Markov Chain Enrollment Model
Presentation Notes
NCAIR 2023

Nathan Hodges

April 4, 2023

Abstract

This presentation will be a summary and extension to an AIR publication on the topic (Article 147) that was released in Fall 2019. The Markov Chain has been a tool used in modeling since the early 1900's. Unlike some modern machine learning algorithms, which can appear more like a "black box" that produces a value; the results of the Markov Chain can be explained through student graduation and drop out rates along with the relative class sizes and new student headcounts. The process is quick and simple to explain relative to machine learning models and requires only two years of enrollment data to yield a projection. During my presentation I will show how the projections from this model have outperformed the UNC system office's enrollment projections for undergraduate students at WCU over the past five years. In addition to an explanation of the model I will also be including the SQL code used to create the projections, which uses connections to SDM data. Thus, any IR office should be able to go home and run the code for themselves to compare to system office projections.

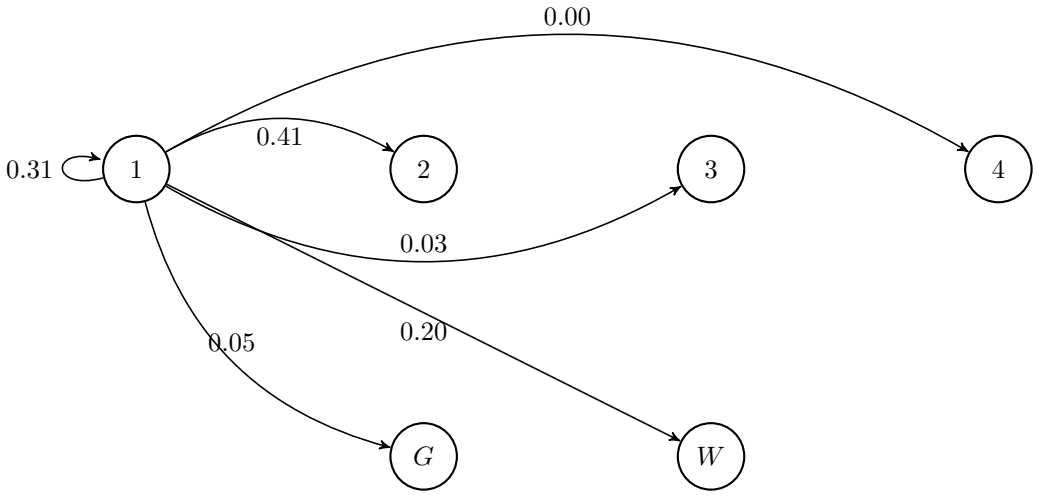
1 Introduction

A Markov chain is a mathematical model used to describe a system that changes over time in a probabilistic way. The model consists of a set of states and a set of probabilities that describe how the system moves from one state to another. At any given time, the system is in one of these states, and the probability of moving to any other state depends only on the current state and not on any previous states. This is called the "memoryless" property, and it allows us to model a wide variety of systems, including weather patterns, stock prices, and even the behavior of people. The Markov chain is named after the Russian mathematician Andrey Markov, who first described the concept in the early 20th century. - Chat GPT (verified/unedited by actual human Nathan Hodges 20230331.)

A usually discrete stochastic process (such as a random walk) in which the probabilities of occurrence of various future states depend only on the present state of the system or on the immediately preceding state and not on the path by which the present state was achieved - Merriam-Webster Dictionary 20230331

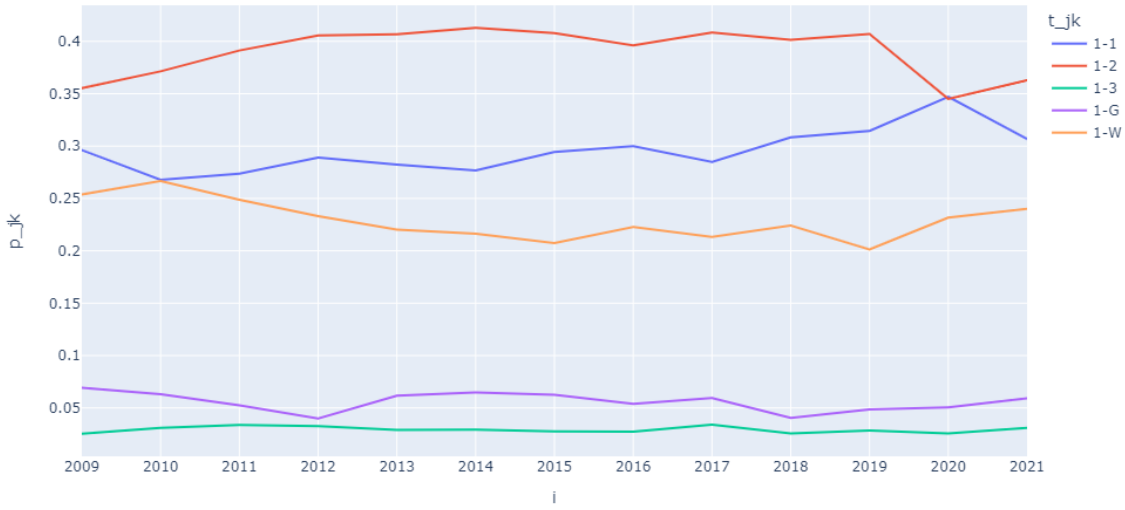
Here is a real example of a transition diagram from fall 2019 to fall 2020 for students who were enrolled in fall 2019 with 0-30 credit hours at WCU. You can see that all possible transitions are shown, the *probability* of transitioning from state 1 to a different state after one year is simply the **proportion** of students from state 1 that transitioned from state 1 to a different state in 2020.

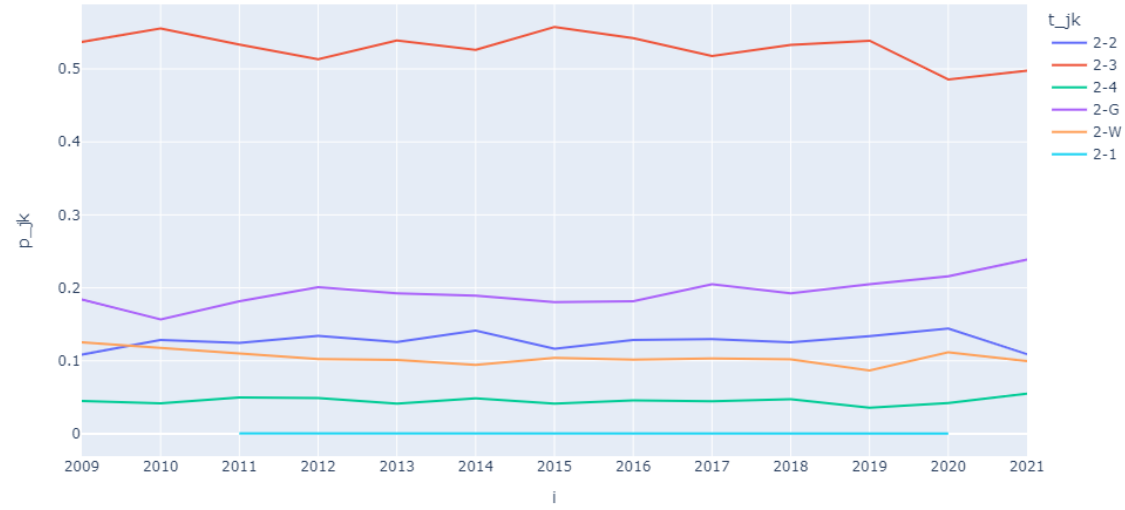
- state 1: 0 - 30 SCH
- state 2: 31 - 60 SCH
- state 3: 61 - 90 SCH
- state 4: $>$ 90 SCH
- state G: Graduate
- state W: Withdraw

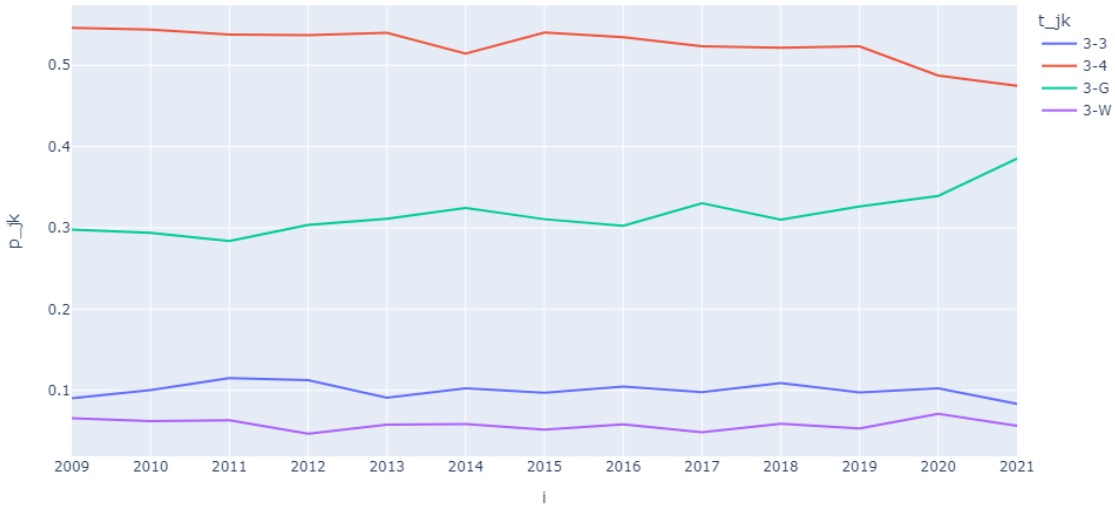


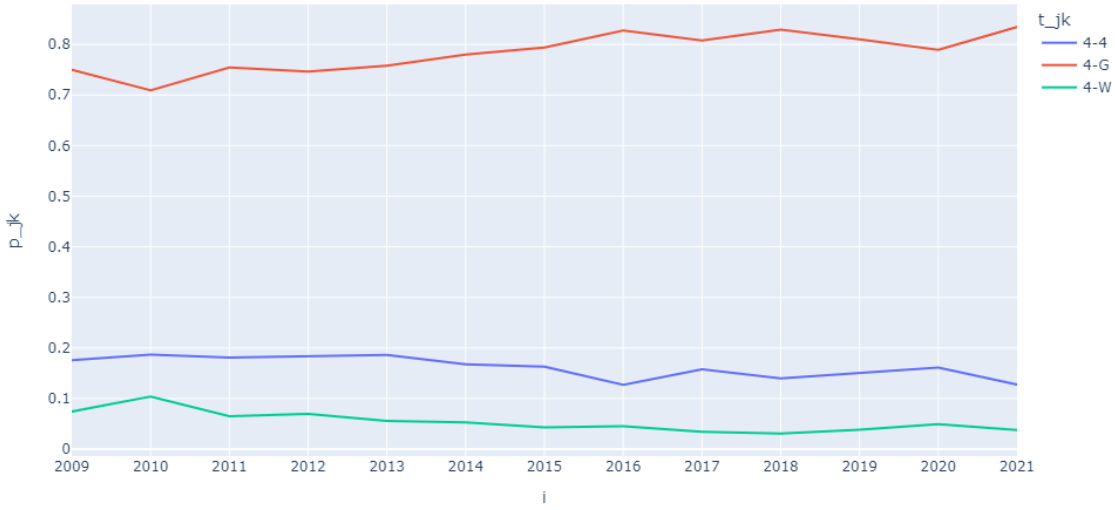
$$p_{11} + p_{12} + p_{13} + p_{14} + p_{1G} + p_{1W} = 1$$
$$0.31 + 0.41 + 0.03 + 0.00 + 0.05 + 0.20 = 1$$

The model simply assumes that for the next year students will progress in the same manner. For example, if there are 100 students enrolled with 0-30 credit hours in fall 2020 than we would expect using this model that next year this cohort of students will consist of $(1, 2, 3, 4, G, W) = (31, 41, 3, 0, 5, 20)$. This is why the AIR publication by Rex, Gandy et. al.[1], titled their paper, *Detecting Leaky Pipes and the Bulge in the Boa*; because it shows you how many students are leaving and how many students are progressing through vaious levels of their career at your institution. The next question you might ask is what are the variances of these transition probabilities. Below is a chart which shows how the probabilities from state one have evolved over time. As you can see from the large swings in year 2020 this model is susceptible to extreme changes in human behavior.

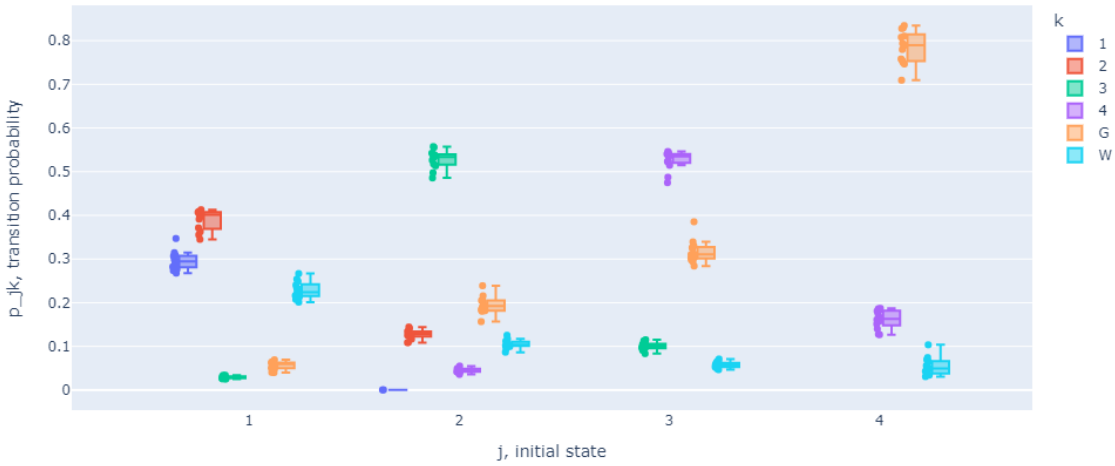








Transition Probability Distribution Summary

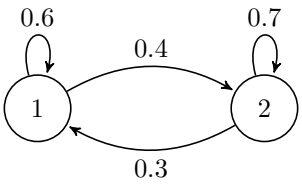


2 Markov Chain

2.1 General Markov Chain Introduction

2.1.1 Basic Markov Process

A Markov chain is a stochastic theory which can be used to model a probabilistic system that has a number, lets say n states or phases. Moving from one state to another is referred to as a transition, the probability of moving from state j to state k can be written as p_{jk}. For a system to be considered stochastic the probability of moving from state j to state k must not depend on the previous state of the system. This property, referred to as the Markov property allows a system to be modeled with a Markov chain. Below is a diagram of a system that satisfies these conditions.



In this system there are two states. The probability of transitioning between the states is labeled on each arrow. So for example the probability of transitioning from state one to state two is 0.4, you could write this as $p_{12} = 0.4$. The probability of transitioning from one state to another is only dependent on which state the system is currently in. All of these transitions can be represented in a matrix like the one below.

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

Note that all of the possible transitions from a specific state are defined and the probabilities add up to 1; that is $p_{11} + p_{12} = 1$. The current state of the system can be defined by a vector v which has n elements and represents the probability of being in a specific state. So if the current state of the system is known to be in state one, than this vector can be written as $(1, 0)$. Where the probability of being in state one is 1 and the probability of being in state two is 0. With the transition matrix defined and the initial state vector known, one can determine the probability of the state after any number of transitions by multiplying the initial state vector by the transition matrix. Here is an example showing how to calculate the final state vector after three transitions.

$$\mathbf{v}_1 = \mathbf{v}_0 P = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix} = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}$$

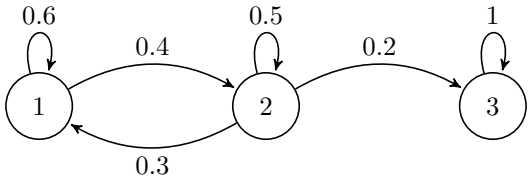
$$\mathbf{v}_2 = \mathbf{v}_1 P = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix} = \begin{pmatrix} 0.48 & 0.52 \end{pmatrix}$$

$$\mathbf{v}_3 = \mathbf{v}_2 P = \begin{pmatrix} 0.48 & 0.52 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix} = \begin{pmatrix} 0.444 & 0.556 \end{pmatrix}$$

You can see that after one step the state vector v_1 is $(0.6, 0.4)$, meaning that there is a probability 0.6 of being in state one and there is a probability of 0.4 of being in state two. Repeating the process for each state vector we see that after three transitions from state one there is a probability of 0.444 that the system would be in state one and a probability of 0.556 that the system would be in state two.

2.1.2 Absorbing Markov Chain

In this section we will discuss a type of Markov chain that has a state in which once the system is in that state it can never leave that state. This is referred to as an absorbing state. Below is a phase diagram and transition matrix for a system with two absorbing states and two transient states.

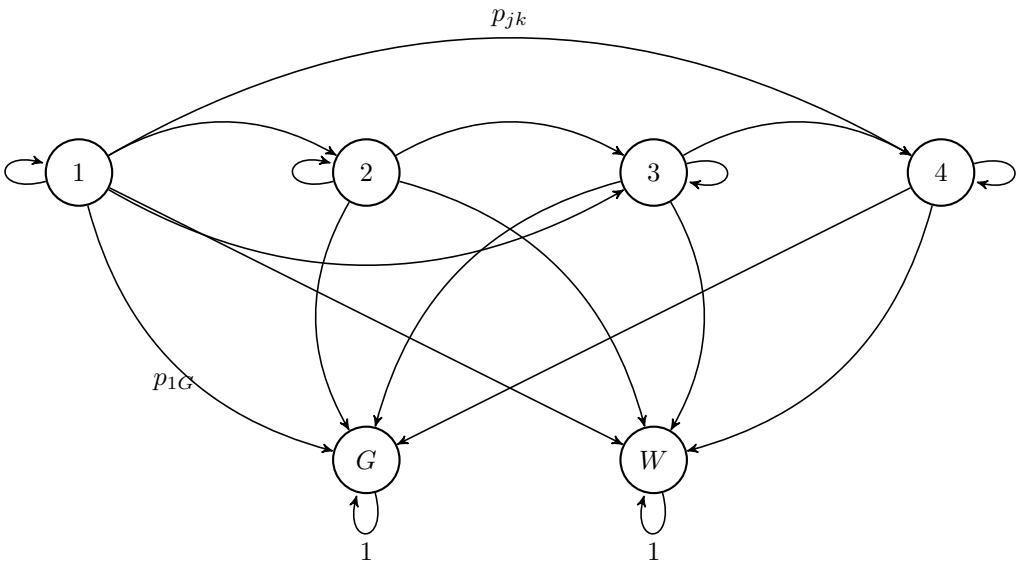


2.2 Applying Markov Chain to Undergraduate Enrollment

The transition period for this modeled will be from one fall term to the next fall term. The different states of a student are determined by the cumulative amount of credit hours that the student has earned while enrolled or whether or not they withdrew or graduated from the university. In this model I chose to use the cumulative institutional credit hours earned (SCH), that is the number of credit hours earned at the university which excludes credit hours earned through AP courses in high school or transfer credit from other institutions. Other works use different credit hour classifications.[1] There are several reasons why I decided to use institutional credit hours. The first reason is that this data seemed more complete and reliable. The second reason is that since we are tracking the progress of students through our institution it made sense to track credit earned at the intuition. The third reason is that after comparing the results of total (transfer and institutional) credit to credit earned through the institution alone; it appeared that using institutional credit hours yielded better results. I chose to use the student credit hour groupings based on the expected amount of credit completion per year, determined by two semesters each completing 15 hours per semester.[1]

- 1: 0 - 30 SCH
- 2: 31 - 60 SCH
- 3: 61 - 90 SCH
- 4: $>$ 90 SCH
- G: Graduate
- W: Withdraw

An important assumption used in this model that should be noted is that I assume that a student who re-enrolls at the institution is treated the same as any other new student with a certain level of credit hours. Put another way, I do not calculate the probability of students re-enrolling after they withdraw. There are several reasons for this, but the most important being complexity. Some papers show how this can be done, but would require you to track the number of terms that pass since the student was last enrolled.[3] Below is a phase diagram illustrating the model.


p_{jk}
1
2
3
4
p_{1G}
G
W
1
1

3 Results and Performance

Once the model is defined you can calculate the probability of a student transitioning between these various states using enrollment data for each term. Note that the transition probabilities will be different each term. This is determined by many factors known and unknown in that time period including changes in admission policies, environmental impacts, government policies etc... There are many reasons that students may decide to continue or withdraw from an institution. To see the process I used in determining these transition probabilities skip to the next section. In this section we will look at the results and performance of the model.

3.1 Transition Matrices

As noted above there will be a transition matrix for each term, in this section I will focus on the matrices for fall terms $2018, 2019, 2020$. As an example, the transition matrix for fall 2019 to fall 2020 is denoted T_{19-20}.

$$\begin{bmatrix} p_{11} & p_{21} & p_{31} & p_{41} & 0 & 0 \\ p_{12} & p_{22} & p_{32} & p_{42} & 0 & 0 \\ p_{13} & p_{23} & p_{33} & p_{43} & 0 & 0 \\ p_{14} & p_{24} & p_{34} & p_{44} & 0 & 0 \\ p_{1G} & p_{2G} & p_{3G} & p_{4G} & 1 & 0 \\ p_{1W} & p_{2W} & p_{3W} & p_{4W} & 0 & 1 \end{bmatrix}$$

$$T_{19-20} = \begin{pmatrix} 0.308 & 0 & 0 & 0 & 0 & 0 \\ 0.402 & 0.125 & 0 & 0 & 0 & 0 \\ 0.026 & 0.533 & 0.109 & 0 & 0 & 0 \\ 0.000 & 0.047 & 0.522 & 0.140 & 0 & 0 \\ 0.040 & 0.193 & 0.310 & 0.829 & 1 & 0 \\ 0.224 & 0.102 & 0.059 & 0.031 & 0 & 1 \end{pmatrix}$$

$$T_{19-20} = \begin{pmatrix} 0.314 & 0 & 0 & 0 & 0 & 0 \\ 0.407 & 0.134 & 0 & 0 & 0 & 0 \\ 0.028 & 0.539 & 0.097 & 0 & 0 & 0 \\ 0.000 & 0.036 & 0.523 & 0.151 & 0 & 0 \\ 0.049 & 0.205 & 0.326 & 0.810 & 1 & 0 \\ 0.201 & 0.087 & 0.053 & 0.039 & 0 & 1 \end{pmatrix}$$

$$T_{20-21} = \begin{pmatrix} 0.347 & 0 & 0 & 0 & 0 & 0 \\ 0.345 & 0.144 & 0 & 0 & 0 & 0 \\ 0.026 & 0.486 & 0.102 & 0 & 0 & 0 \\ 0.000 & 0.042 & 0.487 & 0.161 & 0 & 0 \\ 0.051 & 0.216 & 0.339 & 0.789 & 1 & 0 \\ 0.232 & 0.112 & 0.071 & 0.049 & 0 & 1 \end{pmatrix}$$

$$T_{(i,i+1),j,k} = \begin{bmatrix} p_{11} & p_{21} & p_{31} & p_{41} & 0 & 0 \\ p_{12} & p_{22} & p_{32} & p_{42} & 0 & 0 \\ p_{13} & p_{23} & p_{33} & p_{43} & 0 & 0 \\ p_{14} & p_{24} & p_{34} & p_{44} & 0 & 0 \\ p_{1G} & p_{2G} & p_{3G} & p_{4G} & 1 & 0 \\ p_{1W} & p_{2W} & p_{3W} & p_{4W} & 0 & 1 \end{bmatrix}$$

$$\bar{h_{(i+1),j}} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ 0 \end{bmatrix}$$

$$\bar{n_{(i+1),j}} = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \\ 0 \\ 0 \end{bmatrix}$$

$$T_{20-21} \cdot \bar{h_{21}} =$$

$$\begin{bmatrix} 5154 \\ 2219 \\ 1630 \\ 1142 \\ 0 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 0.347 & 0 & 0 & 0 & 0 & 0 \\ 0.345 & 0.144 & 0 & 0 & 0 & 0 \\ 0.026 & 0.486 & 0.102 & 0 & 0 & 0 \\ 0.000 & 0.042 & 0.487 & 0.161 & 0 & 0 \\ 0.051 & 0.216 & 0.339 & 0.789 & 1 & 0 \\ 0.232 & 0.112 & 0.071 & 0.049 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 3384 & 82 & 54 & 30 & 0 & 0 \end{bmatrix}$$

$$H_{(20,21),j,k} = \begin{bmatrix} 1788 & 1 & 0 & 0 & 0 & 0 \\ 1778 & 320 & 0 & 0 & 0 & 0 \\ 132 & 1078 & 167 & 0 & 0 & 0 \\ 0 & 93 & 794 & 184 & 0 & 0 \\ 261 & 479 & 553 & 901 & 1 & 0 \\ 1194 & 248 & 116 & 57 & 0 & 1 \end{bmatrix}$$

$$= \hat{H}_{(21,22),j,k}$$

$$\hat{h}_{(22,j)} = \begin{bmatrix} 1789 \\ 2099 \\ 1377 \\ 1071 \\ 2194 \\ 1615 \end{bmatrix} + \begin{bmatrix} 3384 \\ 82 \\ 54 \\ 30 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5173 \\ 2181 \\ 1431 \\ 1101 \\ 2194 \\ 1615 \end{bmatrix}$$

$$H_{(20,21),j,k} = \begin{bmatrix} 1769 & 1 & 0 & 0 & 0 & 0 \\ 1759 & 378 & 0 & 0 & 0 & 0 \\ 131 & 1272 & 173 & 0 & 0 & 0 \\ 0 & 110 & 823 & 179 & 0 & 0 \\ 258 & 565 & 573 & 877 & 1 & 0 \\ 1181 & 293 & 120 & 55 & 0 & 1 \end{bmatrix}$$

$$5173 + 2181 + 1431 + 1101 = 9886$$

4 Data Structure

The first issue is the typical retention data problem. Where we need to fill the gaps in enrollment records relative to terms.

Term	ID	Credit Hours	Level
2014	1	15	1
2015	1	45	2
2014	2	30	1
2016	2	30	1
2014	3	15	1
2015	3	20	1
2016	3	61	3
2014	4	10	1
2015	4	20	1
2016	4	61	3

Table 4.1: Enrollment records ordered by ID and Term.

Term
2014
2015
2016

Table 4.2: Available terms.

Term	ID	Credit Hours	Level
2014	1	15	1
2015	1	45	2
2016	1	Null	Null
2014	2	30	1
2015	2	Null	Null
2016	2	30	1
2014	3	15	1
2015	3	20	1
2016	3	61	3
2014	4	10	1
2015	4	20	1
2016	4	61	3

Table 4.3: Joining these tables gives the needed withdraw & exited terms.

Term	Next Term	ID	j	k
2014	2015	1	1	2
2015	2016	1	2	W
2014	2015	2	1	W
2014	2015	3	1	1
2015	2016	3	1	3
2014	2015	4	1	1
2015	2016	4	1	3

Table 4.4: Join table 4.3 to itself by shifting the terms.

Term	Next Term	j	k	n_{jk}
2014	2015	1	1	2
2014	2015	1	2	1
2014	2015	1	W	1
2015	2016	1	3	2
2015	2016	2	W	1

Table 4.5: Group table 4.4 by term and state transition.

Term	Next Term	j	k	p_{jk}
2014	2015	1	1	2/4
2014	2015	1	2	1/4
2014	2015	1	W	1/4
2015	2016	1	3	1
2015	2016	2	W	1

Table 4.6: Divide the number of n_{jk} by the total number of students in that state.

Term	Next Term	j	k_1	k_2	k_3	k_w
2014	2015	1	2/4	1/4	0	1/4
2015	2016	1	0	0	1	0
2015	2016	2	0	0	0	1

Table 4.7: To put this table in a matrix form pivot the table.

As you can see from this example, the data is not complete enough to create a Markov chain.

$$\begin{bmatrix} p_{11} & p_{21} & p_{31} & p_{41} & 0 & 0 \\ p_{12} & p_{22} & p_{32} & p_{42} & 0 & 0 \\ p_{13} & p_{23} & p_{33} & p_{43} & 0 & 0 \\ p_{14} & p_{24} & p_{34} & p_{44} & 0 & 0 \\ p_{1G} & p_{2G} & p_{3G} & p_{4G} & 1 & 0 \\ p_{1W} & p_{2W} & p_{3W} & p_{4W} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ 1 \\ 1 \end{bmatrix} \tag{4.1}$$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \tag{4.2}$$

Resources

[1] Rex Gandy, Lynne Crosby, Andrew Luna, Daniel Kasper, and Sherry Kendrick (2019) *Enrollment Projection Using Markov Chains: Detecting Leaky Pipes and the Bulge in the Boa*, AIR Professional File, Fall 2019 Article 147.

[2] Zachary T. Helbert (2015) *Modeling Enrollment at a Regional University using a Discrete-Time Markov Chain*, Undergraduate Honors Theses. Paper 281. https://dc.etsu.edu/honors/281

[3] Alenka Brezavšček, Mirjana Pejić Bach, Alenka Baggia (2017) *Markov Analysis of Students' Performance and Academic Progress in Higher Education*, Organizacija, Volume 50