Modelling Continuing Load at Disaggregated Levels

EWA SEIDEL

Flinders University

Abstract

The current methodology of estimating load in the following year at Flinders University has achieved reasonable accuracy in the previous capped funding environment, particularly at the university level, due largely to our university having stable intakes and student profiles. While historically within reasonable limits, variation in estimates at the course level is increasing due to the removal of the capped environment, increased competitiveness across universities, and changing student composition, profiles, and study patterns. This translates to uncertainty in funding and how it is distributed across courses. It is now necessary to predict load in a way that accommodates the changing higher education landscape, with greater accuracy at the course level.

This article compares the current method of estimating continuing load in the following year with an alternative method developed by the Planning Services Unit. The current method creates one estimate per course and utilises the previous year's continuation rate unless exogenous information suggests otherwise. The proposed alternative method disaggregates courses according to student academic characteristics that are associated with continuation rates. The method uses a generalised linear statistical model, derived from varying amounts of historic data, to estimate continuing load separately within each course cross-classification. This article will describe the logistics associated with, and the benefits of, applying the new method when predicting continuing load in Funding Group 1 (Commonwealth supported load) in 2013.

Keywords: method, estimate, load, continuing, model

Australia's Government removed an undergraduate funding cap, permitting equilibrium between supply and student demand of higher education (The Commonwealth of Australia, 2009). The funding cap removal has allowed Flinders University to increase its domestic intakes each year, and keeps us on track for reaching a strategic plan target of over 25,000 enrolments by the year 2016 (Flinders University, 2012). Flinders needs to accommodate the growth, while also being prepared for unexpected changes in growth magnitude and direction. We must build a robust planning process as our existing one is unable to streamline changes occurring in the student cohort profile.

Flinders University, a member of the Innovative Research Universities network, currently teaches almost 23,000 students. It is located in the southern suburbs of Adelaide and attracts a relatively large number of non-traditional students such as those who have low socioeconomic backgrounds, are mature-aged, or are not school leavers. Admission of these students continues to grow, which is enabled through alternative entrance pathways and the cap removal. We expect the growth of non-traditional students will affect continuation rates, increasing the need for a more robust planning process.

As Flinders University expands, it introduces more courses, either in areas of new expertise, or specialisations of existing expertise. In addition, it is common for courses to undergo slight changes in context, naming and identification. All these reasons combined make longitudinal analysis, such as load modelling, difficult. Flinders' current planning process deals with these issues manually, and would benefit from a more robust planning process.

The planning process is further complicated when taking into account changes among all students, both traditional and non-traditional. Students are experiencing pressures associated with increases in cost of living; an increased prevalence of seeking and achieving a greater balance between work, study and life; a growing choice of courses; increased flexibility in delivery mode and the flexibility to change courses entirely. The planning process must consider that our entire higher education student cohort is changing.

The Flinders Planning Services Unit (PSU) is developing a data warehouse environment that will allow the planning process to be conducted in the same location as the source data, as well as other data-driven processes. The new environment includes common source data across processes, allows measurement and scenario building, incorporates faculty feedback and streamlines outputs with other budgeting and finance processes. The benefit of the new environment is that it will allow the planning process to incorporate more data and information, enabling a more robust process to exist.

Load projections are measured separately for commencing and continuing load since the uses, inputs, parameters and methods are quite different. Historically, commencing load estimation has been guided by faculty staff and mostly deterministic due to greater demand than supply. The cap removal warrants investigations into utilising additional data sources into the estimation process. While commencing load estimation is important, the source of greater uncertainty that demands review is continuing load estimation, the focus of this article.

Currently, the continuing load estimation method used in the PSU produces a single load estimate per course by funding group. While the method is performed at a relatively aggregated level, it has been suitable and accurate enough for its intended use thus far. However, the aggregation may no longer be suitable as downstream processes and users become more sophisticated and allow or demand more intelligence.

The existing methodology is unable to automatically incorporate information relating to changes in student profiles and characteristics within each estimate, and relies on manual intervention. This methodology has worked well for courses in steady-state and particularly at the university level, where estimate aggregates have historically been within 1–2% of total actual load. However, it does not work well for courses out of steady-state, which is particularly relevant now since we are experiencing and expect more changes in student intakes and characteristics.

Research work within other Australian universities has either acknowledged the need to or achieved improvements by incorporating more detailed enrolment information into the load modelling process (Aitken R, 2010; Lightfoot, 2008; Matulick, 2009). The PSU has been conducting extensive research, seeking an improved method of predicting continuing load in future years. We have developed, tested and are now phasing in a new methodology that improves accuracy, incorporates more student academic and demographic data, deals with changes in student characteristics and cohort sizes, and deals with small and changing courses. This article will discuss the methodology and present the associated results.

Methods

The focus here is on estimating continuing load returning the following year, where estimation is conducted soon after the second semester census date. The body of research thus far has produced estimates for 2012–2014. This article presents results relating to 2013 and makes reference to 2012 results when comparing the existing and new methods. While there is also a growing need for projections for multiple years, the method to deal with this is still under review and will not be discussed here.

Commonwealth supported (Funding Group 1) continuing load, in all but one-year honours Flinders University courses, is within scope of this article. One-year honours courses are excluded due to great associated difficulty in estimating continuing load using the suggested method in this paper. Continuing load in such courses is best estimated using the existing method. In 2013, load in PhD courses switched from Funding Group 1 to Funding Group 4, but remains within scope of this article.

The analysis uses student load and enrolment data from years 2007 to 2012, to predict continuing load in 2013. The data for analysis include student enrolment and demographic information that is likely to capture the changing student profile, has previously been identified as being related to attrition and hence the probability of returning the following year (Adams, 2010; Bone, 2013; Pearson, 2013), and is available at the second semester census date. The characteristics used include enrolled course, progress through the course, equivalent full-time study load (EFSTL), an indicator of whether a student is studying a second degree, first semester GPA (grade point average), age at course commencement, and gender. All of these characteristics were divided into categories that define the boundaries used when cross-classifying load aggregates. While details of the course, EFTSL, age and gender are readily available for all enrolments, the remaining variables are not regularly stored and must be derived using existing enrolment data.

The progress variable categorises students according to their progress within a course, taking into account advanced standing. Based on internal findings and experiences, and external institutional research (Aitken, 2010), students enrolled in the same course were divided into five groups that were designed to minimise within-group continuation rate variation, and maximise between-group continuation rate variation:

- Commencers beginning in semester 1
- Commencers beginning in semester 2
- Continuers who are not near completion
- Continuers who are near completion
- Continuers who are due for graduation.

The progress variable provides flexibility in multiple ways. It accounts for differing composition of students within courses that are either phasing in or out, or that are altering the intake size. While not all categories in the progress variable are relevant to or exist in all courses, the boundary definitions provide increased comparability across courses that have a category in common, allowing new and small courses to borrow strength from similar existing courses.

The current method aggregates all Funding Group 1 enrolments in a course, and applies the previous year's continuation rate to estimate the following year's continuing enrolments. The previous year's load to enrolment ratio is then applied to estimate the corresponding continuing load. These estimates are subject to manual intervention where additional information suggests use of alternative calculations. For example, courses that are phasing out have the enrolment continuation rate reduced to account for an increasing percentage of students graduating.

The proposed method estimates load directly, since internal research has shown it produces more accurate results. The proposed method cross-classifies load in every course by all categorical student characteristic variables mentioned above. All current load within each cross-classified cell, and the student characteristics associated with it, are used as explanatory variables in a generalised linear model to estimate the following year's continuing load corresponding to that cell. Information relating to a single cross-classified cell corresponds to an observation used in model development. When using the model, we assume the associations between continuing load and all explanatory variables remains constant across time.

The saturated model equation takes the following form:

$$Returning\ Continuing\ Load_i = \beta_0 + \sum_{l=1}^{K} \beta_l \times X_{i,l} + \sum_{l=1}^{k} \sum_{j \neq i}^{k} \beta_{lj} \times X_{i,l} \times X_{i,j} + \varepsilon_i$$

Where returning continuing load is assumed to follow a gamma distribution, the 'i' subscript identifies each observation, the X represents a student characteristic, β is the associated regression coefficient, and the ε term corresponds to the model error. All above mentioned variables and associated two-way interactions are included in the saturated model. Backwards stepwise regression is performed to achieve parsimony, creating what will be referred to as the final model.

In total, there are 34 final models predicting load in 215 courses in 2013. Each final model is able to contain a different set of explanatory variables. The amount of historical data used in each model is determined by choosing the model that produces the best model diagnostics. Large courses have sufficient data to form a separate model. Small, new and changing courses are grouped together according to subject matter, student study behaviour, funding group and course level. Each course group has a separate model and includes a course variable in the saturated form, to test whether there is a course effect.

Results

Regression analysis indicated that the progress variable was included in all final models, introduced the most predictive power, and hence provided the largest improvement in estimates. First semester GPAs and EFTSL categories provided equal second largest

improvements. Gender and Age at commencement provided the least gain, and were most often excluded from final models. The number of years of data used to develop the models varied. As expected, courses and course groups that were undergoing large changes in course structure or student study patterns usually excluded older data. While these situations violated the assumption of associations remaining constant across time, use of most recent data was the simplest and most effective solution.

Table 1 presents the total actual and predicted Funding Group 1 load (EFTSL) under both method by broad course level in 2013, the number of courses, and the percentage of courses achieving an improvement from the current method. An improvement was defined by whether the proposed model estimate was closer to the actual value compared with the current estimate.

Table 1

Aggregated Continuing Funding Group 1 Load Diagnostics 2013

Course Level	Actual Continuing Load 2013	New Model Projection	Existing Model Projection	Number of Courses	% of Courses with Improvement
Bachelor (Pass + 4 Year Honours + Graduate Entry)	5824.0	5935.1	5919.0	91	59%
Graduate Certificate	51.8	39.8	44.5	32	75%
Graduate Diploma	52.3	47.1	46.2	16	63%
Masters by Course Work	808.4	734.7	787.0	60	62%
PhD	352.2	290.3	318.6	17	35%
Total	7084.4	7049.0	7093.6	215	60%

Table 1 shows that, overall, the total estimated load under the proposed method for 2013 was around 26 EFTSL further from actual load (EFTSL = 7084.4), compared with the existing method. Both the proposed and existing methods produced overall estimates within 1% of actual load. Exactly 60% of all course estimates achieved an improvement. This is a good result and is consistent with analysis that was conducted for 2012.

Additional analysis advised that, on average, both methods were unbiased and volatility in course residuals (distance of course estimate from actual continuing load) reduced by almost one-third under the proposed methodology. Additionally, overall, the proposed model would prevent funding for 131 EFTSL from being incorrectly distributed across all courses in scope. The direction and magnitude of these results are consistent with those corresponding to 2012 estimates.

Thirty-five courses (16%) experienced a significant improvement in projection of 2 or more EFTSL. The top five of these are shown in Table 2, which presents a variety of measures to assist in comparing the top and bottom five performing unidentified courses. For 158 of the 215 courses (74%), both the current and new methods gave a result within 2 EFTSL of the actual load, indicating that both methods gave very good and comparable results. Twenty-two courses (10%) had worse projections of 2 or more EFTSL. The bottom five are presented in Table 2.

Table 2

Top and Bottom Five 2013 Projections

Course	Continuing Load 2013			Variation		Improvement
	Actual	Projections		Current Model % variation to actual load	New model % variation to actual load	Projected EFTSL improvement (New to Current)
		Current model	New model			
Top 5 (better projections)						
Bachelor Pass 1	138	160	137	16%	0%	22
Masters 1	65	45	67	31%	2%	19
Bachelor Pass 2	72	85	74	17%	2%	11
Bachelor Pass 3	71	57	67	19%	5%	9
Bachelor Pass 4	362	384	376	6%	4%	9
Bottom 5 (worse projections)						
Master's 2	18	16	12	11%	36%	-5
Master's 3	31	35	22	13%	28%	-5
Bachelor Pass 5	166	164	159	1%	4%	-5
PhD 1	49	47	41	3%	16%	-6
PhD 2	76	71	63	7%	17%	-8

Table 2 suggests the two largest gains were in a Bachelor Pass and Master's course. The largest improvement was 22 EFTSL closer to actual continuing load compared with the estimate under the existing method, and the second largest improvement was 19 EFTSL closer to actual continuing load. The top five improved course estimates were at least 9 EFTSL closer to the actuals. All bottom five projections were up to 8 EFTSL further from the actuals.

The final model used to estimate continuing load in the course with the largest improvement included total current load, progress, study load categories and first semester GPA categories as the explanatory variables. The final model predicted larger continuing load for students with high GPAs, and who were continuing but not near completion. Commencers were predicted to return with lower continuing load compared with other students. The course with the largest improvement experienced increases in intakes in earlier years, altering the profile of students with respect to progress. This lowered the load continuation rates at the course level, causing over-estimation using the existing method. Previous associations between the chosen explanatory variables and returning continuing load remained constant across time, allowing the model to achieve an improvement in estimation.

The two PhD courses that benefited least contained continuing load for the first time in 2013. The final model used to estimate continuing load in PhD courses used all PhD courses as observations in a single model. However, the two new courses did not follow the same continuation rates as other PhD courses and would have benefited from manual intervention, similarly to that conducted under the existing method.

Figure 1 displays the magnitude of improvement in EFTSL for each course against the actual continuing Funding Group 1 load within the course in 2013. The improvement measure is the same as that presented in the final column of Table 2. Larger improvements were generally made for larger courses. The points below the zero line show that some

courses did not benefit from the proposed method. However, the effect of this was outweighed by more courses achieving improvements as well as achieving larger positive improvements on average, evidenced by the majority of points lying above and further away from the zero line.

Figure 1. Improvement in EFTSL within each course against continuing load in 2013.
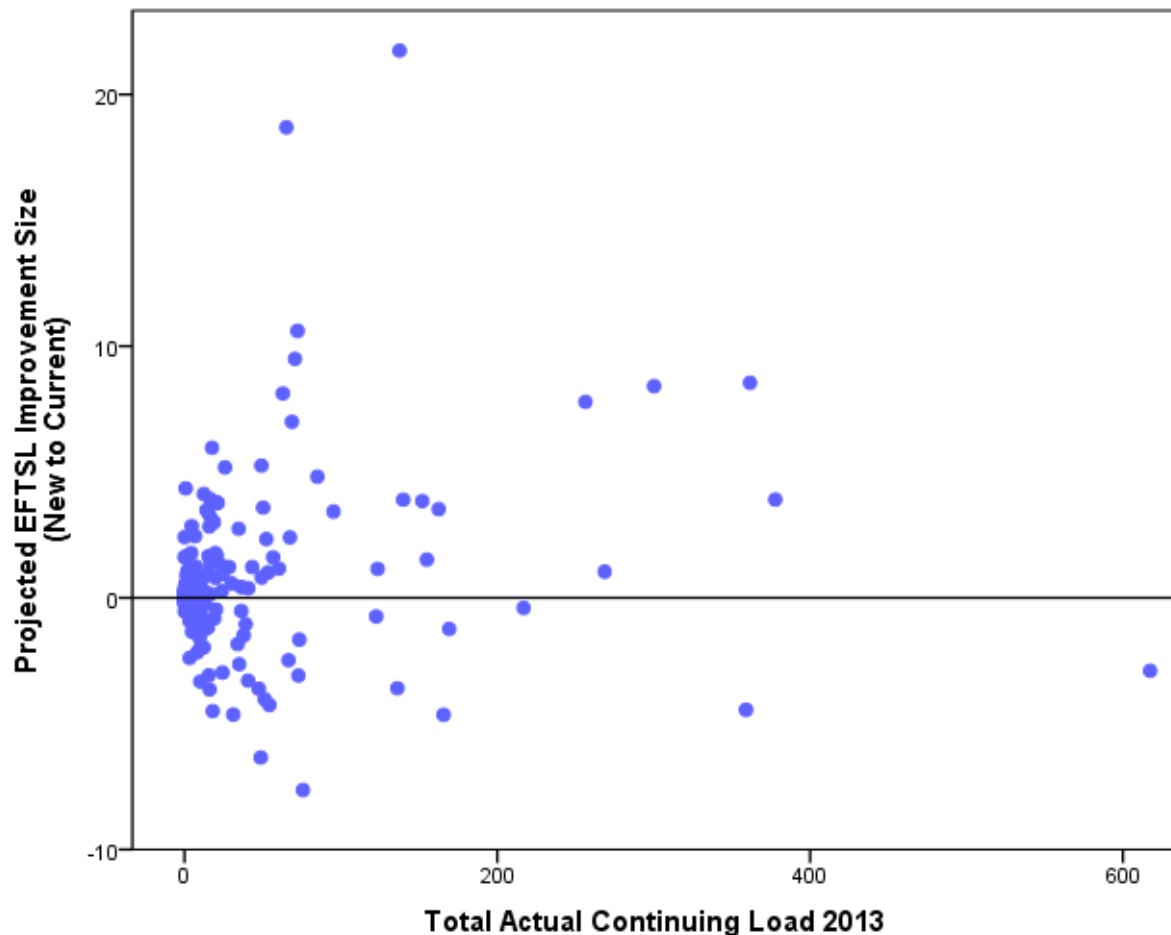
While not in scope of this article, it is useful to consider application of the proposed method earlier in the year. As users and uses of load forecasts become more complex, the need for earlier estimation will rise. Early estimation, such as before second semester census date, introduces multiple complexities to the proposed method. This includes having incomplete current year's data and therefore incomplete continuation rate data for the previous year, and missing variables such as first semester GPAs. Preliminary analysis removed use of the most recent year of continuation rate data, and removed GPA variables from the models. The results have indicated that the projections are not as accurate, but still provide an overall improvement compared with the current method. This demonstrates flexibility in use of the proposed method throughout the year.

Discussion

Analysis showed the proposed estimation method using a generalised linear regression model improved estimates for most courses, and largely reduced the volatility in course level errors. Courses that were either large or changing in composition benefited most.

The improvements were more notable when considering there was no manual intervention in the proposed method, compared with extensive intervention in the current method. However, it was evident that the proposed method would benefit from some manual intervention, particularly as new courses evolve.

Since many courses in Flinders University are currently in steady-state, we did not expect to see improvements across all courses. Rather, in anticipation of a growing and changing student cohort, we were aiming for and generally achieved improvements in courses that were changing in composition each year. An implication of the proposed method is the need to incorporate more student information than is currently used, some of which is not routinely stored. Necessary student information includes enrolled load, first semester GPA, a variable representing a student's progress through a course, gender, age at commencement, and an indicator of whether a student is studying a second degree simultaneously.

Although disaggregated estimation introduces complexity, it allows the university to adapt to changes in student profiles, characteristics and behaviours, which are highly likely in the future. Implementation of the proposed method enables the university to more accurately allocate resources and distribute funds. While the set-up of data and processes uses considerable resources, experience in this project so far has shown that once set-up, it is quick to re-run.

So far, the proposed and existing estimation methodologies have been run in parallel outside the Oracle environment. The existing methodology was also run within the Oracle environment to make sure both environments achieve consistent results. The next stage will involve incorporating the proposed method into the new environment. Even when fully implemented, the proposed method will always be under ongoing review. If we change our approach and aim to simplify the model, we can remove use of all variables but the progress variable, since the results showed that this variable introduces the largest gain. Conversely, if we wish to further improve estimates even at the expense of increased complexity, we may consider the use of additional explanatory variables or completely restructure the model and observations used to build the model.

Conclusion

The PSU at Flinders University is working towards building a more robust planning process that deals explicitly with the inevitable change in student cohort profiles, in addition to evolving courses. This coincides with the development and use of an Oracle-based system that enables the planning process to be integrated with other related data-driven processes such as budgeting and reporting.

This article introduced a new method of predicting continuing load that achieved overall improvement. The method is suitable for use throughout the year, incorporates changes in student characteristics and cohort sizes, and provides users and downstream processes the ability to incorporate more information. The proposed method met the objectives of a more robust and accurate planning process.

References

Adams, T. B. M. (2010, October). *The Hobsons retention project: Context and factor analysis report.* Paper presented at the Australian International Education Conference, Sydney.

Aitken, R. Y. A. (2011). Projecting continuing student enrolments: A comparison of approaches. *Journal of Institutional Research, 16*(1), 25–36.

Bone, E. R. R. (2013). First course at university: Assessing the impact of student age, nationality and learning style. *The International Journal of the First Year in Higher Education, 4*(1), 95–107.

Flinders University. (2012). *Flinders University Strategic Plan 2012–2016: Flinders Future Focus.* Adelaide: Author.

Lightfoot, A. (2008, November). *Predictive modelling of undergraduate student intake.* Paper presented at the Australian Association of Institutional Research Conference, Canberra.

Matulick, A. (2009, November). *The student load forecasting dilemma: Factors influencing prgression rates at higher education institutions before and after the Bradley Review.* Paper presented at the Australian Association of Institutional Research Conference, Adelaide.

Pearson, A. N. H. (2013*).* Identification of at-risk students and strategies to improve academic success in first year health programs. *The International Journal of the First Year in Higher Education, 4*(1), 135–144.

The Commonwealth of Australia. (2009). *Transforming Australia's Higher education system.* Canberra: Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education.