may also suppose that, for general economic reasons, teachers in the first few years of their employment are the most likely to decide they made a poor decision in becoming teachers, and may soon opt for other employment. Thus younger teachers might be more likely to have shorter terms of service. For that reason one might expect a higher ranking for Current Age, yet this is not borne out by the data.

Though the characteristic of Current Age has a noticeably direct connection with the length of a teacher's term of service, with the other characteristics the connection is less direct, and for this reason the current investigation contains some merit. Regardless of *how* the connection is made, the above analysis shows that the procedure employed does have merit: the program establishes that all four of the characteristics under consideration have an impact on predicting the longevity of a teacher in the system. Moreover, looking

at the overall collection of error rates displayed in the graphs, the particular program I have developed correctly predicts the general classification of this longevity (long–, medium–, or short–term) in roughly three out of four teachers.

In order to ensure that the performance and output of the program, and hence the conclusions drawn, do not depend on a particular set of training or test data, I have re-run the program several times. Each time the program runs, the procedure splits the data base into training and test data at random; thus each run uses a different selection of data for the training data base and a different set of data for the test points. The results of some additional runs have been included in appendix B. Here I find something of note: though the point in parameter space given by

$$c_{\text{basepay}} = 0.375, \qquad c_{\text{gender}} = 0.25, \qquad c_{\text{age}} = 0.125, \qquad c_{\text{ethnicity}} = 0.25,$$

does in fact frequently correspond to a minimum on the graph of Current Age vs. Base Pay, it does not always yield a global minimum among the error rates. Thus it is difficult to support the assertion that Base Pay be ranked highest and that Current Age be ranked lowest. Failing to find clear evidence to the contrary, it may be best to simply stick to an equal ranking:

$$c_{\text{age}} \sim c_{\text{ethnicity}} \sim c_{\text{base pay}} \sim c_{\text{gender}}.$$

In addition, the results outlined above show some general correspondence with other investigations found in the literature. For example, Kersaint,

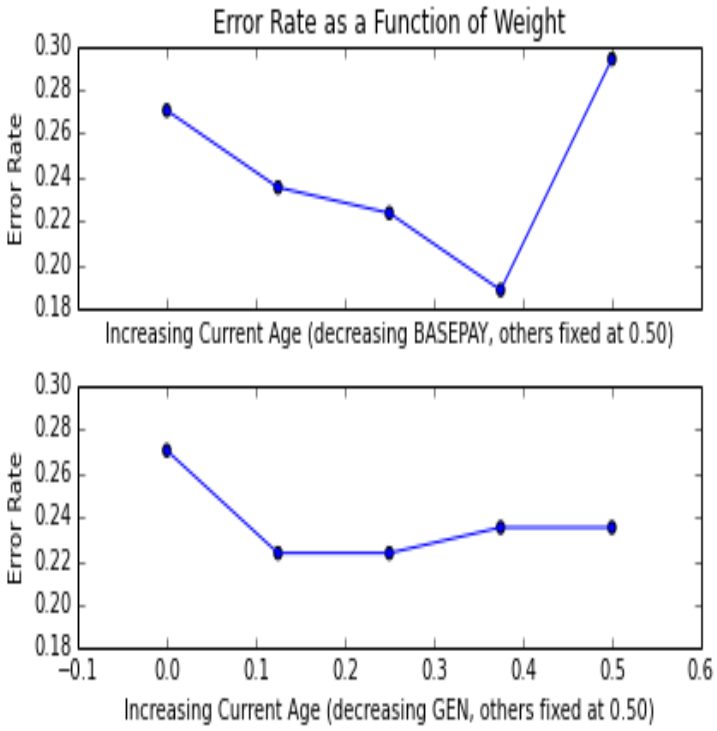
Lewis, Potter, and Meisels (2005) has shown that among the different ethnicities, hispanic teachers tend to be the least influenced by salaries when deciding whether to leave the educational system. This suggests that Ethnicity and Base Pay are both individually important variables: excluding Base Pay from consideration would miss a driving factor in the retention of non-hispanic teachers; while excluding Ethnicity from consideration would imply that Base Pay should be considered equally important for teachers of all ethnicities, which research shows not to be the case. Thus our finding that Ethnicity and Base Pay should have roughly equal weight falls in line with results gleaned from other investigations.

4.3.4 Hispanic Math Teachers: Static

Figures 4.4 and 4.5 describe the results of running the k–Nearest Neighbors algorithm on data including only hispanic mathematics teachers. Since the ethnicity is the same for all teachers in the data, I exclude this characteristic from consideration. This consequently reduces the number of pairs of weights under consideration, and hence the number of graphs produced.

The upper graph of Figure 4.4, whose numeric data is presented in Table 4.32, shows a general tendency whereby increasing the Current Age weight over the Base Pay weight reduces the error rate, denoting better performance. Specifically, when the weight for Current Age is 0.375, and the weight for Base Pay is 0.125 the error rate is at its lowest (holding the weight for Gender at 0.5). This indicates that for hispanic mathematics teachers Current Age might

Figure 4.4: 2003–2004 Hispanic Teachers. Error Rates for Current Age vs. Base Pay and Current Age vs. Gender



be a better indicator for labeling teachers than Base Pay. However completely eliminating Base Pay leads to dramatically worse performance, so we can only say that Current Age should be weighted higher than Base Pay. The same graph shows that when Current Age is eliminated, the program's performance is similarly poor.

The lower graph of Figure 4.4, whose numeric data is presented in Table 4.33, suggests that Current Age should not be eliminated in favor of

Table 4.32: Tabular display of data represented in visual form by the upper graph of Figure 4.4.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.5	0.5	0.0	0.270588
0.375	0.5	0.125	0.235294
0.25	0.5	0.25	0.223529
0.125	0.5	0.375	0.188235
0.0	0.5	0.5	0.294118

Gender. Moreover, weighting Current Age lower than Gender appears to yield slightly better performance of the algorithm.

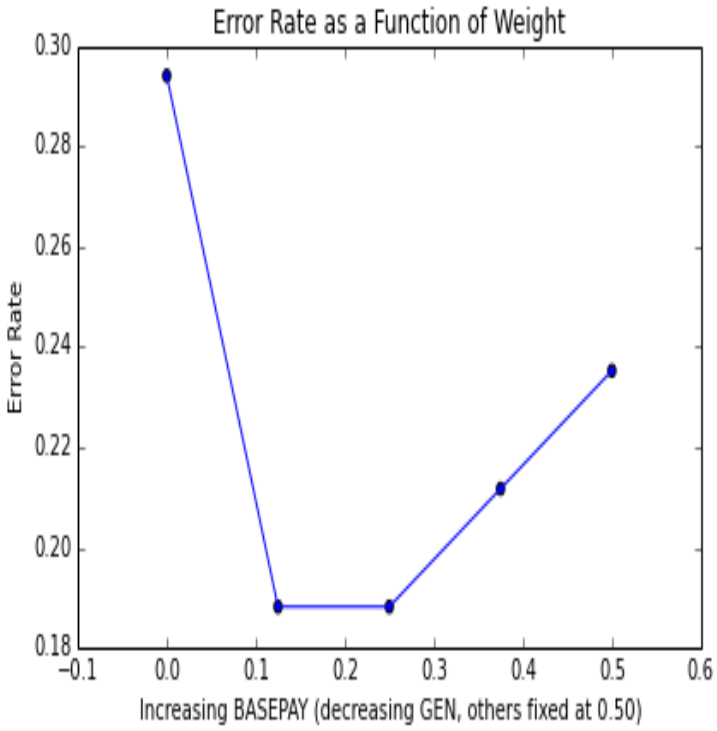
Table 4.33: Tabular display of data represented in visual form by the lower graph of Figure 4.4.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.5	0.5	0.0	0.270588
0.5	0.375	0.125	0.223529
0.5	0.25	0.25	0.223529
0.5	0.125	0.375	0.235294
0.5	0.0	0.5	0.235294

Finally, Figure 4.5, whose numeric data is presented in Table 4.34, shows that Base Pay should not be eliminated, however it should receive a lower weight than Gender. Any weight of Base Pay above 0.25 leads to noticeably poorer performance.

From the above analysis I can propose a ranking of characteristics.

Figure 4.5: 2003–2004 Hispanic Teachers. Error Rates for Gender vs. Base Pay



Based on the weights I might suggest that

$$c_{\text{gender}} \gtrsim c_{\text{age}} \gtrsim c_{\text{base pay}}.$$

conforms to the graphical data listed above. However we may also look directly at the data outlined in the tables. We find that the same minimum error rate occurs for three distinct combinations of weights, shown in Table 4.35. The first and last rows of this table show a collection of weights conforming to the above ranking. The middle row, however, achieves the same minimum in a

Table 4.34: Tabular display of data represented in visual form by Figure 4.5.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.0	0.5	0.5	0.294118
0.125	0.375	0.5	0.188235
0.25	0.25	0.5	0.188235
0.375	0.125	0.5	0.211765
0.5	0.0	0.5	0.235294

case where $c_{\text{age}} \gtrsim c_{\text{gender}}$. We note that, in each of these cases, the algorithm is able to predict the correct longevity for 4 out of 5 hispanic mathematics teachers.

Table 4.35: Weights for minima achieved in Figures 4.4–4.5.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.125	0.5	0.375	0.188235
0.125	0.375	0.5	0.188235
0.25	0.25	0.5	0.188235

As with the preceding discussion, the same procedure was run a total of three times on the data, employing a different randomized split of the data base into training data and test data each time. The results have been included in appendices C and D.

4.4 Math Teacher Characteristics: Variation

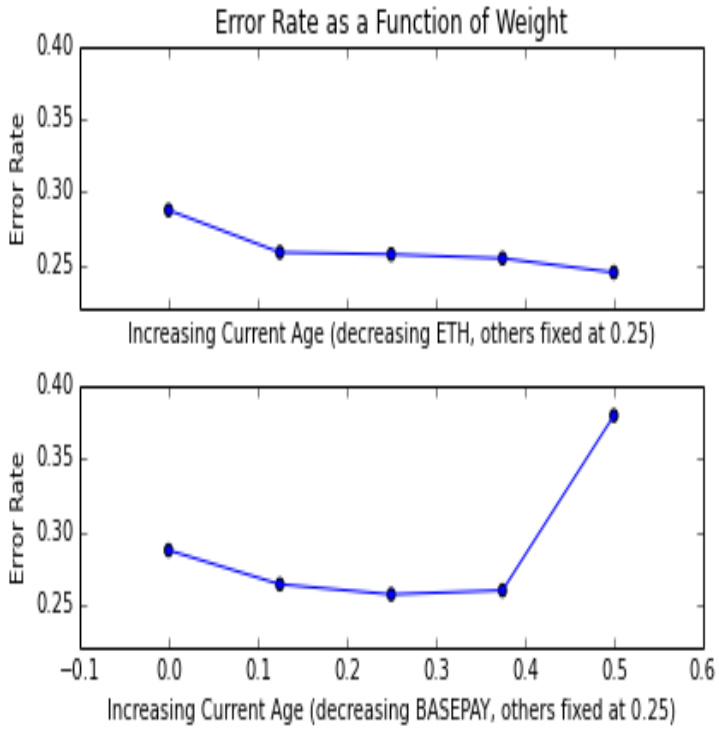
In the preceding section I have seen how the k–Nearest Neighbors algorithm can be applied to the teacher data base to gain insight into which characteristics most impact any determination of a current or future teacher's longevity in the educational system. A natural question which arises is, does the importance of these characteristics change over time?

In order to address such a question, a natural first point of departure is a comparison of one year's data with that of another year. In this section, I discuss such an analysis. In particular, I analyze the data from a second year, in this case chosen to be the 2006–2007 school year, according to the same procedure to determine what kind of ranking of characteristics the data implies for that year. We may then compare this ranking with that determined above to see if there is in fact any change between the two years.

The results of applying the k–Nearest Neighbors algorithm to the mathematics teacher data for the 2006-2007 school year are presented in Figures 4.6–4.8.

Let us consider the upper graph of Figure 4.6, with the accompanying numerical data in Table 4.36. As with the analysis of the data from the previous year, I see an overall trend in which increasing the weight of Current Age relative to Ethnicity yields better performance. What does differ from the previous analysis, however, is that I see no worsening of performance if I eliminate Ethnicity completely (in this case, setting $c_{\text{age}} = 0.5$ and hence

Figure 4.6: 2006–2007. Error Rates for Current Age vs. Ethnicity and for Current Age vs. Base Pay



$c_{\text{ethnicity}} = 0$). However further review of subsequent runs, each with a different random splitting of the data into training and test sets, does not bear out this conclusion. In some runs of the program, I find that removal of Ethnicity leads to distinctly worse performance.

From the lower graph of Figure 4.6, with the accompanying data in Table 4.37, I again see that elimination of either Current Age or Base Pay leads to poorer performance. Moreover, I see a slight tendency for higher weight

Table 4.36: Tabular display of data represented in visual form by the upper graph of Figure 4.6.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.25	0.0	0.5	0.287293
0.25	0.25	0.125	0.375	0.258287
0.25	0.25	0.25	0.25	0.256906
0.25	0.25	0.375	0.125	0.254144
0.25	0.25	0.5	0.0	0.244475

Table 4.37: Tabular display of data represented in visual form by the lower graph of Figure 4.6.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.5	0.25	0.0	0.25	0.287293
0.375	0.25	0.125	0.25	0.263812
0.25	0.25	0.25	0.25	0.256906
0.125	0.25	0.375	0.25	0.259669
0.0	0.25	0.5	0.25	0.379834

given to Current Age leading to marginally better performance, though this preference is so slight it might be better to say that the two might be weighted evenly.

In the upper graph of Figure 4.7, together with the numerical data in Table 4.38, I see that Current Age seems to lie on rather equal footing with Gender. Again eliminating one or the other of the two leads to markedly worse performance of the algorithm.

In the lower graph of Figure 4.7, together with the numerical data in

Figure 4.7: 2006–2007. Error Rates for Current Age vs. Gender and Ethnicity vs. Base Pay

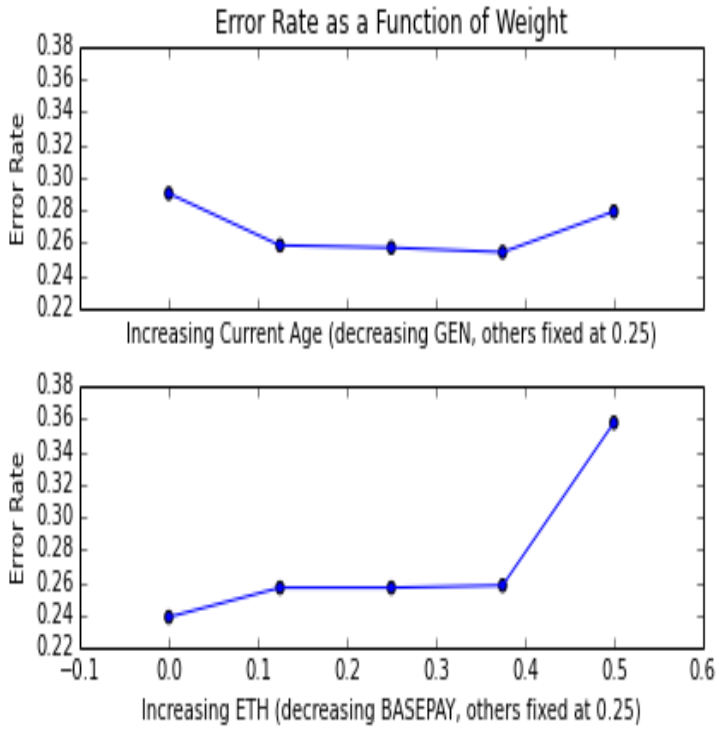


Table 4.39, I again see some improvement in performance upon eliminating Ethnicity. But this does not stand the test of repeated trials. And I see that elimination of Base Pay again leads to worse performance. In the middle range, I see no particular preference for Ethnicity or Base Pay.

Finally, Figure 4.8, together with the accompanying numerical data in Tables 4.40–4.41, show a similar story. Again I find the false friend of improved performance with Ethnicity eliminated. And I see that removing either Gender

Table 4.38: Tabular display of data represented in visual form by the upper graph of Figure 4.7.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.5	0.0	0.25	0.290055
0.25	0.375	0.125	0.25	0.258287
0.25	0.25	0.25	0.25	0.256906
0.25	0.125	0.375	0.25	0.254144
0.25	0.0	0.5	0.25	0.279006

Table 4.39: Tabular display of data represented in visual form by the lower graph of Figure 4.7.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.5	0.25	0.25	0.0	0.238950
0.375	0.25	0.25	0.125	0.256906
0.25	0.25	0.25	0.25	0.256906
0.125	0.25	0.25	0.375	0.258287
0.0	0.25	0.25	0.5	0.357735

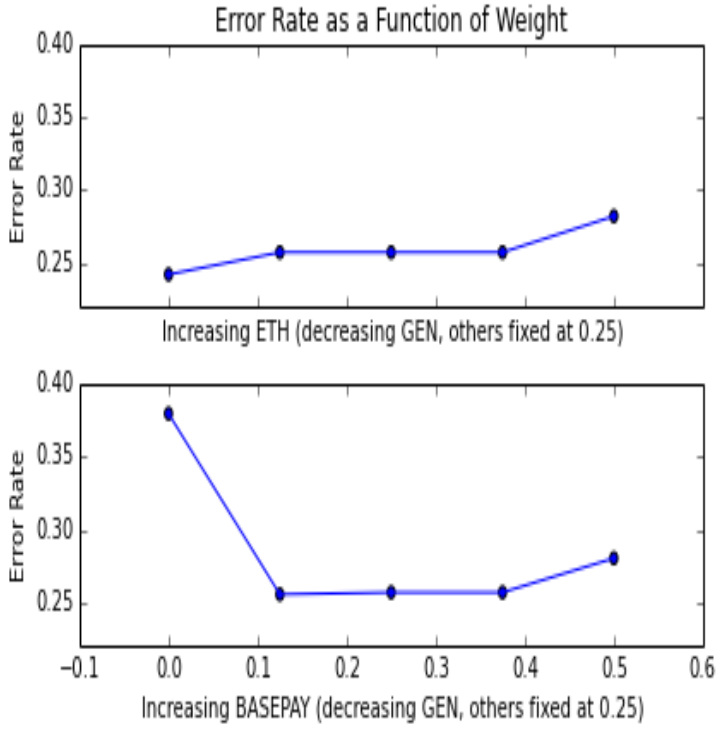
or Base Pay from consideration leads to worse performance.

4.4.1 Summary of Implications

In this section I provide a summary of the inferences drawing from the data outlined in above. In particular, I find the following:

- $c_{\text{age}} \gtrsim c_{\text{ethnicity}}$.
- No reason to eliminate either Current Age or Ethnicity.

Figure 4.8: 2006–2007. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender



- No reason to eliminate Base Pay.
- $c_{\text{age}} \sim c_{\text{base pay}}$.
- $c_{\text{age}} \sim c_{\text{gender}}$.
- No reason to eliminate Gender.
- $c_{\text{ethnicity}} \sim c_{\text{base pay}}$.
- $c_{\text{ethnicity}} \sim c_{\text{gender}}$.

Table 4.40: Tabular display of data represented in visual form by the upper graph of Figure 4.8.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.5	0.25	0.0	0.241713
0.25	0.375	0.25	0.125	0.256906
0.25	0.25	0.25	0.25	0.256906
0.25	0.125	0.25	0.375	0.256906
0.25	0.0	0.25	0.5	0.281768

Table 4.41: Tabular display of data represented in visual form by the lower graph of Figure 4.8.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.0	0.5	0.25	0.25	0.379834
0.125	0.375	0.25	0.25	0.255525
0.25	0.25	0.25	0.25	0.256906
0.375	0.125	0.25	0.25	0.256906
0.5	0.0	0.25	0.25	0.280387

- $c_{\text{base pay}} \sim c_{\text{gender}}$.

In general, I again find that no characteristic under investigation should be eliminated. All continue to maintain some predictive value. If there is any change from the results of the analysis of the data from the preceding year, it seems to be that the characteristics seem more plausibly to lie on an equal footing. I may propose

$$c_{\text{age}} \sim c_{\text{ethnicity}} \sim c_{\text{base pay}} \sim c_{\text{gender}}.$$

Table 4.42: 2006–2007. Weights for smallest error rates achieved in Figures 4.6–4.8.

c_{basepay}	c_{gender}	c_{age}	$c_{\text{ethnicity}}$	Error Rate
0.25	0.25	0.375	0.125	0.254144
0.25	0.125	0.375	0.25	0.254144
0.125	0.375	0.25	0.25	0.255525

as a ranking of weights. The only departure I see from this is a slight possibility that Current Age should receive a higher weight. This does not disagree with the data listed in Table 4.42.

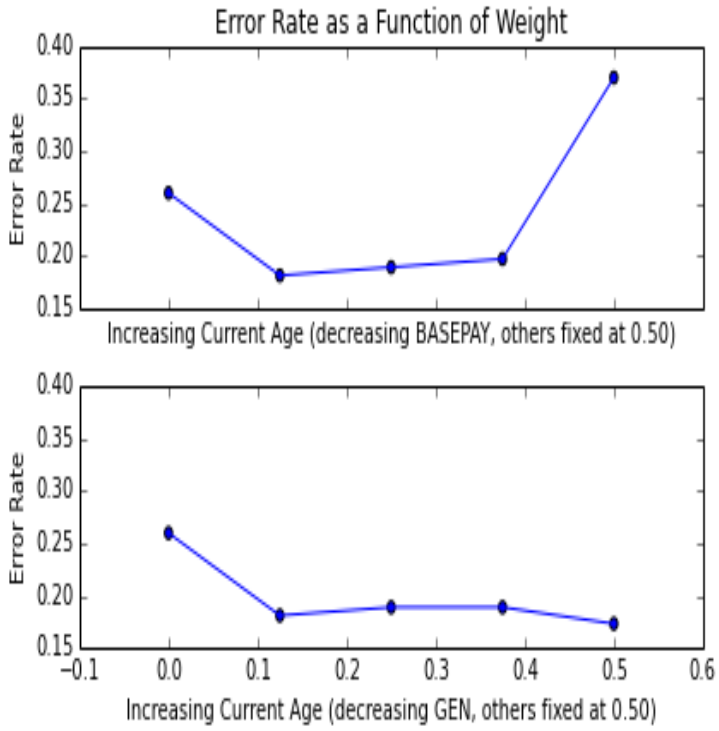
4.4.2 Hispanic Math Teacher: Variation

I may now repeat the analysis of hispanic mathematics teachers, this time concentrating on those teachers present in the 2006–2007 school year. The results of running the program on this set of data are presented in Figures 4.9–4.9.

Table 4.43: Tabular display of data represented in visual form by the upper graph of Figure 4.9.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.5	0.5	0.0	0.259843
0.375	0.5	0.125	0.181102
0.25	0.5	0.25	0.188976
0.125	0.5	0.375	0.196850
0.0	0.5	0.5	0.370079

Figure 4.9: 2006–2007 Hispanic Teachers. Error Rates for Current Age vs. Base Pay and Current Age vs. Gender



From the upper graph of Figure 4.9, together with the accompanying numerical data in Table 4.43, I see that elimination of either Current Age or Base Pay worsens performance of the program. If I may note any tendency from the intermediate points, it is that increasing weight given to Current Age leads to slightly poorer performance.

From the lower graph of Figure 4.9, together with the accompanying numerical data in Table 4.44, I see again that elimination of Current Age leads

Table 4.44: Tabular display of data represented in visual form by the lower graph of Figure 4.9.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.5	0.5	0.0	0.259843
0.5	0.375	0.125	0.181102
0.5	0.25	0.25	0.188976
0.5	0.125	0.375	0.188976
0.5	0.0	0.5	0.173228

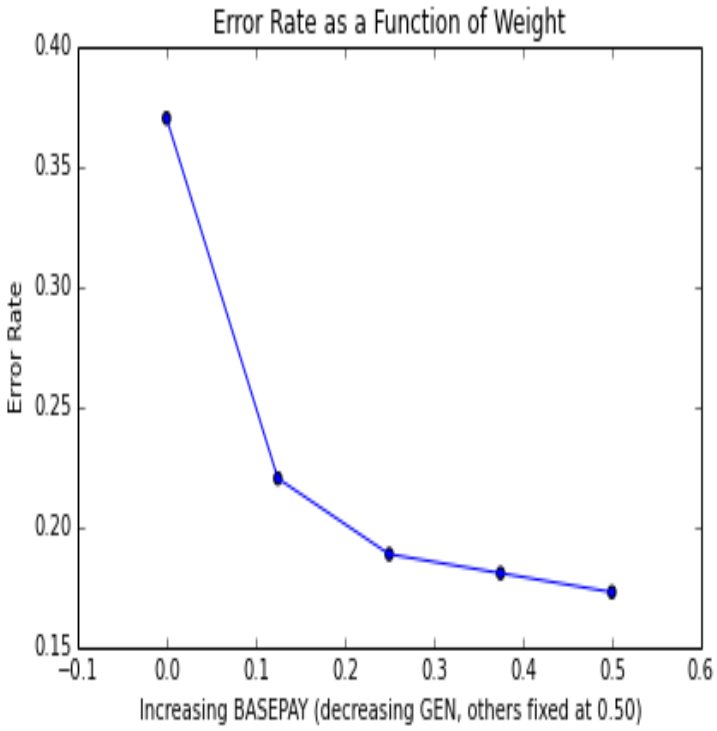
to worsening performance. The shape of the graph also suggests, by contrast, that elimination of Gender might lead to better performance. However I note that the error rate is quite similar to that obtained for $c_{\text{age}} = 0.125$ and $c_{\text{gender}} = 0.375$. This sort of seemingly contradictory behavior suggests that the program is indifferent to the two variables.

Table 4.45: Tabular display of data represented in visual form by Figure 4.10.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.0	0.5	0.5	0.370079
0.125	0.375	0.5	0.220472
0.25	0.25	0.5	0.188976
0.375	0.125	0.5	0.181102
0.5	0.0	0.5	0.173228

The graph in Figure 4.10, accompanied by the numeric data in Table 4.45, shows a distinct tendency toward increasing the weight given to Base Pay. It goes so far as to suggest that Gender could be eliminated altogether

Figure 4.10: 2006–2007 Hispanic Teachers. Error Rates for Gender vs. Base Pay



The data from the various figures is confusing, and it bears taking a look at the minima and near–minima as outlined in Table 4.46. Whereas the absolute minimum is attained when $c_{\text{gender}} = 0$, the next smallest value is attained when $c_{\text{gender}} = 0.5$. This near–smallest value of 0.181102 is attained with $c_{\text{base pay}} = 0.375$, and variously with $c_{\text{gender}} = 0.5$ and $c_{\text{age}} = 0.125$ or vice versa. Such behavior suggests that the program is in fact sensitive in roughly equal measure to the latter two weights. And in the absolute minimum I see

Table 4.46: Weights for smallest values achieved in Figures 4.9–4.10.

c_{basepay}	c_{gender}	c_{age}	Error Rate
0.375	0.5	0.125	0.181102
0.25	0.5	0.25	0.188976
0.5	0.25	0.25	0.188976
0.5	0.125	0.375	0.188976
0.5	0.0	0.5	0.173228
0.25	0.25	0.5	0.188976
0.375	0.125	0.5	0.181102
0.5	0.0	0.5	0.173228

that $c_{\text{base pay}} = 0.5$. Thus, perhaps the safest evaluation of the evidence is to suggest the following ranking

$$c_{\text{gender}} \sim c_{\text{age}} \sim c_{\text{base pay}}.$$

i.e. that each of the three characteristics is roughly of equal predictive value.

Chapter 5

Conclusions

This chapter serves as a brief review of the preceding material with a view toward its understanding and interpretation. I say some words not only concerning what I have accomplished in the production of this work, but also concerning directions for its future development and application.

5.1 Evaluation of Research Questions

Now I will assess what the results in Chapter 4 imply for the research questions of primary interest:

1. What are the characteristics of mathematics teachers that are related to teacher retention in Texas schools?

2. What are the characteristics that most relate to retention of hispanic mathematics teachers in Texas schools?

3. Do these characteristics change over time?

I will treat each one of these in turn in the subsections below.

5.1.1 Characteristics & Retention

Question 1: *What are the characteristics of mathematics teachers that are related to teacher retention in Texas schools?*

Here I address the first of my research questions. In particular, I discuss which of those teacher characteristics under investigation pertain to the determination of the longevity of the teacher in the educational system. The short answer is: *all of them*. Recall from the preceding chapter that I have been studying the impact of the four characteristics *age, gender, ethnicity*, and *base pay* on the determination, via algorithmic means, of the length of time a teacher is likely to stay in the system. In each of the various analyses described above, I found that, overall, when a given characteristic was removed from consideration, the error rate of the algorithm increased. That is, removal of any characteristic led to noticeably *worse* performance of the algorithm. Exclusion of any particular characteristic leads to worse performance of the algorithm (as measured by increased error rate) in almost all cases. Where such does not appear to be the case on a particular run, over successive runs on further randomly selected subset of the data, any putative improvement in performance evaporates. We therefore conclude that Base Pay, Current Age, Gender, and Ethnicity, as exhibited in the data, all play an important role in estimating teacher longevity.

In addition, this is borne out by a look at the descriptive statistics presented for teachers of the various mathematics courses. In particular with

Ethnicity, I see that the particular ethnic background of the teachers leads to implications for the likelihood of retention. I found similar tendencies with Gender and Current Age. With Base Pay, however, the trends were more difficult to discern simply by looking at the tabular data.

It is worth mentioning a notable benefit of the particular computational approach followed in the present investigation. So far, I may characterize the current situation as follows: I studied four characteristics and I found that all four are important. We can imagine that, had I studied five characteristics, I might just as easily have found all five to be important. And so on: I can imagine a situation where most, if not all, relevant information about a teacher can lend itself to bettering estimations of teacher retention. One of the novelties of the approach presented in this investigation is that the program as developed is ready to handle the incorporation of such additional information *without changing a single line of code*. Essentially, the investigator would simply need the names of the new columns added to the data base, and the entire program could be run as–is.

Not only did I identify which factors were important to the determination of teacher retention, but I also made an informed estimation of the relative importance of each of these factors. In the analysis of the data from the 2003–2004 school year, I found that the relative weights of the characteristics appeared to satisfy the following relation:

$$c_{\text{age}} \sim c_{\text{ethnicity}} \sim c_{\text{base pay}} \sim c_{\text{gender}}.$$

Within the data from that school year, there was no clear evidence by which to assert that one factor should be ranked higher than the others. The program produced some results suggesting that Base Pay might be most important among the teacher characteristics, but in successive runs this did not consistently produce the minimum error rate. The lack of any distinct trend as to relative importance may simply be due to the fact that our sample of three error rates for each choice of weights is simply too small for any trend to appear. But the results obtained do appear to fall in line with some findings present in the literature.

The conclusions drawn from the 2006–2007 data show a picture only slightly different, if at all. The most plausible ranking based on the analysis of that data remained

$$c_{\text{age}} \sim c_{\text{ethnicity}} \sim c_{\text{base pay}} \sim c_{\text{gender}}.$$

In this case there was little evidence to suggest that ranking Base Pay higher would substantially improve performance.

I will return to the slight differences in Section 5.1.3 below.

5.1.2 Characteristics & Hispanic Mathematics Teachers

> **Question 2:** *What are the characteristics that most relate to retention of hispanic mathematics teachers in Texas schools?*

In the course of this investigation I have also sought to outline which of the characteristics under investigation most pertain specifically to the determi-

nation of hispanic mathematics teachers' longevity in the system. Again the short answer turns out to be *all of them*. Of course in this case, teacher ethnicity was held constant and therefore removed from determination. But as with the data on the mathematics teachers in general, I saw a notable trend by which removal of a given characteristic led to poorer performance of the program as determined by an increase in error rate.

Again, these findings concur with the descriptive data presented on hispanic mathematics teachers. Each of the characteristics appears to present tendencies relating to retention.

When applied to the data from the school year 2003–2004, the program's output suggested the following ranking of characteristics:

$$c_{\text{gender}} \gtrsim c_{\text{age}} \gtrsim c_{\text{base pay}}.$$

A careful look at the numeric data suggests that the relative importance of Gender and Current Age could be interchanged without worsening importance. But as described in our analysis, when Current Age was ranked higher in importance than Gender, Gender and Base Pay were of equal importance. Thus, though Base Pay seemed overall least important, some data suggests it could be as important as Gender.

By contrast, when applied to the data from the school year 2006–2007, I determined the following ranking:

$$c_{\text{gender}} \sim c_{\text{age}} \sim c_{\text{base pay}}.$$

But in fact the results show some dramatic variation. In one run in particular, I find an absolute minimum in the error rate when Gender is eliminated as a factor altogether. But I find the next smallest error rate when Gender's weight is 0.5, i.e. highest among the three factors under consideration. This suggests that equal weight among the factors is likely the most prudent choice for consistent performance across a range of data bases. This also suggests that we cannot read too much into the lower importance of Base Pay from the 2003–2004 data, since we might not see drastic swings in performance due to the small number of runs.

As a result, if we restrict consideration to hispanic teachers, there might be some tentative evidence for change in the relative importance of factors over time. Thus, whereas the earlier data seemed to suggest that gender and age counted more toward retention than salary, by the time of the later set of data any such difference in impact became harder to discern. But any such assumption of change must be studied under numerous iterations of the procedure to see if the tendency holds up.

Again I defer discussion of the differences in relative rankings between years to Section 5.1.3.

5.1.3 Characteristics & Change

Question 3: *Do these characteristics change over time?*

Finally, I have sought to determine whether the set of characteristics which determines teacher retention changes over time. The short answer here is: *yes and no*. To make this determination, I compared data from the 2003–2004 and 2006–2007 school years.

In the most basic sense, I may say the following with clarity: the collection of factors which influences teacher retention over time *does not change*, at least over the window of time on which I have focused. By this I mean that those factors which are important for the determination of teacher retention in one year are also important for its determination in other years — all of them. In each year, I found that exclusion of a given characteristic would lead to a notable worsening of program performance. By "notable worsening", I mean there was a "large jump" in the graph of the error rate, a change frequently on the order of 100% of the values of the error rate at intermediate values of the characteristic weights.

However what *perhaps* might change is the relative ranking of characteristics between years. That is, whereas a higher Base Pay ranking seemed to have a notably greater impact on program performance for some runs on the data from 2003–2004, its impact on performance did not seem too different from the impact of other characteristics for data from 2006–2007. However, it must be pointed out that a close inspection of the graphs displayed in the preceding analysis shows that, for intermediary values of weights (i.e. values where no characteristic was excluded by virtue of its weight going to zero), variation in error rates remained in a range of roughly 25% of the overall value

of these error rates. In particular this effect is far less pronounced than the jumps which allow us to conclude with relative confidence that no particular characteristic should be excluded from consideration.

Therefore, though one may speculate as to why the difference arises between data from different years, one must be careful to understand the relative imprecision of the ranking to begin with. To properly decide whether or not Base Pay, say, has a larger impact on the classification scheme than other characteristics, one must compare the change in error rate due to its increase to some standard measure of variation. In particular, one must determine in mathematical terms whether or not the increased impact of Base Pay is due to its weight in the distance calculation or simply due to random variation as one randomly selects a new data base. Such a determination would require far more than three iterations of the classification procedure, enough to allow for a statistical measure of the variation. Due to the limitations of computational resources during this study, such large–scale iteration of the procedure was not feasible. This provides a point of departure for further refinements of the study.

In sum there is no clear indication that the relative importance of factors changes from year to year. Admittedly, the current investigation only compares the results based on data from two single years. No sound statistical conclusions can be drawn from this. Several years' worth of data would need to be compared, and one would need to isolate natural variation in the data from true variation due to the putative changing importance of factors.

If we take the lack of change over time at face value, we can potentially use this to improve the algorithm. If there is truly no difference in relative importance of factors from year to year, then the data from any particular year is as good as the data from any other year. We can therefore aggregate *all* data in hand, pooling all teachers from 2000–2012 into one unified data base. We can then split this much larger data base into a training data base and a test data base, and we can see how the program performs with more data in hand.

5.2 The Program

Another product of this investigation has been a program. This program provides an algorithmic way to estimate, given certain characteristics of a teacher (new or already present in the educational system), whether that teacher will likely stay in the system short–, medium–, or long–term. In particular, the program focuses on four teacher characteristics: base pay, age, gender, and ethnicity. The program may include, without modification, other characteristics. And the program may evaluate any data base of teacher characteristics, so long as it is input as a file structured with comma–separated values (CSV). Nothing in the structure of the program ties it to the Texas teacher data to which it has been applied here. It could just as well be applied to data from any other educational system in other regions of the United States, potentially beyond. According to the tests illustrated in the main body of this dissertation, the program correctly estimates the longevity for three out

of four teachers. With optimized combinations of parameter weights and data base, performance occasionally achieves an accuracy of better than four in five.

5.3 Comparison with Traditional Statistical Techniques

This section compares briefly the results of the data mining methodology employed, specifically k–Nearest Neighbors, and the type of results that might be obtained through more traditional statistical methods. Given that there is a plethora of different statistical models available, discussion focuses on the simplest type and makes a few observations on how procedures may or may not differ.

In a typical statistical regression model, such as linear regression, one must first decide which of the various variables available are likely to affect the outcome of the variable one wishes to predict. In this sense, regression differs little from k–Nearest Neighbors: this amounts to deciding which data columns to include in the calculation.

Next a researcher must decide how these variables affect the effect under investigation. For example, in a regression model one must decide if the variables predict the effect in a linear fashion or not. This is a rather non-trivial assumption: relationships between variables can be particularly complicated, and it is not clear that a linear model will encapsulate the nuances of such relations. For example one might suspect that two variables actually interact with one another, and so there should be a nonlinear term involving the product of those two variables. So one finds one in fact wants a regression

model with interaction. Or perhaps several factors work together in combination, within groups, but separately on individual subpopulations within the overall population under investigation. So one should look for a hierarchical regression model, such as HLM.

How is one to decide between these various possibilities? There are two basic elements to the decision:

- A Real–World Model: the researcher must have some *a priori* argument for how the various factors work in the actual environment in which they are found, and then seek the particular statistical technique (if one exists) which best models that particular vision of their interaction;
- Mathematical Consistency: any statistical technique will make certain assumptions on the variables (such as independence, gaussian distributions, etc.) in order to ensure that the conclusions are valid. The researcher must verify that these assumptions are satisfied by the variables under study before the desired technique can be applied cogently.

These criteria are difficult to satisfy in the current study. The dynamics among age, gender, ethnicity, and salary are likely very complicated in the decision of any one teacher to stay or leave the educational system, even more so among all teachers. In particular, the literature has yet to reach consensus on which factors are most important in teacher retention, and so at this stage of investigation it seems sensible to use a methodology which is as flexible as

possible, minimizing the assumptions of how particular variables affect one another.

In some sense the regression model and the k–Nearest Neighbors algorithm parallel one another. At its heart, the regression model draws a line through the data points and moves this line around until it minimizes a specific model of error: the sum total of the squared vertical distances between the data points and the line. The k–Nearest Neighbors algorithm, in a similar fashion, allows the user to tune the weights of the various factors until it minimizes a specific model of error: the error rate in classifying test data. However, whereas the regression model requires that the various factors be independent (that is, uncorrelated) and normally distributed to ensure that the resulting line gives the optimal predictor of the effect under investigation, the k–Nearest Neighbor algorithm has no such requirements. This in fact suits the problem under investigation, since one might expect some correlation between, say, age and salary.

Given the current state of the literature on teacher retention as discussed specifically in Chapter 2 and elsewhere, particularly among mathematics teachers, the results of any statistical method akin to regression would only be noteworthy insofar as the particular real–world model dictating the structure of the regression model applied could be assumed to be credible. Studies of teacher retention show enough variation that any such model would seem premature (see Section 5.4 below). Rather, an exploratory approach seems more apt, one simply trying to identify *which* factors have the greatest im-

pact, without *a priori* stating *how* they manage to have that impact. For these reasons, the current investigation has opted to apply techniques from data mining, starting with one of the most straightforward among these, k–Nearest Neighbors.

5.4 Implications of the Study

As presented in Chapter 2 some studies suggested that the highest rates of teacher turnover are found in the fields of science and mathematics. Chapter 1 has also shown that teacher characteristics influence teachers' decisions to stay or leave the system, and that research on mathematics teacher turnover in Texas has not yet focused on teacher characteristics. My study on Texas teacher characteristics presents a novel contribution to the isolation and study of these crucial factors.

The overriding result of the study is that, in Texas, all characteristics under consideration are in fact equally important for determining the period of time a teacher is likely to stay in the system. This conclusion might seem general, and fairly obvious, but it is not. Some publications have listed low wages as the leading cause for teacher attrition (Buckley, Schneider, & Shang, 2005). Gritz and Theobald (1996) have also suggested that both male and female teachers alike consider income to be the number one reason to remain in the system. There is also research on school conditions and their relation to teacher retention. Hanushek, Kain, and Rivkin (2004) have suggested that better school conditions may be as important as salary for teachers' decision

to stay in the profession. Section 2.2 discusses yet other findings.

Several studies make particular reference to the ethnicity of the teacher, and how it might impact student performance. There is extensive literature in this regard, as described in Chapter 2. Ingersoll and May (2011) have also summarized the research done in this area and highlighted the areas where little attention has been paid: the magnitude, determinants and consequences of minority teacher retention. My investigation has not focused exclusively on minority teachers, but I have been able to identify that, at least in Texas, and for the specific period of time studied in this investigation, there is a small possibility that within the hispanic teachers age and gender might be key components in determining the time a hispanic teacher stays in the system. Although my study is an exploratory study and further analysis is required, adding more years and characteristics, this finding can help researchers identify where to focus their investigations when targeting teacher attrition in the state of Texas.

Ingersoll and Kralik (2004) pointed out that despite the effort put into research for teacher retention there are still many questions that have not been answered, in particular questions related to the design of professional development programs that can contribute to lessening teacher attrition: What teachers are helped most? What should the focus be when providing assistance to teachers of different backgrounds? When working with new teachers, what aspects most influence teacher retention? Although the present study does not offer specific answers to these questions, the results can help to find answers.

Consider some of results of the present study in light of the above questions. The study shows that age and gender might be among those teacher characteristics that most influence retention of hispanic teachers. Given that the study focused on the four characteristics of ethnicity, age, gender, and salary, and given that ethnicity was removed as a variable when considering only hispanic teachers, then one may restate this finding as follows: among hispanic teachers salary seems to be the *least* important factor influencing teacher retention. If one allows some room for interpretation, this suggests that there is some commonality or common experience specifically among hispanic mathematics teachers that overrides salary as a dominant factor in the decision to remain in the education system, an experience that either is not present or does not have the same effect in the teacher population at large. Some sources (Dee, 2005) suggest that this commonality may be a shared frustration in their experience as students, which then drives hispanic teachers to improve the educational system as they become a part of it: this provides a sense of mission which overrides sensitivity to salary to some degree. But this need not be the explanation, or the only explanation, and the findings of the present study hint at the fact that future investigations should focus on ascertaining what might reduce sensitivity to salary among hispanic teachers.

Nevertheless the study highlights that, when looking at the characteristics that most influence teacher retention for the two particular years studied, gender, ethnicity, salary and age are all equally important: despite claims in the literature, salary was not in this case the characteristic that most influ-

enced teacher retention.

Another significant contribution of the present study relates to the methodology used. The methodology described in Chapter 3 provides a tool that will allow the inclusion of more characteristics, and broader data sets, in order to ascertain whether other characteristics are equally, less, or more important in determining the period of time a teacher stays in the system. As discussed in Chapter 2 some researchers have proposed that the condition, location and administrative decisions of the schools might influence teacher retention. With the program used in this investigation, future investigations can now include characteristics such as school location, funds allocated per school per year, number of teachers in a school, and more. One can even analyze the validity of the arguments that tie teacher retention to teacher level of preparation, determined by the certification type of the teacher, highest degree obtained, and number of languages spoken.

It is important to note here that when more characteristics are included, as well as more years, the analysis of the results needs to accompany an analysis of the labor conditions during the years studied. For example, Chapter 2 illustrated that the labor market in education changed significantly when other opportunities were opened to women in the broader labor market. This in fact could be an external characteristic that changes how teachers see the teaching profession. Instead of being viewed a long–term carrer, teaching could be seen as a first or intermediary step toward a different professional path.

5.5 The Future

Here I take a moment to consider avenues for future advances based on the work presented in this dissertation. Many of the avenues open to furthering the investigation surround overcoming the limitations placed on the study due to the time frame and computational resources available.

5.5.1 More Samples

One limitation of the current research surrounds the number of samples of data compared to draw the conclusions. For each choice of weights, I have run the program three times, each time randomly re-splitting the data and checking the resulting error rate. Ultimately I would like to see how much variation in the error rate occurs for a particular choice of weights, but three values form a collection too small from which to derive any meaningful statistical characterization of the error rate's variability for given weights. One would at least like to see 10s or 100s of runs for each choice of weights.

5.5.2 More Weights

Another limitation of the present study is the vanishingly small region explored within the 4–dimensional space of possible weights (which can be reduced to 3 dimensions once one imposes the constraint that the weights sum to 1). In particular, I have only explored along lines in this space, and I have actually only calculated the error rate at five points along each of these lines. A more robust exploratory analysis would select three weights at random

from a uniform distribution, calculate the fourth using the constraint, and run the algoritm (various times) for this selection of weights. Performing such a procedure 100s or 1000s of times would provide a more uniform exploration of the space of weights than the walk–along–lines performed in the current analysis.

5.5.3 More Computer Time

Any of the above analyses requires vastly more computational power and computer time than what was available with the current limited resources and time. In particular, the procedure could easily be parallelized: given a fixed (initially randomly selected) training data base, each point in the test data base could be classified *in parallel* against that data base. That is, the classification of one point does not require the classification of any other point, and so when classifying two points, say, they could in principle be classified simultaneously. Because of the computational resources available, the current program must classify the points one after the other. But once understood, adapting the program to run in parallel would require only minor modifications. (That is one of the reasons why it was initially deemed advisable to develop a program from scratch, rather than simply employing some previously developed k–Nearest Neighbors implementation as a black box.)

5.5.4 More Data

Finally there is the issue of the quantity of data. In data mining, the general trend is that, the more data, the better. In this study we have imposed an arbitrary constraint by focusing on two particular school years: 2003–2004 and 2006–2007. Generally speaking, any procedure in data mining, as with statistics more broadly, improves with additional data. Among the various possibilities of adding more data, two stand out:

- look at variation across several years;
- use data from all preceding years to predict data for a given year.

Any of these methods, however, will suffer from the constraints on computer time. Thus one finds an order of preference: ideally future research should focus first on making such modifications as allow the program to be run in parallel. After this barrier is overcome, then the program should be applied to sets of data of ever increasing size.

Appendices

Appendix A

Complete List of Variables Compiled in PEIMS

1. Region Number
2. County Number
3. County Name
4. District Number
5. District Name
6. District Type Code
7. District Type Name
8. District Charter Type Code
9. District Charter Type Name
10. District Category Code
11. District Category Name
12. District Phone

13. District Fax

14. District Address

15. District City

16. District State

17. District Zip

18. Total District Enrollment

19. African American District Enrollment

20. Native American District Enrollment

21. Hispanic District Enrollment

22. White District Enrollment

23. Total District Economic Enrollment

24. District Bilingual Ed Enrollment

25. District ESL Enrollment

26. District Carrer and Technology Enrollment

27. District Gifted and Talented Enrollment

28. District Special Ed Enrollment

29. Campus Number

30. Campus Name

31. Campus Low Grade

32. Campus High Grade

33. Campus Type Code

34. Campus Type Name

35. Campus Charter Type Code

36. Campus Charter Type Name

37. Campus Grade Group Code

38. Campus Grade Group Name

39. Campus Address

40. Campus City

41. Campus State

42. Campus Zip

43. Campus Phone

44. Total Campus Enrollment

45. Asin Campus Enrollment

46. African American Campus Enrollment

47. Native American Campus Enrollment

48. Hispanic Campus Enrollment

49. White Campus Enrollment

50. Total Campus Economic Enrollment

51. Campus Bilingual Ed Enrollment

52. Campus ESL Enrollment

53. Campus Carrer and Technology Enrollment

54. Campus Gifted and Talented Enrollment

55. Campus Gifted and Talented Enrollment

56. Campus Special Ed Enrollment

57. Unique Id

58. First Name

59. Middle Name

60. Last Name

61. Sex Code

62. Sex Name

63. Ethnicity code

64. Ethnicity Name

65. Year of Birth

66. Experience

67. Degree Code

68. Degree Name

69. FTE

70. Base Pay

71. Other Pay

72. Total Pay

73. Role Code

74. Role Name

75. Role Full Equivalent

76. Role Base Pay

77. Subject Area Code 1

78. Subject Area Name 1

79. Subject Area Code 2

80. Subject Area Name 2

81. Subject Area Code 3

82. Subject Area Name 3

83. Subject Area Code 4

84. Subject Area Name 4

85. Subject Area Code 5

86. Subject Area Name 5

87. Subject Area Code 6

88. Subject Area Name 6

89. Subject Area Code 7

90. Subject Area Name 7

91. Subject Area Code 8

92. Subject Area Name 8

93. Subject Area Code 9

94. Subject Area Name 9

95. Subject Area Code 10

96. Subject Area Name 10

97. Pay Type Code 1

98. Pay Type Name 1

99. Pay Type Code 2

100. Pay Type Name 2

101. Pay Type Code 3

102. Pay Type Name 3

103. Pay Type Code 4

104. Pay Type Name 4

105. Pay Type Code 5

106. Pay Type Name 5

Appendix B

Graphs for Repeated Program Runs All Math Teachers

Figure B.1: 2003–2004. Iteration 1.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

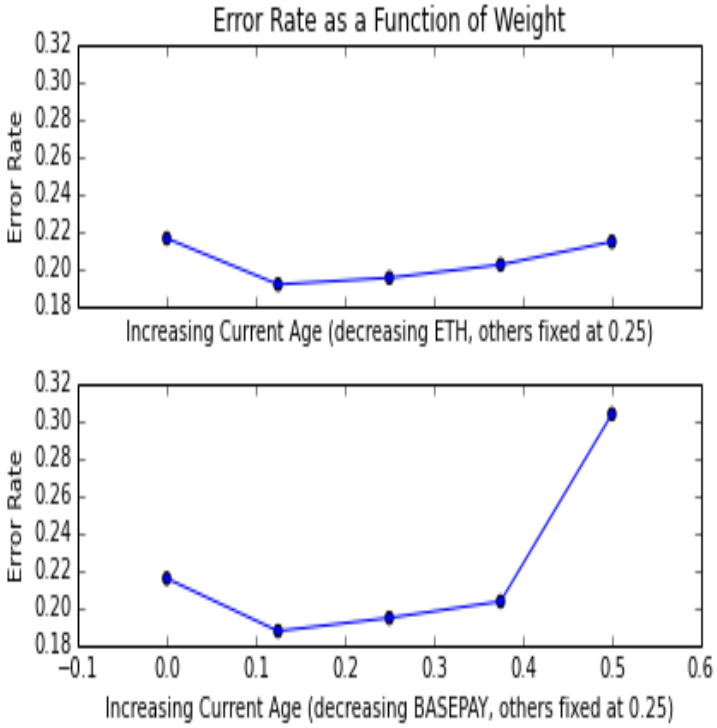


Figure B.2: 2003–2004. Iteration 1.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

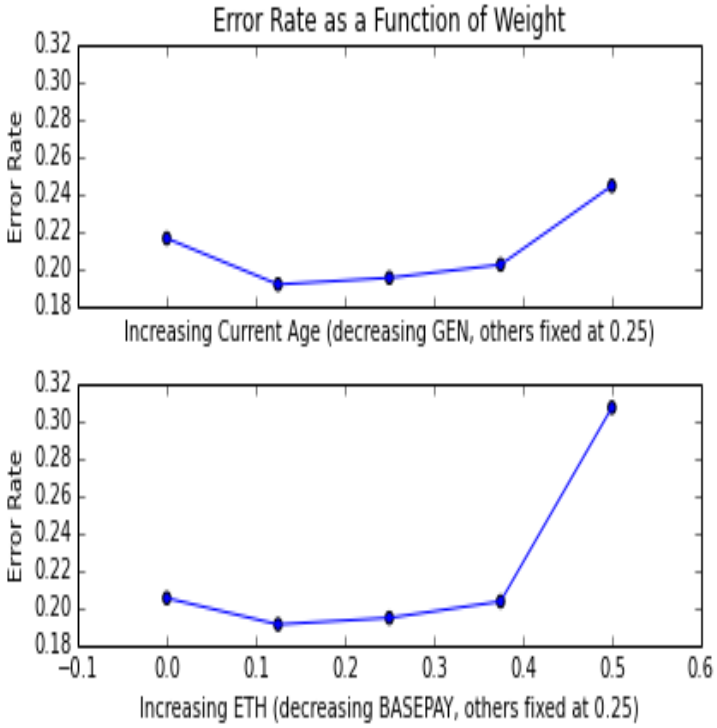


Figure B.3: 2003–2004. Iteration 1.2. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

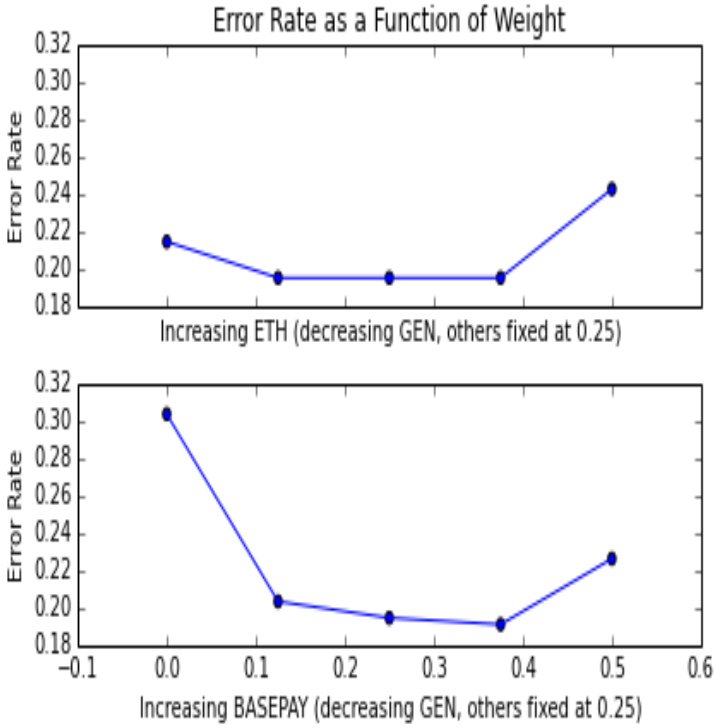


Figure B.4: 2003–2004. Iteration 2.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

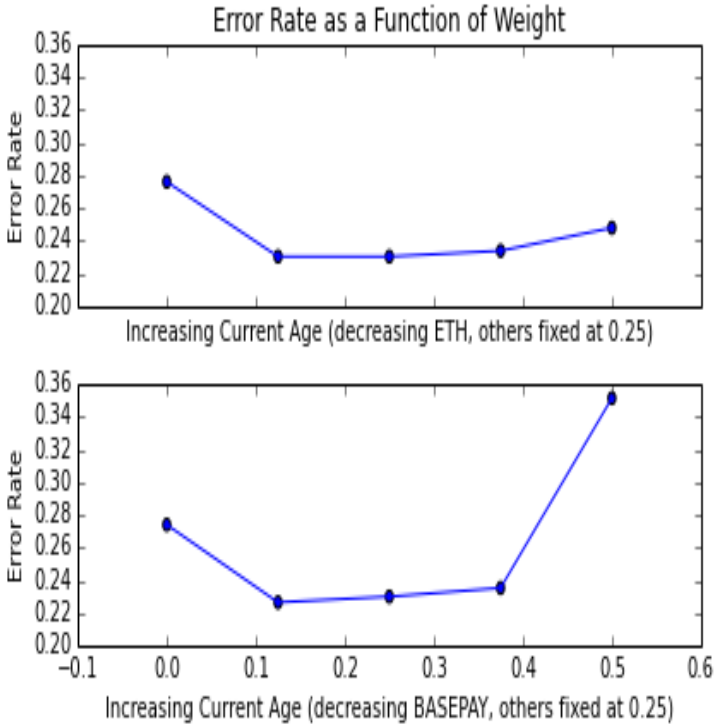


Figure B.5: 2003–2004. Iteration 2.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

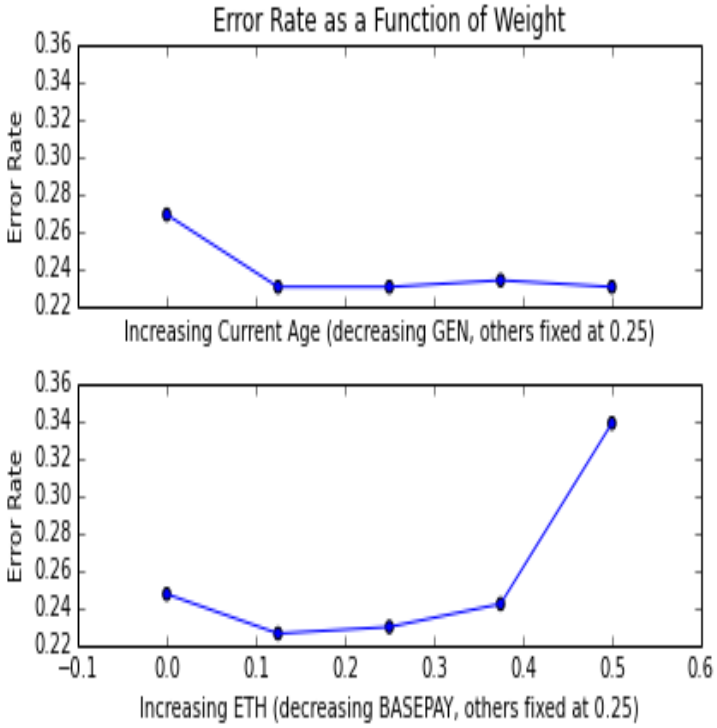


Figure B.6: 2003–2004. Iteration 2.2. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

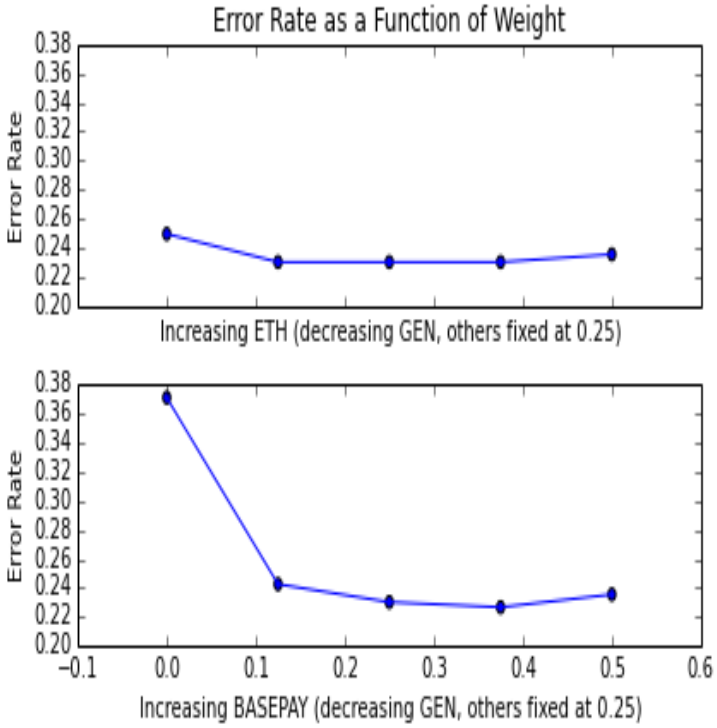


Figure B.7: 2003–2004. Iteration 3.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

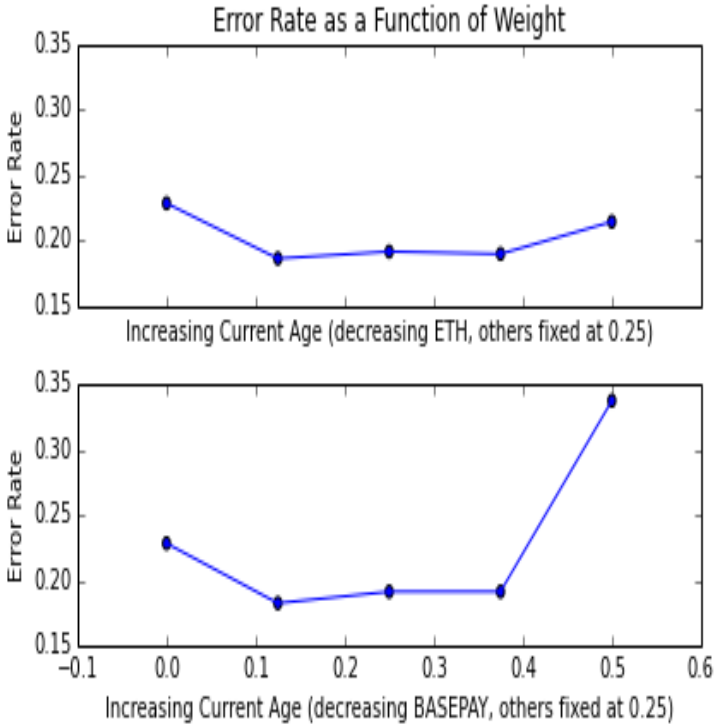


Figure B.8: 2003–2004. Iteration 3.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

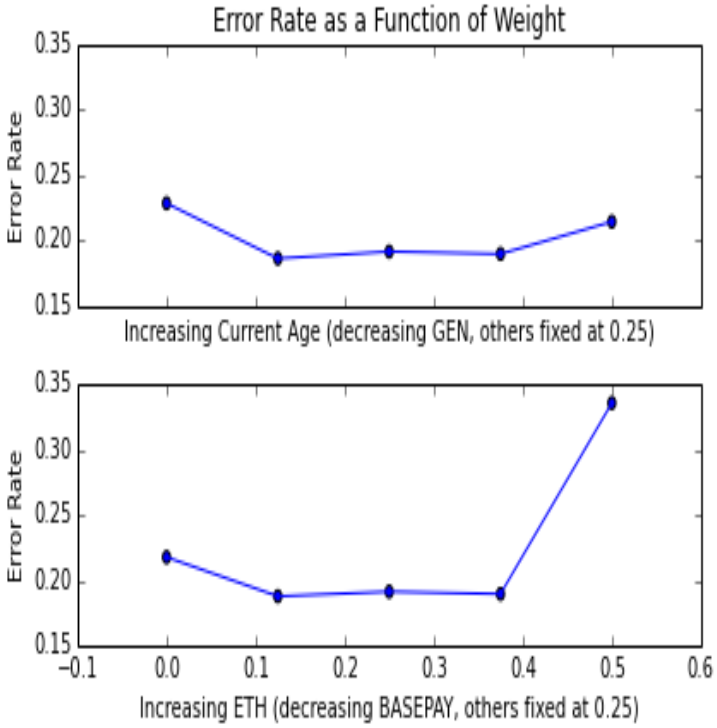
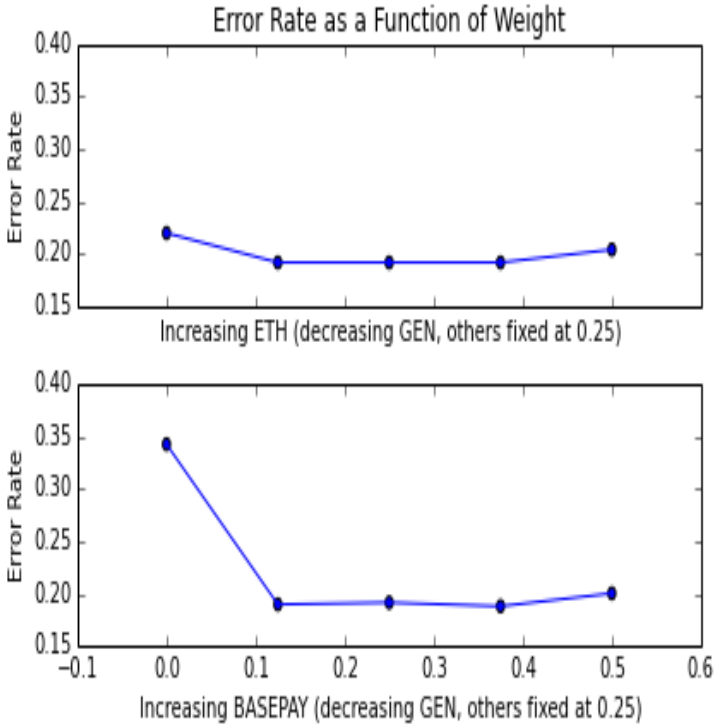


Figure B.9: 2003–2004. Iteration 3.2. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender



Appendix C

Graphs for Repeated Program Runs All Hispanic Math Teachers 2003–2004

Figure C.1: 2003–2004. Iteration 1.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

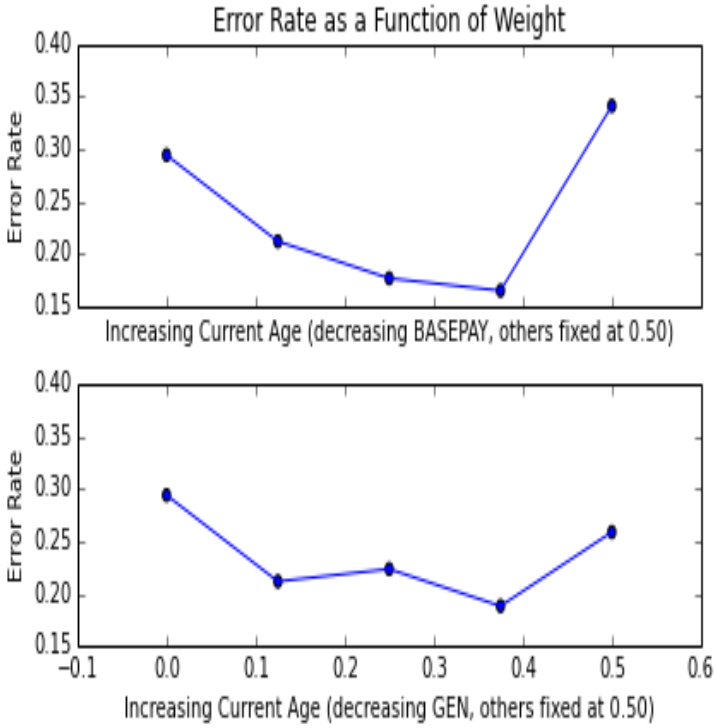


Figure C.2: 2003–2004. Iteration 1.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

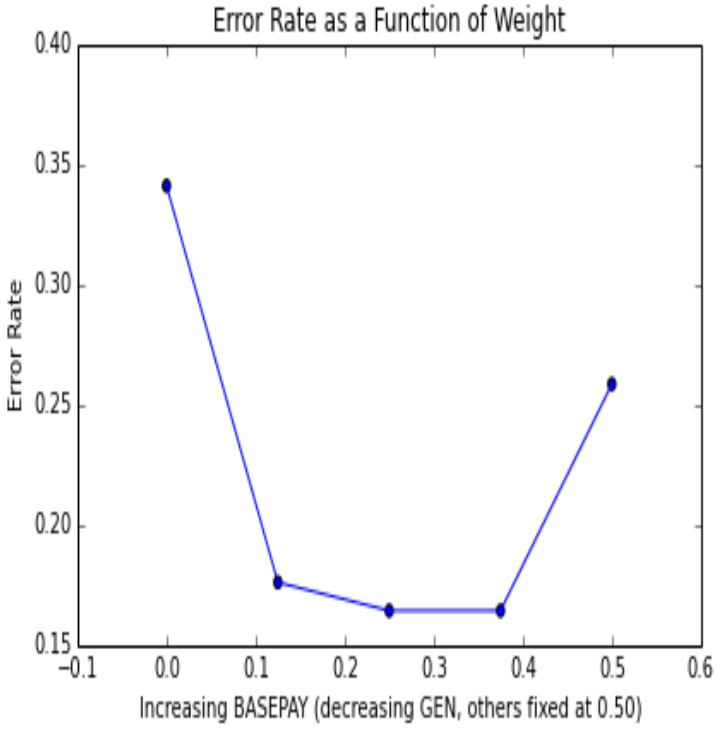


Figure C.3: 2003–2004. Iteration 2.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

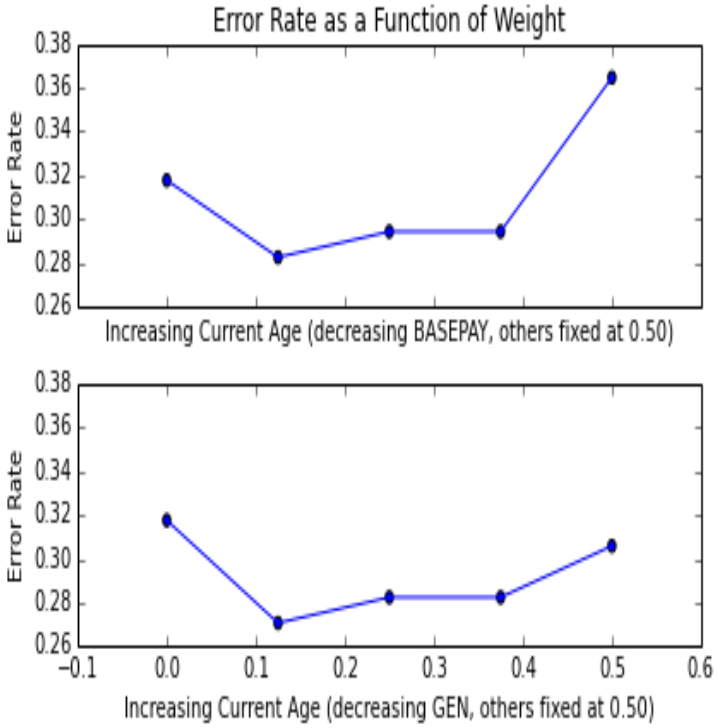


Figure C.4: 2003–2004. Iteration 2.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

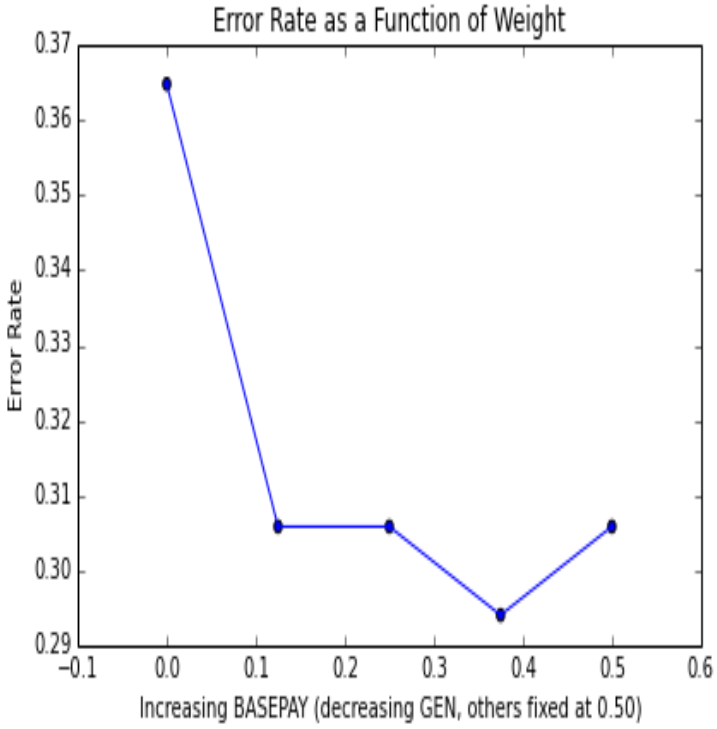


Figure C.5: 2003–2004. Iteration 3.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

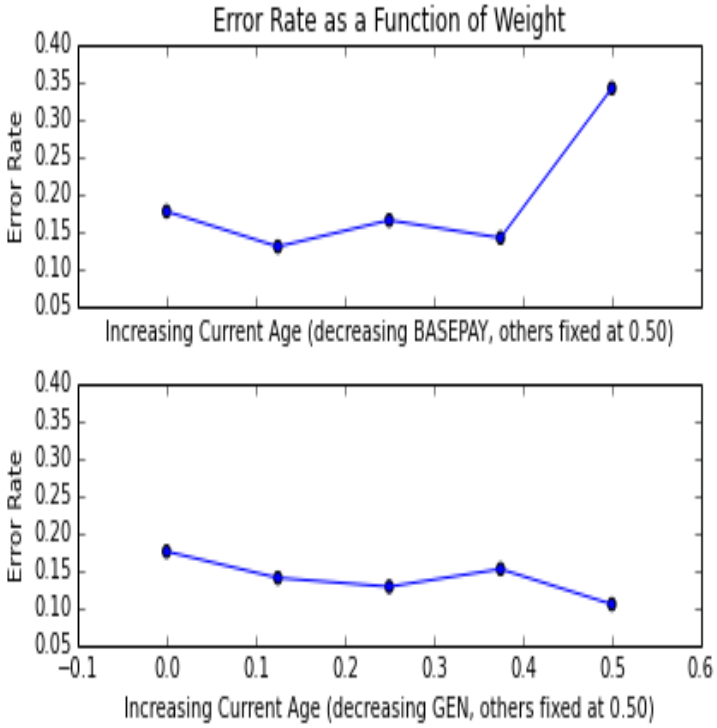
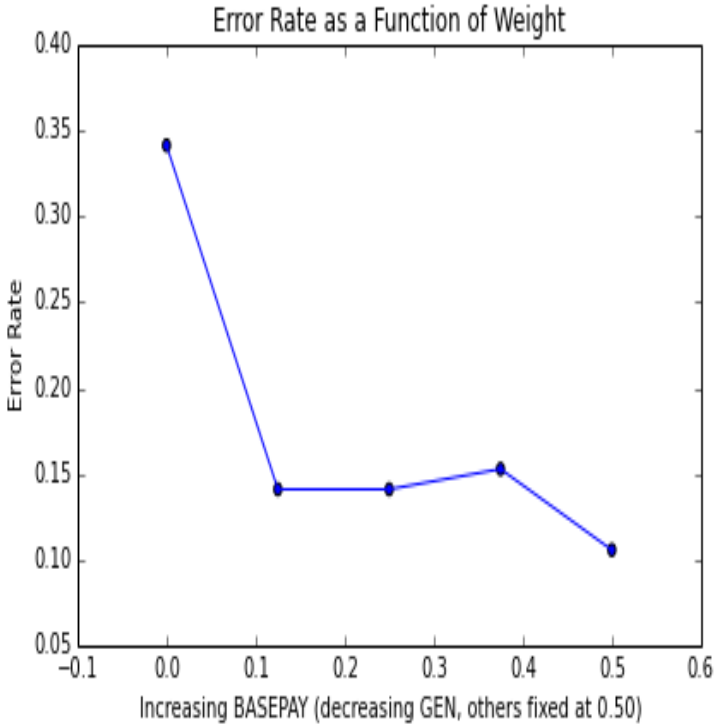


Figure C.6: 2003–2004. Iteration 3.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender



Appendix D

Graphs for Repeated Program Runs All Hispanic Math Teachers 2006–2007

Figure D.1: 2006–2007. Iteration 1.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

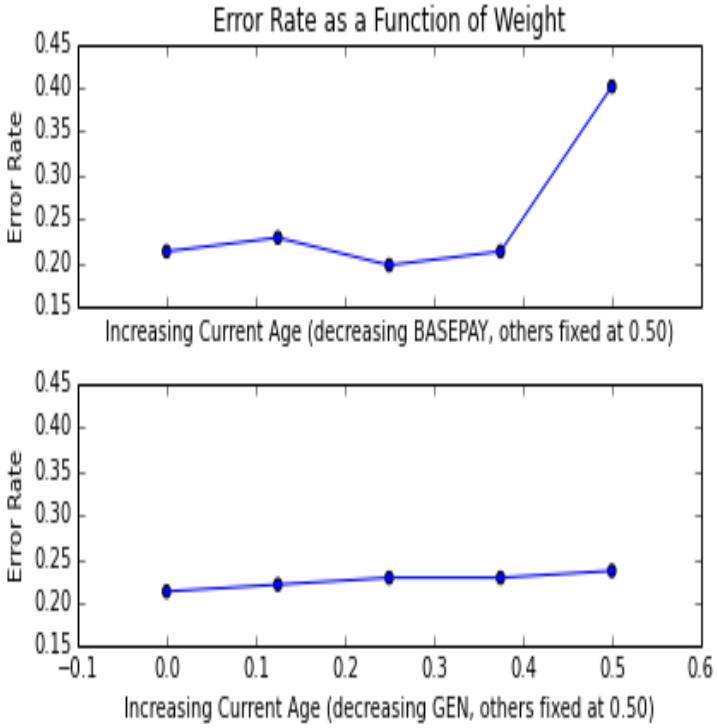


Figure D.2: 2006–2007. Iteration 1.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

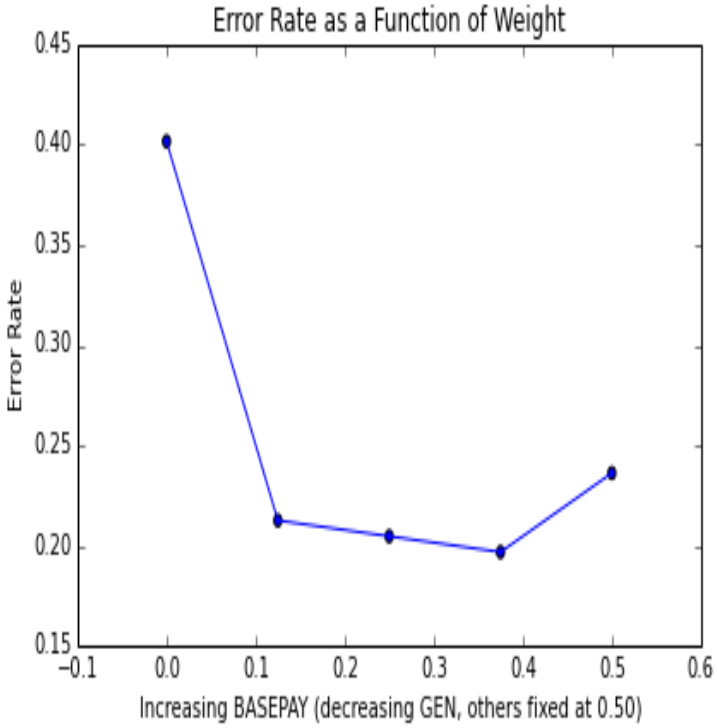


Figure D.3: 2006–2007. Iteration 2.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

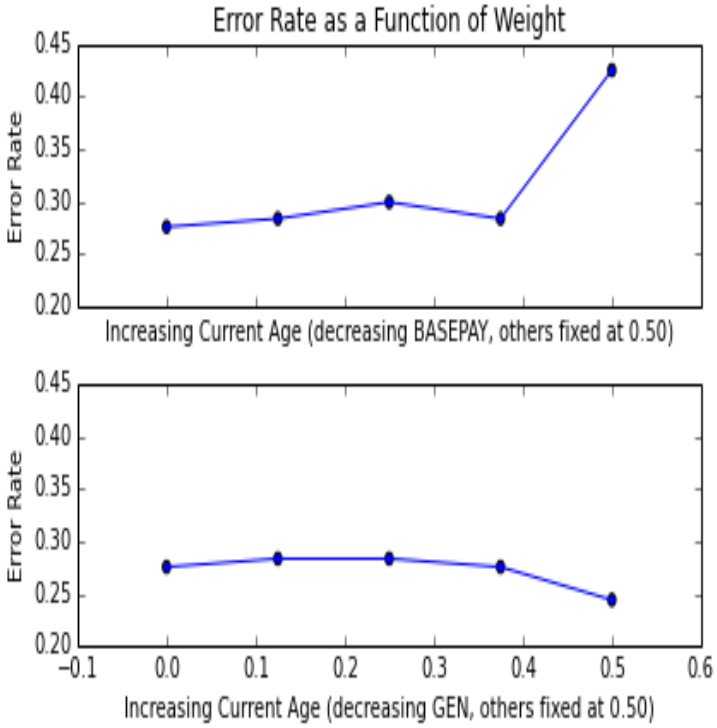


Figure D.4: 2006–2007. Iteration 2.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

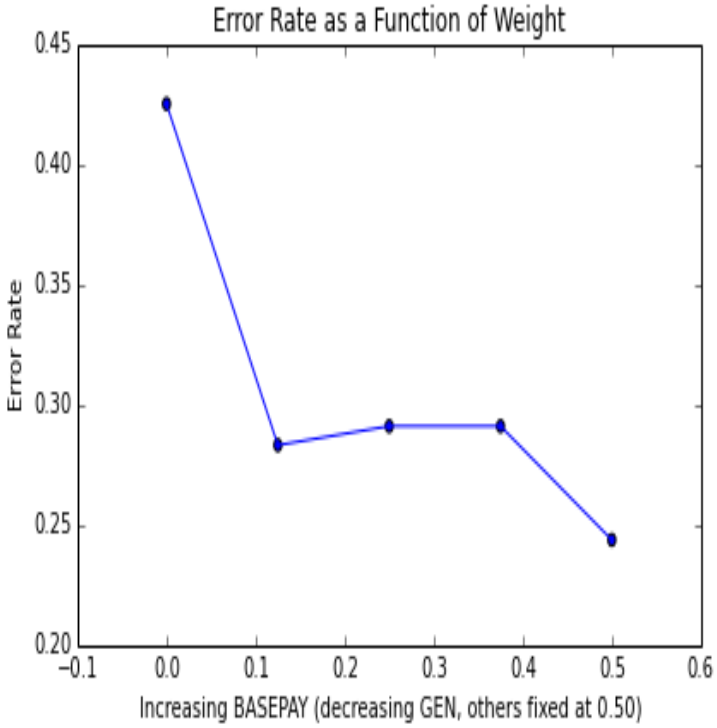


Figure D.5: 2006–2007. Iteration 3.0. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender

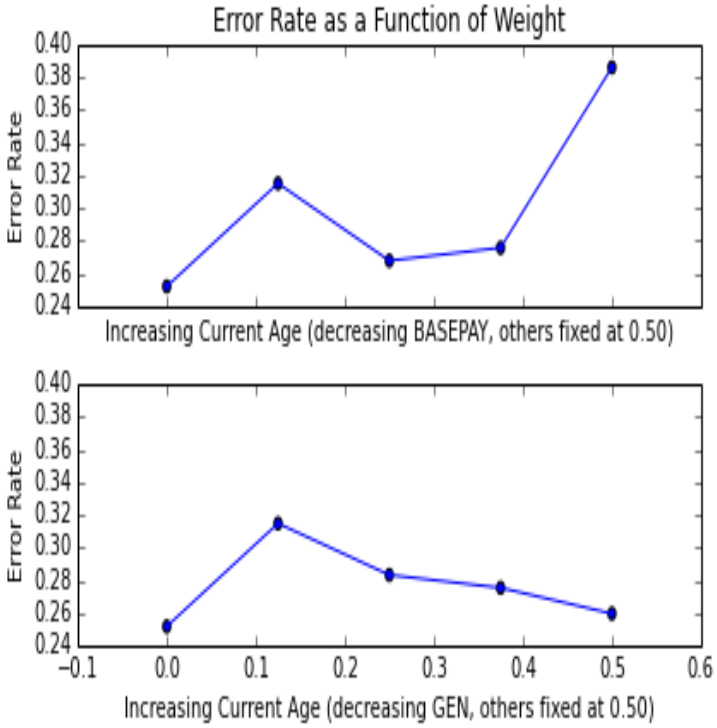
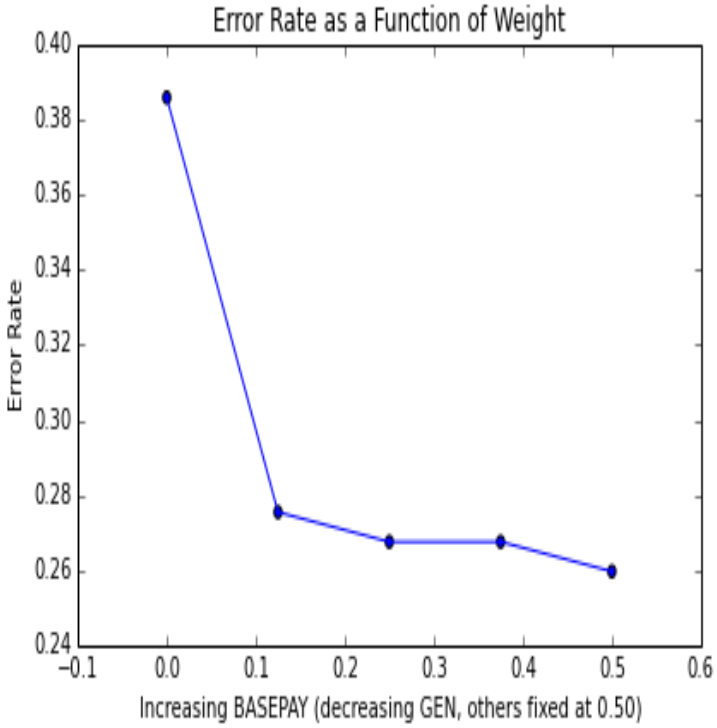


Figure D.6: 2006–2007. Iteration 3.1. Error Rates for Ethnicity vs. Gender and Base Pay vs. Gender



References

Abbott, M., & Joireman, J. (2001). *The relationships among achievement, low income, and ethnicity across six groups of washington state students* (Tech. Rep.). Washington School Research Center.

Allen, M. (2005). *Eight questions on teacher recruitment and retention: What does the research say?* (Tech. Rep.). Education Commission of the States. Available from `http://files.eric.ed.gov/fulltext/ED489332.pdf`

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The cost of teacher turnover in five school districts* (Tech. Rep.). National Commission on Teaching and America's Future.

Buckley, J., Schneider, M., & Shang, Y. (2005, May). Fix it and they might stay: School facility quality and teacher retention in Washington, D.C. *Teachers College Record*, *107*(5), 1107–1123.

Chapman, D. (1984). Teacher retention: The test of a model. *American Educational Researcher*, *21*(3), 645–658.

Cuban, L. (2010). *As good as it gets.* Cambridge: Harvard University Press.

Darling-Hammond, L. (1984). Beyond the commission reports. The coming crisis in teaching. *Rand Corporation*, 1–32.

Dee, T. (2005). A teacher like me: Does race, ethnicity or gender matter? *American Economic Review*, 158 – 165.

Dobbie, W. (2011). *Teacher characteristics and student achievement: Evidence from teach for america.*

Grant, C. (1992). *Research in multicultural education: From the margines to the mainstream.* Bristol, PA: The Falmer Press.

Grissmer, D., & Kirby, S. (1991). Patterns of attrition among indiana teachers, 1965–1987. *Rand Corporation*, 1–28.

Gritz, R. M., & Theobald, N. D. (1996, Summer). The effects of school district spending priorities on length of stay in teaching. *Teachers College Record*, *31*(3), 477–512.

Guarino, C., Santibañez, L., & Daley, G. (2006). Teacher Recruitment and Retention: A Review of the Recent Empirical Literature. *Review of Educational Research*, *76*(2), 173–2008.

Hanushek, E. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*, 466 – 479.

Hanushek, E., Kain, J., & Rivkin, S. (2004, Spring). Why public schools lose teachers. *Teachers College Record*, *39*(2), 326–354.

Harrington, P. (2012). *Machine learning in action*. New York: Manning.

Hemphill, F., & Rahman, T. (2011). *Achievement gaps. how hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress* (Tech. Rep.). National Center for Education Statistics. Available from `http://www.nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf`

Henke, R., Zahn, L., & Carroll, D. (2001). *Attrition of new teachers among recent college graduates* (Tech. Rep.). U.S. Department of Education: National Center for Education Statistics. (`http://www.ed.gov/pubs/edpubs.html`)

Ingersoll, R. (2001, Fall). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research*, *38*(3), 499–534.

Ingersoll, R. (2011). Do we produce enough mathematics and science teachers? *Kappan*, *92*(6), 37–41.

Ingersoll, R., & Kralik, J. M. (2004). *The impact of mentoring on teacher retention: What the research says* (Tech. Rep.). Education Commission of the States. Available from `http://www.gse.upenn.edu/pdf/rmi/ECS-RMI-2004.pdf`

Ingersoll, R., & May, H. (2011). *Recruitment, retention, and the minority teacher shortage* (Tech. Rep.). Consortium for Policy Research in Education. CPRE. Available from `http://repository.upenn.edu/cgi/viewcontent.cgi?article=1232&context=gse_pubs`

Ingersoll, R., & May, H. (2012, April). The Magnitude, Destinations, and Determinants of Mathematics and Science Teacher Turnover. *Educational Evaluation and Policy Analysis*, *2*(1), 2–31.

International Educational Data Mining Society. (2011, July). *International educational data mining society.* Available from `http://www.educationaldatamining.org/about`

IPython Development Team. (2008). *IPython documentation.* `http://ipython.org/`.

Kersaint, G., Lewis, J., Potter, R., & Meisels, G. (2005). A teacher like me: Does race, ethnicity or gender matter? *American Economic Review*, 158 – 165.

Kimmit, R. (2007). Why job churn is good. *Washington Post*.

Krause, G. (2011, May). *Putting the* Earning *back in l*earning*: Does teacher salary affect student performance?* (unpublished)

Krause, G. H. (2013–2014). *kNN.* (unpublished)

Lambda Foundry, Inc., & PyData Development Team. (2012). *Python data analysis library.* `http://pandas.pydata.org/`.

McCornick, R., Carmichael, P., Fox, A., & Procter, R. (2011). *Researching and understanding educational networks.* New York: Routledge.

McKinney, W. (2013). *Python for data analysis.* California: O'Reilly.

Meier, K., & Stewart, J. (1991). *The politics of hispanic education: Un paso pa'lante y dos pa'tras.* Albany, NY: State University of New York Press.

Moreland, A. (2011). *A mixed-methods study of mid-career science teachers: The growth of professional empowerment.* Unpublished doctoral dissertation, University of Texas at Austin.

Mount, J. (2012). *Migration and Attrition Patterns of Texas Secondary Science Teachers.* Unpublished doctoral dissertation, University of Texas at Austin.

National Center on Performance Incentives. (2010). *Teacher pay for performance. experimental evidence from the project on incentives in teaching* (Tech. Rep.). RAND Education.

Python Software Foundation. (1990–2013). *Python language reference.* `http://www.python.org/`.

Rivkin, G., & Hanushek, A. (2007). Pay, working conditions and teacher quality. *Project Muse*, *17*(1), 69–86.

Rockoff, J., Jacob, B., Kane, T., & Staiger, D. (2008). Can you recognize an effective teacher when you recruit one? *Economics of Education Review*, 1 – 56.

Rossum, G. van. (1995, May). *Python tutorial* (Tech. Rep. No. CS-R9526). Amsterdam: Centrum voor Wiskunde en Informatica (CWI).

Rumberger, R. (1987). The impact of salary differentials on teacher shortages and turnover: The case of mathematics and science teachers. *Economics of Education Review*, *6*(4), 389–399.

Stevens. (n.d.). *Texas teacher diversity and recruitment* (Tech. Rep.). Texas Education Agency. Office of Policy Planning and Evaluation. Available from `www.tea.state.tx.us/research/pdfs/prr4.pdf`

Stutz, T. (2010). *Texas needs more minority teachers, experts say.* Available from `http://www.dallasnews.com/news/education/headlines/`

20100802-Texas-needs-more-minority-teachers-9353.ece

Texas Center for Educational Research. (2000). *The cost of teacher turnover* (Tech. Rep.). Texas Center for Educational Research. Available from www.tcer.org/research/documents/teacher_turnover_full.doc

Texas Education Agency. (2010, December). *House Bill 3 Transition Plan.* Available from http://www.tea.state.tx.us/student.assessment/hb3plan/

Texas Education Agency. (2011a, July). *2011–2012 minimum salary schedule* (Tech. Rep.). Available from http://www.tea.state.tx.us/index2.aspx?id=2147501688

Texas Education Agency. (2011b). *PEIMS standard reports* (Tech. Rep.). Texas Education Agency. Available from http://ritter.tea.state.tx.us/adhocrpt/

Texas Education Agency. (2012). *House bill 3 transition plan* (Tech. Rep.). Texas Education Agency. Available from http://www.tea.state.tx.us/student.assessment/hb3plan/

Texas Education Agency, and Texas Higher Education Coordinating Board, and Texas Educators. (2012). *House bill 3 transition plans* (Tech. Rep.). Texas Education Agency, and Texas Higher Education Coordinating Board.

Texas Education Agency Policy Planning and Evaluation Division. (1993). *Teacher supply, demand and quality. policy research project* (Tech. Rep.). Texas Education Agency. Available from http://search.tea.state.tx.us/search?access=p&entqr=0&output=xml_no_dtd&sort=date%3AD%3AL%3Ad1&ud=1&client=default_frontend&oe=UTF-8&ie=UTF-8&proxystylesheet=default_frontend&site=default_collection&q=mathematics%20hours%20per%20week

The Institute for Education in Transformation. (n.d.). *Voices from the inside: a report on schooling from inside the classroom* (Tech. Rep.). The Institute for Education in Transformation. Available from http://www.cgu.edu/pages/3721.asp

Torres, J., Santos, J., Peck, N., & Cortés, L. (2004). *Minority teacher recruitment, development, and retention* (Tech. Rep.). The Educational Alliance at Brown University. Available from http://www.alliance.brown.edu/tdl/minteachrcrt.shtml

Zumwalt, K., & Craig, E. (2005). *Studying teacher education: The report of the aera panel on research and teacher education.* New Jersey: American

Educational Research Association.