19th SEAAIR Conference "Outstanding Paper" Citation

APPLYING MARKOV-BASED FORECASTING IN ENROLMENT PLANNING

Amir H. Rouhi

Analytics and Insights, Finance and Governance, RMIT University, Melbourne (amir.rouhi@rmit.edu.au)

ABSTRACT

The education sector is a multidimensional and complex system, affected by numerous internal and external factors. Institutional planning in such a speculative environment demands appropriate tools, especially when forecasting and modeling the future is necessary. Predictive analytics can help executives to identify the likelihood of future outcomes of their institutions based on past and current data, as well as to consider internal and external influencing factors. Such analyses can utilize several approaches varying from simple statistical techniques, data mining, and predictive modeling tools to advanced machine learning algorithms. Selecting an appropriate yet effective model for two samples of enrolment planning is the goal of the current paper. The Markov Chain is a well-known technique to forecast stochastic time-series data and is used in the current research. The suggested model is a homogenous Markov Chain which is applied to modeling Course-enrolment. Generating the Transitional probability matrix is the core concept of the model. To achieve this, analyzing the historical data to identify all possible valid transitional states is the first essential phase. Calculating transitional probabilities among all states is the second major phase. We have utilized a frequentist approach to achieve the transitional probabilities. The rest is about computing the likelihood of possible future states by implementing different scenarios by way of tweaking the elements of the primary Transitional probability matrix and analyzing the results. In addition to its ability to forecast stochastic processes, another advantage of a homogenous Markov model is its simplicity in implementation.

Keywords: Student headcount prediction, Enrolment projection, Transition probability matrix, Predictive analytics, Homogeneous Markov Chain.

Introduction

International students are a significant cohort in major Australian institutions (DEFAT 2019). Respecting that fact, the quality and quantity growth of institutions in South East Asian countries will finally coincide with significant fluctuation in the demographics of student enrolment in Australian universities. From this will emerge the importance of predictive analyses for strategic and financial planning. This paper introduces a Markov-based model for forecasting institutional enrolment based on historical time-series data.

The result of recording sequential observations in a time sequence is called "*time series*". With this definition, a significant number of data sets can be categorized as time series and is the reason why analyzing time series data is important in a wide variety of disciplines such as engineering, economics, and business. The number of sick-leave requests in each month, the number of bottles of wine sold in a store (over any period), or the number of student enrolments in each semester, regardless of their differences in the context of data, are similar in this concept: all are regularly recordable in a time series.

Forecasting or estimating the future state is an integral part of time series analyses (Box & Ljung 2015). The observations in equal periods or equispaced intervals generate a sequence of discrete data in a balanced time interval: $T-3,\ T-2,\ T-1\ and\ T$ (current time). This characteristic brings the concept of predictability for the time $T+1, T+2$, etc. and it can be considered as an opportunity for more accurate planning and more effective strategies which can be crucial.

Mathematical modeling of physical phenomena is a well-established approach to study the dynamics of a system. If the mathematical model can calculate the exact components of the phenomenon and enable us to predict the future accurately, it is known as a deterministic model. However, the natural phenomena are not deterministic and are mostly under the influence of different external parameters and even unknown factors which can affect the accuracy of the model in the calculation of the component. Such models are known as probabilistic models or stochastic processes. Most behaviors observed in time series are the result of these stochastic models. In such time series, the current moment, including all the states of the internal component as well as the external parameters, plays the most important role in calculating the conditional probability distribution of the next events which leads us to predict the next step.

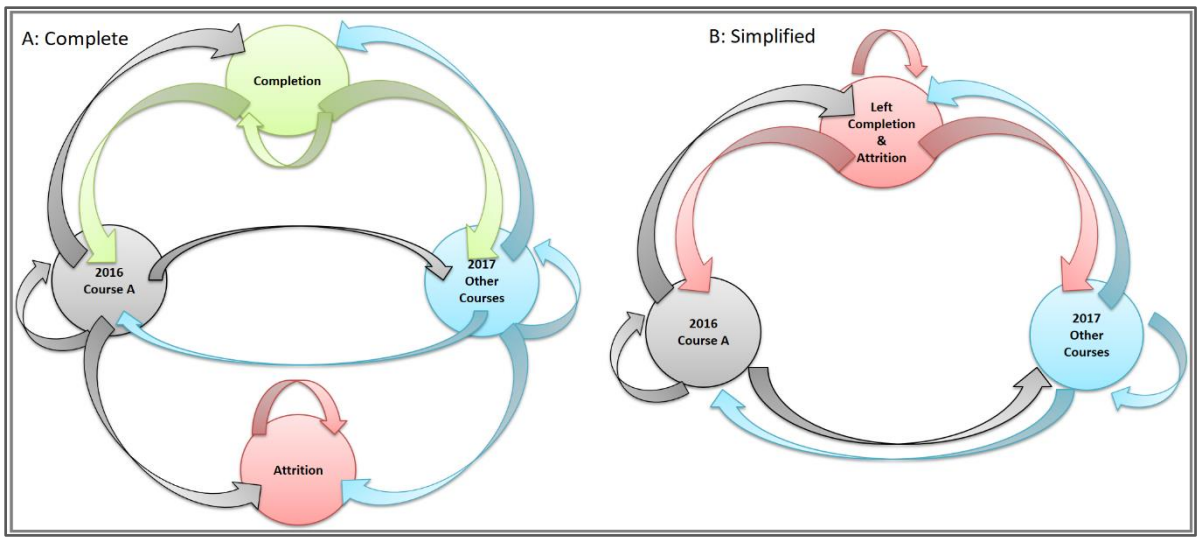


Figure 1: Enrolment states for a sample course: A from 2016 to 2017.

A: All the possible transitions and states. B: Simplified and practical version of the same transitions

In some time series, the conditional probability distribution of the next event is static and does not change over time. This statistical equilibrium, which can be distinguished by constant mean and variance, is the characteristic of stationary stochastic models (Tsay 2005). However, in some systems, where the dynamic of the system is high, and we can detect a moving average in the data set, they are categorized as non-stationary stochastic processes. Some models are involved in both stationary and non-stationary processes in real-world problems.

The Markov Chain is a series of discrete (finite and countable) values generated by the Markov process. This process is a stochastic process where the current state of the system is the only factor to predict the next state (Meyn, 2012). In other words, to generate the next state, past states are irrelevant if the current state is available. Metaphorically, we can label the Markov process as a "spiritual" approach in data analytics because, like most spirituality methods, the only important moment is the current moment and there is no recognizable pattern in the trend of events (stochasticity), as human life.

There are different types of Markov Chain. In a discrete-time Markov Chain, the state of the system changes in discrete time intervals. Most stochastic time series can be considered as a discrete-time Markov Chain (Feinberg & Schwartz, 2012). In such a sequence of random variables, each variable X_t in the chain, in time $t \geq 0$, the next variable in the sequence X_{t+1}, can take different values based on the conditional probability distribution of the current state of the data in the chain: $P(X_{t+1} | X_t)$. If these conditional probabilities remain the same for each sequence, the chain is known as a stationary (homogeneous) Markov Chain (Meyn, 2012).

In a continuous-time Markov Chain, the system condition can change in a continuous-time interval and the conditional probability of the next state is not relevant to the current state. However, we assume that the average time interval among events is known and follows a pattern such as a Poisson Process (Anderson, 2012).

The Markov Decision Process (MDP) is another kind of Markov Chain in which an agent can affect the conditional probability distribution of the next event. The focus of the current paper is using discrete-time homogeneous Markov Chain in institutional research.

Literature review

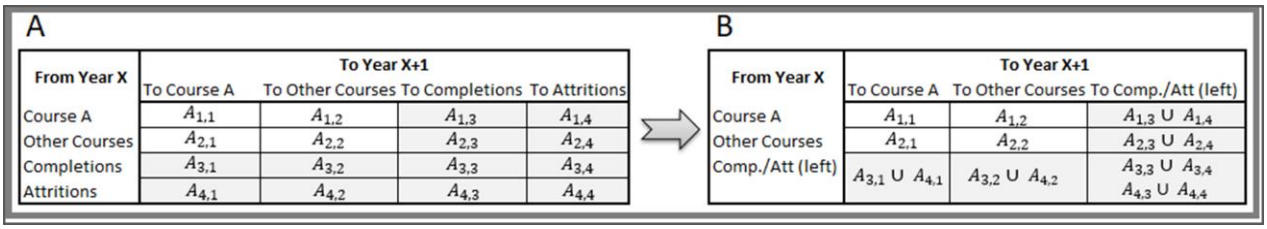
As explained earlier, unlike the deterministic approaches, the essence of the Markov Chain is randomness. Enrolment, as an institutional example, is a random process because we cannot deterministically say that 100% of the transition of a sample course, is reenrolling to the same course. There is always a probability of leakage, re-enrolment in other courses, or leaving. The same concept is applicable for the course intake. Such stochasticity in institutional events makes Markovian-based methods as appropriate tools for institutional applications and grabs the attention of IR researchers.

The University of California is one of the earliest institutions that utilized the Markov Chain for enrolment management (Oliver, 1968). In this research, the grade levels or class statuses construct the probability matrix and Oliver assumed that progress from a grade level to the same or another level is a random process and appropriate to utilize a Markov-based method.

Utilizing Markov Chain for investigating enrolment flow in higher classification levels (freshman, sophomore, junior and senior) was implemented at Stanford University (Hopkins and Massy, 1981). They considered three states for each student progressing to the next iteration and construct their transition matrix based on these states: 1- Number of students that stay in the same class, 2- Number of students that progress to the next class, and 3- Number of students that leave the institution, including attrition or graduates.

Yearly enrolment transition Borden and Delphin (1998) investigated the progression for each class level by their yearly transition matrix. They found that using the Markov Chain model is accurate enough to measure the student progress rate without relying on 6-year graduation rate models which need longer time lags. The current research is like Borden and Delphin (1998) for using yearly transition states. While they distinguished between absorbing states (drop-out or graduation) from non-absorbing states (class levels) in their research, they have been considered

together as non-absorbing states in the current research. In Markov Chain, non-



A

From Year X	To Year X+1			
	To Course A	To Other Courses	To Completions	To Attritions
Course A	$A_{1,1}$	$A_{1,2}$	$A_{1,3}$	$A_{1,4}$
Other Courses	$A_{2,1}$	$A_{2,2}$	$A_{2,3}$	$A_{2,4}$
Completions	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$	$A_{3,4}$
Attritions	$A_{4,1}$	$A_{4,2}$	$A_{4,3}$	$A_{4,4}$

B

From Year X	To Year X+1		
	To Course A	To Other Courses	To Comp./Att (left)
Course A	$A_{1,1}$	$A_{1,2}$	$A_{1,3} \cup A_{1,4}$
Other Courses	$A_{2,1}$	$A_{2,2}$	$A_{2,3} \cup A_{2,4}$
Comp./Att (left)	$A_{3,1} \cup A_{4,1}$	$A_{3,2} \cup A_{4,2}$	$A_{3,3} \cup A_{3,4}$ $A_{4,3} \cup A_{4,4}$

Figure 2: Actual transition matrix for a sample Course-A.

A. Comprehensive transitional states. B. Simplified and practical version of the Actual transition matrix. The two columns: Completions and Attritions are merged.

absorbing states allow transitions to other states but absorbing states do not.

A narrower application of the Markov Chain is implemented to investigate how students in English Language Institutes (ELI) progress through STEM (Science, Technology, Engineering, and Mathematics) programs (Gagne 2015). In this research, transitional states include non-absorbing and adsorbing states (students who left the institute, those who graduated in STEM programs, and those who graduated in non-STEM programs) together. Their research revealed that the progress of ELI students in STEM programs is higher than non-ELI students.

Another application of Markov-based models is in graduation time (Silver, 2016). He assumed that the future probability for the transition from one state, absorbing or non-absorbing (class level), to another state, depends on the present status only and there is no influence from the historical trend involved in this transition.

Recently Austin Peay State University used Markov Chain for enrolment projection (Gandy and Crosby, 2019). They used student credit hours (SCH) to investigate the student flow from one academic term to the next. The objective of their research is to detect the entering and leakage points in the enrolment process. Their transition matrix includes 24 states for 4 classes and 4 SCH groups and they investigated the progress from each SCH group to another for all major classes (freshman, sophomore, junior and senior). Their finding helps administrators to identify enrolment trend and anomalies.

As mentioned, enrolment management is the most popular application for Markov Chain models. The states of the models utilize various student classifications. In the following sections, the big picture of the proposed Markov model in enrolment planning is introduced and then two different applications of enrolment planning are introduced.

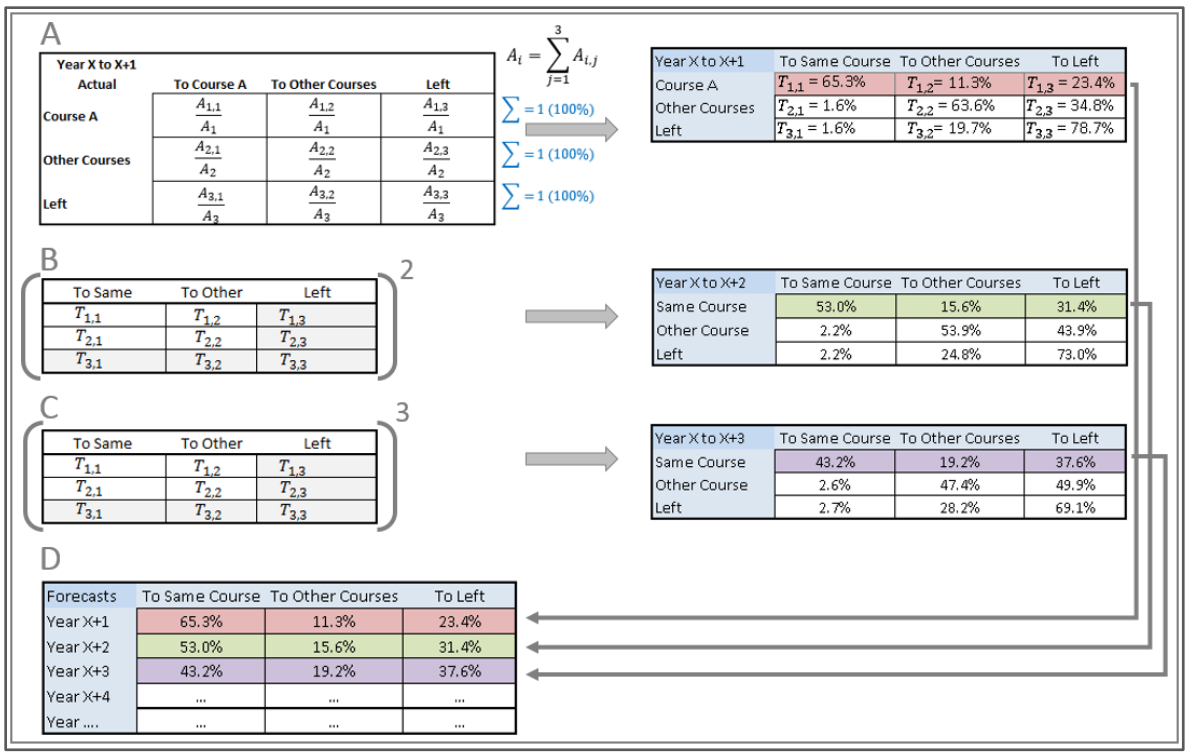


Figure 3: Markov process diagram.

A: Calculating the main Transitional probability matrix from the Actual transition matrix. B: Square of the probability matrix to forecast the 2nd iteration probabilities and C: Cube of probability matrix to forecast the 3rd iteration probabilities. D: Extract the first row of each iteration matrix to construct the final forecasting table.

Enrolment planning utilizing discrete-time Markov model

Mission and vision are the two ends of the analytics spectrum in the institutions which finally develop the strategic plan (Hinton, 2012).

Conventional planning is generally based on what the members of the institutional community believe about the institution. However, at all levels of institutional planning - Strategic, Operational, and Tactical - those beliefs are supported by descriptive data analyses of available historical data (Hinton, 2012). But the question is: what is the role of predictive data analytics in institutional planning? And how can it be utilized for more effective strategic and operational planning? (Calderon & Webber, 2015). Hence the sum of each row should be 1 (100%) because all the possible transitions should be considered in columns and rows. However, this rule is not valid for the column summation.

An Institutional Strategic Plan should provide information about enrolment and student population objectives as well as the impact of changes in the enrolment flow. Hence, any kind of data related to enrolment, which is a periodic event, plays a significant role in institutional planning. Historical enrolment data can be represented as a time series that contains useful information. If it is modeled by an appropriate method to forecast the future, then it would be able to play a significant role in planning the future of institutions.

Enrolment can be considered as a discrete-time stochastic event in which the conditions of the present states if they are not the only effective factor for the next enrolment states, are the most important. Given these assumptions, and enrolment is an event that can be modeled by a discrete-time Markov Chain, and consequently, the future states can be forecast based on the current situation. To implement enrolment planning by predictive analytics based on the Markov Chain, we need to decide three important factors:

- Identifying the transitional states;
- Investigating the historical data for availability of the identified states, and
- Nominating the final transitional states.

The granularity level of the transitions is important. A more detailed transition matrix may provide higher accuracy, but in real-world applications, the availability of historical data at the most detailed granular level may not be possible, and this limitation can impose some constraints on the predictive models. This phenomenon is depicted in Figures 1.A & 1.B. In other words, the accuracy of the model depends on the tradeoff between the granularity of the transitional states and the availability of the data.

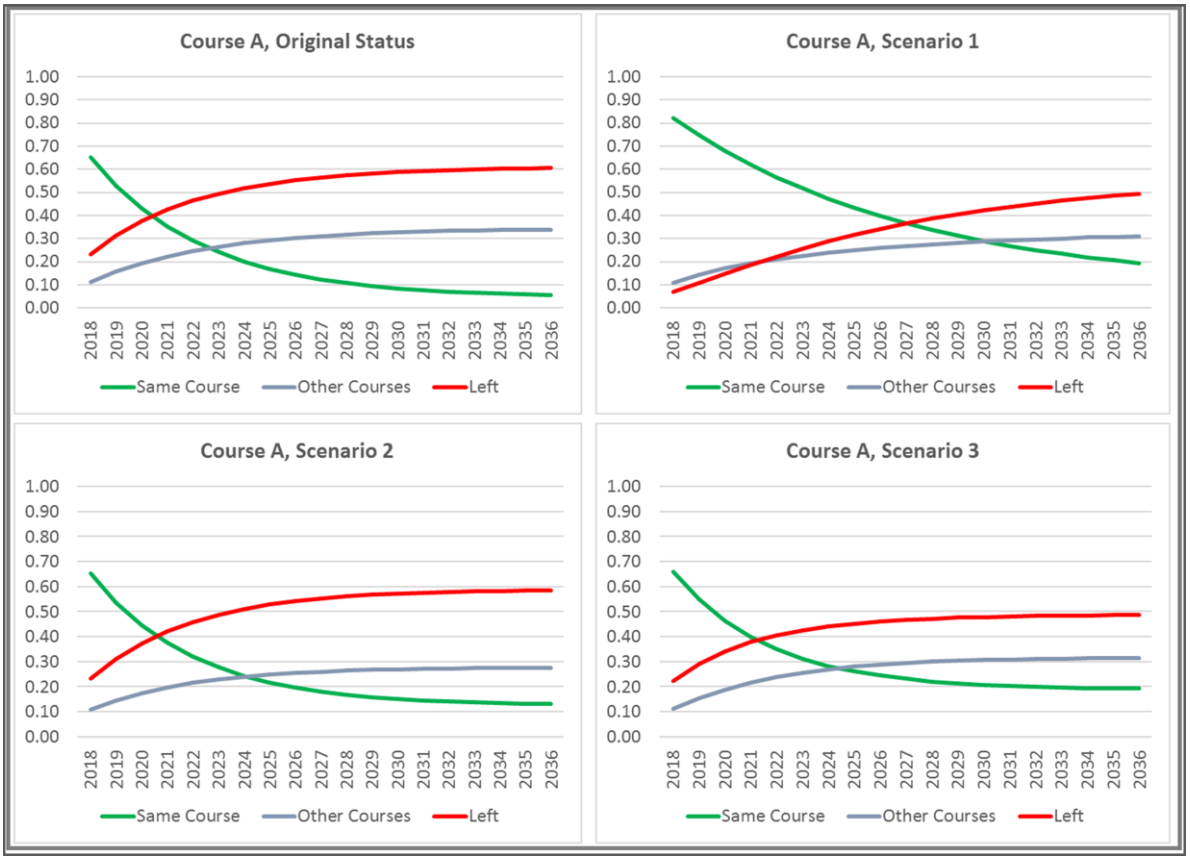


Figure 4: Estimating 20 years of transitional probabilities of enrolment trends.

The original trends are compared with 3 different scenarios. The more changes depict a higher scenario impact.

In the following sections, two samples of institutional applications in enrolment planning are provided. Before considering the details of the two applications, it is necessary to define some terms to identify the transitional states in this context. From the enrolment point of view, a student can possess one of the following four states regarding a sample course, in an academic year:

- *Commencing*: The state of those students who are enrolling in a course for the first time in the institution;
- *Returning*: The state of those students who are already enrolled and are returning to the same course to take the rest of the needed subjects;
- *Completion*: The state of those students who have passed all the subjects and completed the course, or
- *Attrition*: The state of those students who have dropped the course before completion.

The total possible transitions are shown in Figure 1.A. These four states can cover all the transitional states, by a $4\,x\,4$ matrix which covers all the possible distributions, is known as an *Actual transition matrix* and is depicted in Figures 2.A. and 2.B. The elements in the matrix represent the actual headcounts to transit from one state, represented in the row header, to the other states, represented in the column headers, i.e. $A_{1,1}$ represents the actual headcount value that transitions from Course-A in Year X to the same course in the following year. This value will generate the transitional probability $T_{1,1}$ which is represented in Figure 3.B.

Detecting the elements in the Actual transition matrix and computing the probabilities of the transitions *(Probability transition matrix*) is the core calculation of the Markov-based models. In some cases, the ideal transition matrix which covers all the possibilities, either cannot be extracted or is meaningless from a probability distribution point of view.

In such a situation, simplifying the matrix based on the potentials of the real-world data warehouse or realistic probabilities is necessary. Figures 1.B and 2.B represent the realistic transitions and related Actual transition matrix respectively. As can be seen, the two states *Completions* and *Attritions* are merged and labeled *Left*. The reason for this simplification, caused by the difficulties in extracting the required data from the available data warehouse, is the lack of information for the *Attrition* state in the last row of the Actual transition matrix. Any student, who drops a course before completing it, and does not enroll in any other courses, will be categorized in the *Attrition* state. Hence the transition from Attrition to the other states (the last row of the Actual transition matrix depicted in Figure 2.A), would be zero, except for the last column (Transition from *Attrition* to *Attrition*). This means that the probability of transition from students who had dropped their courses in Year X to *Attrition* state in the next year would be 100%, which does not convey any information to the model.

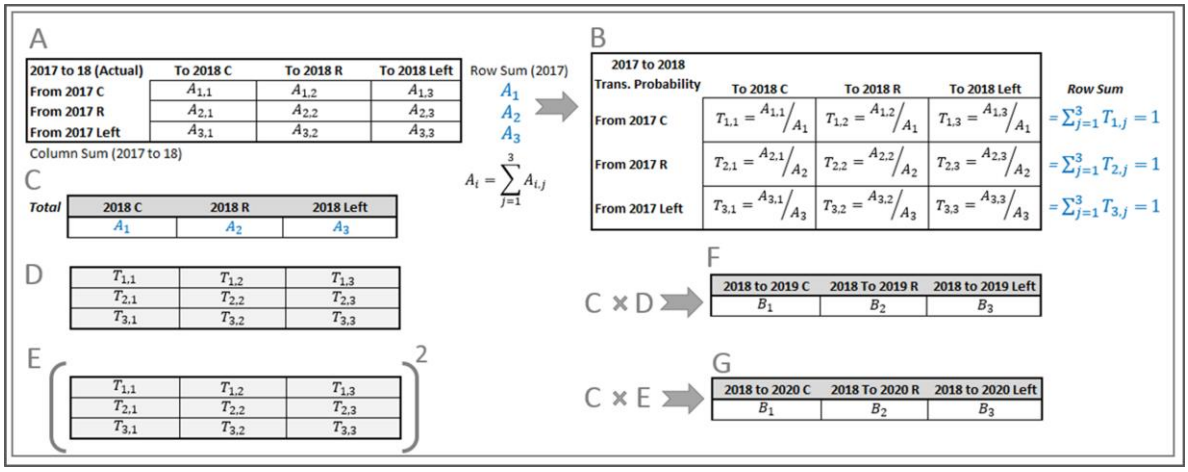


Figure 5: Actual transition matrix (2017 to 2018) (A) and its Transitional probability matrix (B). The multiplication of the actual 2018 data (C) as initial vector to the Transitional probability matrix and its square generates the 1st and the 2nd iteration headcounts: 2018 to 2019 (F) and to 2018 to 2020 (G).

The simplified version is used in this experiment. In this practical version, the two states, *Completions*, and *Attritions* are merged and labeled *Left*. The results would be a 3 x 3 Actual transition matrix shown in Figure 2.B and all the elements can be computed based on the available data.

The following two sections are dedicated to the two planning applications. The enrolment system in both applications is modeled by the homogenous Markov Chain. The subjective in both applications is forecasting enrolment in different planning scenarios and comparing the impacts by statistical significance test.

The first application is trend analysis of three proposed decisions in improving course enrolment strategy and how to detect the most effective method. The trend of changes of transitional probabilities resulting in different scenarios is compared to detect the most significant decision.

The second application is a projection of enrolment headcount to forecast the impact of shrinkage in international student enrolment. Unlike the first application, instead of the conditional probability analysis, the actual headcount is used to forecast the enrolment in different scenarios, compare the results and find the critical point.

Application A: Measuring decision impact on course enrolment planning

In course enrolment planning, a variety of possibilities can be considered in an action plan. However, which one has the maximum impact on the enrolment trend, regardless of the effectiveness of the results? Answering this question is the objective of this section.

To become familiar with the decision impact analyses by a Markov model, three sample scenarios among different decision possibilities are selected and the impact of each approach is compared with the original enrolment trend based on the actual historical data as the basis of the forecasting process. The possible transitional states for a sample Course-A are as follows.

- *Commencing Course-A:* The number of new students who enrolled in Course-A for the first year.
- *Returning to Course-A:* The number of students who re-enroll in the same course in the following year.
- *Returning to other courses:* The number of students who re-enroll but in other courses in the following year.
- *Leaving institution:* The number of students who leave the institution, either by completing or dropping the course

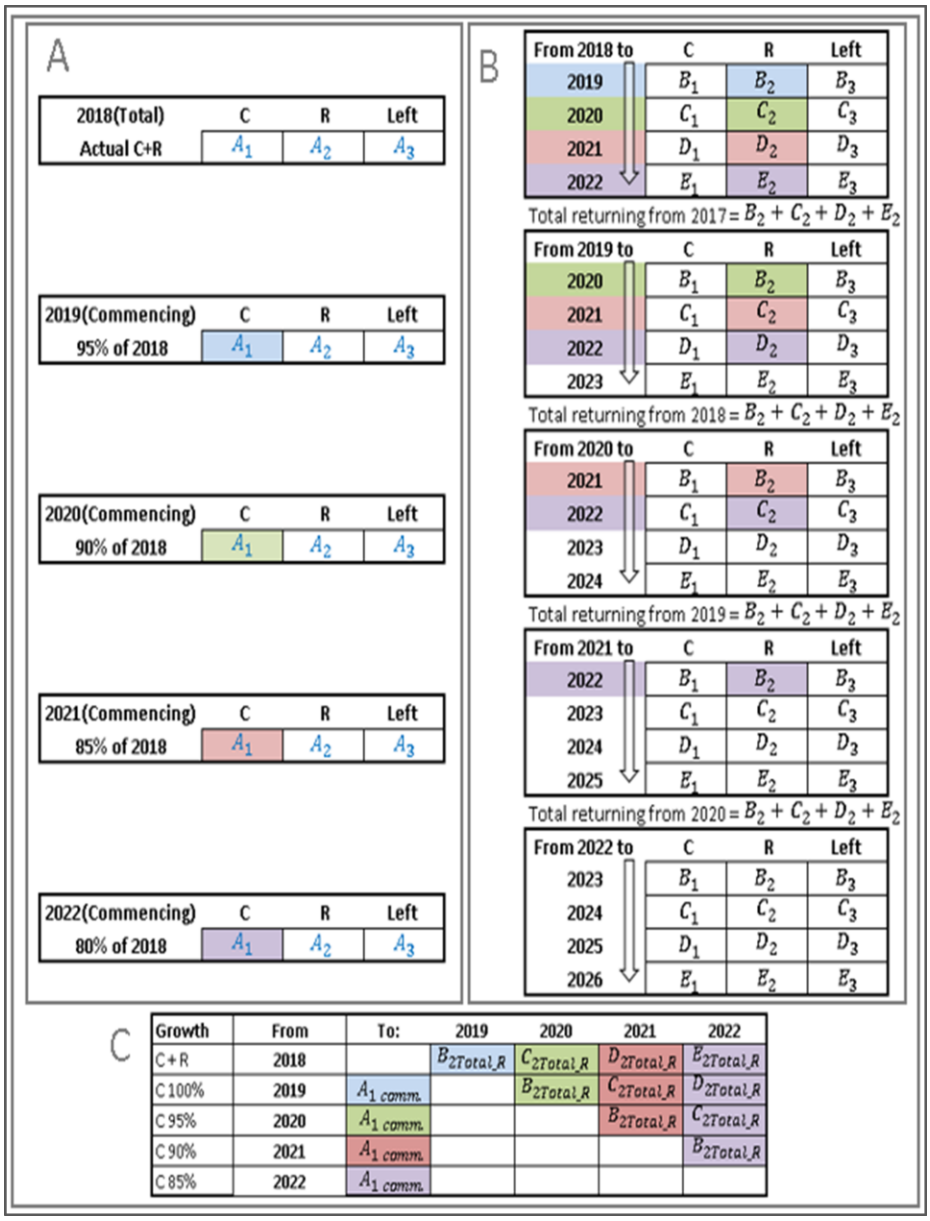


Figure 6: The process of the 1st Scenario. Section (A) shows the initial vector (actual and estimate) for each forecasting period. (B) focuses on the Returning

Based on the described transitional possibilities a set of selected sample scenarios is chosen. The original state and three selected scenarios to improve the course enrolment trend are defined as follows.

- *Original State:* The actual values transitional state based on the original Actual transition matrix elements.
- *Scenario 1:* Increase re-enrolment from the sample Course-A into the same course ($A_{1,1}$), and decrease the attrition ($A_{1,4}$),
- *Scenario 2:* Increase re-enrolment from the other courses into Course-A

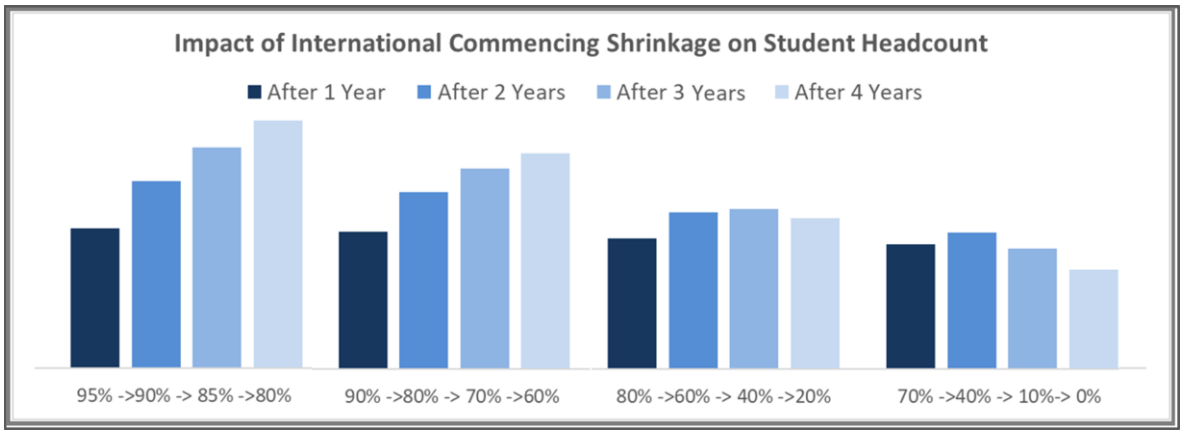


Figure 7: Commencing shrinkage impact on total headcount.

Bar charts showing the impact of shrinkage of commencing international students to the total student headcount in four consecutive years.

($A_{2,1}$), and decrease the attrition ($A_{1,4}$).

- *Scenario 3:* Increase re-enrolment from the completion of other courses into Course-A ($A_{3,1}$) and decrease the attrition ($A_{4,4}$).

As explained earlier, there are numerous possibilities but experimenting on a limited version of the model would be adequate to observe the performance of the method.

The process of calculating the main Transitional probability matrix and predicting the next iterations (enrolment periods or years in this example) is simple. This process is depicted in Figure 3. Based on the homogenous Markov Chain, multiplying the Transitional probability matrix to itself, yielding the Transitional probability matrix of the next periods. Analyzing the probabilities in series of transitional matrices of different periods would be enough for trend analysis. However, multiplying the Transitional probability matrix of each period to the basic actual vector generates the forecasted headcount for that period. Simply put, the square and cube of the Transitional probability matrix provides the forecast for the 2nd and the 3rd year probability matrices (Figures 3.B and 3.C), and so on. The objective of the current experiment is to measure the impact on enrolment planning of a sample Course-A. The transitional probabilities from this course to the other three states are available in the first row of the forecasting probability matrix in each period which is highlighted in Figure 3. Figure 3.D shows the final table which contains all the projected transitional probabilities relating to the enrolments of Course-A.

Each scenario and the original state generate four transition matrices. The next step is applying the Markov process, mentioned above, to the four transition matrices, forecast the n following years, and comparing the results stored in the final forecasting tables (Figure 3.D).

One of the characteristics of a homogenous Markov model is reaching a state that the changes in the following forecast periods are not significant. This state is known as the steady-state. To depict this phenomenon, 20 consecutive periods of each scenario, as well as the original state, are shown in Figure 4. As can be seen, the rate of change will gradually decrease after a couple of periods in all three enrolment states.

To investigate the impact of different scenarios, we should calculate how different they are from the original state. Figure 4 shows a significant change in the shape of the graphs (maximum impact) as can be seen in Scenario 1. With the assumption of normal distribution of enrolment data, a statistical significance test (t-test) is utilized to measure the impact. The *P-values*, shown in Table 1, are calculated over the first five forecasting periods (years) which is a reasonable period for a Bachelor's Degree course. The smaller *P-value* is interpreted as the higher impact. If the *P-value* is less than 0.05, the impact is considered statistically significant and, in this application, belongs to the first scenario.

Application B: Commencing shrinkage impact on student headcount

Commencing enrolments in each year has a long-term impact on the institution's population over the next couple of years. Regarding recent changes in the population of commencing student cohorts, utilizing a model to forecast the student

TABLE 1: MEASURING DECISION IMPACT BY COMPARING THE *P-VALUES*

Decisions	Scenario 1	Scenario 2	Scenario 3
t-test (P-value)	*0.011*	*0.817*	*0.635*

The significant decision impact belongs to Scenario 1.

headcount in the following years is of interest to institutional strategic planners.

The objective of this section is to measure the impact of international student shrinkage on the university population in four years. Unlike the previous application, the major transitional possibilities in this application are not focused on a specific course. As the objective is to investigate the population of a cohort (international students), the major transitional states have been defined as follows.

- *Commencing students:* The headcount of newly-arrived international students

in each year,

- *Returning students:* The headcount of international students who re-enroll in the institution to complete their courses.
- *Leaving students:* The number of international students who leave the institution, having either completed or dropped their courses.

Similar to the previous application, a limited set of possible scenarios have been selected to investigate the commencing shrinkage impact, as follows

- *Base-year State:* The actual values transitional state of a year as base-year, considered as 100% of the cohort enrolment.
- *Scenario 1:* Commencing shrinkage with the rate of 5% in each of the following years.
- *Scenario 2:* Commencing shrinkage with the rate of 10% in each of the following years,
- *Scenario 3:* Commencing shrinkage with the rate of 20% in each of the following years.
- *Scenario 4:* Commencing shrinkage with the rate of 30% in each of the following years.

The above scenarios and the base-year generate five transition matrices that are the core for a homogeneous Markov Chain to forecast the next four years for each scenario. The enrolment year 2018 was selected as the actual base year (considering actual values for *C* and *R* and *Left* students are available at the end of 2019). Two transition matrices were extracted based on the latest available actual historical data (from 2017 to 2018), for total and for commencing students. The transition probability of the three elements: *Commencing, Returning,* and *Leaving* students for the following four years are forecasted based on the first order, square, cube, and the higher power of the Transitional probability matrix (powers 1 to 4). This process partially is depicted in Figure 5 for the 1st and the 2nd year forecast. Multiply the base-year actual values as the initial vector (2018) to the probability matrix provides the year 2018 to 2019 (Figure 5.F). For generating the headcount of the year X+2 (Figure 5.G) the square of the probability matrix is needed (Figure 5.E). The same process is applied for the 3rd and the 4th period after the base year.

The next step is to forecast the total number of students by headcount in each of the four following years after 2018 as the base year. A forecasting process was implemented for each shrinkage scenario to enable a comparison of the effects. This process is depicted in Figure 6 and depicts Scenario 1: 5% shrinkage in the commencing headcount of international students each year.

It is important to consider that the estimated headcount in each year is the sum of *Commencing* students in that year and the *Returning* students of the previous years. This process is shown by highlighted items in similar colors in Figure 6.

The results of the four scenarios can be seen in Figure 7. As can be observed, the increasing trend starts to decrease when the commencing international students reach 60% of the base-year which can be considered as an alarm point. This observation can be investigated and verified by the t-test *P-value* analysis shown in Table 2. The table also shows significant changes to the total headcount when the commencing headcount of international students reaches 60% of the base-year commencing headcount with the highlighted *P-values* in red in Table 2.

TABLE 2: MEASURING SHRINKAGE IMPACT BY COMPARING THE P-VALUES

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Shrinkage	Pvalue	Shrinkage	Pvalue	Shrinkage	Pvalue	Shrinkage	Pvalue
1st	*95%*	*0.808*	*90%*	*0.619*	*80%*	*0.307*	*70%*	*0.121*
2nd	*90%*	*0.619*	*80%*	*0.307*	*60%*	*0.040*	*40%*	*0.003*
3rd	*85%*	*0.449*	*70%*	*0.121*	*40%*	*0.003*	*10%*	*0.000*
4th	*80%*	*0.307*	*60%*	*0.040*	*20%*	*0.000*	*0%*	*0.000*

Accuracy analysis

Accuracy analysis is a *delicate* part of predictive analysis. The word *delicate* is used because accuracy can be measured from different aspects and they can show contradictory results. In the current research, an approach is performed on two categories of datasets. The approach simply extracts the transition of actual commencing headcounts within the base year (2012-13) and the following five years and compares it with the estimated headcounts for the same period. Then the accuracy is measured based on the sum of the all differences between the two vectors (actual and estimate) as the numerator, and the total actual vector as the denominator. The model was applied to two subcategories of commencing international students:

- All levels of tertiary education (Sub degrees, Postgrad, Undergrad, and Research), and
- Postgraduate and Undergraduate only.

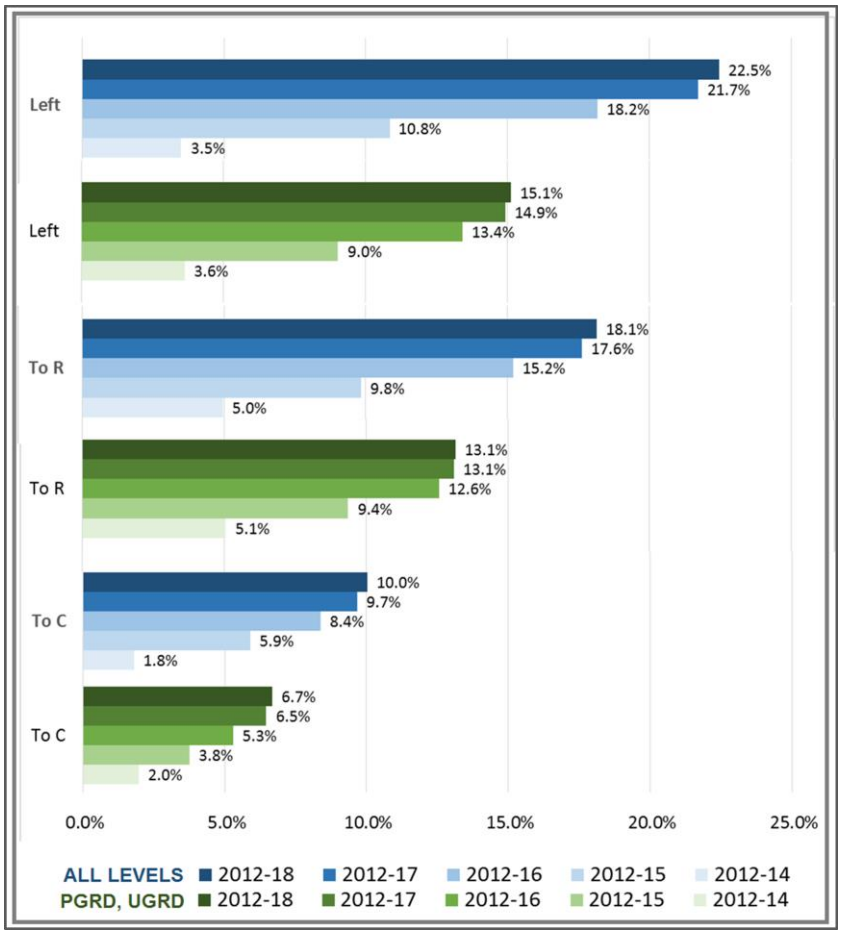


Figure 8: Accuracy rates for two categories.

Blue: All levels of tertiary education, Green: Postgraduate and undergraduate only.

The second category is a subcategory of the first dataset and includes a less diverse group of courses, compared to the highly diverse courses in the first group. The similarities among the courses in the second category were investigated based on the Pearson and Cosine similarity (Rouhi & Calderon, 2017) (Rouhi, 2018). It is designed in this way to investigate if there is any correlation between the diversity among the sub-groups and the accuracy of the predictive model.

The results show that greater accuracy can be seen in the less diverse group of courses. A possible reason for this observation is the correlation of the transition matrix rows and columns *(i.e. Commencing*, *Returning*, *Completion*, and *Attrition)* with the subcategories of data in the main dataset *(i.e. Undergraduate*, *Postgraduate)*. The transitional states among postgraduate and undergraduate courses are more similar to each other than to the other group which includes all levels.

Figure 8 depicts the accuracy values of the two categories for five years. The focus of the chart is on transitions of the commencing international students from the base-year 2012-13, into the three possible states in the following, 2nd, 3rd, 4th and 5th iterations (2012-14 to 2012-18) which are estimated by a homogenous Markov model based on the probability transitional matrix of the base-year (2012-14). Regardless of the more accurate results in the second category, a similar pattern can be seen in both categories. The accuracy is higher in the 1st year estimation compared to the 5th year estimation, and the transitions from commencing to commencing are higher than the transition from commencing to the other states.

Recommendation and implication for future study

This study was conducted on two different scales:

- Course level including all the cohorts (domestic and international students), and
- Total enrolments including all the courses (international students only).

Further investigation conducted by the RMIT/Analytics and Insight team, revealed that the accuracy of homogeneous Markovian-based forecasting depends on the dynamic of the system. The more homogenous behavior of the population in the consecutive years yields more accurate results for homogeneous Markov Chain. Our accuracy analysis revealed that when the international student population was segregated into two major groups: Research and Class-based (Postgraduate and Undergraduates), the accuracy of the model increased, because the dynamic of transitional probability in Research-based courses is not like the Class-based courses.

A Non-homogeneous Markov Chain is an alternate model which is under investigation by RMIT/Analytics and Insight team. Unlike the homogenous model, non-homogeneous models need historical data to extract the most appropriate transitional probabilities for each period. The primary analysis shows non-homogeneous can provide more accurate results for course enrollment forecasts. The reason is rooted in the dynamic of change in sequential periods. The transitional probabilities from the first year to the second and from the first year to the third (and subsequent) fluctuate significantly. Such a dynamic is also related to the length of courses.

Some technical debates in our team encouraged the author to make a technical recommendation about the method for matrix multiplication. Matrix multiplication plays an important role in calculating Markov Chain. Multiplying transition matrices to provide the next iteration transition matrices are always involved with $n \times n$ matrix multiplications. However, in estimating the headcounts we generally need to

multiply an initial vector to the $n \times n$ Transitional probability matrix. The conventional approach considers an $1 \times n$ vector as the initial vector. During the current study, the author noticed that the conventional homogeneous approach can be substituted by multiplying an $n \times n$ matrix of actual values (the source of the first Transitional probability matrix) as the initial matrix to an $n \times n$ Transitional probability matrix. The result of the second approach is an $n \times n$ forecasted value that the sum of the column would be the same as the results of the conventional method ($1 \times n$ initial vector to the $n \times n$ Transitional probability matrix). Although the second approach is more complicated, it provides more details in the $n \times n$ result matrix that may be useful in some applications. However, the current study experiments both approaches and provides the same results. Thus, for simplicity, we explained and demonstrated the conventional approach in the figures. However, the applicability of this approach for the Non-homogeneous model is under investigation.

The last recommendation is about the accuracy analysis method. Since comparing actual and estimated vectors are involved in measuring the accuracy, the conventional approaches consider all the vector elements to provide a holistic error measure similar to the one that used in this study (some of the deviation of all the forecasting vector elements with the historical actual values) or the Mean Square Errors (MSE and RMSE) (NCVER 2016, Mark and Karmel 2010). However, further investigation by our team reveals that the sum of the values is fixed in the estimated vectors; hence the forecasted values are interdependent to each other, and increasing one will affect the other figures in the vector. In such a situation, selecting the maximum error values between the two vectors can be considered as an appropriate representative of the total vector deviation. Compared to the previous method, this approach prevents evaluation of the error and enables us to provide a lower and more realistic error rate.

Conclusion

International students provide a significant cohort among Australian institutions. The quality and quantity growth of South East Asian institutions can cause significant fluctuations in the student demographics in Australian universities. In this situation, robust strategic planning, by utilizing advanced predictive analytical techniques, contrasted with conventional approaches, can provide not only a more realistic organizational vision but also more accurate operational and tactical objectives.

The availability of rich time-series data in institutional data warehouses provides a foundation for a wide range of tools and techniques for predictive analytics. In this paper, a classical artificial intelligence tool, the Markov Chain, is introduced to

estimate the next state based on the available current data. The core concept of Markov-based models lies in the following items:

- Distinguishing the most appropriate *Transitional states* of the current system,
- Extracting *Actual transition matrix* elements from the latest available historical data, and
- Computing the *Transitional probability matrix* which represents the probability distributions among the states.

A homogenous Markov Chain, simply by computing 2^{nd}, $3^{rd,}$ and higher powers of the probability matrix, provides the estimation of the 2^{nd}, $3^{rd,}$ and later periods of the system. In this research, a Markov model is utilized for two applications involved in enrolment planning. The first application shows how we can measure and compare the enrolment planning scenarios. The second application is more complicated and shows how to estimate the impact of international commencing enrolments shrinkage over total student headcount in the next four years. The significance of the impact of different scenarios has been measured by t-test. An accuracy analysis is also provided based on the actual historical data compared to the estimated values provided by the model. The results show the accuracy of the model will decrease when estimating longer periods. However, breaking down the large datasets into cohorts with more homogenous patterns, can reduce the diversity in the data and improve accuracy.

References

Anderson, W. J. (2012). *Continuous-time Markov chains: An applications-oriented approach*. Springer Science & Business Media.

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, May 19.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Calderon, A and Webber, K (2015), *'Institutional research, planning, and decision support in higher education today* in Karen L. Webber and Angel J. Calderon (ed.) Institutional Research and Planning in Higher Education, Routledge, New York, United States, pp. 3-15.

DFAT (*Department of Foreign Affairs and Trades*) (2017-18), https://dfat.gov.au/about-us/publications/Documents/trade-in-services-australia-2017-18.pdf

Feinberg, E. A., & Schwartz, A. (Eds.). (2012). *Handbook of Markov decision processes: methods and applications* (Vol. 40). Springer Science & Business Media.

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

Gandy, R. & Crosby, L. (2019). *Enrollment Projection Using Markov Chains: Detecting Leaky Pipes and the Bulge in the Boa.* The AIR Professional File, Fall 2019 Article 147.

Hinton, K. E. (2012). *A practical guide to strategic planning in higher education*. Ann Arbor, MI: Society for College and University Planning.

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

Mark, K., & Karmel, T. (2010). *The Likelihood of Completing a VET Qualification: A Model-Based Approach.* Technical Paper. National Centre for Vocational Education Research Ltd. PO Box 8288, Stational Arcade, Adelaide, SA 5000, Australia.

Meyn, S. P., & Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

National Centre for Vocational Education Research (NCVER). (2016). *VET program completion rates: an evaluation of the current method*.
Oliver, R. M. (1968). *Models for Predicting Gross Enrollments at the University of California*.

Pierre, C., & Silver, C. (2016). *Using a Markov chain model to understand the behavior of student retention*. The 7th International Multi-Conference on Complexity, Informatics, and Cybernetics/The 7th International Conference on Society and Information Technologies: Proceedings (pp. 248–251).

Rouhi, A., & Calderon, A. (2017). Vector-based Models for Educational Institution Shape Analysis. In *17th Annual SEAAIR Conference 2017* (pp. 1-1). SEAAIR.

Rouhi, A. H. (2018). Exploring New Angles to Analyze Student Load Data. *Universal Journal of Educational Research*, *6*(9), 1950-1961.

Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543). John Wiley & sons.

Acknowledgments

I thank Kate Koch, Nonna Milmister, and Sean Lee from Finance & Governance Department of RMIT University for their considerable support, and Nikhil Sobti, Theo Dufort, and Lily Tsang for guiding deliberating the general concept of Markov chain, accuracy measurement, and *Commencing Impact* issue and its related scenarios.