Benchmarking factor selection and sensitivity: a case study with nursing courses.

Journal:	*Studies in Higher Education*
Manuscript ID	CSHE-2016-0540.R1
Manuscript Type:	Article
Keywords:	Metric benchmarks, machine learning, RandomForest, factor selection, sensitivity

Benchmarking factor selection and sensitivity: a case study with nursing courses

There is an increasing requirement in higher education worldwide to deliver excellence. Benchmarking is widely used for this purpose, but methodological approaches to the creation of benchmark metrics vary greatly. Approaches require selection of factors for inclusion and subsequent calculation of benchmarks for comparison. We describe an approach using machine learning to select input factors based on their value to predict completion rates of nursing courses. Data from over 36 000 students, from nine institutions over three years were included and weighted averages provided a dynamic baseline for year on year and within year comparisons between institutions. Anonymised outcomes highlight the variation in benchmarked performances between institutions and we demonstrate the value of accompanying sensitivity analyses. Our methods are appropriate worldwide, for many forms of data and at multiple scales of enquiry. We discuss our results in the context of higher education management, highlighting the value of scrutinising benchmark calculations.

Keywords: Metric benchmarks, machine learning, RandomForest, factor selection, sensitivity.

Background

There is a sustained and growing impetus worldwide to measure performance in higher education (HE) through the use of comparative quality metrics (Hazelkorn 2015). However, there is a lack of consensus on the choice of factors and methods that should be used to produce such metrics (Tam 2001) and approaches used to 'benchmark' metrics can have significant effects on how institutions are publicly ranked. The term 'benchmarking' is widely used in private and public domains, there are a variety of interpretations and processes used to implement benchmarking in the educational literature (Draper and Gittoes 2004). Formal benchmarking (hereafter referred to simply as 'benchmarking') is a technique originating in the business sector used to analyse

performance between competitors by comparing operating costs, product features and operating capabilities (Asif 2015). Here, benchmarking can be used to compare performance outcomes, but can also be used to inform the management of improvement relative to competitors. The concepts underlying benchmarking in business have undergone a long history of development (Kumar and Chandra 2001), resulting in many different definitions and approaches (Zairi 1998; Kyrö 2003), while the core process typically involves a performance comparison and the identification of factors for improvement in individual businesses (Stapenhurst 2010).

Many potential benefits of HE benchmarking have been proposed, including strengthening institutional ability to self-assess performance, evaluating the reasons for sector differences to inform policy, and to managing improvement in performance (Hazelkorn 2015). The global nature of HE has led to reviews of international benchmarks, highlighting how these measures influence competition for students, staff and research funding (e.g. HESA 2011). Careful consideration of metrics like these are required when they are used for strategic decision-making, particularly to account for the high levels of heterogeneity that exists both within and between universities (Agasisti and Bonomi 2014). Measurement of performance or efficiency in HE usually relies on data relating to the output of degree courses, such as qualification rate, student employment and levels of student satisfaction or engagement. The effective identification and adoption of such performance indicators is necessary to fulfil the first goal of most benchmarking projects, where performance is compared amongst institutions. A second goal of many benchmarking projects is to identify key areas for enhancement, most often in consultation with a range of stakeholders. This should identify best practices and focus resources on particular areas for improvement. While the first goal of merely performing benchmarking is widespread and on the rise, the

latter goal of acting on benchmarking results to improve quality is sometimes challenging (e.g. Tillema 2010).

The education and health sectors have been subject to significant performance comparisons, which has led to league tables of performances and pressures to score highly in a suite of performance indicators (e.g. Draper and Gittoes 2004; Northcott and Llewellyn 2005; Shober 2013; Hazelkorn 2015). Benchmarking in Higher Education (HE) has been described as a process through which performance and practice are analysed to provide a standard measurement (the 'benchmark') of effective performance (e.g. 'league tables' or publication of 'standard performance indicators'; Draper and Gittoes 2004). This is not a simple process and to provide equitability in HE benchmarking systems, direct comparisons of institutions should account for inherent differences between institutions, such as heterogeneity of students, staff and programmes of study. This is a challenge for benchmarking in HE and relies on the principle that there are established patterns of success for any given set of student or institutional typologies. Weighting metrics relative to an expected average of performance across these typologies, rather than simply using absolute measures, can mitigate this issue and is an essential part of benchmarking in the business sector (Stapenhurst 2009). Despite the advantage of adjustments of this type, the use of simple ranked metrics is common in HE despite well-documented shortcomings (Saisana et al 2011). Williams and de Rassenfosse (2016) provide a valuable insights into the current practices and pitfalls of the use of performance measures in higher education systems.

The choice of variables for such calculations and the method of analysis is known to affect the outcome of rankings (Saisana et al 2011; Draper and Gittoes 2004). Asif (2015) points out that the methodologies for selecting processes or factors for benchmarking in HE are sometimes vague or subjective. Benchmarking methodology

should be objective and repeatable, but there is evidence that this is not always the case (Partovi 1994; Garengo et al 2005; De Toni and Tonchia 2001). Thus, because the success of metric benchmarking is governed by the identification of a small set of appropriate factors that influence the measure of performance, factor selection can involve a large number of possible permutations of factors requiring a rigorous process to devise the most appropriate solution. To overcome challenges in factor selection for benchmarking, it has been suggested that factor selection should incorporate data that are: 1) routinely acquired by stakeholders, 2) highly predictive, 3) meaningful to the stakeholders and are 4) relevant and necessary to add value to the benchmark (Hall and Holmes 2003).

In this study, we describe a two phase approach to perform objective and repeatable benchmarking in higher education, using data from nursing courses at nine universities in the UK. In the first phase of our benchmarking, we used machine learning to objectively choose variables of value to predict our measure of success, student progression. Machine learning methods (such as classification and decision trees) are increasingly used as a data reduction method and for variable selection (Rokach and Mainon 2014). This results in a subset of informative and necessary variables to be used in benchmarking modelling. The second phase of our study involved using variable selected via machine learning to construct a benchmarking model. The approach we used is similar conceptually to best-practice using benchmarking, where we used direct standardization of variables to weight variables within and between like student groupings (Draper and Gittoes 2004). Here, direct standardisation is used to benchmark performance of specific student cohorts where the benchmark *per se* is the difference between the actual performance of students at that institution relative to the performance of the same student cohort at a generalized

regional 'average' university (see below for further explanation). We discuss the method we implement here and how this approach can be used to inform quality improvement in higher education.

Metric benchmark development

For the factor selection phase of this study, all analyses were carried out on the open source statistical package 'R' (https://www.r-project.org/). The dependent variable for benchmark calculations was the measure of student qualification, a binomial variable indicating either qualification or withdrawal before course completion. In terms of the accompanying factors available for analysis, routinely captured data were used from 'PETD' (Professional Education Training Database) data that had been provided on a student-by-student basis for nursing courses by Health Education England (formerly Health Education North West) for nine higher education institutions. Our data represented cohorts of students that had finished their degree and who began their degree during the period of 2008-2011and would have qualified by 2014 at the earliest. The database provided a range of factors associated with each individual student, such as student age, gender, registered disability, ethnicity or whether the student had suspended their studies. Home postcode was also provided and this was assumed to represent the (parental) home of the individual. There were also two associated metrics provided as surrogates of socioeconomic status: (i) youth participation rates in 'further' education ('youthEd'), and (ii) adult participation rates in further and higher education ('adultEd').

The approaches to benchmark the dataset drew on methods from Draper and Gittoes who referred to the National Committee of Inquiry into Higher Education (or the 'Dearing Committee') that led to the Performance Indicators Steering Group. Their work focused on student progression data, with an underlying intention to provide more

reliable information on the performance of UK HEIs, in order to inform policy developments and enhance public accountability. Draper and Gittoes (2004) dissected the HEFCE approach, which was based on three main components: inputs (status of individuals as they enter higher education), process (what happens to them when in higher education) and outputs (outcome measures that can be evaluated, such as student success). This framework underpins the input-output (IO) approach, is also known as 'provider profiling' in health areas and has been used in benchmarking for schools and hospitals internationally. The IO approach is used here to provide a schematic overview of our methods (Figure 1).

[Figure 1 near here]

Machine learning for factor selection

In light of the absence of machine learning methods in the literature about benchmarking, we provide some detail to both justify and provide an overview of the approach. Statistical analysis of data sets with high dimensionality (i.e. many potential factors under consideration) is challenging as a consequence of processing and analysing a large number of predictor variables. One approach to managing this challenge is to reduce the dimensionality of the dataset, to reduce redundancy amongst variables and leverage power to resolve the effects in individual variables (Rokach and Maimon 2014). To achieve this, we used machine learning, a computational approach to identify variables with high explanatory power within a modelling framework.

One implementation of machine learning is using classification and regression trees (CART) in creating so-called 'random forests' (Breiman 2001). This approach is robust to strict assumptions of data conformity and has become one of the most widely used data analysis tools involving large and high-dimensional datasets (Liaw and Wiener 2002). Practical applications of this technique can be found in a number of

research fields, including: psychology (Strobl et al 2009), ecology (Prasad et al 2006) and HE student progression (Hardman et al 2013). RandomForest analysis in particular is in wide use as a tool to make predictions based on variable associations, and also to identify variables with predictive value by ranking the predictive importance of these variables (Grömping 2009; Genuer et al 2010). A specific strength of the algorithm is that it can be used to combine factorial and numerical data within a modelling framework and is robust with data sets of high dimensionality, whilst also being free of parametric assumptions of data conformity (Breiman 2001; Strobl et al 2009).

RandomForest was used to identify variables with the highest importance in predicting whether (or not) students in nursing programmes become qualified in their field of study. We calculated the 'variable importance' values of all predictor variables provided in the dataset, generating 10,000 trees. Our first goal was to identify a subset of the most important predictor variables to use in the construction of our benchmarking tool, following Liaw and Wiener's (2002) approach. In this way we ranked predictor variables to inform the benchmarking tool, whilst objectively retaining the highest amount of information possible. To avoid problems due to the uneven distribution of subsets of data (e.g. across universities and programmes) and possible correlations amongst variables, we used an area under the curve (AUC) correction approach with variable and model validation (Strobl et al 2009).

Our second goal was to describe how the variables of high importance (identified in the initial analysis) varied amongst universities and programmes in their association with whether or not students ultimately became qualified. We calculated variable importance using equally weighted measures of mean square error and the Gini index (see Breiman 2001) for each of these data subsets, and partitioned data subgroups

using conditional inference trees. This approach is coherent with the methods in Strobl et al (2009).

Interpretation of the modelling outcomes may be complex. For example, distance of university from home postcode was identified as having high importance in predicting whether a student became qualified. However, larger distances may be an advantage to one university that is highly selecting and draws the best candidates nationwide, whereas larger distances may be a disadvantage to another rural university where this implies difficult commutes from home to placement/university. Thus, future functional interpretations of the outcomes (such as the distance variable) would need to be made by the users of the benchmarks who can account for characteristics of individual universities and their constituent student populations.

Factor selection

There were 52 routinely used factors available from the dataset made available for nine (anonymised) universities. These included variables such as student 'age', 'gender', 'ethnicity', registered 'disability' and 'home address'. The distance of the home postcode to the institution of study was also calculated ('distance'). We note that the dataset available for analysis was limited due to data availability and data protection, which is a practical constraint in any benchmarking project. For example, there is a great deal of literature surrounding the role of student entry qualifications (e.g. Yorke and Longden 2004). This type of data would be available at institutional level, but it was not available from the PETD dataset and thus not available for our analyses. We note that there is underlying debate concerning tariff as being a factor that is 'in the control of the institution' which may erode its value as an input variable for benchmarking (see Draper and Gittoes 2004).

Machine learning analysis using classification and regression trees was carried out using all data for which we had complete cohorts (2008-2011) in order to identify factors of high importance in predicting whether students become qualified in their field of study after graduating. We first included the factors for the institution of study and whether the student had suspended their studies (returning to study at a later date). Both these variables were found to have greater influences than any other factor. The strong 'university effect' was anticipated as there were known differences in the qualification rates between the institutions with associated known differences in entry qualifications. Students that suspended during their studies were also more likely to withdraw. Indeed, suspension was a stronger predictor of whether a student qualified than institution, suggesting the need for future research to explain factors that underpin suspensions. For further analyses, 'suspend' was excluded on the basis that it was an outcome of the educational process (and not an input variable, such as gender) and because of its close association with our dependent variable. Institutions were subsequently analysed separately in order to explore consistency of the factor effects across institutions.

Individual factors to inform benchmarking ideally should represent unrelated characteristics. Both youth and adult participation rates in education were included in the preliminary analyses and both identified as having high importance. However, these factors are strongly correlated with each other, and only youth participation (having a higher importance value than adult participation) was included in the benchmark modelling. As a result of our factor selection analysis, we selected the five factors of highest average importance for inclusion in our benchmark model. These were 'age', 'gender', 'ethnicity', 'registered disability' and 'youth participation rate into higher education'.

Calculation of metric benchmarks using weighted averages

Demographic segments

Here, we refer to a 'segment' as a unique grouping of demographic factors. Segments representing combined groupings of data generated by the analysis are a product of partitioning data between the many combinations of factors. One challenge with benchmarking modelling is to avoid the reduction into too many segments of data. This problem practically limits including a large number of variables in benchmarking. The reason for this is that when data are segmented some may have very few records, for example records which are for Male AND Young AND High Youth HE Participation. We avoided this by overwriting the segment qualification rate with the rate at the higher level in the hierarchy according a lower threshold we set (a standard solution to this problem). For example, instead of using the rate for the segment Male AND Young AND High Participation, the rate for Male AND Young only was used. In the test model, this the low threshold was set at a minimum of 30 students in the starting cohort as this threshold value is commonly used in statistical tests as the minimum number of cases required for the test to be reliable.

Another important consideration was missing data, as not all students provide all the information, particularly concerning protected characteristics such as ethnicity and disability. In the test model, missing data were treated as a characteristic in their own right. For example, gender had three characteristics: male, female and not known (i.e. not reported). This necessary addition resulted in larger number of small segment counts as there were a greater number of attributes.

Factor order

The order that factors are used in benchmark calculation affects benchmark outcomes,

both as a consequence of the segment threshold and because of missing data. In order to measure the impact that factor order had, we used a sensitivity analysis approach. With regards to data segmentation, if the segment count was below our threshold when including all variables in a benchmark, then one factor would be dropped from the analysis to alleviate this (where variable order impacted which variable was dropped). To avoid factor order bias, all possible combinations of factor order were used to compute benchmarks and the variation around resulting benchmark estimates was evaluated in relation to actual performance (Figure 2).

[Figure 2 near here]

The average difference between the maximum and minimum calculated benchmark score for each institution (based on the different orderings of the demographic factors) was 6.52%. For most institutions (except institutions B and G), the actual performance for the programme was either above or below benchmark boxplot whiskers. Thus, regardless of the ordering of the five factors, the actual performance of most institutions can be classified unambiguously as being above (or below) benchmark. For institutions B and G, actual performance lay within the range of likely benchmark results and so is not unambiguously different. This highlights the importance of exploration of the effects of factor order in the development of benchmarks, and the need for informed decision-making on the outcomes of this. A striking feature of our benchmark results is that variation in benchmark estimates is very different for different institutions, where, for example, institution B shows large variation in benchmarks due to factor order while institutions D and H show very little variation in benchmarks. This effect may reflect variation in segmentation between these intuitions, variation in missing data or a genuine differential effect of individual factors on benchmark performance. Finally, we point out that merely ranking actual performance in relation to average actual across

institutions may can be quite different in comparison to ranking performance based on benchmarking. For example, relative to the average benchmark, institutions E, F and G are on or above average in actual performance. However all of these institutions show performance below that predicted by the benchmark. Likewise, while institution B exhibits actual performance below the institutional average, it is performing to expectation based on the benchmark estimates.

Number of factors

We also performed sensitivity analysis to measure the effect of factor number on the outcome of benchmarking. The number of factors included affects benchmark variation and magnitude (Figure 3). The pattern of median values for benchmark estimates is similar to that seen when varying order alone for all five variables. The spread of benchmark estimates in our sensitivity analysis is small compared to that for factor alone (comparing Fig 3 to Fig 2). This suggests that, for these data, the model appears to be (at least) relatively robust regardless of the number of factors included. The interpretation that a particular institution is over- or under-performance appears robust in that most institutions (except institution B) are clearly above or below the benchmark.

[Figure 3 near here]

Which years to include

Another factor that can affect the model is the number of years of data that are used for the calculation. In our models, three years of data were used to inform the benchmark (2009, 2010 and 2011), where year is the starting year for whole cohorts of students followed to degree completion. If fewer years of data were used, for example to specifically calculate benchmarks for each year, then segments would have had smaller

counts that would affect the importance of factor order. When fewer years of data were used, the benchmark range was, in some cases, much broader due to lower sample sizes impacting segmentation issues. While the increased variability resulted in broader benchmark estimates, using only the most recent year meant that the data were specific to individual years which could be useful to inform, for example, year-specific effects such as a change in policy or other measures expected to impact performance.

Overall, when taking into account both factor order and number of factors with sensitivity analysis we describe here, the indirect benchmarking method appears to be robust in that the interpretation of whether an institution is performing well or poorly relative to the benchmark is generally little affected by specific choices for factor ordering. The error around benchmark estimation that is produced using sensitivity analysis in this way is a robust way to aid interpretation of benchmarking results that can be used to overcome challenges involving factor choice, factor order and segmentation.

Discussion

This study outlines the approaches used to calculate metric benchmarks for nursing courses in the Northwest region of the UK, but the findings and approaches described are relevant to benchmarking calculations worldwide and at many scales of enquiry. We describe a conceptual input-output model (aligned to Draper and Gittoes 2004) that identifies clear differences in institutional performances compared to benchmarks and can be used to account for variation in learner compositions and assumptions associated with the benchmarking process. We believe this approach combines a robust factor selection approach with sensitivity analysis to mitigate many of the shortcomings associated with HE rankings. The outcomes provide a valid and reliable method to help both policy makers and universities for the identification of universities that are

performing above or below benchmark predictions, thus helping to target improvement activities. It is noteworthy that this comparative information is potentially valuable but, in isolation, metric benchmarks do not provide an understanding of the explanatory factors that drive performance. Subsequent use of the outputs for performance enhancement requires a layer of interpretation, preferably in consultation with other stakeholders (Jackson and Lund 2000).

Differences in institutional benchmark outcomes were anticipated after the preliminary analysis highlighted 'institution' as the most important explanatory variable in the model (also see the exploration of heterogeneity of institutions in Agasisti and Bonomi 2014). This factor was subsequently removed to allow institutions to be compared and preventing the factor from masking comparatively subtle effects of other factors in predicting student qualification. Similarly, records of student suspensions that were available in the dataset were used in preliminary calculations were also strong predictors of future failure to graduate (as may be anticipated; see Nonis and Wright 2003) and were also not considered in the final benchmark calculations.

The factors selected for the benchmarking process were coherent with other HE benchmarking documentation For example, the UK's National Student Survey (NSS) benchmark has the intention of accounting for the mix of students and subjects at different institutions when reporting on survey outcomes (see Leman 2011). The respondent-related factors included in the NSS benchmarking process are similar to those identified in this study and purported to have demonstrable, consistent effects on survey outcomes and are also considered to be outside of institutional control, namely: subject (not included in our study as limited to nursing courses), ethnicity (included); age (included), mode of study (not included in our study as only full time students are enrolled on these courses), sex (included as 'gender') and disability (included).

However, HESA (2011) also have a benchmarking process that incorporates a similar suite of factors relating to entry qualifications of students, age and subjects of study. The HESA benchmark accounts for the effects of subject profiles and the entry qualifications of students for comparisons of the rates of student qualification. The HESA report (*ibid.*) encapsulates the need for this process, suggesting that benchmark outcomes provide information about the 'sort of values that might be expected for a HE provider's indicator'. This assumes that other factors not used in the benchmark calculation do not have significant effects on the output variable. If this assumption is true, and the factors entered are all 'input variables', it suggests that differences in performance relative to the benchmarks are either a direct result of HE provider performance or are influenced by other factors not included in the benchmark.

The sensitivity analysis used provided a range of potential benchmark values based on potential permutations of the numbers of factors used and the order they are entered into the calculation. Our findings highlight the importance of this stage of benchmark calculation, most markedly the order that factors are used in the benchmark calculation, particularly for institutions that are close to their benchmark values. Our findings concur with other authors that stress the importance of explicit sensitivity analyses for calculating metric benchmarks, although this key component is not always evident in benchmarking systems (see Reichmann & Sommersguter-Reichmann 2006). Our approach contrasts others (e.g. HESA benchmarking of NSS results detailed previously) which provides a single point of comparison. It is suggested that the visual presentation of the actual result against the quartiles of benchmark options provides 'non-statistical' users with a sense of confidence around the interpretation, as it is not an over-simplification. This was an important basis for performance discussions with universities in the current study.

The methodological approaches have been described here in detail, to illustrate the underpinning rationale and highlight approaches such as sensitivity analysis and the debate surrounding the number of factors for inclusion. We suggest that these approaches can be applied more widely, for different subject areas, greater geographical areas (e.g. national datasets) and beyond the educational arena. The novel use of this machine learning technique is proposed as going some way to solve the difficult issue of selecting the most appropriate factors for inclusion to allow for objective benchmarking (e.g. Hall and Holmes 2003). The robust nature of the RandomForest procedure allows objective, data-informed inclusion of many data types, is not limited by data distributions or types, and is seen as a starting point to factor selection that could be adapted year on year as data to include a wider dataset whenever new data acquisition reaches a scale suitable for benchmarking purposes. Of course, the key themes of benchmarking processes go far beyond the calculation and comparison of metric benchmarks. Their importance lies in the subsequent identification and implementation of new processes, for example based around the use of best practices (Camp 1989).

In our case study of nursing programmes at nine UK universities, the choice of factors was determined by the available dataset, in this case data that are routinely collected and available to institutions. We acknowledge that this constraint precludes many other potentially important factors such as aspects of learning and teaching structures. This includes the clinical placement element of nursing courses that appears to be of significant importance in this subject area (Hamshire et al 2011 and 2012). However, we believe this example is a realistic representation of practical limitations in benchmarking and we suggest our method can be used to improve current benchmarking practice in HE, despite unavoidable limitations in data quality or

availability. Because formal benchmarking is recognized as an ongoing process (Stapenhurst 2009), we suggest that our approach provide an opportunity for regular recalculations to incorporate data as they become available and to evaluate their impact on the outcomes through accompanying sensitivity analyses. Potentially, the machine learning techniques described here could also be uses to identify the factors that predict successful completion of other within-degree output variables such as placement completions (e.g. as opposed or in addition to whether a student progresses or becomes qualified in their field of study).

In conclusion, this study has outlined a new approach for factor selection and sensitivity analysis of benchmarking programme performance in HE. This approach is robust and objective, incorporating machine learning for factor selection. The software used is open source and available to all stakeholders and the code used easily shared. The benchmarks created are coherent with others in the educational sector due to the factors included and our outcomes have highlighted the importance of sensitivity analyses to clarify the effects of factor number and order on the outcomes. Our outcomes could be built on and improved, for example by future acquisition of data surrounding the clinical placement component of these courses. Thus, we have provided baseline data and documented processes that could be developed further, and identified issues with current data that are routinely available and valuable for benchmarking processes. Ultimately, we have created and described a benchmarking approach that could be further generalised to explore benchmark performance in other regional or national datasets in order to improve quality in key HE outputs such as successful completion of clinical placements. In a wider context, the approaches described have relevance to the creation of benchmarks on a greater scale, such as those being devised at the time of writing for the proposed 'Teaching Excellence Framework', a component

of the UK government's White Paper intended to reform higher education, that is founded on a selection of benchmarked metrics (BIS 2016).

References

Agasisti, T. and F. Bonomi. 2014. "Benchmarking universities' efficiency indicators in the presence of internal heterogeneity." *Studies in Higher Education* 39: 1237-1255.

Ammons, D.N., C. Coe and M. Lombardo. 2001. "Performance comparison projects in local government: participant's perspectives." *Public Administration Review* 61: 100-110.

Ammons, D.N. and W.C. Rivenbark. 2008. "Factors influencing the use of performance data to improve municipal services: evidence from the North Carolina benchmarking project." *Public Administration Review* 68: 304-318.

Anand, G. and R. Kodali. 2008. "Benchmarking the benchmarking models." *Benchmarking: An International Journal* 15: 257-291.

Asif, M. 2015. "Determining improvement needs in higher education benchmarking." *Benchmarking: An International Journal* 22: 56-74.

Asif, M. and A. Raouf. 2013. "Setting the course for quality assurance in higher education." *Quality & Quantity* 47: 2009-2024.

Besterfield, D.H., Besterfiels-Michna, C., Besterfield, G.H. and Besterfield-Sacre, M. 1999. *Total Quality Management*. New Jersey: Prentice-Hall.

Bhutta, K.S. and F. Huq. 1999. "Benchmarking-best practices: an integrated approach." *Benchmarking: An International Journal* 6: 254-268.

BIS. 2014. *Methodological issues in estimating the value added for further education, higher education and skills: a review of relevant literature*. London: Department for Business, Innovation and Skills, April.

BIS. 2016. Success as a Knowledge Economy: Teaching Excellence, Social Mobility and Student Choice. (Department for Business, Innovation and Skills). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523546/bis-16-265-success-as-a-knowledge-economy-web.pdf

Bowerman, M., G. Francis, A. Ball, and J. Fry. 2002. "The evolution of benchmarking in UK local authorities." *Benchmarking: an International Journal* 9: 429-449.

Braadbaart, O. and B. Yusnandarshah. 2008. "Public sector benchmarking: a survey of scientific articles, 1990-2005." *International Review of Administrative Sciences* 74: 421-433.

Breiman, L. 2001. Random forests. *Machine learning* 45: 5-32.

Camp, R.C. 1989. *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance*. Madison: Quality Press.

Carpinetti, L.C.R. and A.M. De Melo, 2002. "What to benchmark: a systematic approach and cases." *Benchmarking: An International Journal* 9: 244-255.

CHEMS. 1998. *Benchmarking in Higher Education: An International Review*. London: Commonwealth Higher Education Management Service.

Dattakumar, R. and R. Jagadeesh. 2003. "A review of literature on benchmarking." *Benchmarking: An International Journal* 10: 176-209.

De Toni, A. and S. Tonchia. 2001. "Performance measurement systems-models, characteristics and measures." *International Journal of Operations and Production Management* 21: 46-71.

Draper, D. and M. Gittoes. 2004. "Statistical analysis of performance indicators in UK higher education." *Journal of the Royal Statistical Society* Series A (Statistics in Society) 167: 449-474.

Drew, S.A.W. 1997. "From knowledge to action: the impact of benchmarking on organizational performance." *Long Range Planning* Vol. 30, pp. 427-441.

ECPE (2011) *Education Criteria for Performance Excellence*. Gaithersburg: National Institute of Standards and Technology (NIST).

Folz, D.H. 2004. "Service Quality and benchmarking the performance of municipal services." *Public Administration Review* 64: 209-220.

Garengo, P., S. Biazzo, and U.S. Bititci. 2005. "Performance measurement systems in SMEs: a review for a research agenda." *International Journal of Management Reviews* 7: 25-47.

Genuer, R., J.M. Poggi, and C. Tuleau-Malot. 2010. "Variable selection using random forests." *Pattern Recognition Letters* 31: 2225-2236.

Goldstein, H. 2001. "Using pupil performance data for judging schools and teachers." *British Educational Research Journal* 27: 433-442.

Goncharuk, A.G. and J.P. Monat. 2009. "A synergistic performance management model conjoining benchmarking and motivation." *Benchmarking: An International Journal* 16: 767-784.

Grömping, U. 2009. "Variable importance assessment in regression: linear regression versus random forest." *The American Statistician* 63: 308–319.

Hall, M.A. and G. Holmes. 2003. "Benchmarking attribute selection techniques for discrete class data mining." *IEEE Transactions on Knowledge and Data Engineering,* 15: 1437-1447.

Hamshire, C., T. Willgoss, and C. Wibberley. 2012. "Should I stay or should I go? A study exploring why healthcare students consider leaving their programme." *Nurse Education Today* 33: 889-895.

Hamshire, C., T. Willgoss, and C. Wibberley. 2011. "The placement was probably the tipping point' – The narratives of recently discontinued students." *Nurse Education in Practice* 12: 182-186.

Hardman, J., A. Paucar-Caceres, and A. Fielding. 2013. "Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm." *Systems Research and Behavioral Science* 30:194-203.

Hazelkorn, E. 2015. *Rankings and the Reshaping of Higher Education: The Battle for World Class Excellence*. Basingstoke: Palgrave Macmillan.

HESA. 2011. *International Benchmarking in UK Higher Education*. Higher Education Statistics Agency report. https://benchmarking.hesa.ac.uk/wp-content/uploads/2011/10/HESA_International_Benchmarking_report.pdf.

Jackson, N. and H. Lund. 2000. *Benchmarking for Higher Education*. Open University Press, Milton Keynes, UK.

Kouzmin, A., E. Löffler, H. Klages, and N. Korac-Kakabadse. 1999. "Benchmarking and performance measurement in public sectors. Towards learning for agency effectiveness." *The International Journal of Public Sector Management* 12: 121-144.

Kumar, S. and C. Chandra. 2001. "Enhancing the effectiveness of benchmarking in manufacturing organizations." *Industrial Management and Data Systems* 101: 80-89.

Kyrö, P. 2003. "Revising the concept and forms of benchmarking." *Benchmarking: An International Journal* 10: 210-225.

Lema, N. and A. Price. 1995. "Benchmarking: performance improvement toward competitive advantage." *Journal of Management in Engineering* 11: 28-37.

Leman, J. 2011. *Benchmarking and the NSS.* Presentation at the Higher Education Surveys for Enhancement Conference 2011. https://www.heacademy.ac.uk/resource/benchmarking-and-nss.

Liaw, A. and M. Wiener. 2002. "Classification and regression by RandomForest." *R news* 2: 18-22.

Nedwek, B.P. and J.E. Neal. 1994. "Performance indicators and rational management tools: a comparative assessment of projects in North America and Europe." *Research in Higher Education* 35: 75-103.

Nonis, S.A. and D. Wright. 2003. "Moderating effects of achievement striving and situational optimism on the relationship between ability and performance outcomes of college students." *Research in Higher Education* 44: 327-346.

Noordhoek, M. 2013. *Municipal benchmarking: organisational learning and network performance in the public sector*. PhD Thesis, Aston University. Available at: http://eprints.aston.ac.uk/19540/1/Studentthesis-2013.pdf

Northcott, D. and S. Llewellyn. 2005. "Benchmarking in UK health: a gap between policy and practice?" *Benchmarking: An International Journal* 12: 419-435.

Partovi, F.Y. 1994. "Determining what to benchmark: an analytic hierarchy process approach." *International Journal of Operations and Production Management* 14: 25-39.

Prasad, A.M., L.R. Iverson, and A. Liaw. 2006. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." *Ecosystems* 9: 181-199.

Pryor, L.S. 1989. "Benchmarking: a self-improvement strategy." *Journal of Business Strategy* 10: 28-32.

Reichmann, G. and M. Sommersguter-Reichmann. 2006. "University library benchmarking: An international comparison using DEA." *International Journal of Production Economics* 100: 131-147.

Rokach, L., and O. Maimon, 2014. *Data mining with decision trees: theory and applications*. World Scientific. Series in Machine Perception and Artificial Intelligence 69. http://www.worldscientific.com/worldscibooks/10.1142/6604.

Saisana, M., B. d'Hombres, and A. Saltelli. 2011. "Rickety numbers: Volatility of university rankings and policy implications." *Research Policy* 40: pp.165-177.

Shober A. F. 2013. "Debate: Benchmarking inequality – driving education progress in the USA." *Public Money and Management* 33: 242-244.

Spendolini, M.J. 1992. *The Benchmarking Book*. New York: American Management Association.

Stapenhurst, T. 2009. *The Benchmarking Book*. Oxon: Routledge.

Strobl, C., J. Malley, and G. Tutz, 2009. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." *Psychological Methods* 14: 323-348.

Tam, M. 2001. "Measuring quality and performance in higher education." *Quality in Higher Education* 7: 47-54.

Tillema, S. 2007. "Public sector organisations' use of benchmarking information for performance improvement: theoretical analysis and explorative case studies in Dutch water boards." *Public Performance and Management Review* 30: 496-520.

Watson, G.H. 1994. "A Perspective on Benchmarking." *Benchmarking for Quality Management and Technology* 1: 5-10.

Williams, R. and G. de Rassenfosse. 2016. "Pitfalls in aggregating performance measures in higher education." *Studies in Higher Education* 41: 51-62.

Wynn-Williams, K.L.H. 2005. "Performance assessment and benchmarking in the public sector: an example from New Zealand." *Benchmarking: An International Journal* 12: 482-492.

Yorke, M. and B. Longden. 2004. *Retention and Student Success in Higher Education*. Maidenhead: Society for Research in Higher Education/Open University Press.

Zairi, M. 1998. *Effective Management of Benchmarking Projects*. Oxford: Butterworth Heinemann.

Figure Captions

Figure 1. Conceptual model of the input-output processes used to create benchmarks from the available dataset (inputs). The simplified structure does not include the iterations of the computations used for the metrics (adjusted outputs) such as sensitivity analyses.

Figure 2: Performance of adult nursing courses compared with calculated benchmarks when factor order varies. Benchmark outcomes are shown for all possible orderings

using all five factors. Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

Figure 3: Performance of adult nursing courses compared with calculated benchmark when factor number varies. Benchmark outcomes are shown for all orderings for the inclusion of 5, 4, 3, 2 or 1 factor(s). Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

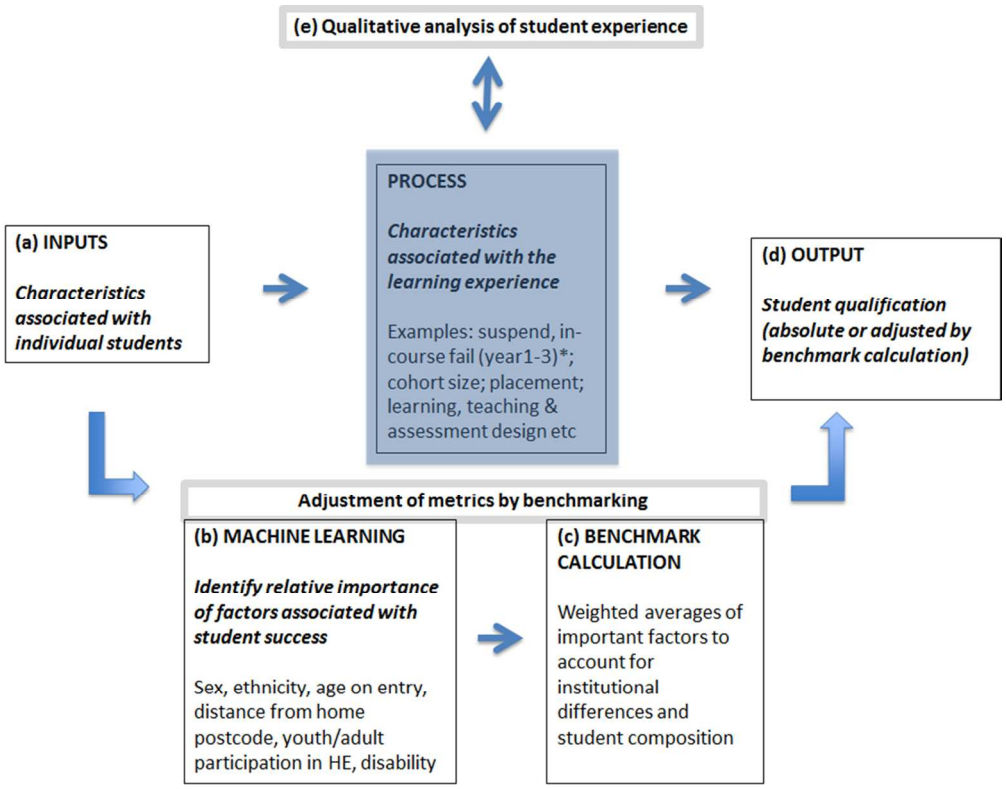

Figure 1. Conceptual model of the input-output processes used to create benchmarks from the available dataset (inputs). The simplified structure does not include the iterations of the computations used for the metrics (adjusted outputs) such as sensitivity analyses.

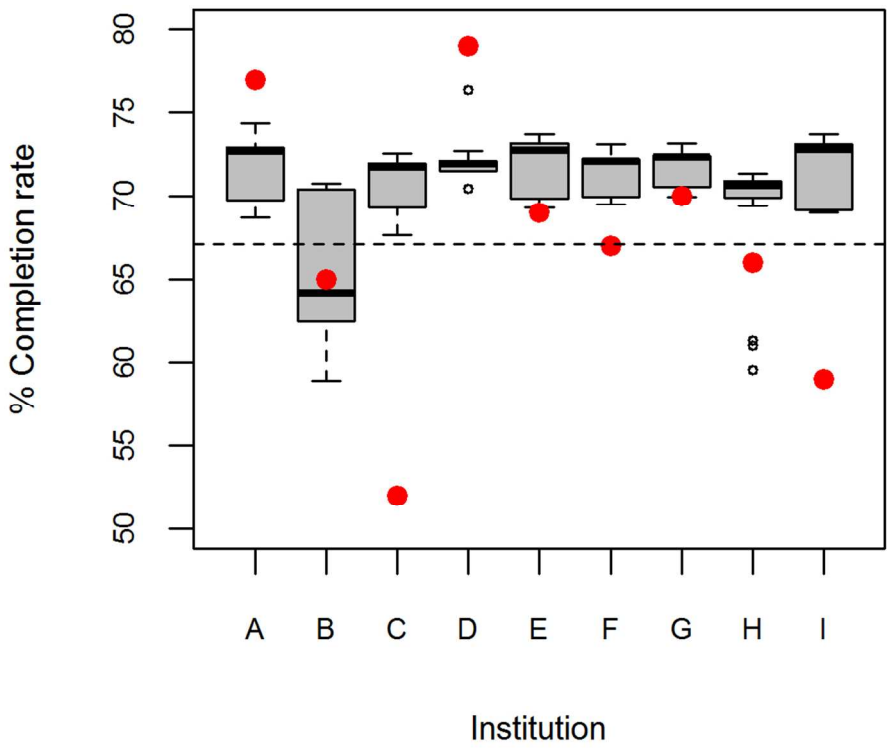

Figure 2: Performance of adult nursing courses compared with calculated benchmarks when factor order varies. Benchmark outcomes are shown for all possible orderings using all five factors. Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

101x101mm (300 x 300 DPI)

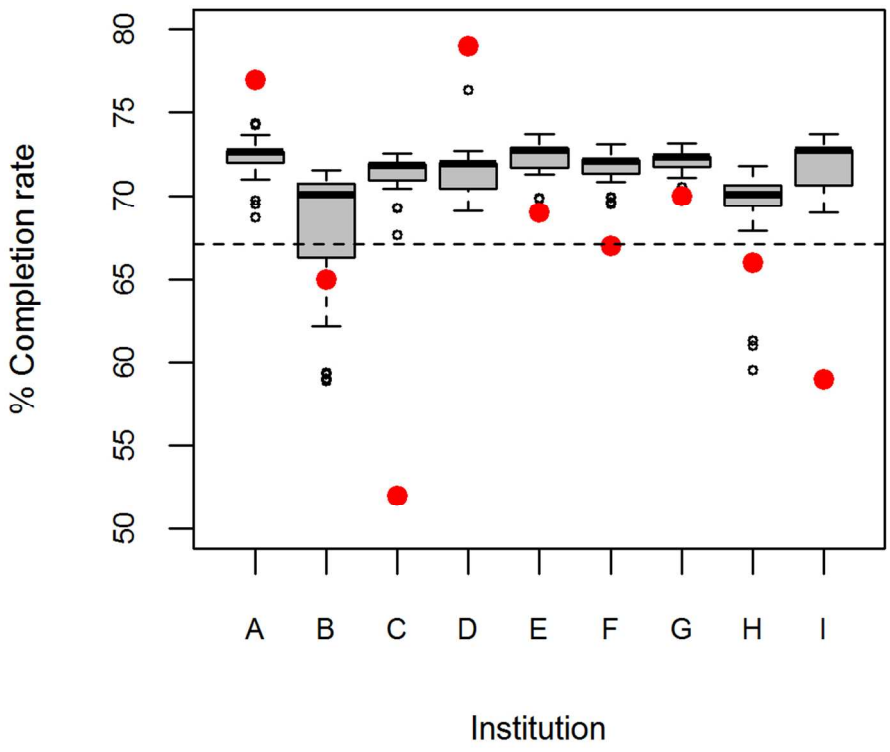

Figure 3: Performance of adult nursing courses compared with calculated benchmark when factor number varies. Benchmark outcomes are shown for all orderings for the inclusion of 5, 4, 3, 2 or 1 factor(s). Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

101x101mm (300 x 300 DPI)