ANALYZING LONG-TERM UNIVERSITY ENROLLMENT DYNAMICS USING A MONTE CARLO METHOD

Michael Quinn
Strategic Analyst, KIMEP University, Kazakhstan
mquinn@kimep.kz

ABSTRACT

The Monte Carlo method is a powerful computational tool for estimating systems of mathematical equations that might be difficult to solve analytically. Since the enrollment dynamic of a university can be modeled with a small number of relatively simple equations, the Monte Carlo can be used as an approximate solver. This process is deterministic, making this method suitable for long-term forecasting only as long as the model's parameters are assumed to be constant over long horizons. Additional tests of the simulated enrolment data are also included, as they illustrate fundamental relationships existing among the variables in the model.

Keywords: Higher Education management, Monte Carlo, forecasting, R, strategy

INTRODUCTION

While universities in developing countries obviously need useful forecasting tools to make strategic decisions, they often lack appropriate reference points for deciding forecasts' accuracy. Data limitations are often a factor, as many universities in developing countries are young in comparison to established universities in America and Europe. At the same time, data systems within these universities might be incomplete or simply underdeveloped, making access to large data sets with reliable measurements difficult.

To help fill this need, this paper presents a Monte Carlo simulation of enrollment at KIMEP University, a North-American style university located in Almaty, Kazakhstan. The Monte Carlo method is a powerful computational tool for estimating systems of mathematical equations that might be difficult to solve analytically. Since the enrollment dynamic at KIMEP can be modeled using a series of equations, the Monte Carlo can be used as an approximate solver. This process allows us to estimate the steady-state population at KIMEP, given a set of parameters that can be calculated from a very limited data set using a bootstrapping algorithm.

The output of this model is generated by a deterministic process that is based a series of stylized facts about recent recruitment and enrollment. Very little data was used in the creation of the parameters; the variables with the most supporting data only had six to eight observations. Nonetheless, the model resulted in an enrollment steady state that was strikingly similar to actual enrollment levels at KIMEP University.

This method is suitable for long-term forecasting only as long as we assume that the model's parameters are constant over long horizons. Considering that many universities might be

interested in growing their student population in the future past this steady state, it rests on the university's administrators to alter current enrollment and attrition rates. While this desire will limit the model's applicability as a long-term forecasting tool, sensitivity analysis using this tool can give possible enrollment outcomes after key strategic decisions.

The rest of this paper will describe the methodology of creating the model and discuss the outcomes of a series of simulations. It will demonstrate how these enrollment scenarios can be tested to illustrate parameter sensitivity, scenario analysis and long-term enrollment planning. The statistical results of this paper are fully reproducible, as they were generated using the statistical programming language R. All annotated R scripts and data used for generating parameters are publicly available for interested researchers.[1]

METHODS: WHAT IS A MONTE CARLO SIMULATION?

"Monte Carlo" is a general class of statistical methods that solve a problem through sampling randomly generated numbers and iteration. The method was developed by scientists working on the atomic bomb project at Los Alamos in the 1940s (Kalos and Paula, 2009), and it has gained in popularity through its easy application in computer simulation, especially since computing power has become cheaper and more accessible (Binder and Heermann, 2010). Stanislaw Ulam is usually credited as its originator, but examples of this methodology date back to the 17th century, when it was used to solve particularly difficult integration problems. Since the method is probabilistic, solutions should always be considered approximations instead of exact answers.

There are two general types of solutions through Monte Carlo methods. In some instances, the problem can be solved through a description of the sample. This is true when the random numbers are generated using a specific algorithm. In others, like the problem we face here, Monte Carlo relies on randomly-generated numbers to approximate the solution to an algorithm. As an example of the former, consider the game of solitaire. It is possible to randomize the sequence of cards in a deck and solve strategy for placing cards into piles (Ulam, 1991). Iterating this process allows us to make general conclusion about the solution to the problem.

The latter type (the one we are interested in this paper) often overlaps with the general category of computer simulation, especially when we are simulating naturally occurring stochastic processes. This has become quite popular in finance, where the method can be used for pricing derivatives and assessing uncertainty (Boyle, 1977). In general, the process works by generating a random number (like a rate of return), applying a transformation and then adding another random variable. The process can occur over a long time series and then iterated to assess possible scenarios.

In the context of higher education, Monte Carlo has a distinct advantage of allowing the forecaster to define levels of uncertainty (Denham and Boston Coll., 1977). It also allows for the interaction of a wide range of variables, as long as the user is able to prescribe the necessary parameters (Denham, 1973). Since this process is purely deterministic and assumes

[1] The R scripts and base data sets can be downloaded here: https://dl.dropboxusercontent.com/u/2517942/MCmodel.zip

that parameters are fixed, this paper will avoid using the model for long-term forecasts. Instead, it will use it as a benchmark for other short- and medium-term forecasts.

METHODS: A DESCRIPTION OF ENROLLMENT

Quinn, Taylor and Kainazarova (2012) previously defined a general algorithm for determining enrollment at KIMEP, and this Monte Carlo simulation takes a slightly expanded version of that definition. Within the model, enrollment in a semester is calculated by using the following two equations. Since the enrollment parameters for undergraduates and graduate students differ substantially, the two groups are kept separate. Nonetheless, the algorithm for calculating enrollment in each group is the same.

The total number of active students equals the new students in a semester plus the number of students continuing from the previous academic year. The number of students in the model is a pseudo-random number that is generated using a normal distribution. The number generator takes the mean and standard deviation of the previous cohorts dating back to 2004. For undergraduate students, the historical enrollment mean was 801 students and the standard deviation was 175 students. For graduate students, the historical enrollment mean was 237 students and the standard deviation was 36 students.

EQUATION 1

$$E_a(t) = (1 - i) * (c(t) + n(t))$$
$$n(t) \sim N(\mu, \sigma^2)$$

where E_a is the number of actively enrolled students
i is the inactive rate for a cohort of students
$c(t)$ is the continuing students from the previous academic year
$n(t)$ is the number of number of new students
$n(t)$ is normally distributed random variable with a mean of μ and a variance of σ^2

Equation 2 provides most of the justification for using a method of approximation instead of solving this system analytically, since the equation contains a delay and an array that refers to previous cohorts. This makes the equation non-linear.

EQUATION 2

$$c(t) = (1 - a - w) * (c_{t-1} + n_{t-1}) - G' * N$$

where a and w are the academic leave and withdrawal rates; together they form the attrition rate
c_{t-1} and n_{t-1} are the number of continuing and new students in the previous year
G is a 4 x 1 array containing graduation rates by year of study
N is a 4 x 1 array of previous cohorts

Equation 2 calculates the number of continuing students using the total number of students in the previous year. The first element in the equation calculates the number of students lost to attrition using preset rates for academic leave and direct withdrawal. This paper considers withdrawal and academic leave as equivalent, since there is limited data on return rates.

In the model, students are given a four-year window within which they can graduate. For undergraduates, this is their third to sixth year at the university. For graduates, it is their first to fourth. The model assesses a probability for students graduating in each one of these years. For example, undergraduate students have a 4 percent chance of graduating in three years, a 41 percent chance of graduating in four years, a 17 percent chance of graduating in five years and a three percent chance of graduating in their sixth year at KIMEP.[2]

The number of graduates is found by multiplying these probabilities by the number of new students in previous cohorts. For example, the number of undergraduate students graduating in a particular year after three years of study is found by multiplying the first graduation rate (4 percent) by the $t - 2$ incoming cohort, and so on for each group of students graduating after a certain length of time. The total number of graduates is the summation of the number of graduates by cohort.

METHODS: CALCULATING PARAMETERS

Two sources of data were used to calculate the parameters that were used in the model. Graduation rates for individual cohorts were found using a report in the Enrollment Management section of KIMEP's website (Department of Enrollment Records, 2012). This report contained four years of information for undergraduates, but a fourth-year graduation rate for graduate students was not available. Instead, a best guess was made. These parameters are summarized in Figure 1 on the following page.

The other elements of the model were obtained using the Attrition Cohort report on the KIMEP Intranet (Department of Enrollment Records, 2012) and a Bootstrapping algorithm. A bootstrap is a tool for estimating statistical parameters when data is limited. To make this estimate, the bootstrap algorithm creates of series of new samples of the data, equal to the length of the original sample, by repeatedly resampling the original data set with replacement (Hesterberg, Moore, Monaghan, et al., 2005). This means that after each observation is randomly taken from the original sample, it is returned to the list so that it could be drawn again. Because of replacement, each new sample differs slightly from the original. A sample statistic is calculated for each of these new samples, and the bootstrap statistic is the mean of these separate sample statistics.

The advantage of using bootstrapping is that it allows us to take advantage of the Central Limit Theorem. The distribution of the sample statistics collected in the bootstrap should almost follow a normal distribution, as long as the algorithm is able to run a large number of times. This allows for additional tests of the data that might not be possible when the original distribution is unknown or the procedure for calculating the statistic is complex. Since the standard error is the standard deviation of the separate sample statistics in the bootstrap, it becomes quite easy to calculate confidence intervals and other descriptions of uncertainty.

FIGURE 1: ENROLLMENT, GRADUATION AND ATTRITION USED IN MONTE CARLO MODEL FOR KIMEP UNIVERSITY. THE NUMBER OF NEW STUDENTS IS AN AVERAGE OF INCOMING COHORTS SINCE 2004. FOR THAT REASON, A STANDARD DEVIATION IS ALSO PROVIDED.

[2] Due to attrition, graduation rates/ probabilities do not equal 100 percent.

	Mean	StDev
UG who Graduate in 3 Years	3.51%	-
UG who Graduate in 4 Years	41.00%	-
UG who Graduate in 5 Years	17.00%	-
UG who Graduate in 6 Years	3.00%	-
GR who Graduate in 1 Year	3.67%	-
GR who Graduate in 2 Years	14.34%	-
GR who Graduate in 3 Years	21.30%	-
GR who Graduate in 4 Years	29.00%	-
New UG	801	175
New GR	237	36

The results of these bootstrap estimations are included in the table below (Figure 2). They are generated from a data set that has a maximum of six observations for each variable. Despite this limitation, almost all of the standard errors are quite small, excluding the estimates for the rate of active graduate students. Fortunately, this rate is not necessary in the Monte Carlo simulation.

FIGURE 2: BOOTSTRAPPED PARAMETERS FOR EXECUTING MONTE CARLO ALONG WITH DESCRIPTIVE STATISTICS. THE FOLLOWING ARE INCLUDED: ESTIMATED MEAN, ESTIMATED STANDARD DEVIATION, STANDARD ERROR OF ESTIMATES (SE), LOWER AND UPPER BOUND OF 95 PERCENT CONFIDENCE INTERVAL (LW, UP).

Parameter	Mean	M-SE	M-LW	M-UP	StDev	SD-SE	SDLW	SD-UP
UG Active	78.48%	3.09%	73.29%	85.38%	7.00%	1.80%	5.37%	12.42%
UG Inactive	4.25%	0.41%	3.63%	5.22%	0.98%	0.31%	0.66%	1.89%
UG Academic Leave	2.84%	0.25%	2.34%	3.31%	0.57%	0.19%	0.39%	1.12%
UG who Withdraw	7.29%	0.43%	6.62%	8.28%	0.94%	0.29%	0.72%	1.86%
GR Active	48.99%	6.86%	34.31%	61.20%	18.98%	4.50%	15.76%	33.41%
GR Inactive	18.95%	1.90%	15.45%	22.88%	4.69%	1.67%	2.62%	9.18%
GR Academic Leave	7.88%	0.94%	5.93%	9.61%	2.12%	0.48%	1.62%	3.52%
GR who Withdraw	7.02%	1.64%	3.79%	10.21%	4.39%	1.09%	3.16%	7.43%

The standard errors for the academic leave and withdrawal rates of graduate students are also worth mentioning, at .94 and 1.64 percent, as is the standard error for the graduate inactive rate at 1.9 percent. These individual standard errors can be looked at as the reflection of the high standard error in the estimate of the average active rate for graduate students. The combined interval for both attrition measurements (academic leave and withdrawal) ranges from 9.72 percent to 19.82 percent. In other words, the mean attrition rate at the graduate level could be as low as 10 percent or as high as 20 percent. Unfortunately, this should be expected when we have limited data with high variance.

This is an important issue to keep in mind in any modeling scenario where data is extremely limited. When variances of the estimates are taken into account, a wide variety of enrollment outcomes becomes possible. This is a useful tool in scenario and sensitivity analysis, as will be shown in Section 4.

METHODS: EXECUTING THE MONTE CARLO

The simulation was written and executed using the R statistical programming language. The author is willing to provide an annotated version of the R script to anyone else interested in experimenting with the model.

The R script consists of five basic sections:

1. **Preliminaries** – This includes defining variables and arraying used in the model and importing the parameters generated by the bootstrap algorithm.

2. **A Function for Calculation Graduation** – This is called by the Monte Carlo simulation. It is separate for clarity purposes and to make sure that the references to the arrays avoid negative subscripts, e.g. referencing the value in the -3rd position in the array. While some programming languages can process these (usually counting from the end of the array), R cannot. This function returns two the values for graduating undergraduate and graduate students separately.

3. **An External Loop** – The first part of the Monte Carlo function creates the data frame to collect the results of the iteration and resets them to zero at the beginning of each new simulation. This external loop allows for the repetition of the interior simulation, which is the essence of any Monte Carlo simulation.

4. **An Internal Loop (The Simulation)** – This part of the function actually executes the enrollment simulation. For a number of periods defined by the user, the loop will calculate the number of new students, continuing students, total students, graduating students and attrition. Each period's calculation is saved to a simulation data frame.

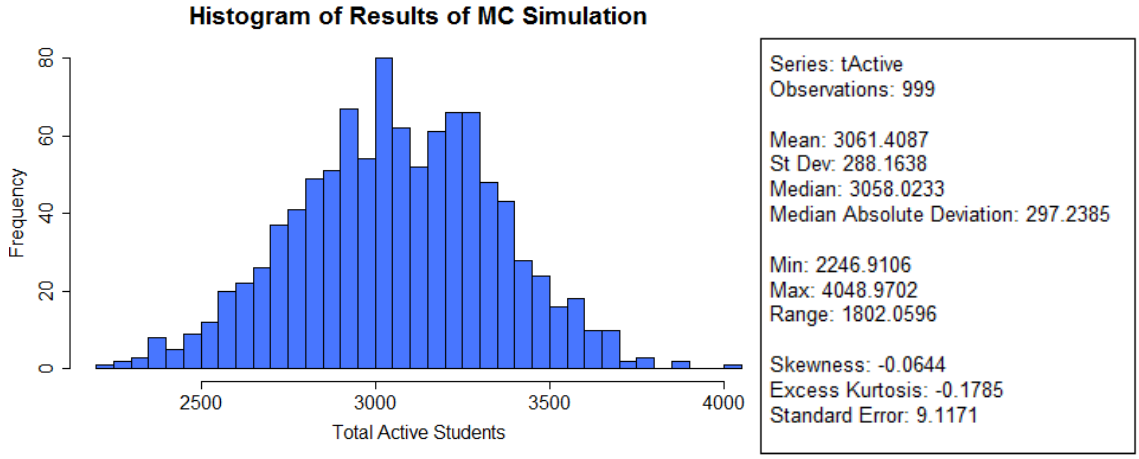
5. **Output** – The data frame for each completed simulation is saved to a list. The final period's enrollment data is saved to the results data frame in a row corresponding to the current iteration of the enrollment simulation. This process is repeated for the number of repetitions defined by the user. After the final iteration of the simulation, the completed results data frame is saved to the first place on the results list, making it easier to access.

While each individual simulation is mean reverting, thanks to the conditions set by the parameters, these simulations are very unstable (see following section). This necessitates the second loop in the Monte Carlo simulation, which executes the simulation multiple times. After a large number of simulations, the approximate solution to the system of equations will equal the mean of the distribution of the final states contained in the results data frame. We can also apply the Central Limit Theorem here, and the distribution of each variable in the results data frame should follow a normal distribution.

RESULTS: INITIAL OUTPUTS

The model resulted in a steady-state mean of approximately 3,075 students. This is not too far from the actual number of students enrolled in the F2012 semester at KIMEP, 3,251 (Department of Enrollment Records, 2012). This is an error of 5.72 percent, and it falls within the estimated interquartile range provided by the steady state (see Appendix). Considering that the model did not rely on any current enrollment data and it calculated its parameters using only six to eight observations of most variables, this is impressive.

FIGURE 3: HISTOGRAM OF ENROLLMENT SCENARIOS GENERATED BY A MONTE CARLO SIMULATION. IN THE RIGHT TABLE, DESCRIPTIVE STATISTICS ARE INCLUDED: THE MEAN OF THE OUTCOMES, THE STANDARD DEVIATION, THE MEDIAN, THE MEDIAN ABSOLUTE DEVIATION, THE MINIMUM, THE MAXIMUM, THE RANGE, SKEWNESS, EXCESS KURTOSIS AND THE STANDARD ERROR OF THE MEAN ESTIMATE.



For this project, 999 simulations of length 40 were generated. The decision for a 40-period simulation was mostly arbitrary. It was assumed that this would be enough time to reduce sensitivity to initial conditions, allowing for each simulation to come close to a general steady state.

The first assumption proved to be mostly true, as outcomes of the simulations rarely reflected initial conditions. This can be seen in the table on the following page (Figure 4). The average correlation between the final total number of students and that during any of the six initial periods (the length of an undergraduate cohort), was zero. On the other hand, correlations between subsequent periods were present, but these grew smaller with each period. As expected, the first period's outcome and the sixth period's were almost entirely uncorrelated.

FIGURE 4: RESULTS OF TWO-SIDED TESTS FOR PEARSON'S PRODUCT MOMENT CORRELATION ACROSS 7 DIFFERENT PERIODS IN 999 ENROLLMENT SIMULATIONS. THE ALTERNATIVE HYPOTHESIS IS THAT THE CORRELATION IS NOT EQUAL TO ZERO. PEARSON'S *RHO* AND STUDENT'S *T* STATISTICS WITH 997 DEGREES OF FREEDOM ARE INCLUDED. STATISTICALLY SIGNIFICANT TESTS ARE IDENTIFIED BASED ON THEIR *P-VALUES*.

	result	year1	year2	year3	year4	year5	year6
result	1	-0.0333	-0.0166	-0.0202	0.0209	0.0617	0.0547
	(Inf)***	(-1.0522)	(-0.5249)	(-0.639)	(-0.6594)	(1.951)*	(1.7289)*
year1	-0.0333	1	0.664	0.5035	0.4284	0.1214	0.0046
	(-1.0522)	(Inf)***	(28.0385)***	(18.4021)***	(14.9717)***	(3.8626)***	(-0.1464)
year2	-0.0166	0.664	1	0.7692	0.6089	0.335	0.0488
	(-0.5249)	(28.0385)***	(Inf)***	(38.0072)***	(24.2354)***	(11.2281)***	(-1.542)
year3	-0.0202	0.5035	0.7692	1	0.7967	0.5522	0.2879
	(-0.639)	(18.4021)***	(38.0072)***	(Inf)***	(41.6203)***	(20.9159)***	(9.4939)***
year4	0.0209	0.4284	0.6089	0.7967	1	0.7611	0.5084
	-0.6594	(14.9717)***	(24.2354)***	(41.6203)***	(Inf)***	(37.0529)***	(18.6411)***
year5	0.0617	0.1214	0.335	0.5522	0.7611	1	0.7634
	(1.951)*	(3.8626)***	(11.2281)***	(20.9159)***	(37.0529)***	(Inf)***	(37.3231)***
year6	0.0547	0.0046	0.0488	0.2879	0.5084	0.7634	1
	(1.7289)*	(-0.1464)	(-1.542)	(9.4939)***	(18.6411)***	(37.3231)***	(Inf)***

* p-value < .1
** p-value < .05
*** p-value < .01

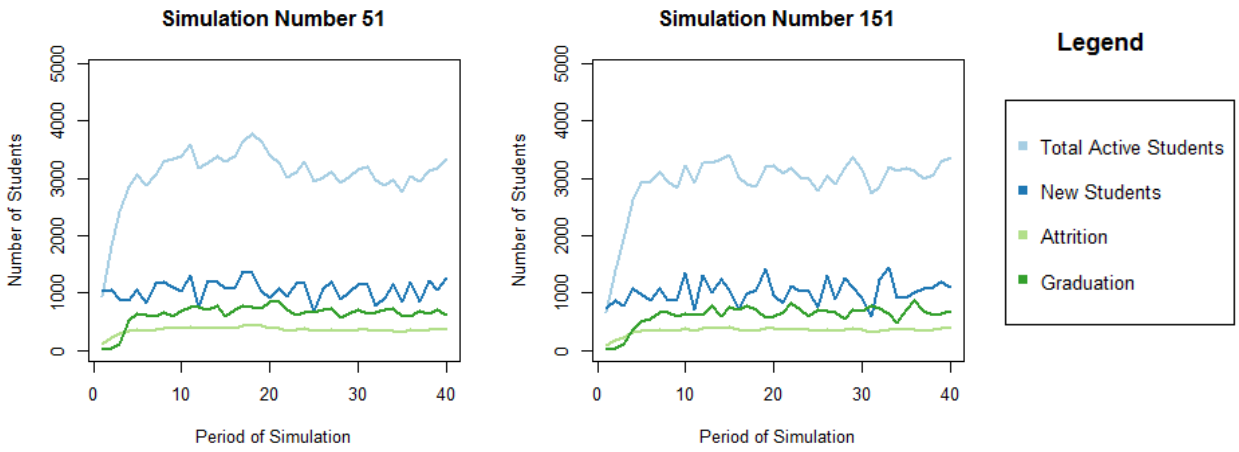
On the other hand, the second assumption proved to be false. Enrollment levels remained highly variable throughout each simulation. The average standard deviation for the total number of active students at any point in the simulation after the tenth period was 269 students;[3] the 95 percent confidence interval for the total number of students at a similar point in any simulation ranges from 2,575 to 3,631. Obviously, this is a very large range for an average population of approximately 3,100 students.

The instability in the enrollment simulations is also not resolved by increasing the length of each individual simulation. This is illustrated in Figure 5 on the following page. The outcomes of a Monte Carlo that relied on simulations with 99 periods were almost indistinguishable from those that used only 40. The average number of total active students still equaled about 3,100 with a standard deviation close to 300. This leads to an important conclusion: each individual simulation does not reach a steady state, regardless of its length. We can only estimate steady-state enrollment by iterating the simulations many times.

The necessary number of iterations is also a worthwhile consideration. For now, the Monte Carlo uses 999 simulations, which were enough to generate a relatively easy-to-see normal distribution (see Figure 3). This was not the case when only 99 iterations were used, which would require more tests to assess where the final distribution was actually normal. Moreover, multiple repetitions of the 999 iteration Monte Carlo yielded similar results: approximately 3100 total active students. This is also a reassuring indication.

FIGURE 5: ARBITRARILY SELECTED ENROLLMENT SIMULATIONS, INCLUDING TOTAL ACTIVE STUDENTS, NEW STUDENTS, ATTRITION AND GRADUATION

[3] The tenth period was chosen to avoid sensitivity to initial conditions.



TESTING THE OUTCOMES OF THE MC MODEL FOR VARIABLE SENSITIVITY

Additional analysis of the results of the Monte Carlo model for variable sensitivity is based on three statistical tests.

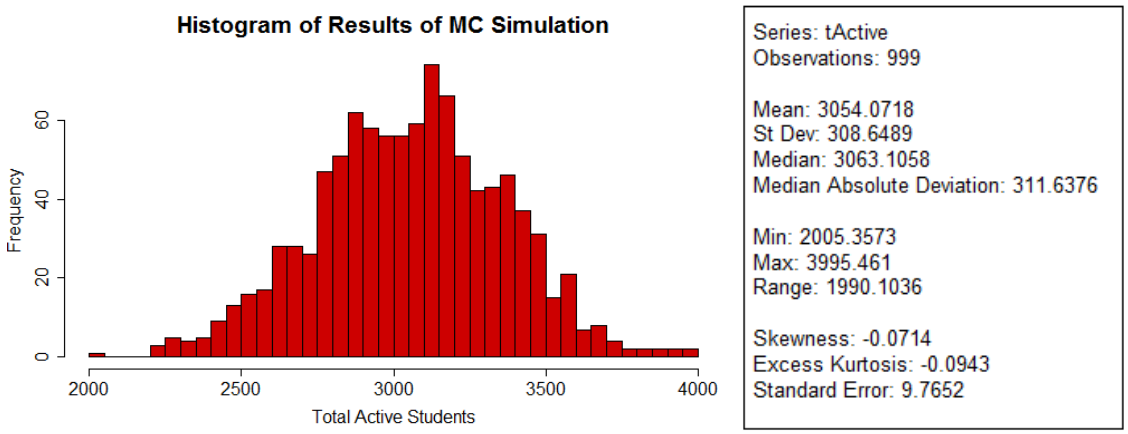
1. A linear regression model to trace variable dependency throughout the model
2. A prediction tree focusing on the impact of individual variables
3. A series of sensitivity tests, in order to assess the effects of specific variables in the model

For each of these additional analyses, the effects on graduate and undergraduate populations will be described separately. For this report, this has the added benefit of providing model comparison, further highlighting issues presented in the additional tests. It will allow us to ask whether these issues are related to specifics of the undergraduate and graduate enrollment dynamics, represented by the parameters used in the model, or whether they are issues specifically related to the way that the model was designed.

DATA TRANSFORMATION

Several data transformations were used to accomplish these sensitivity tests. While all of the enrollment parameters listed above (Figure 1) were fixed during the enrollment simulation presented in the previous section, a second version of the model was generated to treat each parameter as a pseudo-random variable following a normal distribution. In all instances except graduation rates (which were extrapolated from limited observations of undergraduate graduation rates), means and standard deviations were drawn from either the bootstrapping algorithm or KIMEP's key indicators (as was the case for the number of new students).

FIGURE 6: MONTE CARLO SIMULATION, RANDOMIZED PARAMETERS WITH GREATER RANGE AND VARIANCE



As the previous diagram illustrates (Figure 6), these randomized rates resulted in a series of outcomes with a larger range and greater variance that the Monte Carlo with fixed parameters. The model with fixed parameters had a range of 1802; the randomized parameter model had a range of 1990, an increase of 10.4 percent. Similarly, the randomized model's standardized deviation of 308.65 is 7.1 percent larger than that of the original model.

Despite the expanded ranges and standard deviations, the distribution of the results of the model with randomized parameters still followed a normal distribution. A Jarque-Bera test of the number of total active students in the final state of each enrollment scenario resulted in a chi-squared test statistic of 1.1763. This is not large enough to reject the null hypothesis that the results follow a normal distribution. The statistic has a p-value of 0.5553. It is assumed that the additional spread in the randomized results will further highlight the effects of different variables. This will be important for understanding variable sensitivity.

SAMPLING

Each Monte Carlo simulation generated a massive data set when running its enrollment scenarios, but much of these data were not used in calculating a steady state. The calculations refer only to the final state of each scenario, as was noted above (Section 2.4). Each enrollment scenario consisted of 40 states in this version of the Monte Carlo, which means that only 1/40 or 2.5 percent of the available data from the model was being used. The sensitivity analysis would benefit from a larger data set.

Correlation tests (Figure 4) demonstrated that samples starting after the 10^{th} state in any simulation should not show any strong relationships between sampled states and the scenario's initial conditions. In other words, final states can be usefully compared to samples at any other time (past the 10^{th} state) in any other scenario. Following this logic, a much larger data set was built to test the scenarios generated by the model. Limiting samples to only those past the 10^{th} state of each scenario, a sample data frame containing 14,985 observations was generated. This data frame contains 37.5 percent of the enrollment states

generated by the model.

Last, this data frame was not generated by directly pulling from states in the simulations. This is due to the delayed effects of certain variables affecting enrollment. For example, this year's attrition reduces the number of continuing students next year. The same is true for the number of graduates. For that reason, samples of this year's new students should be included with last year's attrition and graduation rates, as each value is partly responsible for the total number of active students this year. Samples were taken to reflect this fact.

A LINEAR REGRESSION MODEL

While the correlation tests showed that after enough time states are independent of each other, further analysis of the effects of variables is merited. In particular, we should know the persistence of individual changes to specific variables. For example, what effect could we expect to the number of students four states from now in the model if we reduced attrition in the current state? This question has a clear overlap with the management of KIMEP University. Is it worthwhile to invest in measures that directly affect enrollment variables?

To answer these questions, a basic linear model was built. It can be described by the following equation:

EQUATION 3

$$E_a(t) = \beta_0 + \beta_1 n_t + \beta_2 n_{t-1} + \beta_3 n_{t-2} + \beta_4 n_{t-3} + \beta_5 ar_t + \beta_6 ar_{t-1} + \beta_7 ar_{t-2} + \beta_8 ar_{t-3} + \beta_9 gr_t + \beta_{10} gr_{t-1} + \beta_{11} gr_{t-2} + \beta_{12} gr_{t-3} + \epsilon$$

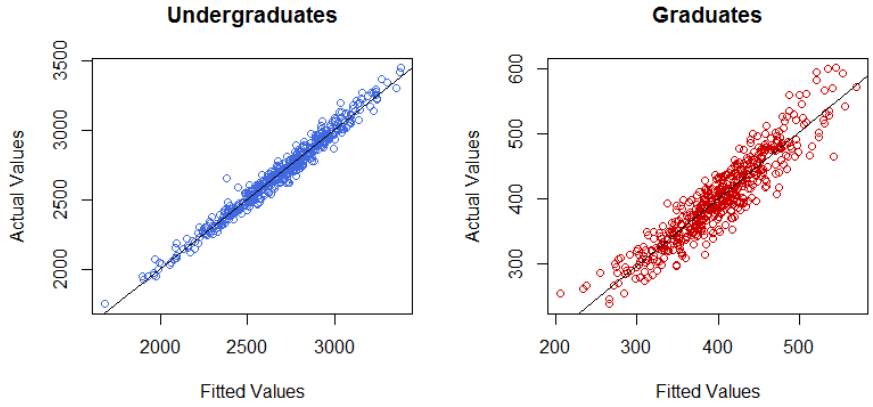
where E_a is the number of actively enrolled students
n is the number of new students in a year
gr is the graduation rate (from the previous year)
ar is the attrition rate (from the previous year)

The linear regression model described above (Equation 3) is based on assumptions of constant expectations. This means that the intercept for each model should more or less correspond to the average total active students at each level, before the effects of the parameters affecting enrollment are considered. In practice, the results do not perfectly follow this assumption, as illustrated in Figures 8 and 9 below. The intercept in the undergraduate model (2420.80) is less than the mean of the number of active undergraduate students (2683.67). On the other hand, the intercept in the graduate regression (417.95) is higher than the average number of graduate students (397.30), although the difference is not nearly as large as that at the undergraduate level, only 5.20 percent.

While one would assume that the graduate model is statistically stronger based on the correspondence between the intercept and the actual mean of the total number of active graduate students, this is not true. The coefficient of determination (R squared) for the undergraduate model (.9487) is quite a bit higher than that of the graduate model (.8418). Regardless, both fits are very strong, as one would expect when working off simulated data. This can be seen in the following chart (Figure 7).

An increase or decrease in attrition rates has an immediate effect that is proportionally greater than the actual change. In other words, a one percent increase in attrition reduces the number of students from the average described in the intercept by more than 27 students. This is a 1.11 percent decrease in the expected number of active students. A persistent reduction, i.e. a decrease in attrition that lasts for all six years described by the lags, would lead to approximately 87 additional students above the expected number of undergraduates in the model. This is a 3.58 percent change in the expected number of active students, which is proportionally considerably less than the accumulated reduction in attrition rates (six percent over six years).

FIGURE 7: THE FIT OF THE TWO REGRESSION MODELS FOR UNDERGRADUATES AND GRADUATES. THE FIRST 500 FITTED VALUES FROM THE REGRESSION ARE PLOTTED WITH THE ACTUAL NUMBER OF TOTAL ACTIVE STUDENTS.



The magnitude of the effect of graduation rates is greater in the first year but less over the full lags. If a persistent one percent decrease in graduation rates at the undergraduate level would occur in the model, we could expect the total number of active students to increase by about 71, a deviation of 2.94 percent from the expected value.

A similar pattern is visible in the regressions for graduate students (Figure 9), with the visible effects of a much shorter enrollment cycle. In the model, only 10.6 percent of new graduate students will still be studying at KIMEP after 4 years. Graduation and attrition rates have a proportionally smaller effect on enrollment at the graduate level. A one percent increase in attrition leads to a decline of 4 students from the expected number, which is a deviation of less than one percent.

FIGURE 8: COEFFICIENTS, STANDARD ERROR AND T VALUES OF HYPOTHESIS TESTS FOR UNDERGRADUATE LINEAR REGRESSION MODEL

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	2422.825	11.447	211.650	0.000
uNew	0.960	0.002	454.640	0.000
uNew.1	0.659	0.003	234.200	0.000
uNew.2	0.474	0.003	156.470	0.000
uNew.3	0.393	0.003	150.730	0.000
uNew.4	0.345	0.004	83.660	0.000

uNew.5	0.244	0.004	58.580	0.000
uAR	-26.954	0.322	-83.800	0.000
uAR.1	-18.758	0.327	-57.420	0.000
uAR.2	-14.308	0.331	-43.240	0.000
uAR.3	-11.488	0.328	-35.080	0.000
uAR.4	-8.254	0.328	-25.200	0.000
uAR.5	-7.056	0.325	-21.690	0.000
uGR	-29.507	0.276	-107.030	0.000
uGR.1	-18.000	0.276	-65.120	0.000
uGR.2	-5.009	0.197	-25.430	0.000
uGR.3	-7.180	0.155	-46.330	0.000
uGR.4	-4.786	0.154	-31.160	0.000
uGR.5	-6.842	0.133	-51.490	0.000

FIGURE 9: COEFFICIENTS, STANDARD ERROR AND T VALUES OF HYPOTHESIS TESTS FOR GRADUATE STUDENT LINEAR REGRESSION MODEL

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**
(Intercept)	409.1338	3.9936	102.4500	0.0000
gNew	0.8024	0.0055	145.9000	0.0000
gNew.1	0.4065	0.0072	56.4500	0.0000
gNew.2	0.2441	0.0070	34.8400	0.0000
gNew.3	0.1077	0.0073	14.7800	0.0000
gAR	-3.9937	0.0429	-92.9900	0.0000
gAR.1	-2.0342	0.0499	-40.7800	0.0000
gAR.2	-1.2164	0.0496	-24.5100	0.0000
gAR.3	-0.6109	0.0493	-12.4000	0.0000
gGr	-3.8572	0.0740	-52.1100	0.0000
gGr.1	-1.6419	0.0811	-20.2600	0.0000
gGr.2	-1.0385	0.0820	-12.6700	0.0000
gGr.3	-1.2995	0.0580	-22.3900	0.0000

Two lessons can be drawn here. First, the effects of an incoming cohort do not last very long, which has the paradoxical effect of making any enrollment outcome highly dependent on the most recent incoming cohort. At the graduate level, most have left by the very next year, and at the undergraduate level, most of the students have left within three years. While neither bad years nor good years have effects that last all that long, the sensitivity to most recent performance can make even a single bad year especially damaging. Second, efforts to change variables affecting enrollment dynamics tend to have a proportionally greater effect on enrollment, at least in the short term. Within the model, one percent decrease in attrition increases the expected number of total active students by more than one percent.

PREDICTING OUTCOMES

While linear regression is often the most typical tool in data analysis for gaining insight from data, classification trees provide an alternative route for understanding dynamics and making

predictions. While predicting outcomes of a simulation of enrollment might not seem like the most pressing issue at first glance, the implementation of a regression tree does have strategic value. In particular, unlike linear regression-based model, a tree does not rely on the application of a formula. This makes it much easier to interpret and apply for planning purposes.

Moreover, the relationship between variables within a tree does not have to be linear. This is usually a more accurate description of real phenomena, and it more accurately describes the series of equations used in the Monte Carlo model. The number of graduating students was calculated by referring to previous cohorts, and the number of students lost to attrition referred to the previous total number of students. Both variables lead to non-linear behavior in the model.

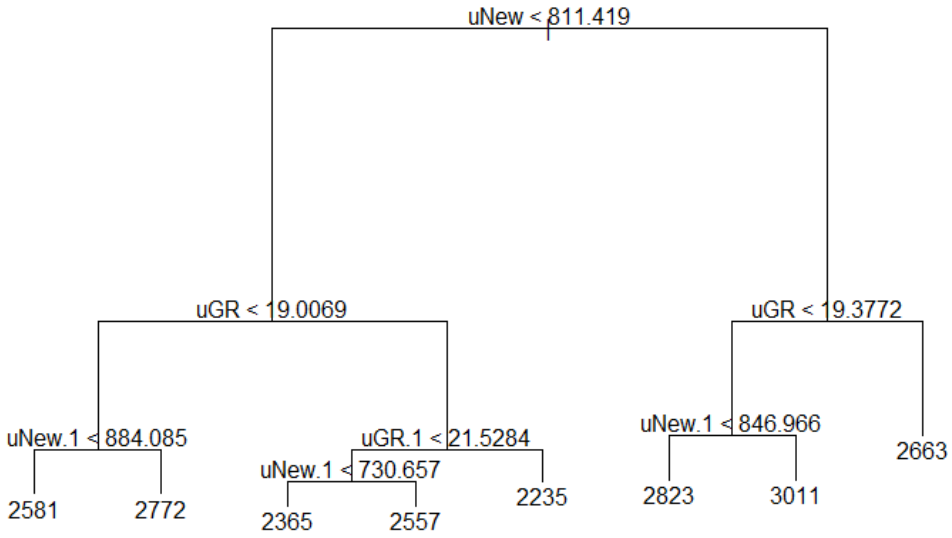
As Hastie, Tibshirani and Friedman (2001) explain, the basic regression tree seeks to minimize the deviance existing between two regions that are divided by a single binary split in a variable and a regression model. This process is iterated in the two splits of the previous data set, and it is iterated again for each new split that occurs. Theoretically, splitting can occur up to *N/2* splits, where *N* is the number of observations of the independent variable, but the tree algorithm in R stops this process once deviance gains fall below a specific threshold. In practice, splitting only occurs up to a predetermined number of terminal nodes (outcomes). Reducing a tree to a certain set of nodes is usually called pruning.

Trees generated with the statistical programming language R can all be interpreted in the same way. Each non-terminal node contains a logical expression. If the expression is true, follow the branch to the left. Follow the branch to the right if the expression is false. Continue this process until reaching the terminal nodes.

A pseudo-coefficient of determination (R squared) can be used to measure the goodness of fit for the models, which has the advantage of familiarity to those that use linear regression models. The undergraduate tree presented above has a pseudo R squared of .4443 while the graduate tree has a pseudo R squared of .5368. Admittedly, this is a quite a bit lower than the coefficients in the linear regression models (ugrad: .9487; grad: .8418).

On the other hand, utilizing a tree provides a simple and easily comprehensible way to assess enrollment dynamics. Just follow the tree through each node. Further, uncertainty is also easy to see, by adding and subtracting the 1st quantile and 3rd quantile of the residuals to each terminal node. Although this measurement is not perfect, since some nodes will slightly overlap at the extremes, it should suffice for most users.

FIGURE 10: REGRESSION TREE FOR TOTAL ACTIVE UNDERGRADUATE STUDENTS IN THE MONTE CARLO MODEL, PRUNED TO 8 TERMINAL NODES, PSEUDO R SQUARED = .4443




```
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-916.900 -153.100   -2.614    0.000  151.300  890.500
```

Many would be surprised to know that enrollment outcomes are predicted in this model without incorporating attrition rates, as is illustrated in Figure 10 above. Instead, only the number of new students in a current year and the graduation rate from the previous year are needed (plus the same variables at one lag in the undergraduate model). If this tool was applied to actual predictions of enrollment, one of the most difficult to forecast aspects of KIMEP's enrollment dynamics (attrition) could be avoided entirely. Instead, we could make reasonably accurate predictions using relatively accurate predictors of new students and graduates.

Moreover, the tool could also serve a useful role in the planning process, by establishing targets in enrollment variables. Working backwards from terminal nodes allows us to set or extrapolate desirable target values for the number of new students and the number of students graduating. In this way, the tree could become a primary benchmark when establishing future strategic plans.

VARIABLE SENSITIVITY

Finally, the last analysis of the Monte Carlo Model will focus on the simulations' sensitivity to underlying variables. To do this, an algorithm was developed that modified the Monte Carlo parameters across a predefined range and then executed the simulation with the changes to the parameters. In this instance, the range was defined as plus or minus 20 percent of the base parameters, with steps of one percent for each simulation.

An example will help. To test the sensitivity of the outcome to the number of new students, the base value is multiplied by .8 and the Monte Carlo simulation is run. The absolute difference from the base outcome is stored to a new array, along with the percentage difference from the base outcome. Next, the base parameter value is multiplied by .81, and the process is repeated. This continues at .01 intervals up to a value of 1.20, i.e. a twenty percent increase over the value of the base parameter. This basic process is repeated for each

parameter in the model, for both undergraduate and graduate outcomes. For the undergraduate level, this algorithm results in the following table (Figure 11).

FIGURE 11: SUMMARY OF SENSITIVITY TEST OF VARIABLES USED IN THE MONTE CARLO MODEL AT THE UNDERGRADUATE LEVEL

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
UGInactive.abs	-64.21	-20.94	-1.72	-3.60	14.56	86.77
UGInactive.per	-0.02	-0.01	0.00	0.00	0.01	0.03
UGAttrition.abs	-442.70	-250.97	-57.98	31.00	287.76	636.95
UGAttrition.per	-0.17	-0.09	-0.02	0.01	0.11	0.24
UGGraduates.abs	-911.66	-507.83	-62.82	-3.43	469.09	936.57
UGGraduates.per	-0.34	-0.19	-0.02	0.00	0.18	0.35
uNew.abs	-529.64	-263.16	-38.12	-3.59	237.42	558.77
uNew.per	-0.20	-0.10	-0.01	0.00	0.09	0.21

The limited responsiveness of the inactive rates is not that surprising. At the undergraduate level, they are already quite low, only 4.31 percent. On the other hand, the high responsiveness of graduation rates was not expected. A twenty percent decrease in the graduation rates led to a 35 percent increase in the total number of students. The effect was 1.75 times greater than the actual change in parameter. Changes in attrition and the number of new students also led to proportional greater effects on the total number of new students, but the magnitude of the effects was larger than the changes resulting from graduation rates.

The sensitivity tests followed a similar pattern at the graduate level (Figure 12), although the magnitude of the effects was larger for all variables except the number of new students. A 20 percent decrease in graduation rates resulted in a number of total active students that was 43 percent larger. This is twice as great as the effect of changing the number of incoming students.

A visualization of these effects is presented below (Figure 13), and it helps drive home the significant disparities resulting from changes in the graduation rate. The plots were generalized using normalized outcomes for the change in the variables, setting the first year to zero and plotting each subsequent change from the first year. The data also had to be reordered to match the steps, since the effects of changes to certain variables are not consistent. For example, an increase in the attrition rate decreases enrollment, while an increase in the number of new students increases the number of active students.

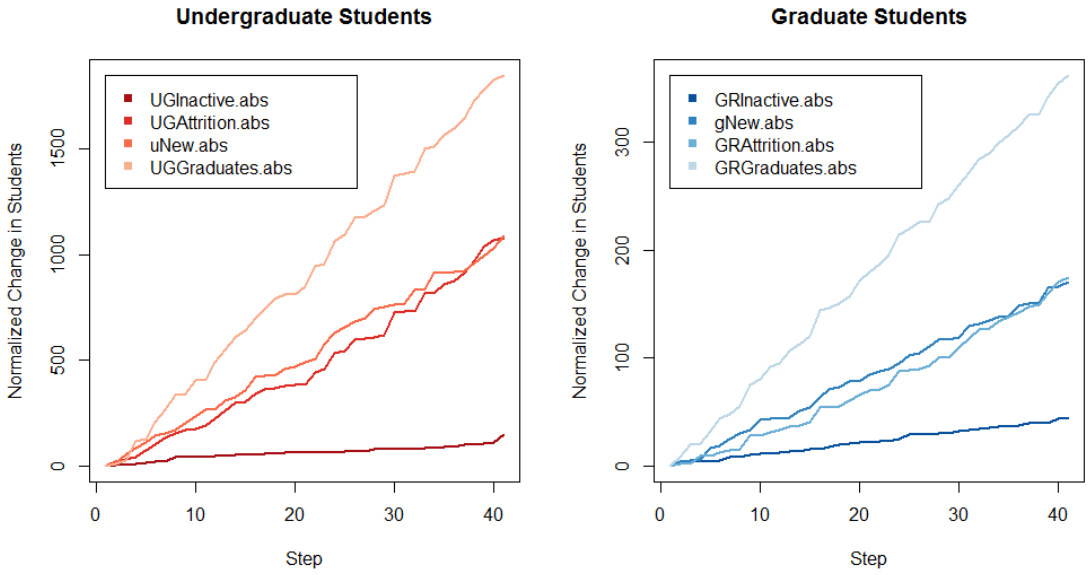
FIGURE 12: SUMMARY OF SENSITIVITY TEST OF VARIABLES USED IN THE MONTE CARLO MODEL

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
GRInactive.abs	-22.95	-11.25	-1.28	-0.89	9.78	20.72
GRInactive.per	-0.05	-0.03	0.00	0.00	0.02	0.05
GRAttrition.abs	-69.50	-38.66	-0.11	5.12	48.42	105.21
GRAttrition.per	-0.16	-0.09	0.00	0.01	0.11	0.25
GRGraduates.abs	-180.25	-88.42	-0.11	-0.87	92.07	181.50

GRGraduates.per	-0.42	-0.21	0.00	0.00	0.22	0.43
gNew.abs	-84.71	-40.89	-0.11	-0.89	45.07	85.79
gNew.per	-0.20	-0.10	0.00	0.00	0.11	0.20

Despite these striking results, there are a couple things to keep in mind. First, the changes are illustrated within a model of enrollment, and there is no guarantee that the same effect would occur if policy were implemented with the same goal in mind. Second, these changes occurred without any other variable changing. There could be positive and negative cross-over effects from attempting to change these variables: reducing attrition my increase the number of new students if it translates to higher quality programs. Reducing graduation rates, essentially forcing students to stay at KIMEP longer, could have a completely opposite effect on enrollment.

FIGURE 13: PLOTS OF THE DIFFERENCES IN ENROLLMENT OUTCOMES DUE TO CHANGES IN UNDERLYING PARAMETERS. ALL OF THE FIRST YEARS WERE SET TO 0, WITH THE DIFFERENCE IN THE TOTAL NUMBER OF ACTIVE STUDENTS PLOTTED FOR EACH SUBSEQUENT CHANGE (OR STEP) IN THE PARAMETER.



CONCLUSIONS

The Monte Carlo model described in this paper was able to provide an alternative prediction for enrollment levels that was based solely on derived statistics describing enrollment dynamics. For KIMEP University, the model produced a steady-state of enrollment that was only 5.72 percent lower than enrollment levels in the previous semester. The model corresponded to most expectations of a Monte Carlo: the results of the simulation followed a quasi-normal distribution after enough iterations, and outcomes of individual simulations did not depend on initial conditions. On the other hand, individual simulations remained highly variable and did not settle towards a mean value, regardless of the length set by the user.

While the output of the Monte Carlo and KIMEP's enrollment levels are quite similar given the constraints of the model, this should not be considered a forecast for future enrollment. The model is deterministic, based on set parameters that would be subject to change under normal conditions. Instead, the output of the model should be thought of as a reference point for current enrollment, given the conditions that the university currently faces. It should be useful for administrators looking to benchmark forecasts based on more suitable methods.

Additional tests of the model illustrated parameter sensitivity, which can be conducted by varying the parameters in the input and applying additional statistical techniques. While this paper is wary of extending the meaning of these results too far, sensitivity tests within the model showed that enrollment outcomes changed at rates that were proportionally greater than the change in the parameter. All three analyses showed that limiting graduation rates increased enrollment considerably, while the sensitivity and linear models showed large benefits to reducing attrition as well. While the real-world effects of reducing attrition might be comparatively smaller, these tests provide incentive for trying to address long-term enrollment variables.

Despite some of this tool's drawbacks, the Monte Carlo model gives a powerful message to university administrators. They should not expect significant changes in enrollment trends without a similar change in the fundamental factors affecting enrollment. In other words, growth is not possible unless attrition and enrollment rates are substantially altered. For many universities, changing these factors requires concerted effort, and development strategies should be drafted accordingly.

APPENDIX

FIGURE 14: SUMMARY OF VARIABLES IN THE FINAL STATES OF 999 ITERATIONS OF THE MONTE CARLO SIMULATION (LENGTH 40), FIXED PARAMETERS

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
New Undergraduates	999	807	174	813	808	176	231	1300	1069	-0.09	-0.11	5.52
New Graduate Students	999	237	36	237	238	39	119	335	216	-0.06	-0.27	1.15
Total New Students	**999**	**1045**	**179**	**1045**	**1045**	**178**	**519**	**1543**	**1025**	**-0.05**	**-0.23**	**5.65**
Continuing Undergraduates	999	1990	251	1988	1991	241	1210	2828	1618	-0.02	0.16	7.93
Continuing Graduate Students	999	256	38	255	256	37	123	380	256	0.01	-0.06	1.19
Total Continuing Students	**999**	**2246**	**254**	**2245**	**2247**	**242**	**1489**	**3089**	**1599**	**-0.03**	**0.11**	**8.03**
Total Undergraduates	999	2797	300	2799	2796	311	1890	3838	1948	0.03	-0.05	9.50
Total Graduate Students	999	493	53	495	493	53	340	663	323	0.01	-0.15	1.68
Total Students	**999**	**3290**	**305**	**3289**	**3290**	**311**	**2418**	**4356**	**1938**	**0.05**	**-0.11**	**9.65**
Active Undergraduates	999	2674	287	2676	2674	297	1807	3669	1862	0.03	-0.05	9.08
Active Graduate Students	999	401	43	402	401	43	276	539	262	0.01	-0.15	1.37
Total Active Students	**999**	**3075**	**290**	**3076**	**3074**	**298**	**2236**	**4090**	**1854**	**0.04**	**-0.10**	**9.19**
Undergraduate Attrition	999	285	31	285	285	32	192	391	198	0.03	-0.05	0.97
Graduate Attrition	999	75	8	75	75	8	52	101	49	0.01	-0.15	0.26
Total Attrition	**999**	**360**	**32**	**360**	**360**	**33**	**271**	**470**	**199**	**0.05**	**-0.14**	**1.00**
Graduating UG Students	999	516	81	519	516	83	241	761	520	-0.12	-0.11	2.56
Graduating Graduate Students	999	161	14	162	161	14	118	202	84	-0.07	-0.28	0.45
Total Graduating Students	**999**	**677**	**83**	**680**	**678**	**81**	**406**	**933**	**527**	**-0.12**	**-0.10**	**2.62**
Undergraduate Attrition Rate	999	10.19%	0.00%	10.19%	10.19%	0.00%	10.19%	10.19%	0.00%	-0.18	0.94	0.00
Graduate Attrition Rate	999	15.18%	0.00%	15.18%	15.18%	0.00%	15.18%	15.18%	0.00%	0.00	17.77	0.00
Total Attrition Rate	**999**	**10.94%**	**0.10%**	**10.94%**	**10.94%**	**0.10%**	**10.68%**	**11.28%**	**0.60%**	**0.39**	**0.21**	**0.00**
Undergraduate Graduation Rate	999	18.51%	2.75%	18.32%	18.47%	2.74%	9.73%	28.58%	18.85%	0.21	0.21	0.00
Graduate Graduation Rate	999	32.90%	2.92%	32.69%	32.79%	2.87%	25.29%	45.34%	20.05%	0.41	0.37	0.00
Total Graduation Rate	**999**	**20.65%**	**2.41%**	**20.53%**	**20.60%**	**2.39%**	**13.81%**	**28.74%**	**14.93%**	**0.22**	**0.10**	**0.00**

REFERENCES

Binder, K., and Heermann, D. W. (2010). *Monte Carlo simulation in statistical physics: an introduction* (Vol. 80). Springer.

Boyle, P. P. (1977). Options: A monte carlo approach. *Journal of Financial Economics*, 4(3), 323-338.

Denham, C. H. (1973). *Probability Distributions of School Enrollment Predictions Using Monte Carlo Simulation*. Journal of Educational Data Processing.

Denham, C. H., and Boston Coll., C. A. (1971). *Probabilistic School Enrollment Predictions Using Monte Carlo Computer Simulation. Final Report*.

The Department of Enrollment Records (2012). Attrition Cohort. Retrieved from http://intranet/infsector/statreport/AttrCohort.htm on March 5, 2013.

The Department of Enrollment Records (2012). Graduation Rates. KIMEP University. Retrieved from http://www.kimep.kz/er/reports/graduation-rates/ on March 5, 2013.

The Department of Enrollment Records (2012). KIMEP Enrollment Data as of Thursday, November 01, 2012. KIMEP University. Retrieved from http://www.kimep.kz/er/files/2012/11/EnrollmentStatistics_F2012.pdf on March 5, 2013.

Hastie, T., Tibshirani, R., Friedman, J. J. H. (2001). *The elements of statistical learning* (Vol. 1). New York: Springer.

Hesterberg, T. C., Moore, D. S. Monaghan, S., et al. (2005). Bootstrap methods and permutation tests. *Introduction to the Practice of Statistics*. David S. Moore and George McCabe, eds.

Kalos, Malvin H., and Paula A. Whitlock (2009). *Monte carlo methods*. Wiley-VCH.

Quinn, M (2013). A Forecast of Enrollment for the 2013-14 AY. The Office of Planning and Development Working Paper Series. February.

Quinn, M., Taylor, L., Kainazarova, M. (2012). A Three-year Economic, Enrollment and Financial Forecast. The KIMEP Department of Planning and Development Working Paper Series. September.

Ulam, S. M. (1991). *Adventures of a Mathematician*. University of California Press.