PROJECT PROPOSAL:

AN ENROLLMENT RETENTION STUDY USING A MARKOV MODEL FOR A REGIONAL STATE UNIVERSITY CAMPUS IN TRANSITION

By

Polly Wainwright

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Masters of Science
in the
Department of Mathematical Sciences
and
Department of Computer and Information Sciences
Indiana University South Bend

June 2006

Committee Members:

Dr. Yi Cheng, Advisor
Dr. Yu Song
Dr. Dana Vrajitoru

TABLE OF CONTENTS

1. Abstract

Enrollment retention rates are an accepted indicator of a university's success in providing quality degree programs and learning environments which will lead to a student's continued enrollment and timely graduation from that university campus.

This project will analyze enrollment data from five consecutive years from a state university regional campus, and construct an enrollment retention model using a Markov process.

The model will be used to analyze enrollment retention rates for commonly overlooked segments of the student population, as well as the retention rates for specific degree programs, rather than just the retention rates for aggregate incoming freshmen.

This additional information will provide an opportunity for enrollment retention committees to focus on improvement for all classifications for students, rather than just the commonly observed, but non-representative freshmen.

2. Introduction

The university from which the data will be collected is a regional campus of a state university which has, over the past 40 years, transitioned from an off-campus site, to a two-year community college, and recently, to a four-year-degree granting institution with graduate programs in planning and student housing under construction.

Retention figures tend to be used for comparison against national averages, and against the rates of local, as well as similar peer institutions.

This project seeks to compare various retention rates within the schools of this campus for the purpose of gaining greater insight into areas where retention might be improved.

3. Literature Review

3.1 Retention Rates

There are two types of commonly accepted retention rates: the freshman-to-sophomore retention rate, and the six-year retention rate.

The freshman-to-sophomore retention rate is based on the number of first semester freshmen enrolling in a fall semester who then enroll as sophomores the following fall at the same campus. This retention rate is widely accepted as the key retention figure since students are most likely to drop out during their freshman year. The assumption is that students who progress to their sophomore year are then less at risk, and will likely go on to graduate from that campus.

The six-year retention rate is based on the number of first semester freshmen enrolling in a fall semester that go on to graduate from that campus, with a four-year degree, within a six year period. This retention rate seeks to measure timely graduation.
In both instances, the retention rates apply only to full-time students.

Retention literature stresses that students who attend full-time and do not have outside employment have a 70-75% graduation rate, nationally. Students who attend only part-time or who work full-time have an approximately 50% graduate rate. Further, only about 50% of students graduate from the institution where they began as freshmen. Of students whose college experience includes campus transfers, fewer than 10% progress to graduation.

Retention rates are of interest not only to universities hopeful of maintaining or increasing enrollment, but act as a socio-economic indicator of well-being for the community as a whole. It is, therefore, considered to be in the best public interest to maintain high retention rates.

3.2 Enrollment Retention Models

Markov processes, regression models, and cohort flow models are three common techniques used for enrollment prediction.

The Markov model makes use of readily available cross-sectional historical enrollment data and can be initiated at any point for which data is available. (Armacost and Wilson)

Markov models tend, however, to comprise a small number of states derived from aggregate student populations, specifically all students in all degree programs, or freshmen in all degree programs, rather than more specifically define student classifications. (Kraft and Jarvis)

Further, Markov models have specifically appropriate uses. The Markov model assumes no trend in data, that the probability of flow from one state to the next will remain consistent for the life of the model. Yet, the university environment naturally evolves with changes to and additions of degree programs, changing student populations, and the growth of distance learning options. (UCF) The Markov model is best reserved for projections of no more than 10 years.

4. Statement of Problem

The retention rate definitions described above apply to full-time first semester freshmen enrolling in a fall semester, and generally apply only to those students pursuing a four-year degree.

A preliminary analysis of the data for the campus being studied indicates that at least one-third of the students are registered for 9 or few credit hours, with one-half of the students registered for fewer than 12 credit hours.

Slightly fewer than 25% of new students at this campus are first-time freshmen. About 10% of new students are non-degree seeking students. This means that about 65% of the newly enrolled students are transfer students who are not represented in typical retention rate analysis.

Finally, about half of all enrolled students are over the age of 25, with the average student age being approximately 29. Students over the age of 25 are classified as non-traditional or adult students. While the available data does not include employment status, it would be safe to assume that adults are typically employed at least part-time.

Clearly, the normally accepted retention rate descriptions summarized in the previous section do not adequately account for the student demographics at this university campus. Therefore, a new model, using a Markov process, will be used to analyze available enrollment data with respect to appropriate rather than the commonly accepted student classifications.

5. Development of Model

The Markov model is based on an underlying stochastic process in which a system in one state, s_i moves to a subsequent state, s_j. The states are commonly referred to as the Current state and the Next state. The act of moving from one state to the next is referred to as a *step* or *transition*.

The states are from a finite set of possible states $S = \{s_1, s_2, \ldots, s_n\}$ with each state having a transition probability p_{ij} of occurring. The total probability of *S* is 1.

When the conditional probability of a Next state occurring, given the Current state, is dependent only on the Current state, or independent of states previous to the Current state, the process is said to have the *memorylessness*, or Markov Property.

$$\Pr\left[s_{j+t} = y \mid s_i, i \le j\right] = \Pr\left[s_{j+t} = y \mid s_j\right]$$

A stochastic process with the Markov Property can be called a Markov Process or Markov Chain.

The transition probabilities for this model will be obtained from the relative frequencies of students in Classification Index *i* moving to Classification Index *j* in the time period of one semester. Since only the Classification Index of the Current semester, and the Classification Index of the Next semester will be noted, the probability of moving from any Current state to any Next state will be independent of states previous to the Current state. This independence from previous states meets the criteria for the Markov Property, so this model can be justifiably called a Markov Process.

The Markov Process relies on two matrices: an initial state matrix, and a transition matrix.

The transition matrix is an n-by-n matrix containing the transition probabilities for moving from the Current state in the process to the Next state. In this model, the Current states will be indicated by the matrix columns, and the Next states by the matrix rows. For transition matrix *T*, each cell contains the following transition probability:

$$a_{ji} = p_{ji} = \Pr\lfloor s_j \mid s_i \rfloor$$

where *i* indicates the row, or Next state, and *j* indicates the column, or Current state, and $1 \le i, j \le n$.

Following is a simplified version of the transition matrix for this model.

$$T = \quad \text{Next State} \quad \begin{array}{l} \\ \text{Freshman} \\ \text{Sophomore} \\ \text{Junior} \\ \text{Senior} \end{array} \begin{array}{c} \text{Current State} \\ \begin{array}{cccc} \text{Freshman} & \text{Sophomore} & \text{Junior} & \text{Senior} \end{array} \\ \left| \begin{array}{cccc} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{array} \right| \end{array}$$

The purpose of the transition matrix is to represent the probability of movement between states in a single time period. In this case, with what probability does a student achieve a particular Classification Index by the end of the current semester?

The possible states into which students could be classified are determined by the logical sorting criteria described below. The most basic states will include classification index, academic standing, and assumptions regarding non-enrollment.

Initially, there will be 10 classification indexes, 2 levels of academic standing, and 3 reasons for leaving the university. All reasonable combinations of these factors will produce 23 individual states into which student data may be grouped, thus, necessitating a 23-by-23 transition matrix. The following table lists these states and provides a description of each.

STATE		DESCRIPTION	
ND	G	non-degree seeking,	good academic standing
ND	P	non-degree seeking,	academic probation
FF	G	first semester freshman,	good academic standing
FF	P	first semester freshman,	academic probation
1	G	1 - 15 credit hours,	good academic standing
1	P	1 - 15 credit hours,	academic probation
2	G	16 - 30 credit hours,	good academic standing
2	P	16 - 30 credit hours,	academic probation
3	G	31 - 45 credit hours,	good academic standing
3	P	31 - 45 credit hours,	academic probation
4	G	46 - 60 credit hours,	good academic standing
4	P	46 - 60 credit hours,	academic probation
5	G	61 - 75 credit hours,	good academic standing
5	P	61 - 75 credit hours,	academic probation
6	G	76 - 90 credit hours,	good academic standing
6	P	76 - 90 credit hours,	academic probation
7	G	91 - 105 credit hours,	good academic standing
7	P	91 - 105 credit hours,	academic probation
8	G	106+ credit hours,	good academic standing
8	P	106+ credit hours,	academic probation
X	D	no longer enrolled,	academically dropped
X	GR	no longer enrolled,	graduated
X	L	no longer enrolled,	lost

Further sorting and grouping will be done by school or degree program, student age, and full- or part-time status.

The initial state matrix for the retention model will be a column matrix with a row for each state, or a 23-by-1 matrix. Each cell will contain the probability that a newly enrolled student will be classified in this state.

The data for students who are enrolled in the second of two consecutive semesters will be used to determine the initial state matrix.

6. Data and Technology Requirements

The 10 most recent consecutive semester's enrollment data will be obtained from the university campus being studied. The data used for sorting and determining states will include an anonymous student identification key, classification by credit hours earned, cumulative grade point average, number of credit hours registered, date of birth, and school or degree program. Only undergraduate student data will be considered for this model.

6.1 Sorting Criteria

Student data will be sorted by classification index, levels 1 through 8, which is determined by credit hours earned. A classification index of 1 indicates 1 to 15 credit hours earned; a classification index of 2 indicates 16 to 30 credit hours earned, and so forth. Special classifications include FF for first-time freshmen and ND for non-degree seeking students.

Data will also be sorted by cumulative grade point average in order to further group students by academic standing. Students will be determined to be in good academic standing, or on academic probation based on the criteria in the university's student handbook.

Academic standing is based on a sliding scale which becomes more stringent for higher classification indexes. For example, a student with CI 1 is placed on academic probation with the cumulative GPA falls below 1.5, with a student with CI 8 is placed on academic probation if the cumulative GPA falls below 2.0.

Classification Index	Cumulative GPA
1	1.5
2	1.6
3	1.7
4	1.8
5	1.9
6	2.0
7	2.0
8	2.0

Further sorting will be based on school or degree program, student age, and full- or part-time status.

6.2 Assumptions

Students who are enrolled in the first of two consecutive semesters but not the second became non-enrolled because they had been academically dropped by the university, graduated, transferred to another university, or stopped attending college altogether. No direct information about these states is available in the data. Assumptions are therefore made based on classification index and cumulative GPA.

No distinction can be made between students who have transferred to other universities and students who have stopped attending altogether. They will simply be given the classification (L)ost.

6.3 Technology

The data is made available from the campus registrar in the form of Excel spreadsheets. Sorting will be done using the Excel sorting function.

The transitioning of the Markov model will also be done in Excel using the matrix multiplication function, mmult(array1, array2).

This function is a tool for finding the ordinary matrix product of two arrays. The arrays may, and in this case, are two-dimensional arrays, or matrices.

The product of m-by-n Matrix A and n-by-p Matrix B will be placed in m-by-p Matrix C. Each element in Matrix C will be calculated as follows:

$$C_{ij} = A_{ik} \bullet B_{kj} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \ldots + a_{in} b_{nj}$$

where $1 \leq i \leq m$ and $1 \leq j \leq p$.

The Excel function mmult(MatrixA, MatrixB) is applied to each cell in the defined range for the destination Matrix C.

7. Application of Model

A Markov process is performed through matrix multiplication of the initial state matrix and the transition matrix, resulting in, for this application, a distribution of student states after one semester. In order to obtain a distribution matrix after many semester transitions, the transition matrix need only be applied the desired number of times.

Specifically,

$$TX_0 = X_1$$

where T is the transition matrix, X_0 is the initial state matrix, and X_1 is the resulting distribution matrix after one transition.

Since the *memorylessness* property has been shown to be applicable, subsequent transitions, and subsequent distribution matrices can be obtain by using the previous distribution matrix as the current initial state matrix.

$$X_1 = TX_0$$
$$X_2 = TX_1 = T(TX_0) = T^2 X_0$$
$$X_3 = TX_2 = T(T^2 X_0) = T^3 X_0$$

This process can be generalized to say that the distribution matrix, after m time periods or transitions is given by

$$X_m = T^m X_0$$

Since one objective of this model is to produce a six-year retention rate, the transition matrix will be applied (2 semesters) x (6 years) = 12 times, or

$$X_{12} = T^{12} X_0$$

8. Objective

The goal of this project is to re-examine retention data, and to build a new Markov retention model using student classifications that are more representative of the demographics at this campus.

The one-year freshman-to-sophomore retention rate is a traditional indicator that can be expanded for all student classification, specifically the transfer student classifications that comprise 65% of new students at this campus. Therefore, one-year retention rates will be determined for transfer students, by Classification Index, and then by traditional and non-traditional age grouping.

Similarly, a six-year retention model will be constructed for all student classifications using a Markov Process.

These additional retention rates are being sought in order for retention committees to address the unique needs of transfer and non-traditionally aged students, especially given that this particular overlooked student segment represents the majority of enrolled students at this campus.

9. Conclusion

Enrollment retention rates are a widely accepted indicator of university success. However, the most commonly used retention measures only address the progress of first-time freshmen attending full-time, in pursuit of a four-year degree. This student demographic does not adequately describe the university campus in this study.

By developing and applying a new Markov model to the student data, information can be obtained to create education strategies that will benefit all categories of students.

10. References

R. L. Armacost and S. Archer (2005): **Spreadsheet models for program enrollment planning.** *Paper presented at 2005 SAIR Conference.*

R. L. Armacost and A. L. Wilson (2002): **Three analytical approaches for predicting enrollment at a growing metropolitan research university.** *Paper presented at the Annual Meeting of the Association for Institutional Research (42nd, Toronto, Ontario, Canada, June 2-5, 2002.)*

R. L. Armacost and A. L. Wilson (2004): **Using Markov chain models to project university and program level enrollment.** *Paper presented at the 2004 SAIR Annual Conference.*

R. L. Armacost, S. Archer, J. Pet-Armacost and J. Grey (2006): **Using Markov chains to assess enrollment policies.** *Paper presented at the 2006 AIR Annual Forum.*

J. B. Dworkin (2005): *Retention: How Important Is It Really? Retention and Graduation Rates:*
http://www.pnc.edu/co/Retention.pdf

Grinstead, Charles M. and J. Laurie Snell (1997): *Introduction to Probability.* Providence (RI): American Mathematical Society.

C. R. Kraft and J. P. Jarvis (2005): *An Adaptive Model for Predicting Course Enrollment*
http://www.math.clemson.edu/reports/TR2005_11_KJ.pdf

Lawler, Gregory F., (1995): *Introduction to Stochastic Processes.* Norwell (MA): Chapman & Hall.

Microsoft Office Online (2003): *MMULT:*
http://office.microsoft.com/en-us/assistance/HP052091811033.aspx

Purdue University North Central (2006): *Academic Probation Information:*
http://www.pnc.edu/sa/probation.html

Ross, Sheldon M., (1995): *Stochastic Processes.* New Jersey: Wiley.

University of Central Florida, University Analysis and Planning Support (2005): *Overview of Enrollment and Degree Projections (August 12, 2005):*
http://uaps.ucf.edu/enrollment/overview.html