ENROLLMENT PROJECTION USING MARKOV CHAINS: DETECTING "LEAKY PIPES" AND THE "BULGE IN THE BOA"

Rex Gandy, Lynne Crosby, Andrew Luna, Daniel Kasper, and Sherry Kendrick

Abstract

While the use of Markov chains is widely used in business and industry, their use within higher education has been sporadic. Furthermore, when used to predict enrollment progression, the majority of these models use student level as the classification variable. This study uses grouped earned credit hours to track the movement of students from one academic term to the other to better identify where students enter or leave the institution. Results from this study indicate a high level of predictability from one year to the next and the use of the credit hour flow matrix can aid administrators in identifying trends and anomalies within the institution's enrollment management process.

Introduction

The current challenges facing higher education administrators create myriad reasons to find a crystal ball of sorts in order to effectively forecast enrollments, predict how many current students will stay at the institution, forecast new students, and adequately estimate revenues. These challenges have only become more pressing.

Over 20 years ago when public college and university revenues were ample, administrators were not readily concerned about the future of college enrollments or student persistence. State appropriations were healthy and usually made up more than half of an institution's revenue source. Moreover, with lower tuition, more students could afford to obtain a degree without going into significant financial debt (Coomes, 2000).

Currently, however, the costs to run higher education have skyrocketed causing institutions to seek out scarce resources within an ever-diminishing financial pool. As states tackle other pressing issues such as infrastructure, entitlements, and prisons, the amount they give to higher education naturally wanes. Decreased state revenue, therefore, compels institutions to increase tuition to make up the difference. According to Seltzer (2017), for every $1,000 cut from per-student state and local appropriations, the average student can be expected to pay $257 more per year in tuition and fees. He further noted the rate is rising.

In addition to decreases in state revenues, higher education administrators are under increasing pressure to be accountable to federal and state governments as well as regional and discipline-based accreditors. This accountability is increasingly seen in tougher reporting standards, outcomes-based funding formulae, and mandated student achievement thresholds.

The closest resource to a crystal ball available to administrators is a set of mathematical prediction tools. These prediction tools range from simple formulae contained in spreadsheets to much more complicated regression, ARIMA (autoregressive integrated moving average), and econometric time series models.

According to Day (1997), current predictive tools that are statistically based rely on the ability to access and manipulate large datasets as well as individual student-record data. While more complicated statistical models incorporate variables such as tuition cost, high school graduate numbers, economic factors, and labor-market demand, other models look more specifically at institutional indicators such as high school grade point averages of entering freshmen, as well as retention, progression, and graduation rates of students.

One such model, the Markov chain, has been relatively under-utilized as an enrollment projection tool in higher education but, when used properly, can aid institutions in determining progression of students. Specifically, Markov Chains are unique from more traditional ARIMA and regression prediction tools in that:

1. They are capable of giving accurate enrollment predictions with only previous year's data which can be helpful when large longitudinal databases are not available.
2. They can be used to generate predictions on segments of a group of students rather than the entire population which are often required for other models.
3. The almost intuitive nature of the Markov Chain lends well to changes in student flow characteristics which can oftentimes not be explained by a complex statistical formula.

Moreover, Markov chains might be particularly helpful in determining progression of students during benchmark years when enrollments vary significantly due to state mandates/policies or institutional changes in admission standards. The purpose of this study is to show how a public, southeastern, Masters Level (Larger Programs) institution utilized the unique properties of this model to create a tool to better understand credit hour flow and student persistence.

Enrollment Management's Problem with "Leaky Pipes" and the "Bulge in the Boa"

While enrollment management has clearly evolved since its inception in the 1970's, some fundamental processes have essentially stayed the same. Institutions have always wanted to attract the right students who fit well within the institution's role, scope, and mission. Once matriculated into the institution, there is also a strong desire for students to adequately progress through their program and graduate within a reasonable amount of time (Kurz &

Scannell, 2006). As enrollment management developed through time, however, administrators became increasingly aware that college-age students were more difficult to enroll, higher tuition was causing some students to forego their degree, and institutional loyalty was waning as students transferred to similar or different institutions. Furthermore, institutions have seen an increasing number of students who are not fully prepared for the rigors or college work, putting greater enrollment strain on institutions (Johnson, 2000).

After more than 40 years of enrollment management within higher education, it is not surprising that metaphorical associations have entered into the lexicon of the profession as administrators try to better understand and predict student matriculation, persistence, and graduation. For instance, Ewell (1985), referred to students progressing and moving throughout their program as student flow while Clagett (1991) discussed following the flow of student cohorts through to graduation. Luna (1999) used the concept of student flow to explain the various pathways by which students may be retained at the institution, and Torraco & Hamilton (2013) discussed the student flow of selected groups of minority students. Furthermore, many software companies have exploited the student flow metaphor to describe use of data to identify areas where leakage is present in student flow pipelines. It is easy, then, to see how the management of student retention can be associated with a pipeline and how administrators are busy trying to plug the leaks.

Markov chains are uniquely suited to identifying these leaks, because they are able to model student flow as a set of transitions between a number of states, much like a set of pipes with various inflows, outflows, and interconnections. In addition to projecting enrollments using the model, it is also possible to observe from year to year where students enter into the absorbed

state (do not return to the institution). Leakage within the SCH (student credit hour) flow pipeline occurs when students withdraw or stop out due to reasons that are academic, non-academic, or both. If the model can isolate where the major leaks occur, the institution can identify causes and work to retain and maintain the flow of students within the pipeline. These leaks in the student flow pipeline can be detected and monitored from term to term so that strategies may be developed to maintain a healthier flow.

Another colorful bit of jargon among enrollment management professionals is the idea of bulging enrollments. For example, Fallows & Ganeshananthan (2004) used the term "bulging of enrollments" to describe a significantly larger share of students needing financial aid or when, due to rising tuition costs, students "bulge" into less expensive two-year colleges. Herron (1988) used the term "bulge in the boa" to define instances of oversupply in student populations quickly entering the student flow pipeline much like a large meal enters a boa constrictor. Liljegren & Saks (2017) added that these bulges can significantly affect higher education and its future. These bulges occur when large groups of students suddenly enter higher education putting a strain on the student flow pipeline. As the bulge dissipates, it may redefine student flow for the future. With Markov chain models, these bulges in the system can also be monitored so that issues such as course offerings and instructor availability may be addressed.

Markov Chains and Higher Education

A Markov chain is a type of projection model, which was created by Russian mathematician Andrey Markov around 1906. It uses a stochastic (random) process to describe a sequence of

events in which the probability of each event depends only on the state attained in the previous event.

The Markov chain is a stochastic rather than a deterministic model. Unlike a deterministic process where the output of the model is fully determined by the parameter values and by sets of previous states of these values, a stochastic process possesses inherent randomness and the same set of parameter values and initial conditions can lead to different outputs.

Take, for example, the scenario of an individual returning home from work. In a deterministic process, there is only one route (route A) from work to home, and the amount of time to get home depends only on the variable speed of the driver. In a stochastic process, the individual will have multiple routes (A, B, and C) from which to choose, and each of the routes intersects the other routes at various points. The randomness of the process occurs when the individual combines routes to go home, if the choices at each intersection are made randomly. For example, the driver may take route A part of the time, followed by route C, then to route B, and back to A again, or take some completely different path. There are many random possibilities the individual may take to get home, leading to a variety of possible driving times.

Markov chains utilize transition matrices that represent the probabilities of transitioning from each possible state to each other possible state. These states can be absorbing or non-absorbing; non-absorbing states allow future transitions to other states while absorbing states do not.

Markov chains have been widely and successfully used in business applications, from predicting sales and stock prices to personnel planning and running machines. Markov chains also have been used in higher education, albeit with much less frequency.

In most studies where Markov chains were used in enrollment management, the various transitional states were categorized either by student classification or other simpler dichotomous measures. Given the strength of the Markovian stochastic process in generating student flow probabilities using data only from the previous year, the process of classifying students into other kinds of states could be appealing. Such states could include credit hours, student debt, and (on a more system-wide level) the transitioning from one institution or program to another. The possibilities are diverse.

One of the first to use Markov chains in determining enrollment projections was Oliver (1968) when he compared Markov chains to the much more established use (at that time) of grade progression ratios to predict enrollments at the University of California.

According to the study, enrollment forecasting made a prediction on the basis of historical information on past enrollment and admission trends. In determining a stochastic process, Oliver demonstrated that the fraction of students who leave one grade level (class status) *i* and progress to class status *j* is a fraction p_{ij} which could also be time dependent. These fractions p_{ij} can also be interpreted as random transition probabilities. He determined that the process allowed for contributions in one grade level which were identified by their origins such as prior grade level, returning to the same grade level, and new admissions (Oliver, 1968).

According to Hopkins and Massy (1981), the use of Markov chains allows the researcher to observe the flow of students from one classification level (i.e. freshman, sophomore, junior, and senior) to the next class level. The chain also incorporates students who stay at the same class level from one year to the next. Therefore, the Markov chain for class level, as studied by Hopkins and Massy, can be described as follows:

1. The number of students in class level *i* who progress to class level *j*;
2. The number of students in class level *i* who stay in the same level;
3. The number of students who leave the institution (dropout, stop-out, or graduate).

Similarly, Borden and Dalphin (1998) used Markov chains to develop a one-year enrollment transition matrix to track how students of each class level progressed. The authors found that unique Markov chain models were valuable in measuring student progression without having to rely on six-year graduation rate models, which could be ineffective due to the large time lags. Specifically, the model was built around a transition matrix where student flow was tracked from one year to the next, and the rates of transition from four non-absorption states (i.e. freshman to sophomore) were placed into a separate matrix than were the two absorption states (i.e. dropout, graduation).

Using the percentages in the two matrices, those students who continue in non-absorption states were processed through the matrix using the established rates of transition until, asymptotically, all students reach the final absorption state.

Additionally, Borden and Dalphin (1998) developed discrete Markov chain processes to simulate the effect of changes in student body profile on graduation rates. In these models, the

authors incorporated credit-load and grade performance categories. In their study, the results indicated that, while a strong association between grade performance and persistence existed, it took very large changes in levels of student performance to impact retention and graduation rates modestly.

In a more narrowly focused study, Gagne (2015) used Markov chains to predict how English Language Institute (ELI) students progressed through science, technology, engineering, and math (STEM) programs. Specifically, the model created transitional (non-absorbing) states based on classification level and three absorbing states to include those students who left the institution, who graduated from a STEM program, or who graduated from a non-STEM program. Findings from their study indicated that the ELI students tended to progress at a higher rate in STEM programs than non-ELI students and that ELI students who repeated the freshman year are more likely to repeat again than transition to the sophomore year.

Correspondingly, a study by Pierre and Silver (2016) used Markov chain models to determine the length of time it took students to graduate from their institutions. As with previous studies, students were divided into non-absorbing transitional states (i.e. freshman, sophomore, junior, and senior) and absorbing states (i.e. graduate and non-returning). Using the Markovian property, the future probability of transitioning from one state to another depended only on the present state of the process and was not influenced by its past history. The study found that it took 5.9 years for a freshman to graduate and 4.5 years for a sophomore to graduate from the institution.

Brezavšček, *et al* (2017), successfully used Markov chain models to investigate the pattern of students' enrollment and academic performance at a Slovenian higher education institution. The model contained five transient or non-absorbing states and two absorbing states. Using student records, a total of eight consecutive academic seasons were used and the students' progression towards the next stage of the program was estimated. From those transition percentages, progression, graduation, and withdrawal probabilities were obtained.

As mentioned earlier, most Markov chain models involving enrollment management and prediction use student classification to create the various states of the model. Using student classification in model specification, however, could create states that are overly broad in nature since, at most semester-based colleges and universities, student classification varies by 30 hours.

Ewell (1985), who also used Markov chains to predict college enrollments, noted two limitations of the models. First, because the estimation of the probabilities rests on historical data, Markov chains may be sensitive to the time the data were collected. This could be especially true with significant enrollment gains or declines from one year to the next. Second, according to Ewell, different subpopulations may behave in different ways necessitating the need to disaggregate into smaller groupings.

However, the Markov chain's attributes may allow a unique ability to detect the "leaks" and "bulges." Because this type of projection model uses the stochastic process to describe a sequence of events in which the probability of each event depends only on the state attained in the previous event, changes to student flow are immediate and are not subject to potentially

skewed results of the past. In short, the limitations mentioned by Ewell (1985) can be utilized when building the student flow matrices to detect significant shifts in enrollment and which groups of students are leaving the institution at a higher rate.

Methodology

The current study used Markov chains to predict fall enrollment at a southeastern, Masters Level (Larger Programs), public institution based on annual fall semester enrollment for degree-seeking undergraduates. The process involved obtaining data from the institution's student information system and separating students into groupings based on their cumulative credit hours earned. Student flow was measured from fall of year *i* to fall of year *i+1* based on whether students stayed within their credit hour category, moved into another credit hour category, or did not enroll at the institution. These student flow changes for each category were then summed and applied to year *i+2* as a prediction of enrollment.

Within the model, at a given point in time, each student has a particular state, and each student is treated as having a particular probability of transitioning to each other state or staying within the same state. Most of these states are based on the number of credit hours the student has accumulated, (i.e. the student's credit hour category). Because the SCH category of a student was determined by the number of cumulative credit hours a student earned, most of the credit hour flow scenarios included students advancing to a higher credit hour category or students withdrawing or graduating. While it is rare that a student will move from a particular credit hour category to a lower category, it can happen through the transfer process when certain

credit hours from a former institution may not be accepted at the current institution after the student has enrolled.

The characteristic that makes this model a Markov chain is the fact that a given student's transition probabilities between states are assumed to depend only on that student's current state and not on any of the student's previous states. This is a simplifying assumption, which allows all students within a given state to be treated similarly regardless of their histories; otherwise, the model would become much more complicated and difficult to apply.

The main parameters of the model are estimates of these transition probabilities. These transition probabilities are estimated by calculating the fractions of students that transitioned from each state to each other state relative to the number of students initially in that state in past years' enrollment data. The other parameters of the model are the fractions of new incoming students by credit hour category. The total number of new incoming students is assumed to be fixed, thus the estimated number of incoming students by credit hour category follows from these fractions.

The model process is recursive in that predictions for Fall X are produced from the enrollment data from Fall X-2 and Fall X-1 and the subsequent flow rates from Fall X-2 to Fall X-1.

The basic assumptions can now be described that were used to construct the predictive models:

1) Each model models flow from one year to the next and is named accordingly, e.g. Fall 2013 to Fall 2014 is known as the 13_14 model and is based on the starting data for Fall 2013 and the new student data from Fall 2014.

2) As the model is applied, the output headcount by SCH level for the $(i+1)^{th}$ year becomes the input headcount for the next iteration of the model.
3) When the model is applied to a "future" year, the total number of new students is assumed to be constant and the same as the number of new students for the $(i+1)^{th}$ year. The distribution of new students by SCH level is also assumed to be constant.
4) When the model is applied to a "future" year, it is assumed that the fractional student loss and fractional student continuation ratios are fixed by SCH level.
5) When the model is applied to a "future" year, it is assumed that the fractional flow from SCH level to SCH level is the same as for the year used to construct the model.

The model divides the undergraduates into 6-Student Credit Hour (SCH) groupings. This method uses historic ratios of SCH student subsets gathered from the student information system to predict future fall headcounts.

The 6-SCH grouping method was developed using SCH levels spaced by 6 SCH and a greater than 162 SCH level, creating a total of 28 levels. Using interval notation, the 6-Credit Hour categories were [0, 6]; (6, 12]; (12, 18]; (18, 24]; (24, 30]; (30, 36]; (36, 42]; (42, 48]; (48, 54]; (54, 60]; (60, 66]; (66, 72]; (72, 78]; (78, 84]; (84, 90]; (90, 96]; (96, 102]; (102, 108]; (108, 114]; (114, 120]; (120, 126]; (126, 132]; (132, 138]; (138, 144]; (144, 150]; (150, 156]; (156, 162]; and (162, ∞). From these data, six basic models were constructed that generated predictions which can be verified against actual enrollment data: 1) 10_11 model, 2) 11_12 model, 3) 12_13 model, 4) 13_14 model, 5) 14_15 model, and 6) 15_16 model.

The 6-SCH groupings used in this model are individually less broad than the more familiar student classification levels. However, it is possible to aggregate the 6-Credit Hour bins into a version of these student levels, which we define as:

Freshmen	≤30 SCH
Sophomore	>30 SCH and ≤60 SCH
Junior	>60 SCH and ≤90 SCH
Senior	>90 SCH

Note that these classification level definitions do not exactly match the institution's definitions. In using student credit hour (SCH) groupings, the enrollment pipeline may be much more finely observed and enrollment patterns among students may be more precisely distinguished. While it is the goal of this study to develop a model to predict the coming fall enrollment once the previous fall enrollment is known, the model will not address enrollment by major, academic department, or college.

Model Description

Once the groupings or states were created, student information system data from a given fall semester were used to parse out students into the various student credit hour categories. These students were then tracked during the following fall semester to determine student flow percentages. Within this study, student flow states are defined as:

1. Students in credit hour group *j* who stayed within that group
2. Students in credit hour group *j* who moved to a different credit hour group
3. Students in other credit hour groups who moved to group *j*
4. Students who were no longer enrolled at the institution

Within this model, the following terms and symbols are used:

1) *n* is the number of SCH levels in the model (n=28 for the 6-SCH groupings).

2) h_{ij} is the i^{th} Fall semester headcount for the j^{th} SCH level.
3) H_i is the total undergraduate headcount for the i^{th} semester.
4) l_{ij} is the number of the h_{ij} subset students not enrolled the next Fall semester.
5) L_i is the total number of undergraduates enrolled in the i^{th} Fall semester that are not enrolled in the $(i+1)^{th}$ Fall semester.
6) $c_{ij} = h_{ij} - l_{ij}$ is number of continuing students in the j^{th} SCH level.
7) C_i is the total number of undergraduates that enrolled in the i^{th} Fall semester that are also enrolled in the $(i+1)^{th}$ Fall semester.
8) d_{ijk} is the number of the continuing c_{ij} subset students that move from SCH level j to SCH level k from the i^{th} Fall to the $(i+1)^{th}$ Fall.
9) w_{ij} is the number of the C_i subset students that flow from all other levels into level j.
10) o_{ij} is the number of the c_{ij} subset students that flow out of level j into all other levels.
11) $s_{(i+1)j}$ is the number of the new incoming students for the $(i+1)^{th}$ Fall semester where j is the SCH level.
12) $N_{(i+1)}$ is the total number of incoming new undergraduate students for the $(i+1)^{th}$ semester.

With this terminology in place, the previously stated assumptions of the models can now be described algebraically:

1) When applying a model to a "future" period from Fall *(i+1)* to Fall *(i+2)*, the total number of incoming students is assumed be the same as it was for the period used to build the model, so it is assumed to have the value N_{i+1}. The fraction of new students by SCH level for that upcoming year is also assumed to be the same as it was in the period used to train the models, so each is assumed to be $s_{(i+1)j}/N_{i+1}$. Therefore, the estimated number of new students for a particular SCH level in that "future" year can be obtained by multiplying the value of this fraction by the estimated total number of students in the "current" year. That is, the estimate for the number of new students in the future year for that particular SCH level is given by $s_{(i+1)j}/N_{i+1} \times N_{i+1} = s_{(i+1)j}$.

2) The fractional loss and fractional continuation ratios are also assumed to be fixed by SCH level, i.e. for a future year, these ratios are assumed to be l_{ij}/h_{ij} and c_{ij}/h_{ij}, the same as what they had been in the year used to build the model. Therefore, for the upcoming "future" period from Fall *(i+1)* to Fall *(i+2)*, the estimated number of lost and continuing students for the *j*th SCH level are obtained by multiplying these ratios by the number of students $h_{(i+1)j}$ in that SCH level in the "current" Fall *(i+1)*. This multiplication is $l_{ij}/h_{ij} \times h_{(i+1)j}$ to estimate lost students in the *j*th SCH level and $c_{ij}/h_{ij} \times h_{(i+1)j}$ to estimate continuing students in the *j*th SCH level.
3) Finally, the fractional flow from a particular SCH level to another SCH level is assumed to be fixed; i.e. for a future year, these ratios are assumed to be d_{ijk}/c_{ij}, the same as what they had been in the year used to build the model. Therefore, for the upcoming "future" period from Fall *(i+1)* to Fall *(i+2)*, the estimated number of students transitioning from SCH level *j* to SCH level *k* is given by the value of this ratio d_{ijk}/c_{ij} multiplied by the estimated number of continuing students in the *jth* SCH level.

The processes described above can be applied iteratively to obtain estimates for years even further into the future by using the estimated values from one iteration as inputs into the next iteration.

Using the terms and formula, a spreadsheet matrix was created (**Table 1**) that includes the various credit hour classifications as well as the non-absorbed transient student states as well as the absorbed state of no longer enrolled.

Table 1

Basic Structure Matrix of the Markov Chain Model

SCH Levels	Fall *i* Headcount	Lost	Continuing			Movement from SCH Level 1	2	.	.	*n*	Static	Inflow	Outflow	Fall *i*+1, New Students	Fall *i*+1 Headcount
1	h_{i1}	l_{i1}	c_{i1}	Movement to SCH Level	1	d_{i11}					d_{i11}	w_{i1}	o_{i1}	$s_{(i+1)1}$	$h_{(i+1)1}$
2	h_{i2}	l_{i2}	c_{i2}		2	d_{i12}	d_{i22}				d_{i22}	w_{i2}	o_{i2}	$s_{(i+1)2}$	$h_{(i+1)2}$
.	.	.	.		.	.	.	.			.	.	.	.	.
.	.	.	.		.	.	.	.	.		.	.	.	.	.
n	h_{in}	l_{in}	c_{in}		*n*	d_{i1n}	.	.	.	d_{inn}	d_{inn}	w_{in}	o_{in}	$s_{(i+1)n}$	$h_{(i+1)n}$

This table shows the basic structure matrix of the headcount SCH flow associated with the Markov chain model that connects the undergraduate headcount in the i^{th} Fall to the headcount in the $(i+1)^{th}$ Fall.

From this SCH flow structure, one can observe the relationships of credit hour flow between and among the various states, including flow into non-absorbing states (staying or moving into another credit hour state) or into absorbing states (not enrolling at the institution). Below, the relationships among the variables are listed:

1) $c_{ij} = \sum_{k=1}^{n} d_{ijk}$ represents those current students who were in SCH level *j* who stayed at the institution.
2) $o_{ij} = \sum_{\substack{k=1 \\ k \neq j}}^{n} d_{ijk}$ represents those current students who were in SCH level *j* who moved to all other SCH levels.

3) $w_{ik} = \sum_{\substack{j=1 \\ j \neq k}}^{n} d_{ijk}$ represents those current students who were in SCH levels other than *k* who moved to SCH level *k*.

4) $H_i = \sum_{j=1}^{n} h_{ij}$ represents semester headcount at fall semester *i*.

5) $L_i = \sum_{j=1}^{n} l_{ij}$ represents those students at fall semester *i* who did not re-enroll.

6) $C_i = \sum_{j=1}^{n} c_{ij}$ represents those students at fall semester *i* who did re-enroll.

The following relationship,

$$\sum_{k=1}^{n} w_{ik} = \sum_{j=1}^{n} o_{ij}$$

shows two equivalent ways of expressing the collection of students who remain at the institution and move from any SCH level to a different SCH level during the year. Conservation of student flow is obtained only when students from level *j* stay in SCH level *j*, move to other SCH levels, or when students from other SCH levels move into SCH level *j*.

Given these relationships, the number of undergraduates by level in the second Fall semester can be calculated using the following formula:

$$h_{(i+1)j} = h_{ij} - \ell_{ij} - o_{ij} + w_{ij} + s_{(i+1)j}$$

This is the number of total transient students in one of the SCH levels after one year, who were not absorbed by withdrawing or graduating. Therefore, the total number of students in the *(i+1)*th fall semester is simply given by:

$$H_{i+1} = H_i - L_i + N_{i+1}$$

since the inflow and outflow terms cancel upon summation.

Results

The model used actual data from a southeastern, Masters (Large Programs), public institution for fall 2010 through fall 2017. The enrollment for these eight years are displayed in **Table 2**.

Table 2

Annual Enrollment Data

Fall *i*	Fall *i* Headcount	Lost	Continuing	New	Fall (*i*+1) Headcount
Fall 2010	9,652	3,773	5,879	3,957	9,836
Fall 2011	9,836	4,082	5,754	3,721	9,475
Fall 2012	9,475	3,965	5,510	3,761	9,271
Fall 2013	9,271	3,843	5,428	3,574	9,002
Fall 2014	9,002	3,685	5,317	3,598	8,915
Fall 2015	8,915	3,792	5,123	3,993	9,116
Fall 2016	9,116	3,945	5,171	3,919	9,090
Fall 2017	9,090	not known	not known	not known	not known

Total enrollment data used, combined with more fine-grained student transition data, to construct various models.

In developing the Markov chain matrix for each year, the total number of students within each category were noted and tracked to the following year. Within this matrix, one can observe the various student states by each category to determine who is moving into transitional (non-absorbing) states and who is graduating or not returning. These more granular data within the matrix offer clues as to when students may be leaving the institution and where there are potential bulges in the system coming from new or transfer students.

Table 3 represents one such matrix, the 6-Credit Hour matrix from fall 2016 to fall 2017. The 28 6-SCH groupings are labeled down the left with the same corresponding 28 groupings across the center of the matrix. This table also contains headcount by groupings, how many within each grouping did not return, how many graduated and how many new students enrolled in fall 2017 but not fall 2016. Matrices such as this one can be examined to identify the aforementioned leaks and bulges in the enrollment pipeline.

The following labels are used in Table 3:

1) HC1: Fall 2016 census undergraduate enrollment excluding special groups.
2) HC2: Fall 2017 census undergraduate enrollment excluding special groups.
3) Lost: Enrolled fall 2016 but not in fall 2017. This includes students that graduated without re-enrolling, as a subset. When determining if the student returned fall 2017, only undergraduate students, excluding special groups, were considered.
4) Continuing: Enrolled in fall 2016 and fall 2017.
5) GradA16: Awarded an associate degree in Fall, Spring, or Summer of Academic Year 2016-2017. Note that only one degree is counted per student to avoid double-counting, with bachelor's degrees given precedence over associate degrees.
6) GradA16E: Awarded an associate degree and enrolled in next fall term in another degree program. These students are a subset of GradA16.
7) GradB16: Awarded a bachelor's degree in Fall, Spring, or Summer of Academic Year 2016-2017.
8) GradB16E: Awarded a bachelor's degree and enrolled in next fall term in another degree program. These students are a subset of GradB16.

9) Columns in the center indicate movement of continuing students from the fall 2016 SCH categories to the fall 2017 SCH categories. Note that the central portion of Table 3 does not include counts for students who enrolled both semesters but remained in the same SCH level; these counts are instead separately labeled *Static*.

10) Static: Enrolled in fall 2016 and fall 2017 and stayed in the same SCH level.

11) New: Enrolled in fall 2017 but did not enroll in Fall 2016. (NewUnder30Hrs and Transfer are subsets of New).

12) NewUnder30Hrs: New students with fewer than 30 hours.

13) Transfer: Transfer students.

Table 3

Fall 2016 to Fall 2017 6-Scredit Hour Matrix

SCH Level Number	SCH Level Definition	HC1 (Fall 2016)	Lost	Continuing	GradA16	GradA16E	GradB16	GradB16E	Movement to SCH Level Number	Movement from SCH Level Number																											Static	Inflow to	Outflow from	New	NewUnder30Hrs	Transfer	HC2 (Fall 2017)
										1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28							
1	[0, 6]	1589	597	992	0	0	0	0	1																													15	0	977	1606	1606	32	1621
2	(6, 12]	264	81	183	0	0	0	0	2	14																												5	14	178	239	239	23	258
3	(12, 18]	245	110	135	1	0	0	0	3	60	5																											6	65	129	195	195	46	266
4	(18, 24]	322	141	181	1	1	0	0	4	158	6	5																										5	169	176	161	161	37	335
5	(24, 30]	501	147	354	11	9	0	0	5	408	10	9	7		1																							3	435	351	130	113	60	568
6	(30, 36]	386	121	265	17	13	0	0	6	258	42	18	13	11																								6	342	259	116	0	57	464
7	(36, 42]	332	108	224	13	8	0	0	7	60	85	35	31	19	5																							5	235	219	103	0	45	343
8	(42, 48]	316	94	222	16	13	0	0	8	15	21	45	58	57	19	9																						3	224	219	79	0	32	306
9	(48, 54]	291	92	199	19	17	0	0	9	3	5	13	46	103	31	17	11		1																			4	230	195	85	0	38	319
10	(54, 60]	399	114	285	31	21	0	0	10		3	4	16	115	82	33	17	10																				1	280	284	116	0	74	397
11	(60, 66]	446	105	341	37	27	0	0	11	1	1		4	38	89	59	39	13	9																			2	253	339	141	0	103	396
12	(66, 72]	369	85	284	16	13	1	0	12				1	5	24	70	43	27	21	9																			2	200	282	105	0	63	307
13	(72, 78]	346	84	262	18	12	14	1	13					3	6	19	69	49	26	23	12																		1	207	261	107	0	56	315
14	(78, 84]	347	121	226	13	12	34	1	14						1	10	29	56	58	28	13	7																	1	202	225	96	0	45	299
15	(84, 90]	347	165	182	9	5	100	0	15						1	1	8	31	84	79	25	18	8															3	255	179	77	0	25	335
16	(90, 96]	400	248	152	11	6	175	0	16								3	7	71	123	79	34	17	2														5	336	147	87	0	24	428
17	(96, 102]	341	204	137	8	6	147	1	17							1			9	53	85	64	17	10	5				1								3	245	134	62	0	21	310
18	(102, 108]	323	230	93	5	1	188	5	18									1	5	20	46	79	84	28	13	5											3	281	90	46	0	19	330
19	(108, 114]	271	188	83	3	3	159	2	19											4	18	45	62	52	23	16	8										3	228	80	25	0	7	256
20	(114, 120]	237	171	66	5	5	135	3	20									1			4	5	28	60	49	19	10	6									2	182	64	50	0	19	234
21	(120, 126]	179	131	48	3	2	98	3	21													5	8	18	38	44	17	11	7								2	148	46	45	0	17	195
22	(126, 132]	154	116	38	1	1	90	2	22													2	1	8	16	27	18	11	12	3							4	98	34	41	0	12	143
23	(132, 138]	135	89	46	1	1	56	3	23															1	1	16	25	23	8	6	4						2	84	44	32	0	12	118
24	(138, 144]	113	79	34	0	0	48	1	24													1			2	4	12	22	21	6	4	5					1	77	33	20	0	2	98
25	(144, 150]	96	78	18	2	0	56	0	25													1				2		4	10	14	4	6	5				1	46	17	22	0	5	69
26	(150, 156]	61	45	16	4	2	32	0	26																			1	4	10	10	9	5	3			0	42	16	14	0	2	56
27	(156, 162]	49	25	24	1	1	21	1	27																			2	1	6	8	6	3	1			0	27	24	10	0	4	37
28	>162	257	176	81	4	4	124	2	28																	1				1	4	18	20	13	16	24		81	97	0	109	0	34	287

According to the table, in fall 2016, there were 1,589 students contained in the [0, 6] SCH group. Out of these, 597 did not return the next fall semester. A total of 408 of these students transitioned into the (24, 30] SCH group indicating that they were progressing normally while 232 transitioned into groups of 24 or fewer SCH. With a quick examination of the flow, it is clear to see that the majority of students are leaving within the SCH groupings that make up the freshmen and sophomore years. In the (84, 90] SCH grouping, 109 students graduated, and 5 of the students who graduated re-enrolled in fall 2017, meaning that 104 of the students who graduated did not re-enroll. A total of 165 students in the (84, 90] SCH grouping were lost (did not re-enroll), and subtracting the aforementioned 104 students leaves 61 students who neither graduated nor re-enrolled.

A total of 914 new transfer students entered for fall 2017, indicating a significant number of students who took some type of transfer credit. Many of these new transfers could constitute dual enrolled students who took both high school and college classes. The bulk of the new transfer students, however are entering with above 54 and less than 84 credit hours.

In observing the higher groupings, the table indicates that 865 students had accumulated more than 126 SCH and 448 (52%) graduated. 608 of the students who earned more than 126 hours did not re-enroll in the institution.

While this table only represents one of the six matrices created for this study, the possibilities of tracking student flow by groupings, classifications, or years are numerous. Moreover, it can be argued that the process of tracking student flow through transitional states within the

Markov process is somewhat intuitive and indicative of the strong predictive properties of the model.

Table 4 shows the predictions for the next three years, along with the actual data. The model was built using the flow of students over a particular academic year. There were six such academic years used for construction of the models. The columns of **Table 4** show the years for which an enrollment prediction was generated. As can be seen in the table, predictions for the 10_11 Model for both methods were over-specified by about 5% for fall 2012 and about 11% for fall 2014. The 11_12 Models produced better projections, coming within less than 1% of the actual values for all three years. The prediction of the 12_13 Model differed from the actual enrollment by an average of -0.2%. Results from the 13_14 model indicate that the prediction differed by an average of 3.1%. In most cases, predictions further into the future from the years used to train the models have greater residuals, which is to be expected in any forecasting problem.

Table 4

Actual Enrollment and Predictions

Model		Fall 12	Fall 13	Fall 14	Fall 15	Fall 16	Fall 17
Reality	Actual Headcounts	9,475	9,271	9,002	8,915	9,116	9,090
Model 10_11 6SCH	Predicted Headcounts	9,948	9,999	10,002			
	% Diff. from Actual	4.99%	7.85%	11.11%			
Model 11_12 6SCH	Predicted Headcounts		9,244	9,076	8,958		
	% Diff. from Actual		-0.29%	0.82%	0.48%		
Model 12_13 6SCH	Predicted Headcounts			9,105	8,980	8,903	
	% Diff. from Actual			1.14%	0.73%	-2.34%	
Model 13_14 6SCH	Predicted Headcounts				8,839	8,745	8,694
	% Diff from Actual				-0.85%	-4.07%	-4.36%
Model 14_15 6SCH	Predicted Headcounts					8,874	8,865
	% Diff. from Actual					-2.65%	-2.48%
Model 15_16 6SCH	Predicted Headcounts						9,258
	% Diff. from Actual						1.85%

The model predictions for the next three years (when actual data is available for comparison) for each of the models using the 6-Credit Hour methods.

Using the actual and predicted enrollment from **Table 4**, averages of the absolute values of the percentage differences between the actual and predicted values for enrollment were calculated. The percentage difference between the predicted and actual value is defined as:

$$\%\, difference\ = \frac{predicted\ value - actual\ value}{actual\ value} \times 100\%$$

Using the average value of the absolute values of these percentage differences, the predictive ability of the models can be examined, as these values show on average how far off the models were, regardless of sign. In a mathematical sense, the absolute value between two numbers is known as the standard Euclidean distance between two points and indicates the real distance

between two numbers (Bartle & Sherbert, 2011). The results as shown in **Table 5** clearly indicate that the predictive ability of the model decreases as number of years out from the years used to build the model increases, which is expected, much like the way weather forecasts further into the future are less accurate.

Table 5

Mean Absolute Value of Percent Differences by Years Out for 6-SCH Models

Prediction Timeframe	Mean Absolute Value of Percent Difference
One Year Out	1.96%
Two Years Out	3.19%
Three Years Out	4.57%

Based on the results from **Table 5**, the study will only examine one year out predictions, as these were the most accurate. The actual values are compared with those one year out predictions in **Table 6**. The predicted enrollment for Fall X in **Table 6** is produced from the enrollment data from Fall X-2 and Fall X-1 and subsequent flow rates from Fall X-2 to Fall X-1.

Table 6

The 6-SCH Models' One Year Out Predictions Compared to Actual Enrollment

		Freshmen	Sophomores	Juniors	Seniors	All Levels	Mean Absolute % Difference of Class Levels
2012	Actual	2,876	2,035	1,871	2,693	9,475	
	Predicted	3,114	2,090	1,966	2,778	9,948	
	% Difference	8.28%	2.68%	5.10%	3.17%	5.00%	4.81%
2013	Actual	2,729	1,890	1,870	2,782	9,271	
	Predicted	2,817	1,875	1,834	2,718	9,244	
	% Difference	3.23%	-0.79%	-1.92%	-2.30%	-0.29%	2.06%
2014	Actual	2,644	1,803	1,870	2,685	9,002	
	Predicted	2,709	1,800	1,789	2,807	9,105	
	% Difference	2.47%	-0.16%	-4.36%	4.53%	1.14%	2.88%
2015	Actual	2,533	1,944	1,738	2,700	8,915	
	Predicted	2,574	1,809	1,816	2,640	8,839	
	% Difference	1.60%	-6.93%	4.51%	-2.24%	-0.85%	3.82%
2016	Actual	2,921	1,724	1,855	2,616	9,116	
	Predicted	2,543	1,885	1,801	2,644	8,874	
	% Difference	-12.93%	9.36%	-2.89%	1.07%	-2.65%	6.56%
2017	Actual	3,048	1,829	1,652	2,561	9,090	
	Predicted	3,053	1,788	1,766	2,651	9,258	
	% Difference	0.17%	-2.24%	6.91%	3.51%	1.85%	3.21%
	Mean Absolute % Difference	4.78%	3.69%	4.28%	2.80%	1.96%	

Note that the six-year average of the absolute values of the percentage differences by class range from 2.8-4.7%. The 2016 freshman percent difference of -12.9% represents an outlier due to a major university initiative to increase new freshmen enrollment. This influx of new freshmen was significantly different than past years and clearly signals the bulge in the student flow pipeline as mentioned above. By utilizing the iterative process of producing Fall X projections from the enrollment data from Fall X-2 and subsequent flow rates from Fall X-2 to Fall X-1, the effect of this bulge in the system can be tracked into the future in order to plan upcoming course offerings.

By observing the predictive capabilities of the model, it is clear to see how administrators and enrollment managers can use these results to plan for classes and instructional personnel. Here, both annual projections and classification average projections for the five-year period were off by no more than 6.6%, which should fall within the margin of error for most larger institutions.

Furthermore, Monte Carlo simulation could be used to obtain enrollment predictions that give a range of plausible values instead of a single point estimate for a future year's enrollment. Monte Carlo simulations have been used in the context of higher education by Torres et al. (2018) to examine degree plans for potential bottlenecks. In applying these methods to this enrollment model, the fractions of students transitioning between specific levels would be treated more like the result of many coin flips than as fixed fractional values, and the ranges of predicted values could be obtained by repeated random simulation. This level of simulation was not performed in this study.

Conclusion

The use of Markov chains in projecting enrollment and the management thereof has gained popularity among higher education professionals. The short-term projections created by this stochastic process are unique to other time-tested forecasting tools used in enrollment management. When used properly, Markov Chains can aid institutions in determining progression of students that are different from more traditional ARIMA and regression prediction tools in that:

1. They are capable of giving accurate enrollment predictions with only two previous years' data which can be helpful when large longitudinal databases are not available
2. Can be used to generate predictions on segments of a group of students rather than the entire population which may be required for other models.
3. The almost intuitive nature of the Markov Chain lends well to changes in student flow characteristics which can oftentimes not be explained by a complex statistical formula.

By creating groupings and tracking students within those groupings by the state they transition into, the researcher can also get a better picture of what type of students are leaving and when they are leaving.

As shown in this study, the strong predictability of Markov chains allows administrators to better plan course scheduling and instructor demand while managing tight budgets. In this study, several predictive headcount models were developed using SCH flow as the annual driver. Eight years of fall enrollment data from the institution were used to develop the models. When applied to historical data each gives one year out predictions within a calculated level of

uncertainty. The models can easily be modified to change the new student input data, the continuation rates, and the inter-level flow rates, should that be desired. Furthermore, similar models could be used to track fall to spring retention as well as spring to fall retention.

References

Agevall, O. (2017). Social closures: on metaphors, professions, and a boa constrictor. In Liljegren, A. & Saks, M. (Eds.), Professions and metaphors: understanding professions in society. London: Routledge, Francis, and Taylor Group.

Bartle, R.G. and Sherbert, D.R. (2011). Introduction to real analysis. Urbana-Champaign, IL: University of Illinois.

Brezavšček, A.; Bach, M,; and Baggia, A (2017). Markov Analysis of students' performance and academic progress in higher education, 50(2), 83-95.

Borden, V.M.H. & Dalphin J.F. (1998). Simulating the effect of student profile changes on retention and graduation rates: A Markov Chain analysis. Paper presented at the Annual Forum of the Association for Institutional Research: Jacksonville, FL.

Clagett, C.A. (1991). Institutional research: the key to successful enrollment management. Office of Institutional Research. Largo, MD: Prince George's Community College.

Coomes, M.D. (2000). The historical roots of enrollment management. In Coomes, M.D. (Ed.). *The role student aid plays in enrollment management.* New Directions for Student Services, 89. San Francisco: Jossey-Bass.

Day, J.H. (1997). Enrollment forecasting and revenue implications for private colleges and universities. In Layzell, D.T. (Ed.) *Forecasting and managing enrollment and revenue: An overview of current trends, issues, and methods*. New Directions for Institutional Research, 93. San Francisco: Jossey-Bass.

Ewell, P.T.(1985). Recruitment, retention and student flow: A comprehensive approach to enrollment management. National Center for Higher Education Management Systems Monograph 7: Boulder, CO.

Fallows, J. & Ganeshananthan, V. (October, 2004). The big picture. The Atlantic. Retrieved from URL: https://www.theatlantic.com/magazine/archive/2004/10/the-big-picture/303520/.

Gagne, L (2015). Modeling the progress and retention of international students using Markov Chains. University of Akron Honors Research Projects 3: Akron, OH.

Herron, J. (1988). Universities and the myth of cultural decline. Detroit, MI: Wayne State University Press.

Hopkins, D.S.P, & Massy, W.F. (1981). Planning models for higher education. Stanford, Cal.: Stanford University Press.

Johnson, A.L. (2000). The Evolution of Enrollment Management: A Historical Perspective. *Journal of College Admission*, n. 166, p4-11.

Kurz, K. & Scannell, J. (2006). Enrollment Management Grows Up. *University Business*. Retrieved from URL: https://www.universitybusiness.com/article/enrollment-management-grows.

Luna, A. (1999). Using a matrix model for enrollment management. *Planning for Higher Education*, 27(3), 19-31.

Oliver, R.M. (1968). Models for predicting gross enrollments at the University of California. Ford Foundation Program for Research in University Administration: Berkeley, CA.

Pierre, C. & Silver, C (2016). Using a Markov Chain model to understand the behavior of student retention. In Callaos, N.; Chu, H.; Ferrer, J.; Fernandes, S.; & Belkis Sánchez, B. (Eds). The 7th International Multi-Conference on Complexity, Informatics, and Cybernetics/The 7th International Conference on Society and Information Technologies : Proceedings (pp. 248-251). International Institute of Informatics and Systemics: Winter Garden, FL.

Seltzer, R. (July 24, 2017). State Funding Cuts Matter. *Inside Higher Ed*. Retrieved from URL: https://www.insidehighered.com/news/2017/07/24/new-study-attempts-show-how-much-state-funding-cuts-push-tuition.

Torraco, R.J. & Hamilton, D.W. (2013). The Leaking U.S. Educational Pipeline and Its Implications for the Future. *Community College Journal of Research and Practice*, 37(3), 237-241

Torres, D., Crichigno, J., & Sanchez, C. (2018). Assessing Curriculum Efficiency Through Monte Carlo Simulation. *Journal of College Student Retention: Research, Theory & Practice.* Retrieved from URL: https://doi.org/10.1177/1521025118776618.