5-2015

Modeling Enrollment at a Regional University using a Discrete-Time Markov Chain

Zachary T. Helbert
East Tennessee State University

Modeling Enrollment at a Regional University using a Discrete-Time Markov Chain

Zach Helbert[1],
Michele Joyner[1,2] (Supervisor)

[1] Department of Mathematics and Statistics
Box 70663
East Tennessee State University
Johnson City, TN 37614-0663

[2] Institute for Quantitative Biology
East Tennessee State University
Johnson City, TN 37614-0663

April 27, 2015

Abstract

A discrete time Markov Chain is used to model enrollment at a regional university. A preliminary analysis is conducted on the data set in order to determine the classes for the Markov chain model. The semester, yearly, and long term results of the model are examined thoroughly. A sensitivity analysis of the probability matrix entries is then conducted to determine the overall greatest influence on graduation rates.

1 Introduction

Between the years 2001 and 2011 the overall enrollment of universities across the nation has increased by 32 percent, from 15.9 million to 21.0 million [4] from 2011 to 2013 college, however, enrollment dropped 930,000 [5]. The dropout rates of United States colleges are at an alarmingly high rate. Many people are beginning college only to find themselves quitting before they are finished. Many papers address that the dropout rates of the United States are high and things need to be done to improve them [6, 7]. The papers explore a number of explanations for the high dropout rates including lack of preparation [1], race [8], class [9], and even neighborhoods surrounding the college [2]. This thesis focuses on a regional university which seems to have maintained a steady dropout rate over the past few years, but which is still at a disappointing high rate.

In this thesis, we will be using data obtained from a regional university to create a discrete-time Markov chain model for the transitioning of students from freshman through graduation including dropout rates as well. This has been done by others, in particular a group at the University of Memphis [3]. A discrete-time Markov chain is a stochastic process, or family of random variables, $\{X_n\}, n \in \{0, 1, 2, ...\}\}$ which satisfies the Markov property. In other words,

$$Prob\{X_{n+1} = j | X_0 = x_0, ..., X_{n-1} = x_{n-1}, X_n = i\} = Prob\{X_{n+1} = j | X_n = i\}.$$

Thus the next state only depends on the current state and is independent of the past states. Discrete time Markov Chain models are used to model many things including physical systems [10], speech recognition [11], and epidemics [12]. Using this type of model for our data involves utilizing a probability matrix in order to predict the future enrollment and graduation rates at the university, and the long term effects of changing one or more of the probabilities by a certain percentile. We chose to use a discrete-time Markov chain, because transitioning between classes at a university only occurs at discrete times, namely at the end of the semester. Furthermore, which class you go to next depends only on the class you are currently in; hence, the Markov property is satisfied. Furthermore, some of the advantages of using a discrete-time Markov chain model are that they are relatively easy to derive from successive data, it does not require deep insight into the reasons for the change but can give insight into the process, and the results from a Markov chain model are easily interpretable.

Our goal is to create a model that mimics the trends in our actual data. If possible, we can then use the model to potentially predict future trends depending on current data or trends if changes were implemented somewhere within the university system. We might also be able to predict the long-term effect of the current dropout rates and where changes might need to be implemented in the university system to have the most direct impact on student success. By altering certain entries in the probability matrix we can observe the long term effects caused by those slight changes. The focus for improving enrollment or increased gaduation rates can then be centered on that group.

In Section 2, we describe the data set and perform a preliminary statistical investigation of the impact of the given variables on the graduation rates and semester-to-semester drop-out rates. In Section 3 we develop the discrete-time Markov Chain model and discuss the variability across time. In Section 4 we explore the output of our model by looking at the mean time to absorption, the state of the system after four and five years, the steady state, and address any discrepancies between our model and actual data. In Section 5 we conduct a sensitivity analysis on our model. We alter the probability matrices and observe the change in long-term enrollment, drop-out rates, and graduation rates caused by this change. In Section 6 we summarize our work and discuss what can be done in the future to improve the model based on our findings.

2 Analysis of Data Set

Our data consists of fall and spring semester data starting in Fall 2008 until Spring 2013. Upon receiving the data, we first eliminated any data not pertaining to this study. For example, we only focus on undergraduates in this thesis; therefore, data for both graduate or post-doctorate students were eliminated. We also chose to eliminate data in which there were discrepancies in the data with the typical progression of a student from freshman to graduation in a university system. For example, we have some students who were classified in a certain class, say sophomore, and then they went up to the next one, junior, and then returned to the previous one, sophomore. We believe this occurred because some students may have been categorize incorrectly due to transfer hours, then properly placed after all of their hours were sorted and accounted for. We chose to not include these types of "backward" transitions in the model; therefore, the data for these students were removed. There are 83 students who fell under this category, which is a very small portion of our data set of approximately 41,000 students.

In the data, each student was represented by a student ID number; therefore, we were able to follow them for the time period in which we have data. We have many different variables and information about the students, which we display in the Table 1.

Table 1: Data set parameters

Parameter	**Explanation**
Descriptive Data	
Age Today	The student's age this year
Gender	Male or female.
Residential Status	In-state resident or out-of-state.

Continued on next page

Table 1 – *Continued from previous page*

Parameter	**Explanation**
Lottery Scholarship Residential Status	In-state or out-of-state as defined for the lottery scholarship (The student's residency status for the scholarship requires that they are in-state for one year before the application deadline).
United States Citizenship	Whether or not the student is a U.S. citizen.
Zip Code	The zip code for the student's official home residence.
County	The student's home county of residence.
Race	The student's race.
Ethnicity	The student's ethnicity.
Term Data	
Age during term	The student's age during the corresponding term.
Class Level	Freshman, Sophomore, Junior, or Senior
Hours Acquired	The number of credit hours the student has earned.
Major	The students current major.
Term GPA	The student's GPA for the current term.
Overall GPA	The student's GPA for their entire college career.
Enrollment Data	
Registration Type	If the student is a returning student or first-time student.
Transfer Student	Whether or not the student transferred to this college or not.
ACT Score/SAT Score	The scores that the student made on either the ACT or the SAT (individual and cumulative scores).
High School Attended	The high school from which the student graduated
High School Graduation Year	The year that the student graduated from high school.
High School GPA	The student's GPA upon graduation.

Continued on next page

Table 1 – *Continued from previous page*

Parameter	**Explanation**
Diploma Type	The students high school diploma type such as regular diploma or GED.
College Admission Type	Classification for the student's admission type such as first-time freshman or transfer student.
Parents Highest Level Attended	The highest level of school each parent attended. i.e. some, high or college.
Family Income	The student's family's income.

Though we have access to all of this information, we do not utilize all of it. To create a model containing all of these variables would be extremely tedious and most likely unnecessary. We simply didn't include some of the variables because of missing data for the majority of students. For example, the students income was missing a large portion, thus we did not include it. This occurred for many variables for a portion of the students simply because it is not required on the official application.

We chose to use Matlab for simulating our model; therefore, we needed to be able to import the data into Matlab. To do this, we chose numerical representations for some of our categories. We substituted freshman, sophomore, junior, and senior with 1, 2, 3, and 4 respectively. We also changed each semester from Fall 2008 up to Spring 2013 with numbers 1-12 in chronological order. After making these substitutions in the data we transfer it into a separate Excel worksheet and imported it into Matlab. We then plotted the data in order to observe trends in it, see Figure 1.

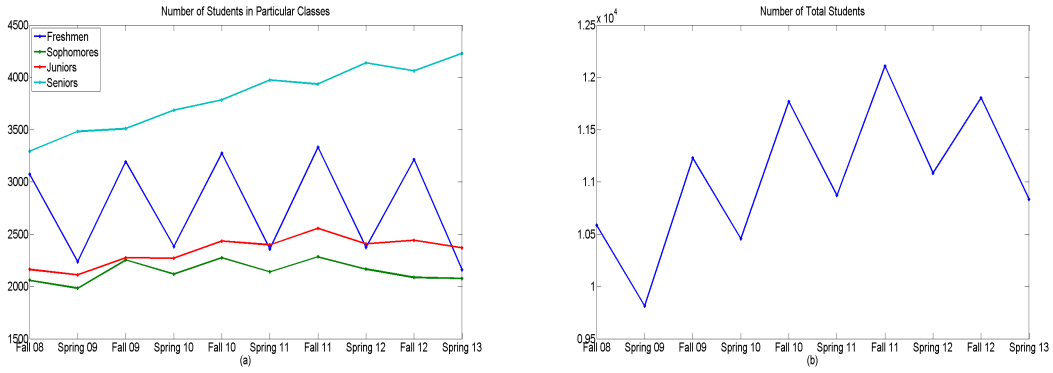

Figure 1: Plot of the number of total students in the university, and the total number of students in each class level, for the entirety of our data set.

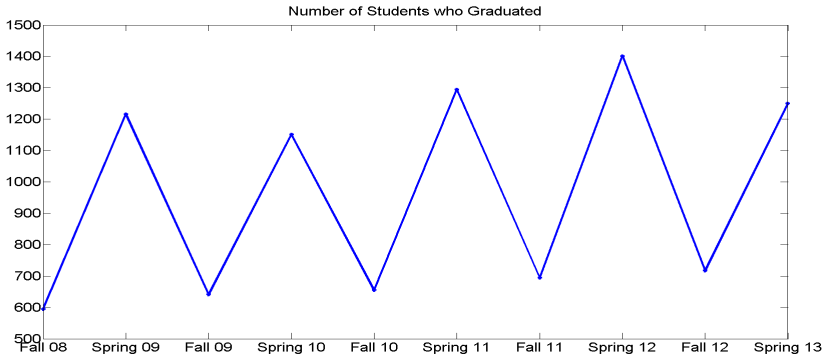

Figure 2: Plot of the number of students who graduate in our data set.

In Figure 1(a) for the trend in the number of students in each class, we notice that the number of seniors continues to rise. We believe this is because many students are not graduating in four or five years as they should. We also notice that the number of freshman in the Spring terms is much less than the fall terms. Some of this can be caused from students coming in from high school with university credit hours and transitioning to sophomore status after their first semester; however, the sophomore trend does not indicate this is the only reason for a dramatically lower number of freshman in the spring. We believe a large part of this trend is due to the high dropout rate indicated in our data. The dropout rate can also be seen in the total enrollment plot for the spring semesters. We also plot, in Figure 2, how many people graduate throughout our data set. Clearly more students graduate in the spring, which could also be affecting the total enrollment in the fall.

Before taking a deeper look into the effects of our many variables statistically, we eliminated those which we thought would be either insignificant or redundant in our analysis. We focused on the student's ACT scores, race, high school graduation year, overall GPA, whether the student's parents attended college or not, the college they are enrolled in, and the student's current class. We made these decisions for a few reasons. We wanted to determine if the variables we chose affected enrollment and dropout rates. We also had some incomplete data, as mentioned previously, thus we had to eliminate some variables we would have liked to use like income. At this particular regional university, it is believed that a student's need to work while taking large class loads is at least part of the reason for higher than usual drop out rates - both during freshman and senior year. However, without more thorough data, we could not investigate this avenue. Though all of these variables could affect success, we wish to only focus on some of them. The parameters we include are ones that could be potential subclasses for our Markov chain model.

After deciding which variables we would focus on, Minitab was used to analyze the data for potential predictors for both graduation and/or returning the following term.

We import the data into Minitab, and used the built in logistic regression to "model" our data as a function of the predictor variables. The ideal situation for using logistic regression is if there is a balance between successes and losses. Our data did not have this balance. Therefore, we do not use logistic regression to create a predictive model based on the variables. Instead, we use it as a method for getting *a priori* information about which classes should be included in the Markov chain model. More precisely, we wanted to determine if we could simply look at classes of students as a whole or if we needed to break the classes down into subclasses, such as Freshman Arts and Science Major, or Freshman Business major, etc. If the p-value of the variable was less than 0.05, then we can assume that it plays a significant role. For those that play a significant role we consider the variables odds ratio. If an event A has probability P, the odds against A are $A = (1-P)/P$, and obviously the odds in favor of A are $A = P/(1-P)$.

First we ran the logistic model for graduating in five years. We use the ID numbers of all the students who were freshmen during Fall 2008. The goodness of fit for the logistic regression was not adequate using only these variables which indicate there is a lot more affecting graduation than simply the potential predictors we examined. From the analysis, we determined whether or not these students graduate in five years; if so, we consider this a success. We have 3079 students in the data set. Of the 3079 students, 1085 of them graduate within 5 years. The overall GPA variable has an odds ratio of 19.85. For students enrolled in the College of Business and Technology, they have an odds ratio of 2.20. Thus for each additional point of GPA a student has they have a 19.85 times change of graduating. Also if they are enrolled in the College of Business and Technology they are 2.20 times more likely to graduate. From this analysis, there does not seem to a true need to separate the classes finer than the freshman, sophomore, junior and senior level.

We also ran a logistic regression model for each semester transition, from Fall 2008 to Spring 2009, and then from Spring 2009 to Fall 2009, and so on. If the student is still enrolled the next semester, then we consider that a success. In Table 2 we show how many students we have at the beginning of each semester and at the end of each semester.

We use the same variables for the logistic regression analysis and determine that the only variable that was consistently significant in these simulations is the overall GPA of students. The odds ratio for this is different for each semester, but is still very high. This result seems self-explanatory, because if the GPA of students is lower, then they are failing classes, thus they end up on academic probation and/or lose their scholarships. Since we did not have any categorical significant variables in both analysis, we did not split our model into any separate classes. In this Markov chain model, we only include those students present during a semester but do not take into account new or transfer students specifically.

Table 2: Number of students for each term simulation. We have the number of students who return and who do not.

Beginning Term	Beginning Students	Returning Students	Non-returning Students
Fall 2008	10652	8939	1713
Spring 2009	9880	7118	2761
Fall 2009	11286	9521	1765
Spring 2010	10510	7708	2802
Fall 2010	11809	10018	1791
Spring 2011	10909	7908	3001
Fall 2011	12138	10198	1940
Spring 2012	11112	7834	3278
Fall 2012	11828	9920	1908
Spring 2013	10854	7640	3214
Fall 2013	11399	9450	1949

3 Development of the Model

In this section, we discuss the discrete-time Markov chain model which has the state-transition diagram given in Figure 3. We include all of the transitions which we see in the data. Though it is highly unlikely, there is data that reinforces the fact that a freshman can go straight to being a senior in one semester. This is believed to be caused by an error in the data due to a student transferring college credit. We assume that there are two absorbing states: graduation and drop out. We also assume that once a student enters an absorbing state, they do not return. We will elaborate on this assumption later and the potential need to explore if this is a valid assumption in future models.

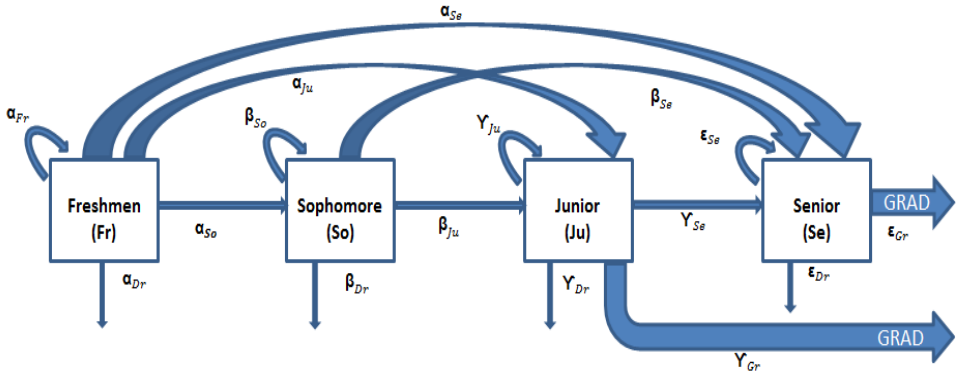

Figure 3: Markov chain model state diagram.

The probability transition matrix for this Markov chain can be represented by

$$P = \begin{bmatrix} \alpha_{Fr} & 0 & 0 & 0 & 0 & \\ \alpha_{So} & \beta_{So} & 0 & 0 & 0 & 0 \\ \alpha_{Ju} & \beta_{Ju} & \gamma_{Ju} & 0 & 0 & 0 \\ \alpha_{Se} & \beta_{Se} & \gamma_{Se} & \epsilon_{Se} & 0 & 0 \\ \alpha_{Dr} & \beta_{Dr} & \gamma_{Dr} & \epsilon_{Dr} & 1 & 0 \\ \alpha_{Gr} & \beta_{Gr} & \gamma_{Gr} & \epsilon_{Gr} & 0 & 1 \end{bmatrix} \tag{1}$$

where we assume $s_{n+1} = Ps_n$ where s_n is the state vector for the given semester n. Here each matrix entry represents the probability of transitioning from one class to another the following semester. In order to calculate these values in the probability transition matrix, we use the student id number to determine the students in each class during a semester and their class the following semester. If a student is not present the following semester, we assume they have dropped out. We then divide each class by the total number of students for the students that fall under each classification. For example, α_{Fr} is given by α_{Fr} =freshmen next semester/total feshmen this semester. We do this for each entry in the matrix, for each of our terms, resulting in the following nine matrices.

$$F8S9 = \begin{bmatrix} 0.63 & 0 & 0 & 0 & 0 & 0 \\ 0.22 & 0.53 & 0 & 0 & 0 & 0 \\ 0.00 & 0.36 & 0.54 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.36 & 0.75 & 0 & 0 \\ 0.15 & 0.10 & 0.09 & 0.08 & 1.00 & 0 \\ 0 & 0.01 & 0.01 & 0.17 & 0 & 1.00 \end{bmatrix} \tag{2}$$

$$S9F9 = \begin{bmatrix} 0.32 & 0 & 0 & 0 & 0 & 0 \\ 0.40 & 0.42 & 0 & 0 & 0 & 0 \\ 0.00 & 0.40 & 0.40 & 0 & 0 & 0 \\ 0.02 & 0.01 & 0.48 & 0.55 & 0 & 0 \\ 0.26 & 0.17 & 0.11 & 0.11 & 1.00 & 0 \\ 0 & 0.01 & 0.01 & 0.34 & 0 & 1.00 \end{bmatrix} \tag{3}$$

$$F9S10 = \begin{bmatrix} 0.63 & 0 & 0 & 0 & 0 & 0 \\ 0.22 & 0.55 & 0 & 0 & 0 & 0 \\ 0.00 & 0.34 & 0.57 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.35 & 0.75 & 0 & 0 \\ 0.15 & 0.11 & 0.08 & 0.08 & 1.00 & 0 \\ 0 & 0.00 & 0.00 & 0.17 & 0 & 1.00 \end{bmatrix} \tag{4}$$

$$S10F10 = \begin{bmatrix} 0.34 & 0 & 0 & 0 & 0 & 0 \\ 0.40 & 0.40 & 0 & 0 & 0 & 0 \\ 0.00 & 0.40 & 0.41 & 0 & 0 & 0 \\ 0.00 & 0.02 & 0.47 & 0.59 & 0 & 0 \\ 0.26 & 0.18 & 0.11 & 0.11 & 1.00 & 0 \\ 0 & 0.00 & 0.01 & 0.30 & 0 & 1.00 \end{bmatrix} \tag{5}$$

$$F10S11 = \begin{bmatrix} 0.63 & 0 & 0 & 0 & 0 & 0 \\ 0.23 & 0.55 & 0 & 0 & 0 & 0 \\ 0.00 & 0.34 & 0.57 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.36 & 0.75 & 0 & 0 \\ 0.14 & 0.11 & 0.07 & 0.07 & 1.00 & 0 \\ 0 & 0.00 & 0.00 & 0.18 & 0 & 1.00 \end{bmatrix} \tag{6}$$

$$S11F11 = \begin{bmatrix} 0.35 & 0 & 0 & 0 & 0 & 0 \\ 0.38 & 0.41 & 0 & 0 & 0 & 0 \\ 0.00 & 0.41 & 0.42 & 0 & 0 & 0 \\ 0.02 & 0.00 & 0.46 & 0.58 & 0 & 0 \\ 0.27 & 0.18 & 0.12 & 0.10 & 1.00 & 0 \\ 0 & 0.01 & 0.00 & 0.32 & 0 & 1.00 \end{bmatrix} \tag{7}$$

$$F11S12 = \begin{bmatrix} 0.62 & 0 & 0 & 0 & 0 & 0 \\ 0.22 & 0.56 & 0 & 0 & 0 & 0 \\ 0.00 & 0.33 & 0.55 & 0 & 0 & 0 \\ 0.00 & 0.01 & 0.37 & 0.75 & 0 & 0 \\ 0.16 & 0.10 & 0.08 & 0.07 & 1.00 & 0 \\ 0 & 0.00 & 0.00 & 0.18 & 0 & 1.00 \end{bmatrix} \tag{8}$$

$$S12F12 = \begin{bmatrix} 0.35 & 0 & 0 & 0 & 0 & 0 \\ 0.34 & 0.39 & 0 & 0 & 0 & 0 \\ 0.00 & 0.40 & 0.40 & 0 & 0 & 0 \\ 0.00 & 0.01 & 0.48 & 0.57 & 0 & 0 \\ 0.31 & 0.20 & 0.12 & 0.11 & 1.00 & 0 \\ 0 & 0.01 & 0.00 & 0.33 & 0 & 1.00 \end{bmatrix} \tag{9}$$

$$F12S13 = \begin{bmatrix} 0.58 & 0 & 0 & 0 & 0 & 0 \\ 0.24 & 0.53 & 0 & 0 & 0 & 0 \\ 0.00 & 0.36 & 0.56 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.37 & 0.76 & 0 & 0 \\ 0.18 & 0.11 & 0.07 & 0.06 & 1.00 & 0 \\ 0 & 0.00 & 0.00 & 0.18 & 0 & 1.00 \end{bmatrix} \tag{10}$$

The variation between the fall to spring transitions and the spring to fall transitions are quite extensive while spring to fall transitions are similar to one another and similarly for the fall to spring ones. We can see that the graduation probabilities are much higher for the spring to fall transitions, due to the majority of students who graduate doing so in the spring. We also see that the dropout probabilities are higher for the spring to fall transitions. This may be because many students finish out a year and then lose their financial aid due to bad grades, decide to transfer to another university the following academic year, personal or financial reasons, etc. The dropout to dropout and graduate to graduate entries in the matrix are set to 1.00 because once a student falls into one of these classes, it's an absorbing class, and therefore, they do not return to the system and remain either graduated or dropped out. The entries also indicate that not very may students make large jumps in class such as going from a freshman to a senior in one semester, but the data does indicate that it is possible, as we see in the $S11F11$ matrix.

We also use the median matrices of all of the fall to spring transitions and all of the spring to fall transitions in our calculations. We found the median representations for our entries, but then had to adjust them proportionally to maintain a column sum of one. We include the probability transition matrices in equations (11, 12) as well as the exact number of students in each category in equations (13, 14).

$$PmedFS = \begin{bmatrix} 0.63 & 0 & 0 & 0 & 0 & 0 \\ 0.22 & 0.54 & 0 & 0 & 0 & 0 \\ 0.00 & 0.35 & 0.56 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.36 & 0.75 & 0 & 0 \\ 0.15 & 0.11 & 0.08 & 0.07 & 1.00 & 0 \\ 0 & 0.00 & 0.00 & 0.18 & 0 & 1.00 \end{bmatrix} \tag{11}$$

$$PmedSF = \begin{bmatrix} 0.34 & 0 & 0 & 0 & 0 & 0 \\ 0.38 & 0.41 & 0 & 0 & 0 & 0 \\ 0.00 & 0.40 & 0.40 & 0 & 0 & 0 \\ 0.00 & 0.01 & 0.47 & 0.58 & 0 & 0 \\ 0.28 & 0.18 & 0.12 & 0.10 & 1.00 & 0 \\ 0 & 0.00 & 0.01 & 0.32 & 0 & 1.00 \end{bmatrix} \tag{12}$$

$$medFS = \begin{bmatrix} 2024 & 0 & 0 & 0 & 0 & 0 \\ 739 & 1229 & 0 & 0 & 0 & 0 \\ 8 & 758 & 1374 & 0 & 0 & 0 \\ 5 & 9 & 877 & 2856 & 0 & 0 \\ 451 & 228 & 178 & 272 & 1 & 0 \\ 0 & 1 & 2 & 648 & 0 & 1 \end{bmatrix} \tag{13}$$

$$medSF = \begin{bmatrix} 816 & 0 & 0 & 0 & 0 & 0 \\ 900 & 856 & 0 & 0 & 0 & 0 \\ 2 & 855 & 957 & 0 & 0 & 0 \\ 3 & 13 & 1085 & 2307 & 0 & 0 \\ 616 & 381 & 277 & 405 & 1 & 0 \\ 0 & 0 & 12 & 1238 & 0 & 1 \end{bmatrix} \tag{14}$$

Here for the freshman class in equation (11), there is a probability of .15 that they will drop out at the end of fall semester and a .28 probability of dropping out after spring semester. This probability goes down as students progress through classes. When a student is a senior, for instance, there is only a .07 probability of dropping out in the fall semester and a probability of .10 during spring. Even though the probability is much lower, the number of students who are classified as seniors is larger than that for the freshman class, so this still represents a fairly large number of students as seen in equation (13), where there are 3776 seniors and 3227 freshmen.

We also created a boxplot of all the entries of all of the probability transition matrices so that we could see the variation among them. We exclude entries that are mostly zero. This gives us the plots below. We can see that some values have a greater variation than others. For example, there is a large variation in the dropout and graduation rates for seniors in the plot for spring to fall transitions, while there is little variation in the number of freshman returning to being freshmen for the spring to fall transitions.

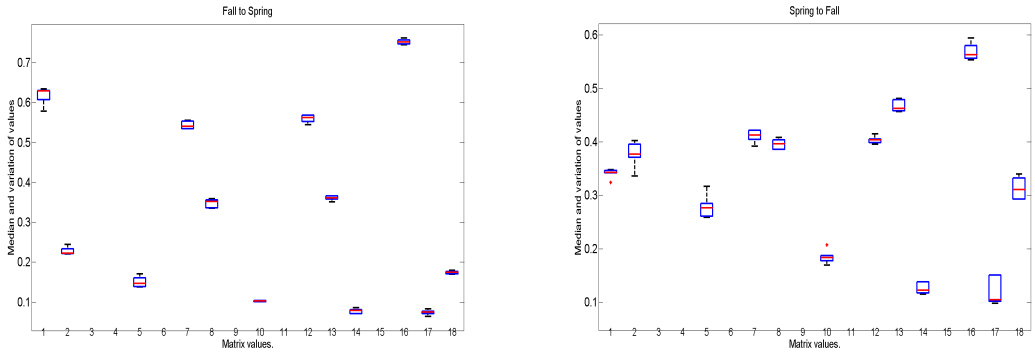

Figure 4: Box plot of the entries of our probability transition matrices. Each value on the x-axis is an entry in the matrix. They are in chronological order from top left to bottom right and proceeding by going down the column then to the next one such as $(1,1),(1,2)\ldots,(6,5),(6,6)$.

Next we move on to attempting to observe the year effects of our model. In our particular example we have to multiply two probability matrices together in order to advance

one year. We must first go from fall to spring, then from that spring to the next fall. Multiplying these two matrices together in reverse order gives us our matrix for a one year advancement.

$$YR = medSF * medFS = \begin{bmatrix} 0.22 & 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0.23 & 0 & 0 & 0 & 0 \\ 0.09 & 0.36 & 0.23 & 0 & 0 & 0 \\ 0.00 & 0.17 & 0.47 & 0.44 & 0 & 0 \\ 0.36 & 0.24 & 0.18 & 0.15 & 1.00 & 0 \\ 0.00 & 0.00 & 0.12 & 0.41 & 0 & 1.00 \end{bmatrix} \tag{15}$$

We can see that if a student starts as a freshman at the beginning of the year, there is a .22 probability, that student will still be a freshman at the end of the year and a .33 probability that they will be a sophomore. There is a much smaller probability that these students will become juniors by the end of the year. However, this is a possibility if the student came into college with a large number of credits and then took overloads during their first year. The more worrisome probability is having a .36 probability of not returning after one year. For the students beginning the year as sophomores we see that there is a .23 probability of remaining a sophomore and a .36 probability of becoming a junior. There also is a slightly smaller probability of becoming a senior. The probability for dropping out decreases, but still remains at .24 The junior students have a probability of .23 to remain a junior and an outstanding probability of .47 of becoming seniors. Here we see an even lower dropout probability of .18 and the first probability of graduating at .12. Finally, the senior students have a probability of .44 to remain in college as seniors. They also have the lowest dropout probability of .15 and the highest graduation probability of .41.

4 Predicted Long-Term Trends

One of the goals is to use the model to determine long-term trends and especially the probability of successfully graduating in four to five years. We can use our YR matrix in Eq. (15) to determine the probabilities of dropping out or graduating after four or five years. To determine what is likely to happen after four years given this model, we can simply take

$$YR^4 = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.02 & 0.00 & 0 & 0 & 0 & 0 \\ 0.04 & 0.03 & 0.00 & 0 & 0 & 0 \\ 0.11 & 0.11 & 0.08 & 0.04 & 0 & 0 \\ 0.68 & 0.51 & 0.38 & 0.25 & 1.00 & 0 \\ 0.14 & 0.34 & 0.54 & 0.70 & 0 & 1.00 \end{bmatrix}. \tag{16}$$

This version of the four year matrix is based off of a median of all of the transition matrices for spring and fall. We can also calculate the four year matrix by using the exact transition matrices. This is given by

$$F12S13 * S12F12 * F11S12 * S11F11 * F10S11 * S10F10 * F9S10 * S9F9 * F8S9 =$$

$$FourYr = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0.00 & 0 & 0 & 0 & 0 \\ 0.03 & 0.01 & 0.00 & 0 & 0 & 0 \\ 0.10 & 0.09 & 0.06 & 0.03 & 0 & 0 \\ 0.72 & 0.55 & 0.40 & 0.27 & 1.00 & 0 \\ 0.15 & 0.34 & 0.54 & 0.70 & 0 & 1.00 \end{bmatrix}.$$

The difference in this matrix and the one in equation (16) is that this one is calculated using the exact matrices for each semester rather than the median matrices over and over. We see there are no drastic differences in the exact matrix entries for the four year transition and the four year transition matrix using median approximations. Here we have that in four years freshmen have a probability of .72 of dropping out, while there is a probability of .15 to graduate. There is also a few small probability of still being enrolled in various classes. In the case of sophomores we can see that there is a probability of .55 to dropout and a .34 probability to graduate. Once again there are a few small probabilities to be elsewhere. Nearly all of the juniors either dropout or graduate with probabilities of .40 and .54 respectively. The seniors have the best chance of graduating with a probability of .70 and the lowest chance of dropping out with a probability of .27. We can also determine the probability transition matrix for five years

$$YR^5 = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.06 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.07 & 0.07 & 0.04 & 0.02 & 0 & 0 \\ 0.71 & 0.54 & 0.39 & 0.26 & 1.00 & 0 \\ 0.20 & 0.39 & 0.57 & 0.71 & 0 & 1.00 \end{bmatrix}.$$

Once again using the same idea as before with the exact matrices we obtain

$$F13S14*S13F13*F12S13*S12F12*F11S12*S11F11*F10S11*S10F10*F9S10*S9F9*F8S9 =$$

$$FiveYr = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.01 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.06 & 0.04 & 0.02 & 0.01 & 0 & 0 \\ 0.72 & 0.55 & 0.40 & 0.27 & 1.00 & 0 \\ 0.21 & 0.40 & 0.58 & 0.71 & 0 & 1.00 \end{bmatrix}.$$

One can see that there is a slight difference between the two, but that comes with estimating things using the median. In the exact matrix we have very small variations in the probability values, such as the freshman dropout probability being higher by .01. These differences are not extreme, so we use the median matrices in our other calculations. More students simply fall into the absorbing states, graduating and dropping out. It seems more fall into the dropout category than graduating, as we see a rise in freshmen dropping out. We can take this method even further to determine the steady state of the matrix. We have

$$YR^{50} = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0 & 0 \\ 0.73 & 0.56 & 0.40 & 0.24 & 1.00 & 0 \\ 0.27 & 0.44 & 0.60 & 0.76 & 0 & 1.00 \end{bmatrix}.$$

Since the top portion of the matrix all has zeros, the steady state is given in rows 5 (for dropout) and 6 (for graduation). This matrix shows the steady state of the model. Here we have that freshman have a probability of 0.73 to dropping out and a probability of 0.25 of graduating. While sophomores have a probability of 0.55 to dropping out and a probability of 0.43 of graduating. Juniors have a probability of 0.40 to dropping out and a probability of 0.59 of graduating, and seniors have a probability of 0.26 to dropping out and a probability of 0.72 of graduating.

One way we observe long-term trends is by looking at the mean time to absorption. If a Markov chain has an absorbing state, then one would be interested in the expected time to absorption and the probability of absorbing in a particular state if there is more than one absorbing state. In order to find the mean time to absorption we must first rearrange the matrix into the form

$$P = \begin{bmatrix} I & A \\ 0 & T \end{bmatrix}, \tag{17}$$

where I is the identity, A contains the rows with the absorbing states, and T is the rest of the probabilities. We must then find what is known as the fundamental matrix of our Markov chain. It is found by the equation $F = (I - T)^{-1}$, where T is as before [13]. The expected time to absorption vector, M, is found by summing up the columns of F. When performing these calculations we use median matrices for the transition from fall to spring (denoted FS) and spring to fall (denoted SF). Thus our one year matrix is found by $YR = SF * FS$ as in Eq. (15).

Using the matrix YR for our model, we can see that T and A are given by

$$T = \begin{bmatrix} 0.22 & 0 & 0 & 0 \\ 0.33 & 0.22 & 0 & 0 \\ 0.09 & 0.36 & 0.23 & 0 \\ 0.00 & 0.17 & 0.47 & 0.43 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 0.36 & 0.24 & 0.18 & 0.15 \\ 0.00 & 0.00 & 0.12 & 0.41 \end{bmatrix}$$

Therefore, the mean time to absorption vector can be calculated as

$$M = \begin{bmatrix} 2.71 & 2.74 & 2.35 & 1.75 \end{bmatrix}. \tag{18}$$

As a typical freshman, the expected time to graduation would be somewhere between 4 or 5 years or possibly longer; however, in our model the mean time to absorption for freshmen is approximately 3 years. This means that the mean time to either graduation or dropping out is about three years which would indicate the drop out must have a large effect on this mean time to absorption. The mean time to absorption for the sophomores and juniors seems fit the appropriate time to graduate, but the senior value of nearly 2 suggests that the majority of students do not graduate in four years. Thus the dropout rate also affects the other classes. Below we actually explore the probability of absorption into the graduation class versus the dropping out class. If we take the first through fourth elements of the fifth and sixth rows from out one year matrix, we have the dropout and graduation probability vectors. Thus let

$$PD = \begin{bmatrix} \alpha_{Dr} & \beta_{Dr} & \gamma_{Dr} & \epsilon_{Dr} \end{bmatrix} = \begin{bmatrix} 0.36 & 0.24 & 0.18 & 0.15 \end{bmatrix} \tag{19}$$

$$PG = \begin{bmatrix} \alpha_{Gr} & \beta_{Gr} & \gamma_{Gr} & \epsilon_{Gr} \end{bmatrix} = \begin{bmatrix} 0.00 & 0.00 & 0.12 & 0.41 \end{bmatrix}. \tag{20}$$

Then we can multiply these vectors by F and obtain the probability of dropping out or graduating. These are

$$DR = \begin{bmatrix} 0.73 & 0.55 & 0.40 & 0.26 \end{bmatrix}$$
$$GR = \begin{bmatrix} 0.25 & 0.43 & 0.59 & 0.72 \end{bmatrix},$$

respectively. So we have that the probability for a freshman, sophomore, junior, and senior to dropout is 73%, 55%, 40%, and 26% respectively. We also have that the probability of a freshman, sophomore, junior, and senior graduating is 25%, 43%, 59%, and 72% respectively.

From this long term analysis of our year matrix we can see that the values in the vectors DR and GR above are the same as those of our steady state. However, analysis of the actual data indicated that 1085 students graduated in the five years out of the 3079 total freshman. This indicates that the data shows that 35% of the freshman present in Fall 2008 graduated by Spring 2013 (5 years). This is different than what the model indicates; the model indicates a much lower graduation rate of only about 21% in five years. There are a number of things that could have caused this. It may be that at this regional university students do tend to drop out for one or two semesters and either work or go to a larger university, but some do return. We might also need to choose an alternative method for

calculating the entries in the probability transition matrix such that the five year trend is a better match instead of the semester-to-semester trend. Another possibility is that more classes need to be added to the model, or that this model simply is not robust enough to handle the complex reasons which must be included in order to accurately describe university progression. We did expect the values to be of this nature, leaning toward a higher probability of dropping out than graduating, but we did not expect it to be so drastic.

5 Sensitivity of Parameters

Even though there are discrepancies in the model which must be accounted for in future models, we can still investigate the sensitivity of the results on the specific transition probabilities given in the model. In our sensitivity analysis the outputs we consider are the four and five year trends, and the steady state. In order to do this we first alter entries in the probability matrix by a certain percent value. When altering the values we only change one at a time such as the freshmen dropout rate. We must keep the sum of the column equal to one, so we alter all of the other values in the column proportionately to their current fraction of the sum. Next we evolve the matrix, or raise the year matrix to a high degree, until it reaches its steady state. We then compare the altered matrix to the original matrix. We perform this analysis with the percent values of 1%, 2%, 3%, 5%, and 10%. The probability change of each output is calculated by the equation $(new - old) * 100$. We use Matlab to calculate all of these probability change values for the dropout and graduate probabilities after four years, five years, and in the steady state. We also find the probability change of the mean time to absorption. We conduct this analysis by changing the dropout rate by a certain percent, then we alter all other values in the column proportionately. We use our change which had the most influence as an example. For example the original steady state graduation matrix is

$$ORG = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0 & 0 \\ 0.73 & 0.56 & 0.40 & 0.24 & 1.00 & 0 \\ 0.27 & 0.44 & 0.60 & 0.76 & 0 & 1.00 \end{bmatrix}$$

We then show the same matrix with a 2%, 5%, and 10% respectively change in the spring-to-fall senior dropout rate only.

$$P2 = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0 & 0 \\ 0.73 & 0.54 & 0.38 & 0.24 & 1.00 & 0 \\ 0.27 & 0.46 & 0.62 & 0.76 & 0 & 1.00 \end{bmatrix}$$

$$P5 = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0 & 0 \\ 0.72 & 0.52 & 0.35 & 0.21 & 1.00 & 0 \\ 0.28 & 0.48 & 0.65 & 0.79 & 0 & 1.00 \end{bmatrix}$$

$$P10 = \begin{bmatrix} 0.00 & 0 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0 & 0 & 0 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0 & 0 \\ 0.69 & 0.48 & 0.28 & 0.15 & 1.00 & 0 \\ 0.31 & 0.52 & 0.72 & 0.85 & 0 & 1.00 \end{bmatrix}$$

We do this for the dropout rates for all classes in both the median matrix for the fall to spring transition and the median matrix for the spring to fall transition. We include the plots for a 2% change of the effects on the five year graduation rates and the steady state graduation rates in Figures 5 and 6.

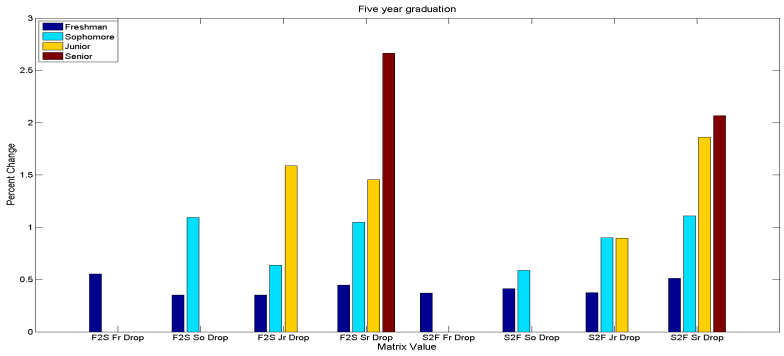

Figure 5: Bar plot of the probability change for the graduation rates after five years.

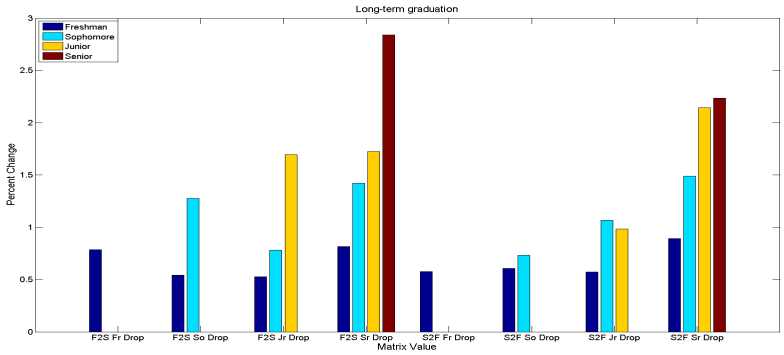

Figure 6: Bar plot of the probability change for the graduation rates for the steady state.

In the long term graduation the three changes that are the most influential on freshman success, in descending order, are decreasing the spring-to-fall senior dropout rate, the fall-to-spring senior dropout rate, and the fall-to-spring freshman dropout rate. All of these increase the graduation rate for freshman in the long-run by approximately 2%. For the five year graduation rate, the biggest effect is about a 1.5% increase in graduate rates for freshman when the fall 2 spring freshman drop-out rate is decreased. Once again, for sophomores, the spring-to-fall senior dropout rate has the largest long-term effect on the sophomore dropout rate followed closely by fall-to-spring senior dropout rate and fall-to-spring sophomore dropout rate (with 3-4% increases in the graduation rates). In the five-year, these same three have the most effect on five-year increase graduation success by over 2.5%. For juniors the most influential change in the long-term and five-year graduation rates is also if the spring-to-fall senior dropout rate is decreased (with almost 5% increase in the five-year rate and 5.5% increase in the long-term. In the long-term, both decreases in fall-to-spring senior and junior drop-out rates results in about a 4% increase in graduation rates. For the five-year, both of these have an effect of slightly less than 6% increase in graduation. For seniors, the only effect which would influence them is the senior dropout rate with fall-to-spring having more impact on success than spring-to-fall (7% increase versus about 5.5% increase).

6 Summary and Future Work

In summary, we were able to use data from a regional university to examine graduation and drop out rates. We first performed a preliminary analysis to determine potential classes to include in the model. It indicated that the only parameters that had any significance in whether or not students graduated was their overall GPA and being a member of the College of Business and Technology. The semester to semester preliminary analysis allowed us to further investigate these findings. This led to the conclusion that the only significant

thing was the overall GPA. We then developed a discrete-time Markov chain model in which we saw great variation in the probability transition matrix from fall to spring when compared to the transition matrix from spring to fall. However, the overall trend when comparing fall semesters or spring semesters only was relatively stable. We were able to use these matrices to analyze four year and five year graduation and dropout rates of the students, the steady state of the probability matrix, and the mean time to absorbtion for each class. We then performed a sensitivity analysis of the effect of variation in transition probabilities on both the drop out rate and graduation rates in four and five years as well as long term trends. We determined the parameter with the largest effect on the graduation rates of each class was the spring to fall senior dropout rate.

Based on our sensitivity analysis, one might explore ways to decrease the dropout rates of seniors to try to increase success toward graduation. However, during our analysis, we noticed discrepancies between the graduation rates that our model predicted and those observed exactly in the data. Therefore, in the future it is necessary to explore the reasons for this discrepancy in order to improve the model and its predictive capabilities. Moreover, future models should also directly include new and transfer student enrollment and its impact on the success towards graduation.

References

[1] John Bound, Michael Lovenheim, and Sarah Turner "Why have college completion rates declined? An analysis of changing student preparation and collegiate resources.", *American Economic Journal: Applied Economics*, American Economic Association, 2010

[2] Thomas P. Vartanian and Philip M. Gleason "Do neighborhood conditions affect high school dropout and college graduation rates?" *The Journal of Socio-Economics*, Elsevier Inc, 1999

[3] University of Memphis Office of Institutional Research, "Enrollment Data", www.memphis.edu/oir/enrollment/index.php, May 2014

[4] Institute of Education Sciences, "FAST FACTS", http://nces.ed.gov/fastfacts/display.asp?id=98, October 2014

[5] Tim Omarzu "College enrollment slides: Nationwide, student numbers are down by nearly 1 million", http://www.timesfreepress.com/news/2014/oct/13/college-enrollment-slides-nationwide-student/, Chattanooga Publishing Company, Inc, 2014

[6] Betsy O. Barefoot, "Higher education's revolving door: confronting the problem of student drop out in US colleges and universities.", *Open Learning: The Journal of Open*, 2004

[7] William G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis.", *Interchange*, Kluwer Academic Publishers, 1970

[8] Erica J. Gosman, Betty A. Dandridge, Michael T. Nettles, and A. Robert Thoeny "Predicting student progression: The influence of race and other student and institutional characteristics on college student performance." *Research in Higher Education*, Kluwer Academic Publishers, 1983

[9] John P Bean "Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome." *American Educational Research Journal*, 1985

[10] C. T. Haan, D. M. Allen, and J. O. Street "A Markov Chain Model of daily rainfall." *Water Resources Research*, 1976

[11] L. Rabiner "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, IEEE, 1989

[12] Philip D. ONeill "A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods." *Mathematical Biosciences*, Elsevier Science Inc, 2002

[13] Robert P. Dobrow "Introduction to Stochastic Processes with R" John Wiley & Sons, Inc., 2015