Predicting Enrolments in University Units

ITEC810 Final Report

Macquarie University

Author: George Gemayel

Student No.: 30629438

Supervisor: Steve Cassidy

Date Submitted: 5th June, 2009

Abstract

In this paper we present a prediction model capable of forecasting undergraduate unit enrolments at the Faculty of Science, Macquarie University. With limited data, we were able to define a model for 88% of units in computing, mathematics and physics at 200 and 300 levels. Our model was able to predict 2009 figures for 60% of all the defined units within 9 enrolment counts from actuals. In this paper, we examine the many external and internal factors that may affect unit enrolments, and their feasibility in our model. We also show how our model, which was only dependent on institutional data, was able to auto correct predictions when effected by external factors.

Acknowledgments

I gratefully acknowledge Dr Steve Cassidy for his support of this project, the supplying of data, the reviewing of deliverables and suggestions regarding the analysis.

I also like to thank Dr Robert Dale for his guidance and reviews of relevant deliverables. I also thank my fellow colleagues for their feedback and input during the life of the project.

Table of Contents

1 Introduction

The need for flexible programs of study at Universities has complicated the dependencies and hierarchies between units. What motivates a student to enrol into a unit may depend on a number of factors, predominantly:

- If it is a core unit
- If it is a pre-requisite unit
- If it is a popular unit

The first and the second factors are relatively straightforward to analyse since we can track the movement of cohorts of student from one unit to the next. The last factor however is dependent on many variables both internal and external to the university.

Study patterns also have an effect on unit enrolments. For example, some students choose to study full time while others study part time. Some students may choose evening classes while the majority prefer daytime classes. Some core units are also offered in both semesters of an academic year for those who have failed or those who have an unorthodox program of study. These students are more difficult to track.

To add to the complexity of unit dependencies, students are not limited to units in the faculty they are enrolled in. For example, a student enrolled in the Division of Science with an IT undergraduate degree can legitimately enrol into 100 level law units.

Postgraduate units at Macquarie University are less complex in their structure. There are no unit pre-requisites and so unit enrolments are mostly dependent on the following factors:

- It is a core unit
- It is a popular unit
- The units offering does not clash

Forecasting unit enrolments is therefore a complicated exercise. Ideally, the model would take into consideration external factors such as the economy, the unemployment rate, tertiary study trends and internal movements of students within degrees. However the aim of any good prediction model is to keep it simple. Therefore in this paper, we aim to address notable factors of unit enrolments and explore their purpose, if any, in our prediction model.

The rest of the report is structured as follows. In Section 2 we will discuss the aim of the project, the need to predict unit enrolments, and the requirements of Macquarie University. Section 3 will provide details of related work in the area of prediction, which will be used to determine the necessity and importance of various predictors. Section 4 will report on the data available in a form of a data audit. Section 5 will explore generic forecasting techniques and our approach in training the model. Section 6 will describe the details of training the model, and we introduce the dependency maps for three science streams. In Section 7 we report the results of the model and compare our enrolment forecast to actual figures. Finally in Section 8 we draw some conclusions from these results and outline any further work.

2 Problem Specification

Each year, Divisions at Macquarie University are required to predict enrolment counts for each unit per semester. The predictions are required to prepare timetables, to allocate lecture rooms, and to allocate other resources. It can also be used to determine expected gross income for that division that year.

Currently there is no accurate method used to predict these enrolment figures. Generally the process is a best guess, calculated as a flat percentage increase of 10% from the previous year. For the past four years however, enrolment figures for most science units have been decreasing, causing the predicted estimates to be way off the mark. This come as no surprise, since these rough estimates do not factor in the effects of university wide drops in course enrolments, failure rates of units, and most importantly student transitions from related units such as pre-requisites.

The advantage of the current process is that they are quick simple calculations. A unit conveyor is currently not required to apply some complex equation to estimate the forecasted count. Only some of the conveyors across Macquarie University have the skills to do this and fewer still have the time. The ideal scenario is to have a system that can crunch the appropriate data, apply a predefined model of best fit, and then output the forecasted figures for that academic year.

The goal of this project is therefore three parts. Firstly it must be determined wether a prediction model is feasible, given the complexity of the unit system and the amount of data that can be obtained. Secondly, a best fit model must be applied, which is flexible enough to account for the various factors and not too loose that it is not reliable. Thirdly, if a model can be proven to work, a system should then eventuate that implements this model and can use current data to forecast unit enrolments across the Faculty of Science, and if possible, the rest of Macquarie University.

3 Background and Related Work

There are no published works on predicting university unit enrolments; however there are a few papers on forecasting university enrolment figures. These projects are significant as they expose the various external factors that may determine changes in university course enrolments and hence changes in university unit enrolments. Other interesting aspects of these models are their approach and technique towards forecasting and the final outcome.

There are three aspects that categorise the success of a prediction model. The first is the identification and obtainment of data sources, the second is the model techniques and methods, and the third is the lifespan, or refresh rates, of the model.

The key to a successful prediction model is good quality data. If good quality data is not available, then the modelling process cannot be undertaken. In this section, we review the types of factors, external and internal, that may effect enrolment prediction. We will delve into how significant a variable they are and how complex the data can be obtained. Data elicitation is a juggling act between the eliciting and using of data and the significance of the data. We will discuss how in some successful models less is more, and in other cases, where data quality was dubious, how the results were affected.

The heart of a prediction model is its technique. We will investigate four different models with three different approaches to predicting university enrolments. These include linear regression, logistic regression and rule based prediction. We will identify which methods yielded good results and which didn't and why. Analysis of past successful techniques will give us a way forward. We will also consider the possibility of aggregating separate approaches and how they may affect future predicting models.

The lifespan of the model is indicative of its true value. Some models aim to project decades in the future, others are content with a short term projection of five years. We will investigate the accuracy of long term prediction with short term prediction and identify the need to refresh the model. We will also look into unique techniques to improve the lifespan of a prediction model.

3.1 Data, Parameters and Variables

If we knew every variable and parameter associated with every problem, then forecasting would not be a complex exercise. We wouldn't need to make assumptions or generalizations, and we would not need to apply in-depth statistical models or mathematics to solve a problem. Unfortunately, many factors need to be considered when predicting enrolments in university units, university programs or any other post secondary institution. When investigating predictors of enrolments, it is not just a question of what affects enrolment

figures, it is also a question of what data is available to us, and whether this data contributes significantly to the model.

Throughout the various prediction models assessed, there were common environment variables considered. These included the population count of the city, the strength of the economy and the value of education. Demographic variables of students were also commonly used such as age, gender and financial situation.

Herb O'Heron (1997), as senior analyst at the Association of Universities and Colleges of Canada, performed an in-depth study of forecasting undergraduate enrolment figures. His paper focused on two distinct parts of enrolment. The first analysis tackled the factors that influenced university enrolments, by looking at demands for a higher education, primarily for full time students at the typical age of enrolment (18 to 21 years old). The second analysis looked at the institution and movement of students between different parts of the post secondary system.

O'Heron (1997) goes on to mention that university enrolments are not predictable in the same way as primary or secondary enrolments. He explains that university enrolments based solely on population figures are usually wrong. Hence his analysis included demographic trends, population rates, economic cycles, social values, institutional fees and policy changes.

3.1.1 Population Ratios

O'Heron (1997) had found it difficult to establish any solid relationship between population and enrolment in the 18 to 21 year old range. Between the years of 1972 to 1997, there were three spurts of enrolment growth, but only one had occurred at the same time the 18 to 21 year old population was growing.

A participation rate is another method used to analyse demographic information by calculating the ratio between student enrolments and the population of the same age group. O'Heron had calculated a ratio of 10.5, which means 10.5% of the population of that age group were enrolled in Universities in the 1970s. He also found that after 10 years, in the 1980s, the ratio was identical. Although there was a large increase in enrolments in the mid 1990s, O'Heron attributed it to a large increase in Women participation, which levelled shortly after. The rise could not be linked to population and instead was attributed to demand.

Dan Anderson (2006) described a traditional enrolment forecasting model. His data sets included existing university population by retention rates; graduating and newly enrolled students by ethnicity; and ratios of non-resident and graduate students to undergraduate students.

Anderson (2006) made three comparable scenarios. The first was an aggressive analysis where the college going rates rose until 2010 then flattened. The second was a moderately aggressive analysis where there were smaller increases in college rates until 2010 then flat. The third scenario was a trend analysis where

the college going rates were kept constant. It was found that the trend analysis yielded more accurate results. In fact, the deviation between the aggressive forecast and the actual figures were double that of the deviation between the trend forecast and the actual figures. This result complements O'Heron's findings that changes population rates are not strong predictors of enrolment figures, and better results are achieved when the population rates is considered a constant.

3.1.2 Economic Cycles

According to O'Heron (1997), growth in the economy generally had a negative impact on university enrolments. The reverse is also true; enrolment growth was observed when the economy was at a stall. This can be attributed the wider range of choices a prospecting student may have when the economy is improving. These observations are conclusive, however such trends a very difficult to include in a model, due to the difficulty in obtaining relevant data.

Economic cycles will continue to have a push-pull effect on participation rates and thus economic changes will be expected to have a slightly negative effect on any forecasting model.

3.1.3 Financial Standing

Dr Helena Lim, Dr Rhod Davies and Dr Steve Jackson (2008) developed a predictive model of student enrolments. The model had a heavy reliance on Siefert and Galloways (2006) probability model of students' financial 'tipping point'. The aim was to determine how likely a prospecting student was to enrol based on the amount of awards. To calculate a student's financial 'tipping point', the model required financial data and admission data of the institution. Using logistic regression, the model was at best able to predict 20.2% of enrolled students based purely on financial controls. The low figure was attributed to the quality of the data available and not to the low significance financial standing may have in enrolment projections.

O'Heron (1997) theorised that the cost of attending university seemed to have little impact on participation. In the 1970s, there was a decline in tuition costs in Canada; however participation rates in universities also declined. In the 1980s, tuition cost increased, and likewise did the participation rates. The 1990s contradicted both decades, where participation rates levelled off, but tuition fees continued to grow.

There is no concrete evidence showing a connection between financial standing and enrolment levels. The cost of attending university is not limited to tuition fees; there are added costs of books and materials. Government grants and university awards also have an impact on a student's decision to enrol. Eliciting financial information is also a tricky task. Research, questionnaires and surveys must be conducted to get a realistic view of the student population, and even these cannot be assumed to be accurate. The difficulty in attaining financial

information is one of the reasons Lim, Davies and Jackson attributed their low prediction rates to the quality of the data. There may be merit in including financial information in a prediction model; however the data must be at a high quality, otherwise, it cannot be regarded as a significant control.

3.1.4 Institutional Data

The University Analysis and Planning Support division in The University of Central Florida (UCF) utilises a prediction model for course enrolments. The approach they employed during development of the model was to limit their data to internal university figures. That is, no external sources were considered. The model builds student headcount by starting with the returning students, based on the previous two years. Also the undergraduates are estimated using cohort retention from the previous years 10.

By limiting the control variables to just the last two years for returning students, and the last decade for new students, the model almost self-adjusts when external factors take effect. If for example the economy falls into a recession, O'Heron (1997) theorises that enrolment rates will rise. If this was to happen, it would be a gradual change from year to year, and since the control variables are only limited to the previous few years, the projections are automatically adjusted for an economic recession.

This model requires revision every few years, however at the moment; it is successfully predicting head count accurately within 0.5% for a one year prediction, and 2% for a five year prediction.

Since this model depends solely on institutional figures, the data is relatively easier to obtain. The input data is split between new enrolments, graduate enrolments, and continuing student. With such strong results, this approach to a prediction model puts a question mark of the validity of including external factors such as those described earlier.

In his analysis, O'Heron (1997) makes mention that due to the increased retention of university students and the delayed entry of older students into university, the causes of enrolment growth were difficult to determine.

O'Heron (1997) also concludes that changes in enrolment figures were traced back to simultaneous factors, which included the population of 18-21 year olds, economic growth or recession, the value of a tertiary degree and tuition cost. He continues to highlight that enrolment patterns will only change if there is an intervention from government, universities or other sources, to tip the scale in any of these factors.

3.2 Forecasting Techniques

In this section we will take a deeper look into three approaches to forecasting as used by four different enrolment prediction models. Although each model had differing purposes, the approach and techniques used are relevant in any prediction exercise. Extra relevance will be placed on undergraduate unit enrolment and the how the forecasting techniques may be applied.

3.2.1 Linear Regression

Whenever a linear regression model is fit to a group of data, the range of the data should be carefully observed. Attempting to use a regression equation to predict values outside of this range is often inappropriate, and may yield incredible answers. This practice is known as extrapolation

Dan Anderson (2006) utilises traditional enrolment forecasting as one of his techniques for predicting university enrolments. The analysis was limited to Arizona in USA and Arizona's population growth was included as a control variable.

As discussed earlier, Anderson takes three approaches in his projection model, each approach representing a different scenario for comparison. The first was an aggressive model, where college going rates increased until 2010, then flattened. The second was a moderately aggressive approach, where college going rates increased steadily until 2010, then flattened. The third was a trend approach, where college going rates remained constant through the projection (Anderson, 2006)

Although it was found that the trend approach was twice as successful as the aggressive approach, the overall results is what is interesting. The trend scenario at best was accurate within 8%. This is a large deviation when compared to the other models and is indicative of how risky or error prone classic extrapolation could be. However given the nature of predicting undergraduate units, extrapolation could be useful for 100 level units only, where the majority of enrolled student are new and enrolment figures may be more affected by external factors.

The purpose of the UCF enrolment prediction model, as mentioned earlier, was to provide a means of estimating headcount by student classification and semester for up to six years (a prediction year and subsequent five years). The model was developed using a base year in order to predict enrolments for the following year. The base year is defined as the preceding year. The model made use of cohort based retention fractions. This was an indication of the observed surviving fraction of undergraduate students, based on the student's annual entering cohort enrolled in a given classification.

Since predictor variables can vary, UCF had implemented a set of adjustment parameters so that the model fit the actual enrolment figures of the previous

years perfectly. These adjustment parameters were built in as part of the model, so that prediction for the current year had a stronger effect from the figures of the preceding year. This technique of interpolation is less risky than others. It ensures the model is a perfect fit to current data before predicting future counts (UCF, 2009).

The technique of using figures of a base year and segmenting the predictors into separate classifications can also be applied to unit enrolments. Students can be classified as new, returning or continuing. The most interesting feature of this model however is the multiplicative adjustment parameters. They ensure that the deviation of predicted figures from actual figures are corrected each time, so that projection errors are not carried through and compounded. This increases the lifespan of the model, and reduces the need for continual revision every few years. This is especially important since the model has no external predictors, as discussed earlier, and hence is susceptible to external factors such as the economic climate (UCF, 2009).

The results of this analysis were accurate within 0.5% in the prediction year, and within 2% for the five subsequent years (UCF, 2009).

3.2.2 Logistic Regression

Lim, Davies and Jacksons (2008) predictive model of student enrolment sought out to use logistic regression to calculate the actual amount of award required to positively influence a students decision to enrol in university.

Logistic regression is a method of predicting the outcome of a dependent variable based on a series of independent variables. In this case, the independent variables included socio economic and demographic variables such as age, gender and ethnicity, application variables such as open-day attendance, date of offer and level of offer, and financial variables such as household income and bursaries.

Logistical regression consists of fitting data to a logistic curve; that is, a curve that grows exponentially, before slowing down then ultimately plateaus. Lim, Davies and Jackson (2008) took two approaches in the model, the first was a univariate analysis, the second was using variables where the p-value < 0.25.

Lim, Davies and Jackson (2008) concluded that when dealing with data loss or considering the representativeness of the data, a univariant approach worked best. When the importance of the variables in the analysis was the main concern, then the approach of using variables with p-value < 0.25 should be used.

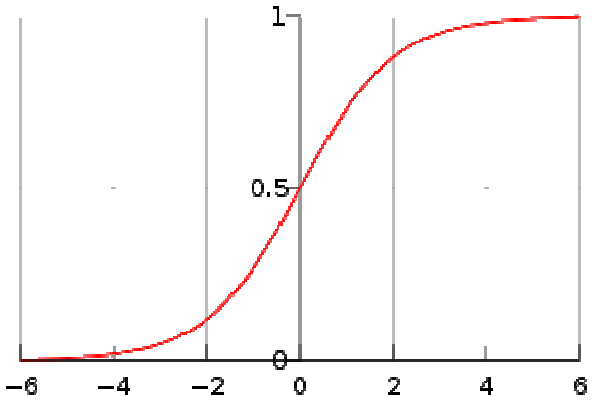


Figure 1: A logistic curve.

The steps taken during the analysis where:

a) Identify variables in the model

b) Run logistic regression

c) Identify and remove variables that are collinear

d) Identify and remove applicants that have unusual values

e) Rerun logistic regression.

(Lim, Davies & Jackson, 2008).

Lim, Davies and Jackson (2008) concluded that the results of the logistic regression was feasible but was very limited to the quality of data. The model did explain enrolment patterns but better results were expected in the future as more data and cleaner data is made available.

3.2.3 Rule Based Prediction

Aksenova, Zhang and Lu (2006), conducted an analysis of enrolment prediction through data mining. In this paper, support vector machines and rule based predictive models were employed to help predict total enrolment head counts, which is comprised of new, continued and returning students. By segregating the students, Aksenova, Zhang and Lu were able to implement different rules for each classification. Using these rules, the system was able to compute separate predictions for each segment. The approach was to build a predictive model for each student type independent of each other, then aggregate their results to generate the total headcount model.

The types of data included in this model were population, employment, tuition fees, household income, and historical enrolment figures. The processes were split into two phases; the support vector machines produced an initial prediction,

which was then imported into a tool to generate rule-based predictive models (Aksenova, Zhang & Lu, 2006).

The results of this model were encouraging, however the effort and resources required here exceeded that of the other models. The approach of segregating the students into new, continuing and returning groups and projecting enrolments for each does show some merit. It allows different rules, variables and adjustments to be made to each segment in isolation, meaning the analyst can be more granular and specific for each group. Merging the predicted results at the end will give the total enrolment projection across all students.

3.3 Lifespan of Model

The third aspect of a prediction model is the lifespan of the model, in other words, how far into the future can a model accurately predict and how often must it be revised or recompiled. This of course differs depending on the amount of data chosen, the relevance of the data chosen and the quality of the data chosen. It also heavily relies on the type of analysis and technique employed in the model. For example, if the model is self adjusting then the model may last a while longer, compared to a model that relies on external data sources.

Arun C. Mehta, from the National University of Educational Planning and Administration in New Delhi, developed a prediction model for student enrolments and flows. His approach in projection was to apply linear and non-linear models to fit time series data. Predictions were derived from enrolment data of the previous two years of both new enrolments and repeating students. The predictions were accurate; however it was found that the model constantly required revision on a periodic basis with the latest refresh of enrolment and repeaters data.

To increase the life span of an extrapolation model, it should be able to dynamically adjust the projection as variables change. Similar to the UCF prediction model where adjustments are made to fit the data to updated historical data. If the model cannot be dynamically modified, then it requires constant refreshing.

However, even though a model can dynamically adjust itself, it does not remove the need to constant revisit and refresh the model. It only extends the time period between refreshes. The UCF model, designed to predict up to six years in the future, requires constant refresh every few years (the precise time period was not given).

Although we have determined that the lifespan of a prediction model can theoretically last many years, we need to determine if any environmental effect can drastically alter the actual figures, skewing the projection and forcing the model to be refreshed. After building a predictive model for university enrolments in Canada, O'Heron (1997) concluded that enrolment patterns were only expected to change due to the intervention of government, universities or

other sources. This implies that unless major changes occur from the government or institution or both, then other factors are unlikely to alter enrolment figures significantly, and hence a prediction model can still accurately function.

3.4 Key Learnings

We have identified three significant attributes of a prediction model, which dictate the complexity, the accuracy and the relevancy of the model and resulting projections. We first focused on the data aspect of a model, and the external sources that may or may not act as control variables. These included population and its effect on participation rates, economic cycles and the financial standing of prospecting students. It was found that population growth does not directly correlate with an increase in enrolments. It was also found that by assuming a constant ratio between the population of the typical age group (18 to 21 years old) and enrolment rates yielded better results than a changing ratio. When dealing with unit enrolments, this ration may be a significant variable when dealing with new students only.

There was no concrete evidence that linked financial standing with a prospecting student's decision to enrol into university. This was attributed to the lack of data and does not necessarily mean that no correlation exists. The economic cycle however was found to have a reverse effect on university enrolments. This must be taken into consideration within a prediction model, however measuring such a variable is very complex and requires historic data that span multiple economic cycles.

We also had a look at the necessity to include external sources and data in a prediction model by analysing a model that only relied on internal institutional data. The model was found to work very accurately, and due to its self adjusting mechanisms, external control variables such as economic cycles had a reduced effect.

Data is just the first piece of the puzzle, the second piece is the forecasting technique used. We had taken a look at three techniques; linear regression, where a model makes an estimated projection based on derived rules; logistic regression, where projected data is fitted to a logistic curve; and finally rule based predictions, where the dependent variables are segmented with each having separate rules calculated and applied.

No technique was found to be superior when predicting university or unit enrolments. Segmenting students into new, returning and continuing student allowed each segment to have its own rules. For example, new students may be dictated by population participation rates; however returning students are heavily controlled by the curriculum. Both segments must be treated separate, with separate rules applied to each and separate projections for each. By incorporating self adjusting parameters, the projections can be altered to best fit

the base year's results before projecting future figures. Finally, all segmented projection should be aggregated to give the final prediction across all segments.

Lastly, we looked at the lifespan of prediction models. In most cases, the projections were accurate within five years before discrepancy with actual figures became an issue. We have concluded that any model, regardless of technique requires constant revision, especially since external sources, such as government and institutional policies, have a direct and sometime indirect effect on enrolment figures. We have also seen how implementing self adjusting parameters within a model can increase the life span and reduce the need to refresh the data.

4 Data Audit

Enrolment figures from 2001 through to 2009 were provided for the Faculty of Science, which encompasses units in computing, mathematics, electronics, physics and information systems. The files provided were in inconsistent formats, with unit counts having to be joined on unit name, study period and availability.

Various units are no longer offered and so these were ignored. Other units, specifically those that began in or after 2006, are too recent to have significant historical data and these too were excluded. Finally, other units had too small a number to extrapolate any significant findings. These units had less than 100 enrolments in total for the previous six years. These units may be revisited after the results have been validated; however the numbers are too small to use to train the model.

The trends of unit enrolments across the 8 years showed a strong decrease, with most computing units having 2008 enrolment figures a quarter of what they were in 2003. This was found to be general trend in the sector coupled with the many degree changes involving these units. It was also found that actuarial studies had deferred 100 level computing units to the second year, which made their enrolment trends inconsistent with other 100 level units.

A handful of other prominent and core computing units were also found to be approaching zero. Comp225 for example had only 49 enrolments in 2008, down from 410 in 2003. This was attributed to changes in programs of study which now offers alternatives to Comp225. This drop was also attributed to the extra 200 level units introduced with the ISYS stream, which, as expected, cannibalised some of these students.

Other inconsistent trends that required closer examination belonged to the 100 level ISYS units. Defying the general decreasing trend, enrolment figures for these units increased in 2005 and 2006. It was later verified that this strong increase is widely due to the number of Business students taking up these units.

Enrolment figures for courses were also requested to aid in the prediction of 100 level units. Unfortunately, data for enrolments in programs were never made available, and so historical data could not be obtained.

5 Approach

Forecasting is the ability to predict the future by examining trends and dependencies. The models are data driven, and typically have one target variable, which is the independent variable, and multiple dependent variables, which are the predictors. As discussed in section 3, background and related work, prediction models are best kept simple, where only a handful of significant variables are used. The most common types of prediction models fall into the following categories:

a) Regression

b) Decision Tree

c) Neural Network

Regression analysis is the ability to extrapolate numerical information based on response variables, and this is the most common approach used. Regression models are powerful for predicting a score based on other dependent scores, and generally involve linear transformations of the predicting variables. Regression models can be used in two ways, either to predict an exact score, within a degree of error, or to estimate an interval of scores, generally referred to as interval estimates.

Decision trees model possibilities and outcomes for sequences of event, and is represented as a tree like graph, with branches denoting the variables or events, and the leaves of the tree denoting the outcomes. It can be used to calculate probabilities for each likely scenario, e.g. how likely is a student to take up Comp226 if they satisfy the appropriate requirement. Although decision trees are simple to understand and implement, they require data from multiple sources to in order to calculate the various probabilities and possibilities.

Neural networks can automatically learn dependencies from numerical data with varying success. Its strengths lie in discovering hidden dependencies that can be used to predict the future that regression modelling may fail to find. The disadvantage to this is that neural networks are not represented by an explicit model. It is analogous to a black box, and hence is extremely difficult to implement the model into a working system.

With all types of prediction models, there are three data sets that should be used. The training data is used to calculate any correlations and build the model. The validation data is used to validate the model to ensure the correct variables and constants were used. Lastly, the test data which is used to test the success rate of the model. The test data is not compulsory since the model is already validated using the validation data. Given the limited amount of data available to build this model, all historical figures, up to and including 2008, were assigned as the training data, with 2009 semester 1 figures being defined as the validation data.

6 The Solution

6.1 Comparison Techniques

We began the analysis with comparing raw enrolment figures over time for all units. The data, after transformation, was inputted into excel and graphs were computed in an attempt to discover initial correlations. It was found that the enrolment figures varied considerably and so consistent trends were very difficult to determine.

The following figure compares the enrolment figures for two units, Comp330 and Comp343, both offered in the first half of the year.

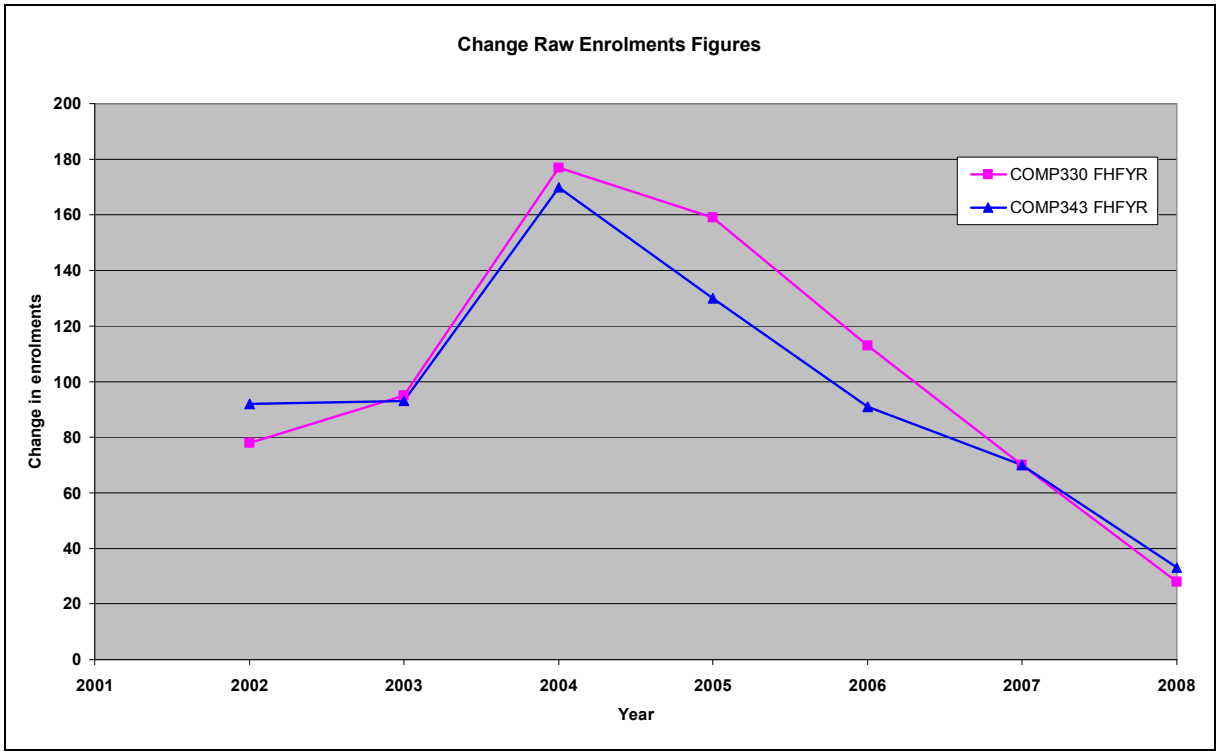


Figure 2: Comparison of raw figures between Comp330 and Comp343

As shown in figure 2, linear regression is not possible, as enrolments rose to 2004 then continuously fell until 2008. Also, these are not faculty wide trends, as other units showed a completely reverse pattern to Comp330 and Comp343. Figure 2 does provide some key points. Firstly, and most noticeably, the trends follow a similar pattern, that is, they both rose until 2004 then dropped until 2008. The other important point is that they seem to rise and fall at almost identical rates.

This key finding motivated the graphical analysis of how enrolment figures changed from year to year, and how these changes compared to each other.

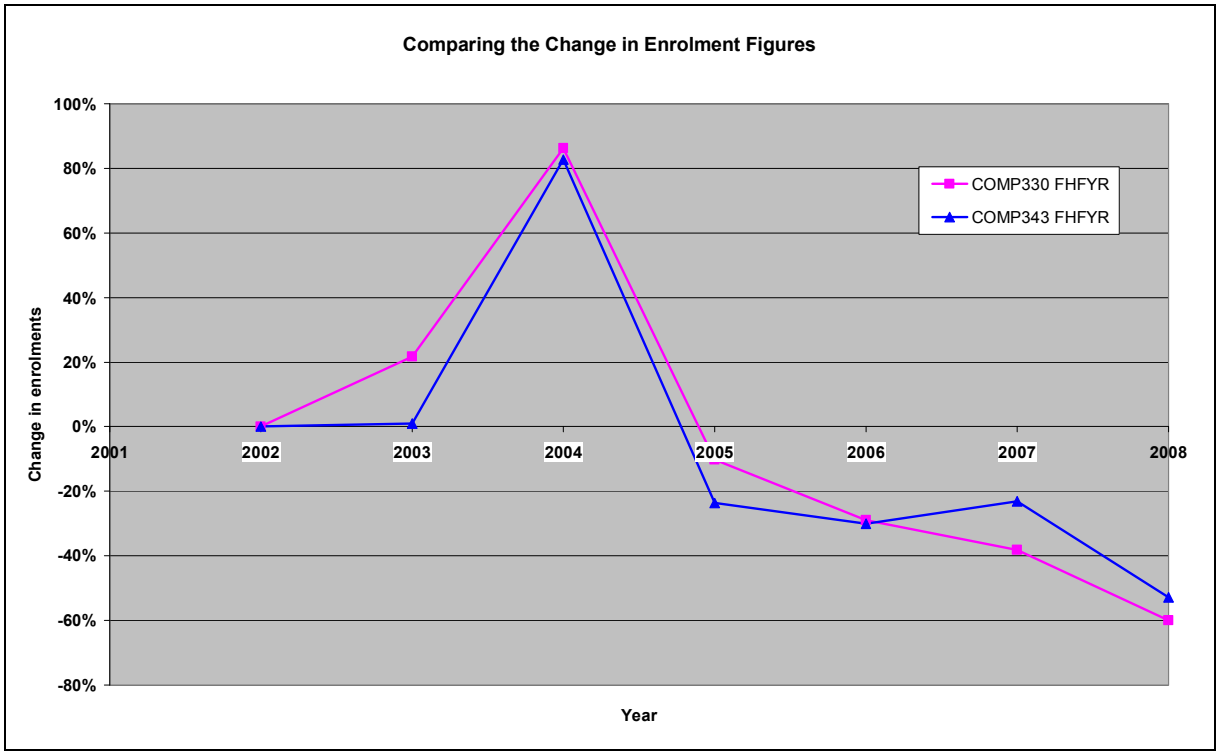


Figure 3: Comparison of deltas between Comp330 and Comp343

Figure 3 is a graph of the changes in enrolment figures, referred to as deltas, for Comp330 and Comp343. This graph shows more promise, as now we can clearly see a closer correlation between the two units, that is, when one rises the other follows, and when one falls so does the other.

It was found, by applying this technique across all other science units, that solid patterns could be found. As an example, Comp226 enrolment figures have decreased since 2003, however the rate of change was not constant; in 2005 it decreased by 33% and in 2006 it decreased by 40%. Comp229 enrolment figures are slightly higher than that of Comp226, however in 2005 it decreased by 22% and in 2006 it decreased by 25%. A clear correlation was found when the changes in enrolments were graphed, where Comp226 trends following that of Comp249 by a factor of 1.6. That is, if Comp249 figures decrease by 20% in 2009, as it did in 2007, then Comp226 figures for 2009 are expected to drop by 41%.

What we have now is a technique where we can cluster units based on the patterns of their deltas. This method also makes realistic sense, because all we are doing is grouping together units that act like one another from year to year. An example is Math136 enrolments following the patterns of its pre-requisite, Math135.

An interesting discovery is that in some cases, there is no direct correlation between units and their pre-requisites. For example, Comp330 has Comp225 as a pre-requisite however its delta pattern does not emulate that of Comp225. There are many reasons to this, but predominantly, Comp330 is not a core unit

and so it is not taken by all computing students. Comp225 however is a pre-requisite for most Comp 300 level units and is also a core unit in most programs of study. This finding however suggests that we cannot rely on pre-requisites to define enrolment dependencies between units, and so the analysis must be done without regarding relationships between units as specified in the handbook.

Discovering correlations between unit deltas was a graphical exercise. Unit deltas were calculated, graphed then analysed to determine consistent trends. Let us define x_{09} as the enrolment figure for unit x in year 2009. Let us also define x_d, the change in enrolment figure for unit x as:

$$x_d = (x_{09} - x_{08})/x_{08}$$

Let us also introduce unit y, where y_d is dependent on x_d by a factor of k (in our Comp249 and Comp226 case, k was equal to 1.6). Therefore we have:

$$y_d = k.x_d$$

$$(y_{09} - y_{08})/y_{08} = k.(x_{09}-x_{08})/x_{08}$$

$$y_{09}/y_{08} = k.(x_{09} - x_{08})/x_{08} + 1$$

$$y_{09} = k.y_{08}.(x_{09} - x_{08})/x_{08} + y_{08}$$

The final equation above allows us to predict the final enrolment figures of unit y in 2009 as a function of variables x_{08}, x_{09}, y_{08} and of course the constant k. Enrolment figures for x_{08} and y_{08} are known. The figure for x_{09} is to be determined from other units using the same technique or by using course enrolment figures for 100 level units. This is explained in section 6.2. Hence the role of this analysis is to determine the unit clusters and calculate the constant k.

An important note is that constant k was bound by upper and lower limits in most units. This is because in some cases, units grew by more than 100% or dropped by a factor close to 100%. Although the magnitude of the delta was considered unreliable, the direction of the delta, that is, did enrolments rise or fall, was still valid. Hence the value k was limited by |60%|.

$$-60\% \leq k \geq 60\%$$

To simplify this task, each unit was compared to units at the same level (e.g. 300 level) to find a core subject that dictates the rise and drop in all the other units at that level. This technique worked for the majority of units, with other unmatched units having too many outliers, or too few data points to extrapolate anything relevant.

It must be noted that a unit in the second half of the year can depend on units offered in the first half of the year. Likewise, a unit in the first half of the year

can depend on a pre-requisite which is offered in the second half of the previous year. For these cases, delta calculations must be shifted back one year. For example, Phys301 is predicted by Phys201 of the previous year. In these cases we have the association:

$$y_d = k.x_{d-1}$$

Comparing deltas of unit enrolments against similar units has many advantages, most notably; it replaces the need to track student transitions from one unit to another. Tracking clusters of students throughout their courses is a thorough but tedious method to build probabilities around a student's tendency to enrol in a unit. By clustering units depending on their deltas, as shown in figure 3, we are generalising at a high level that if Comp330 increases then there is a strong likelihood that Comp343 will also increase, without even considering the attributes of the student enrolling.

Another advantage of comparing deltas is that it replaces the need to look at external effects on enrolment figures. From related work, we have seen how factors such as the economy and population may affect overall enrolment counts, however these effects are generally institutional wide, or at least faculty wide. Therefore if our independent units, that is, the units that our model will depend on to predict other unit counts, are affected by external factors, then it passes on this effect throughout the model.

Another benefit of comparing deltas (x_d) of enrolment figures is that the accuracy of the model is constantly being corrected. i.e., an error in prediction for one year is not carried through to the prediction the following year, as the new deltas are applied to the actual figures of the preceding year.

What we have at this point is what we aimed for; a model dependent solely on internal institutional figures. This approach reduces the models susceptibility to external factors; for example, a increase in Comp225 will predict a rise in Comp248, without having to take into consideration why that increase happened.

6.2 Clustering of Units

Our analysis of deltas was applied to the majority of units, where data was sufficient and patterns were found. Units were clustered based on the delta patterns they exhibited, however there were also special consideration needed with respect to the following unit categories:

a) 100 level unit

b) Core units

c) Elective units.

100 level units cannot be treated the same way as 200 or 300 level units when determining dependencies and the constant *k*. This is because most 100 level units do not rely on pre-requisites. 100 level units have the most diverse students in regards to which faculty the students are enrolled in. Also they are more dependent on course enrolments, and so these units are more susceptible to external factors.

However this does not mean that trends of some 100 level units do not correlate with other 100 level units; for example, Math135 semester 1 is a predictor for Math136 semester 2. For those 100 level units that had no dependencies, these must be predicted using course enrolment figures. An example is Comp 125 enrolments may increase or decrease as enrolments in BCompSc increase or decrease. Unfortunately, no data was able to be sourced to follow through with this analysis, and so predictions for 100 level units remain unresolved.

As for core units, these were preferred as predictor units wherever possible, mainly due to the high number of enrolments figures which made calculating x_d more precise and significant.

Elective units tend to follow the trends of core units; however changes in enrolments of electives are more susceptible to popularity and offering of the unit. Hence, some were found to have some outliers, data points completely against the norm, for a random year, while the rest of the data points followed the general trend of the core units. These volatile units were identified as their predictions are more likely to deviate from the actual enrolment counts.

6.3 Dependency Maps

Dependency maps were drawn to illustrate the determined relationships between units, and it demonstrates the order in which predictions should be made for a target unit. These maps are sufficient if the predictions were to be replicated and they also play a vital role if the model was to be implemented as a system. The maps cover Comp, Math and Phys units and the constant *k* was included for each dependency.

6.3.1 COMP Units

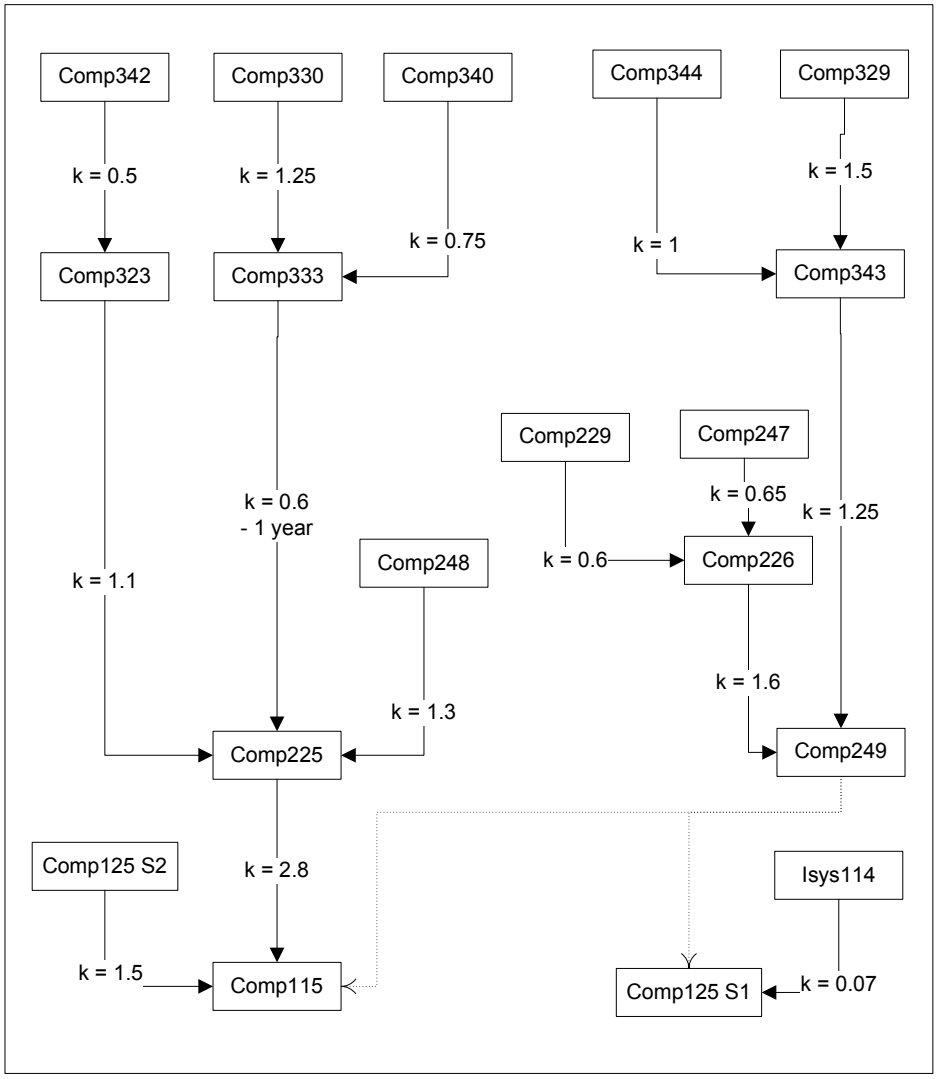


Figure 4: Dependency map for Comp undergraduate units.

The dependency arrows denote a unit being dependent on the target unit. For example, Comp248 is dependent on Comp225 with a value *k = 1.3*. Computing units were found to follow two trends, those of Comp225 and those of Comp249 and these two units also correlated with core 100 level units.

The dotted lines stemming from Comp249 to both Comp125 and Comp115 indicates that a correlation was made and is dependent on both units. However this association was found to be a bit dubious and requires further analysis with more data points. Since six units depend on an accurate prediction of Comp249, it was better to err on the side of caution.

Comp348 and Comp332 were found to have no correlations between other units and so these were not represented in the dependency map. Out of the 19 Comp units that were not excluded during the data audit phase, 15 units had dependencies identified, with the remaining 2 units, Comp125 and Comp115,

denoted as the bottom most dependent units. Isys units did not have enough data points to extrapolate any key dependencies.

6.3.2 MATH Units

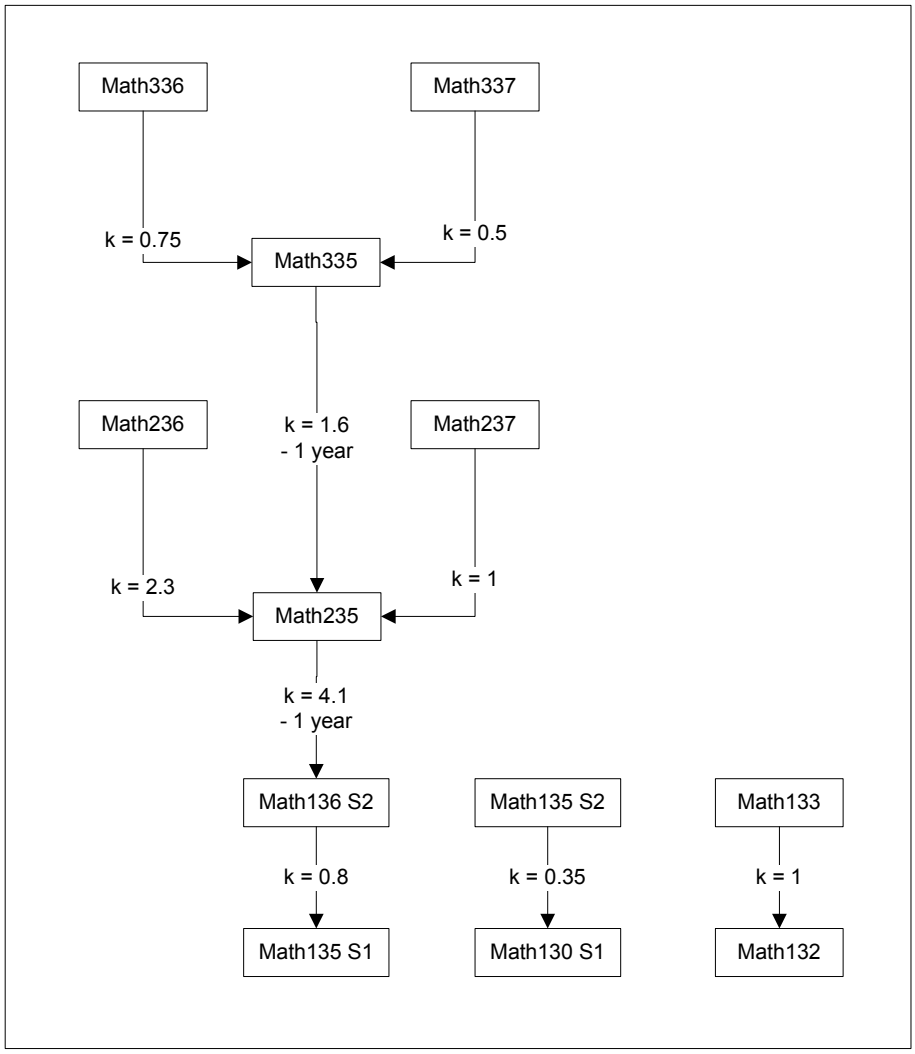


Figure 5: Dependency map for Math undergraduate units.

Dependencies in Math units were found to be very similar with all branches merging back to Math235. An interesting finding is that Math130 semester 1 predicts Math135 semester 2, likewise Math135 semester 1 predicts Math136 semester 2, which are the expected transitions of first year maths. This supports our assumption that tracking dependency units will replace the need to track cohorts of students and their transition between units.

Math132 and Math133 are advanced mathematic units, which explains why they are not connected to the rest of the map. After Math133, students tend to move into Math235, however given that enrolment counts of Math136 are much more significant than that of Math133, the latter is not included in Math235 forecasts.

Only 12 out of the 17 significant Math units were included in the map. However the missing 5 are all 100 level units, which as we discussed earlier, were limited due to the absence of course enrolment figures.

6.3.3 PHYS Units

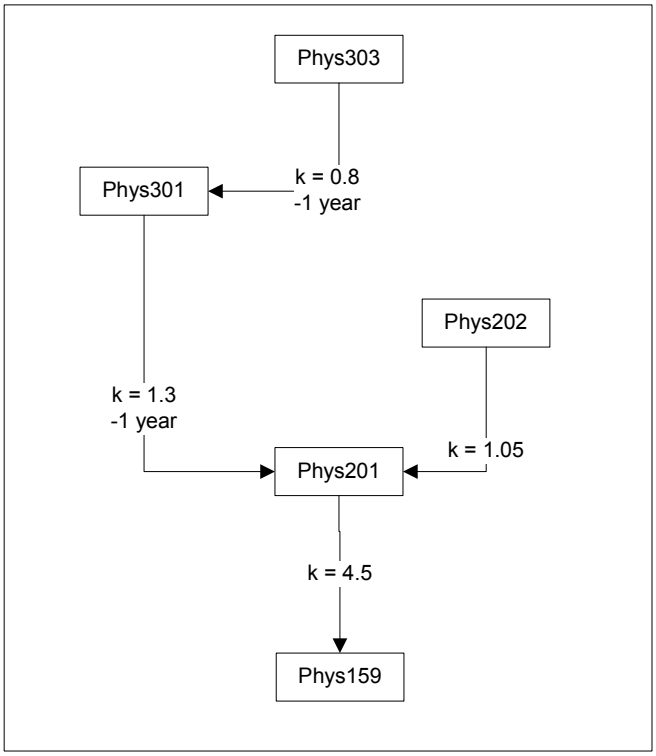


Figure 6: Dependency map for Phys undergraduate units.

Physics units are less complex in their associations, especially since there is so few of them. Physics courses generally require you to participate in all physics units which simplified the analysis of the model. Another factor was that physics did not attract many students from other faculty as we see in computing.

5 of the 7 significant Phys units were included in the map, with the exception of Phys149 and Phys270. Phys149 is a 100 level semester 1 unit and so our predictions were limited without course enrolment details. Phys270 is a specialised unit only for students enrolled in BSc Astronomy and Astrophysics. Because of its unique requirement, no correlation could be found between Phys270 and other Phys units. However, because of its unique requirement, predictions are expected to be possible based on course enrolment figures.

6.3.4 Rolling Up Dependencies

These dependency maps also provide the opportunity to roll up the dependencies to one or two master dependent units, with the aim of simplifying the number of associations. The result is that all units in a division are dependent on the same

predictor. From our maps, if Comp247 is dependent on Comp226, and Comp226 is dependent on Comp249, then Comp247 is dependent on Comp229.

From our equation, where y is dependent on x, we introduce z, which is dependent on y with a constant j. Therefore we have the two prediction equations:

$$y_{09} = k.y_{08}.(x_{09} - x_{08})/x_{08} + y_{08}$$

$$z_{09} = j.z_{08}.(y_{09} - y_{08})/y_{08} + z_{08}$$

We use the following substitution of y_d to remove the dependency between z and y:

$$(y_{09} - y_{08})/y_{08} = k.(x_{09}-x_{08})/x_{08}$$

Therefore we now have:

$$z_{09} = j.k.z_{08}.(x_{09}-x_{08})/x_{08} + z_{08}$$

Resulting in z being solely dependent on x with a constant ($j.k$).

7 Results

2009 enrolment figures were predicted for those units defined in the dependency maps in the Comp, Math and Phys streams. The baseline prediction was calculated as a 10% increase across all units and is indicative of current forecasting processes. The predicted figures were those calculated using our model as defined in the dependency maps. The actual figures listed are those obtained from Macquarie University as at the 23rd of March 2009. At that point, semester one figures were close to exact and semester two figures were only indicative. Delta values, which are the change in the enrolment figures from the previous year, are included in the result tables for baseline, predicted and actual figures.

Since we could not include course enrolment figures in our model, 2009 actual figure of bottom line dependent units were used to kick off the prediction processing. Bottom line dependent units are those units on the bottom line of the dependency map which has no predictor but which all other units dependent on. These are Comp115, Comp125 semester 1, Math135 semester 1, Math130 semester 1, Math 132 and Phys159.

The aim is to verify that our prediction methods are feasible for those units defined. We aim to identify where predictions were accurate and where further work is required.

7.1.1 COMP Units

Unit	Baseline Figure (delta)	Predicted Figure (delta)	Actual Figure (delta)	Baseline Deviation (%)	Model Deviation (%)
COMP125	131 (10%)	187 (57%)	216 (82%)	-85 (39%)	-29 (13%)
COMP225	54 (10%)	101 (106%)	108 (120%)	-54 (50%)	-7 (6%)
COMP229	74 (10%)	83 (24%)	81 (21%)	-7 (9%)	2 (2%)
COMP247	149 (10%)	170 (26%)	168 (24%)	-19 (11%)	2 (1%)
COMP226	28 (10%)	35 (40%)	38 (52%)	-10 (26%)	-3 (8%)
COMP323	36 (10%)	53 (61%)	15 (-55%)	21 (140%)	38 (253%)
COMP330	31 (10%)	26 (-7%)	23 (-18%)	8 (35%)	3 (13%)
COMP342	14 (10%)	17 (31%)	13 (0%)	1 (8%)	4 (31%)
COMP343	36 (10%)	43 (30%)	24 (-27%)	12 (50%)	19 (79%)
COMP340	6 (10%)	5 (0%)	13 (160%)	-7 (54%)	-8 (62%)
COMP329	25 (10%)	33 (43%)	19 (-17%)	6 (32%)	14 (74%)
COMP333	24 (10%)	21 (-5%)	23 (5%)	1 (4%)	-2 (9%)
COMP344	55 (10%)	65 (30%)	41 (-18%)	14 (34%)	24 (59%)
ISYS114	257 (10%)	241 (3%)	280 (20%)	-23 (8%)	-39 (14%)

Figure 7: Table of prediction results for Comp units.

A chart of deviation of baseline and predicted counts from actual figures has been included in Appendix A.

Figure 7 shows that 8 out of the 14 predicted units were accurate within 8 enrolment counts. Of the other 6 where the predicted value was incorrect, the baseline value was likewise inaccurate.

An interesting point was that our model was able to predict the decline in Comp330 figures when all other unit enrolments increased. This is because we defined Comp330 as dependent on Comp225 figures of the previous year. Since Comp225 figures dropped in 2008, we accurately forecasted a similar drop in Comp330 in 2009, even though Comp225 rose in 2009. This supports our assumption that some units follow the trends of another unit of the preceding year.

Comp229 illustrates how our prediction model is not dependent on external factors that affect enrolment figures. Since 2005, Comp229 enrolment figures have dropped consistently, i.e. 20% in 2005, 24% in 2006, 28% in 2007 and 22% in 2008. If our model was dependent on time trends, then it would have forecasted another drop in enrolment. Instead, our model has forecasted an increase of 24% in 2009 to raise the enrolment count to 83 which was only 2 counts away from actual, which was 81. Comp247 is another example where a rise was accurately predicted against declining historic data.

The decrease and subsequent increase in Comp229 figures could be attributed to many factors, including a general trend in university enrolments; however this does not concern the model. What is important to the model is that we have successfully linked Comp229 enrolment changes to that of Comp226, which rose 52%, and in doing so was able to not only accurately depict if the Comp229 counts would rise or fall, but also by how much.

Other predictions, specifically for units Comp125 semester 2, Comp323 and Comp343 were not so accurate. For Comp125 semester 2, it was only offered in the evenings in 2009, which may have affected the results. For Comp323 and Comp343, the dependencies and constant *k* appear to be chosen incorrectly. The training data consisted of declining figures throughout the past five years, and now that enrolments are on the rise, it is expected that dependencies on predictors may vary.

7.1.2 MATH Units

Unit	Baseline Figure (delta)	Predicted Figure (delta)	Actual Figure (delta)	Baseline Deviation (%)	Model Deviation (%)
MATH133	167 (10%)	155 (2%)	166 (9%)	1 (1%)	-11 (7%)
MATH135 S2	87 (10%)	74 (-6%)	88 (11%)	-1 (1%)	-14 (16%)
MATH136 S2	58 (10%)	64 (21%)	79 (49%)	-21 (27%)	-15 (19%)
MATH235	58 (10%)	21 (-60%)	41 (-23%)	17 (41%)	-20 (49%)

MATH237	77 (10%)	28 (-60%)	83 (19%)	-6 (7%)	-55 (66%)
MATH236	42 (10%)	15 (-61%)	23 (-39%)	19 (83%)	-8 (35%)
MATH335	15 (10%)	22 (57%)	13 (-7%)	2 (15%)	9 (69%)
MATH337	12 (10%)	14 (27%)	14 (27%)	-2 (14%)	0 (0%)
MATH336	14 (10%)	19 (46%)	10 (-23%)	4 (40%)	9 (90%)

Figure 8: Table of prediction results for Math units.

Unlike the predictions in computing, forecasted Math figures were more accurate at 300 level than they were at 200 level. There are two main explanations for this. Firstly, the model was trained on enrolment figures of daytime classes. However 2009 has seen strong enrolment counts for their evening offerings, which of course affect the outcomes of the predictions. Secondly, 300 level units were defined to be dependent on Math235 figures of the preceding year, so the inaccuracy of predictions at 200 levels did not carry through to 300 levels, yielding better results.

4 out of the 9 predicted units were accurate to within 10 enrolment counts, with the other 5 units being inaccurately forecasted. 6 out of the 9 units were better predicted by baseline than our model; however this seems to be more a result of our model incorrectly predicted a drop at 200 level where in fact it rose.

7.1.3 PHYS Units

Unit	Baseline Figure (delta)	Predicted Figure (delta)	Actual Figure (delta)	Baseline Deviation (%)	Model Deviation (%)
PHYS201	17 (10%)	17 (13%)	17 (13%)	0 (0%)	0 (0%)
PHYS202	14 (10%)	15 (15%)	9 (-31%)	5 (56%)	6 (67%)
PHYS301	17 (10%)	9 (-40%)	8 (-47%)	9 (113%)	1 (13%)
PHYS303	15 (10%)	13 (-7%)	6 (-57%)	9 (150%)	7 (117%)

Figure 9: Table of prediction results for Phys units.

Physics predictions were accurate for all 4 units defined in our model. In all 4 cases, deviations from actuals were below 8. Also only in 1 unit was baseline predictions more accurate than the predictions of our model. The percentage deviation is however still high and this is because all physics units have very low enrolment counts compared to mathematics and computing units.

8 Conclusion

The current method to predict unit enrolment figures at Macquarie University is inadequate as it is an exercise of simply increasing the previous years figures by 10%. The impact of inaccurately predicting enrolment figures includes the incorrect allocations of rooms, material and resources, and in general complicates the forward planning of University faculties.

By reviewing related work in the field of predicting University course enrolments, we were able to identify external factors, such as population ratios and economic cycles, which have an effect on course enrolment figures. However since eliciting relevant data for these sources is difficult, we also reviewed a prediction model, from The University of Central Florida, which was based on institutional data only. Its results were comparable to the models that incorporated external factors but also showed the attributes of being self adjusting, reducing the frequency to retrain the model.

Unit enrolment data was obtained for years 2001 through to 2009, with the first 8 years of data allocated to train the model, and the 2009 figures allocated to validate the model. After excluding units with insignificant data, we chose to focus on undergraduate units from the Department of Computing, Mathematics and Physics.

Since most enrolment figure were on the decline for the past 6 years, and other units showed inconsistent behaviour, linear regression could not be applied. Instead, we looked at how unit enrolments changed from year to year, which we labelled deltas. Each units delta patterns were then compared to delta patterns of units in the same stream, resulting in clusters of units that shared similar delta patterns. Using these patterns, we were able to define a relationship between units in the same cluster. Therefore we were able to predict if a unit enrolment should rise or fall, and the magnitude, based on the change of its dependent unit. Our dependency maps illustrate these relationships and the strength of the dependencies.

The results varied from stream to stream. In computing we were able to predict 8 out of the 14 defined units within an accuracy of 8 enrolment counts. For the other 6 units, baseline predictions were similarly incorrect.

The predictions for mathematic units were more accurate at 300 level than they were at 200 level. This was due to 300 level units being dependent on 200 level units of the preceding year which strengthens the prediction. But also, evening classes at 100 and 200 level have a stronger pull than previous years. The model would require retraining to take this into consideration.

Only 4 units were forecasted in the physics stream, and all four predictions were accurate within 8 enrolment counts and were slightly more accurate than

baseline predictions. Since physics units attract far less people than computing and mathematic units, these results could be misleading.

We have presented a model capable of defining dependencies between undergraduate units, and with these dependencies, we were able to show how trends in one unit can predict enrolment figures for other units in the same faculty. We defined the model free from external sources, and only relied on institutional data, hence simplifying the model, the variables and the data required.

The model is expected to improve pre semester as more data is collected, hence further work is necessary in two areas. Firstly more work is required to better define those units with inaccurate predictions, specifically mathematics. Secondly course enrolments must be taken into consideration for 100 level units, which in turn, will improve the model definitions for all undergraduate units. Given the simplistic nature of the prediction algorithm and its reliance on a readily available data source, implementing the model as an automated system is relatively easy. The model can be rebuilt using any programming language or a spreadsheet application such as Excel. However these issues must first be addresses and the predictions must cover more units more reliably.

References

Aksenova, S.S., Zhang, D., Lu, M., 2006. Enrolment prediction through data mining, *Information Reuse and Integration*, pp.510-515

Anderson, D., 2006. *Enrolment prediction techniques*. [PowerPoint slides]. Arizona Board of Regents

Lim, H., Davies, D., Jackson, S., 2008. *'Hark who goes there?': Developing a predictive model of student enrolment.* [PowerPoint slides]. Southampton Solent University

Marek Obitko., *Prediction using neural networks*. [Online] (Updated 3 June 2009) Available at: http://www.obitko.com/tutorials/neural-network-prediction/prediction.html [Accessed 3 June 2009]

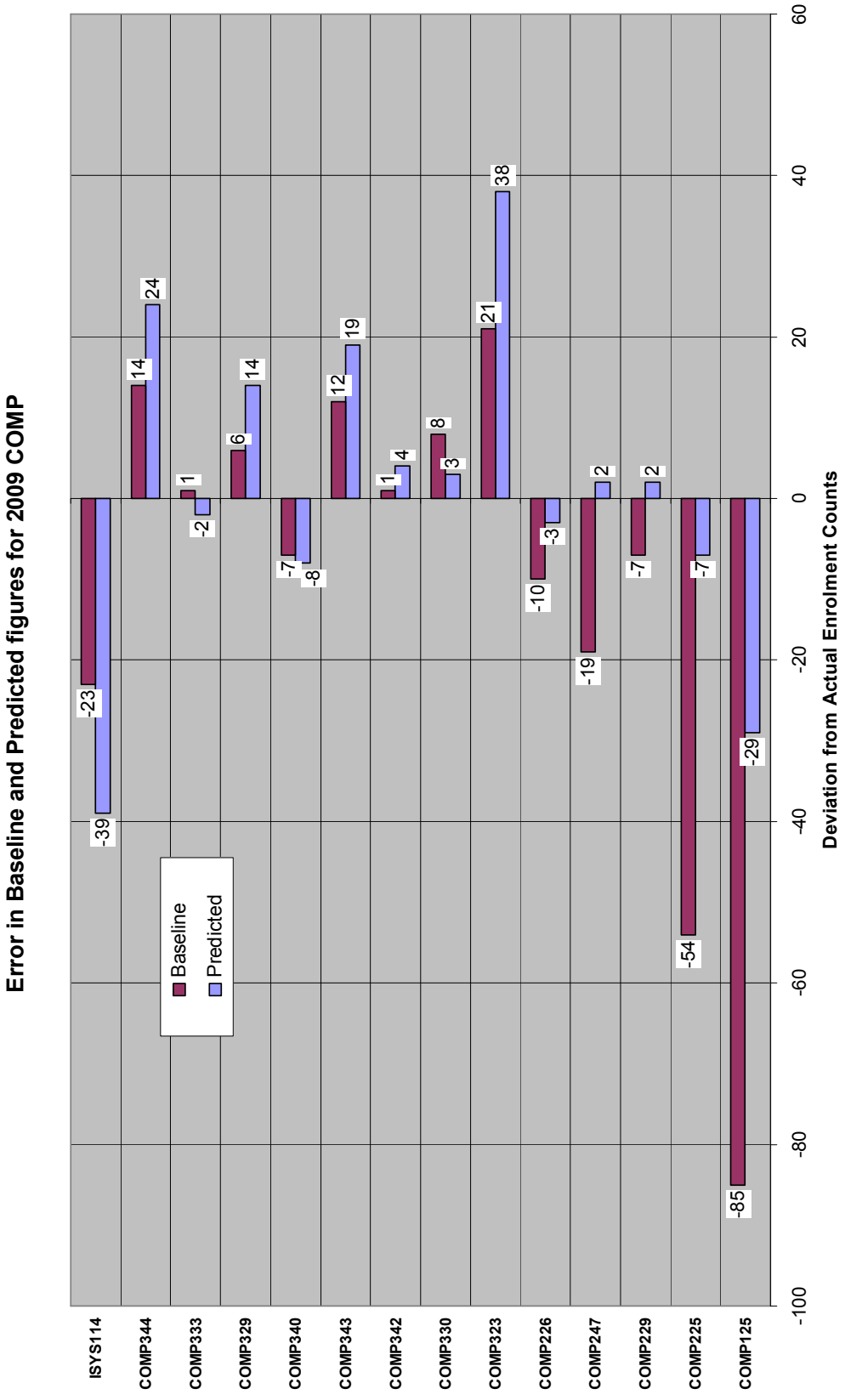
Mehta, C.A., *Projections of student enrolments and flows*. [Online] Available at: http://www.educationforallinindia.com/page24.html [Accessed 23 March 2009]

Stockburger, David W., *Regression Models*. [Online] (Updated 18 Feb 1998) Available at: http://www.psychstat.missouristate.edu/introbook/sbk16.htm [Accessed 26 May 2009]

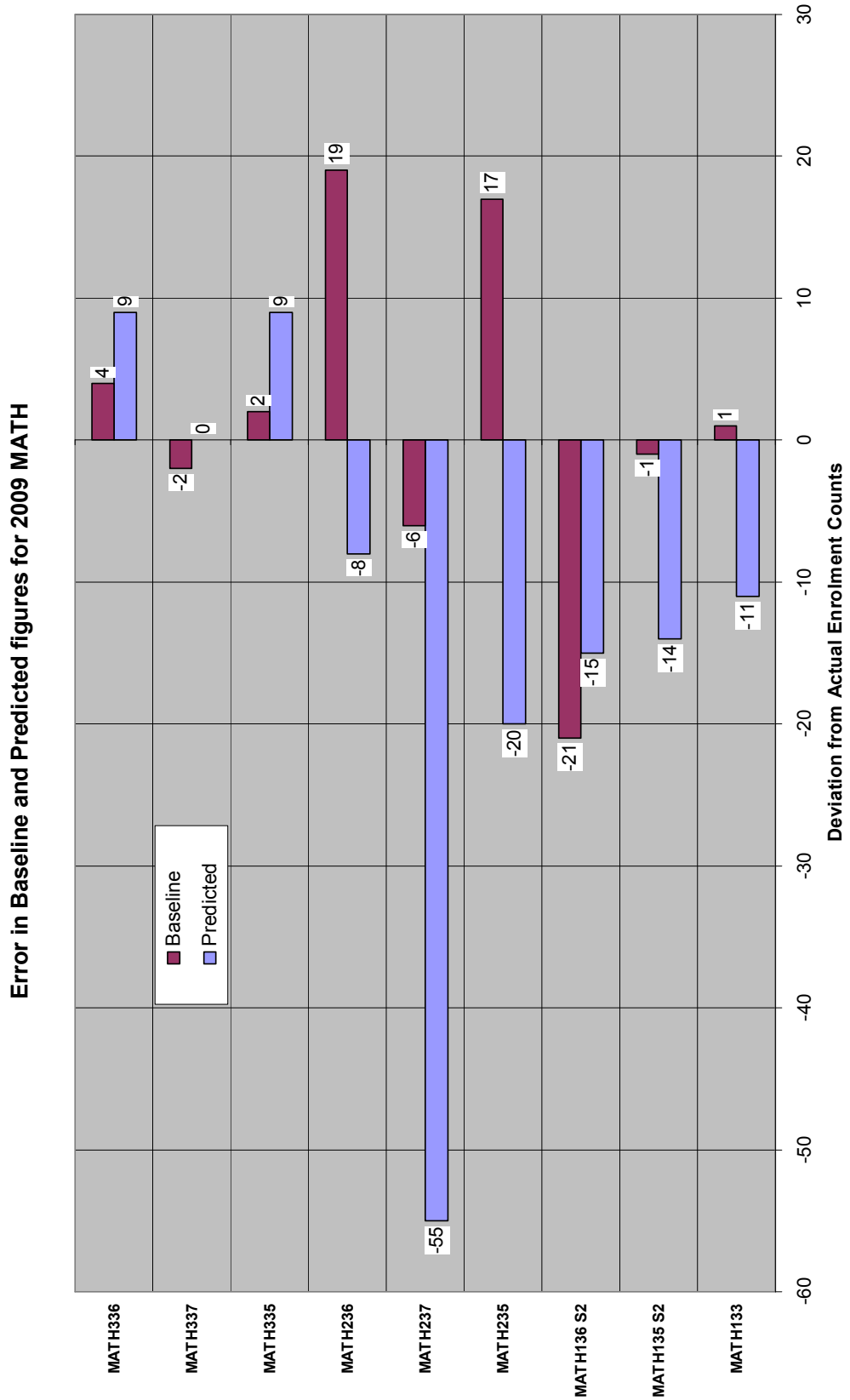
O'Heron, H., 1997. Undergraduate enrolment forecasts: A tricky science. *Research File*, 2(1), pp.1-15

University Of Central Florida., *Detailed enrolment prediction modelling method*. [Online] Available at: http://www.uaps.ucf.edu/enrollment/methods_detailed.html [Accessed 21 March 2009]

Appendix A – Deviations in Predictions: COMP



Appendix B – Deviations in Predictions: MATH



Appendix C – Deviations in Predictions: PHYS