University of Sheffield

MSC Data Analytics

Industrial Team Project

Analysing REF

Authors

Eleanor Woodhead
George Chrysostomou
Jagoda Karpowicz
Rawaida Kamarudin
Weijiang Lin

Supervisors

Thomas Hain
Eleni Vasilaki

December 11th, 2017

Contents

List of Figures

List of Tables

1 Introduction

The United Kingdom (UK) Government invests millions of pounds each year to improve and maintain the quality of higher education in the UK, in fact, based on HEFCE (Higher Education Funding for England) "Guide to funding 2017-18" published report [HEFCE, 2017], a total of 3,536 million pounds has been allocated for academic year 2017 – 18. This allocated money supports funding for teaching, research, capital grant, knowledge exchange and national facilities and initiatives, with 45.1% of the funding, a major chunk amounting to 1,595 million pounds, goes to research. These amounts will then be distributed into HEFCE-funded higher education institutions (HEIs).

It is therefore imperative to monitor and measure the effectiveness of research grants allocated to HEIs. To achieve this, an overarching framework for both assessment and funding of research has been developed, called The Research Excellence Framework (REF). The REF was created following the UK Government's announcement in December 2006 [Eastwood, 2007] to replace previous framework, the Research Assessment Exercises (RAE) which was last used in 2008. In short, REF is a system designed to analyse and evaluate the quality of research at various HEIs based in the UK and allocate funding based on which HEIs display the most outstanding quality of work [REF, 2014a].

The REF is conducted every seven to eight years, with the last REF having been conducted in 2014, where a total of 191,150 datasets had been submitted by HEIs [REF, 2014a]. The submission period generally opens 14 months prior to the assessment exercise, where the REF committee will invite HEIs to submit their datasets into 36 different disciplines segregated by units of assessments (UOAs). For the 2014 REF, after the submission period had ended, the expert review process was then conducted by the four main panels throughout 2014, from January to November, and concluded with publishing of REF results in December 2014. This process in whole took 26 months to complete.

1.1 Project Aim

In view of the lengthy processes and estimated high man-hours required to conduct REF assessment exercises by HEIs and REF committee, a team of five has been tasked to systematically analyse REF 2014 data. By knowing what factors do or do not contribute to getting a outstanding REF rating score for HEIs, a more focused application of man-hours on factors that increase the REF score could be taken by the HEIs.

In summary, this project aims to discover potential features that could affect the scores given by the REF, based on REF 2014 data, by using exploratory data analysis [NIST SEMATECH, 2013]. This would allow the team to not only find which features affect the REF score, but also which features might not have had any significant impact on the overall rating score.

2 Background

2.1 Research Excellence Framework (REF) 2014

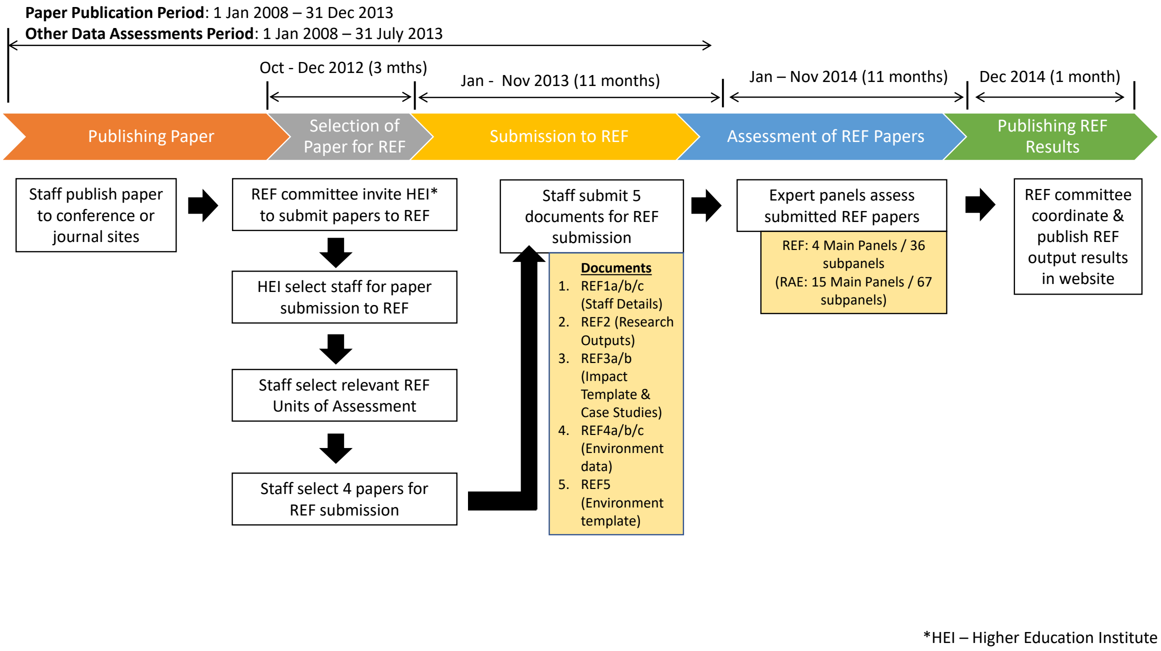

Paper Publication Period: 1 Jan 2008 – 31 Dec 2013
Other Data Assessments Period: 1 Jan 2008 – 31 July 2013
Oct - Dec 2012 (3 mths)
Jan - Nov 2013 (11 months)
Jan – Nov 2014 (11 months)
Dec 2014 (1 month)
Publishing Paper
Selection of Paper for REF
Submission to REF
Assessment of REF Papers
Publishing REF Results
Staff publish paper to conference or journal sites
REF committee invite HEI* to submit papers to REF
HEI select staff for paper submission to REF
Staff select relevant REF Units of Assessment
Staff select 4 papers for REF submission
Staff submit 5 documents for REF submission
Documents
1. REF1a/b/c (Staff Details)
2. REF2 (Research Outputs)
3. REF3a/b (Impact Template & Case Studies)
4. REF4a/b/c (Environment data)
5. REF5 (Environment template)
Expert panels assess submitted REF papers
REF: 4 Main Panels / 36 subpanels
(RAE: 15 Main Panels / 67 subpanels)
REF committee coordinate & publish REF output results in website
*HEI – Higher Education Institute

Figure 1: The overview process involved in REF 2014 from publishing paper by HEIs staff members up to REF 2014 committee members publishing the assessment results.

Following an invitation from REF committee to participate in the REF, HEIs must submit a dataset containing information on various articles such as information on the staff, outputs published by those staff, information of awarded degrees as well as a description of research environment, with each article in the dataset relating to a specific time period [REF, 2011]. For example, for output specifically, up to four papers per staff and published only in the assessment period of 1st January 2008 (after the end of RAE) to 31st December 2013 were accepted, unless there are special circumstances which allow staff to submit less than four outputs [REF, 2011].

The datasets submitted are then examined by four main expert panels; Main Panel A (Medicine, health and life sciences), Main Panel B (Physical sciences, engineering and mathematics), Main Panel C (Social sciences) and Main Panel D (Arts and humanities) [REF, 2014b], these panels are then further divided to sub-panels. The sub-panels review submissions according to a set of assessment criteria and level definitions. Each sub-panel has appointed members with a specific role to oversee and participate in the assessment of interdisciplinary research submitted in that UOA, to ensure its equitable assessment. All sub-panels include research users who participated in the assessment of Impact in particular. Specialist advisers are tasked to assist with outputs in languages that the sub-panel members are unable to assess.

The panels will judge the submissions based on three assessment criteria: Output, Impact, and Environment. Output is the basic quality of the submitted work, and is conducted according to international research quality standards. Impact measures how much of a societal, cultural or economical effect the submitted work has had on the world, or if the research could have any potential effect. Environment is a measure of how sustainable the research environment is, and if it contributes to the sustainability of a wider discipline or not. Each criteria has five rankings, 4*, 3*, 2*, 1* and Unclassified, with 4* being the highest ranking, and unclassified being the lowest, as it represents a submission that failed to reach the criteria for even a 1* grade. Once the grades for each assessment criteria are finalised, an overall score is calculated by making each assessment criteria a percentage of the overall score. These are weighed as 65% for Output, 20% for Impact, and 15% for Environment [REF, 2011]. A higher overall score would mean more funding would be allocated by the bodies to the institution.

3 Literature Review

3.1 Previous Work on REF 2014

3.1.1 Analysis of Impact Case Studies

A particular report was found documenting the approaches used by the main five Funding Councils in analysing the impact case studies submitted to REF, namely The Nature, Scale and Beneficiaries of Research Impact. Impact case studies are part of the submissions to REF made by institutions, outlining how past researches benefit "the economy, society, culture, public policy and services, health, environment and quality of life".

According to the report, the approach to assessing impact was a combination of text mining and qualitative analysis. In particular, three approaches were used during text mining - topic modelling, keyword searches and information extraction. In topic modelling, a cluster of words that frequently occurred in documents of similar contexts was defined as the "topic". Latent Dirichlet Allocation (LDA) algorithm was used during this process. Keyword searching made it easier to look at specific content in documents. The case studies were also matched against third party information through information extraction [HEFCE, 2015].

3.1.2 Analysis on Citations

The citation scores of each submission were provided by Scopus after request of the REF for input, but not for all the UOAs [REF, 2014a]. According to the panel criteria and working methods, as supplied by the panel, [REF, 2012c] the citation counts for Panel A, Panel B and Panel C where used as a supplementary tool as part of the indication of academic significance of the published work. Panel D on the other hand clearly specified in the assessment criteria [REF, 2012c], that they would neither receive or make use of citation data to assess the submitted work.Panels A and B appeared to have given more emphasis on the citation count than Panel C, with Panel B even utilising additional citation data from Google Scholar as an indicator. Panel B criteria [REF, 2012a] have specified that Google Scholar was accessed in a systematic way to observe if there was work highly cited outside of Scopus, which raises some issues. It does not specify if the Google Scholar citation data was retrieved during the same period for all papers and to what extent was this data used. Nevertheless , what one can interpret from the published criteria from these panels, is that the citation data could partly be responsible for the output assessment, except for Panel D.

Previous work on how citations could have influence the outcome of the REF 2014 scores have been performed by Alan Dix [Dix, 2015a] and Morris Sloman [Sloman, 2014]. Morris Sloaman, who was also the deputy chain in sub-panel UOA 11 [Sloman, 2014]. He was responsible for assessing the work submitted under the Computer Science and Informatics UOA. Therefore, the author attempted to analyse the REF output results by using citation-based metrics. Citation data from Google Scholar, as well as from Scopus were used to observe the percentage of REF ranked output against the citation data from both of the sources. Interpreting Morris Sloman's results,it is observed that there is an overall correlation between the REF ratings and citations.

Alan Dix, who was a member of the sub-panel responsible for assessing the Computer Science and Informatics UOA, also examined the work done by Morris Sloman [Sloman, 2014] and by others for the sub-areas in this UOA and found a systematic pattern. The pattern was that in several analyses performed the theoretical sub-areas of Computer Science and Informatics were benefiting over the practical ones [Dix, 2015b].

In Alan Dix's work *Citations and Sub-Area Bias in the UK Research Assessment Process* [Dix, 2015a], he has also produced work to assess the fairness of the REF process again explicitly in the Unit of Assessment he was responsible for assessing. The author's work focused mainly on assessing whether the citation counts have influenced the outcome of the scores. He also examined further the use of citation data to uncover if there was any bias within the sub-areas of Computer-Science and Informatics. The author used publicly available data for this analysis including raw Scopus data for the period of the process, the normalised Scopus data submitted to the REF and Google Scholar data. The Computer Science and Informatics UOA had the citation data supplied by Scopus. The results suggested that there was bias between sub-areas leading several universities to benefiting more than others and possibly also introducing gender-bias [Dix, 2015a].

3.1.3 Analysis on Income / Journal Ranking

Although Morris Slomans work [Sloman, 2014] mainly focused on citation based bias, the author also researched other metrics that could correlate with the REF output ratings. Examining Income per FTE against the REF scores, the author uncovered that there is strong correlation between the two, which can suggest that a higher research income per staff can produce higher REF scores. Additionally, the author used journal ranking data from Scopus and Thomson as metrics to assess the REF 4* percentages. It appears that there is no correlation between the journal ratings and the percentage 4*.

3.1.4 Existing Policy and Study Done on Selective Staff Submission Process

Existing policy has been developed by REF on Equality & Diversity (E&D) to ensure staff selection was done on equitable manner. In fact, a part of the submission guideline was dedicated to explicitly mentioning the principles of staff selection, and HEIs were requested to produce a code of practice and be transparent in the process.

On top of the existing policy and submission guideline, in an investigation by Higher Education Funding Council of England (HEFCE), which is one of the funding bodies behind REF, they analysed biases in terms of quantitative measures, such as disability, age, sex, ethnicity, and nationality from processes of staff selection. However, the report does not include scope of investigating whether there were biases in terms of research quality from this process alone.

4 Methodology

4.1 Programs Used

RStudio Version 1.0.153 and Python 3.6.1 with the Spyder Version 3.1.4 environment were used to view, clean, analyse, and visualise the data.

4.2 Data Used

The base datasets were downloaded from the REF 2014 website [REF, 2014b] for each UOA and were originally in the .xlsx format. Along with this dataset there were three other datasets used in this project, the Excellence in Research for Australia (ERA), Higher Education Statistics Agency (HESA), and the SCimago Journal and Country Rank.

4.2.1 Excellence in Research for Australia (ERA)

Similar to the REF, the ERA is a process initialised by the Australian Government to assess the quality of work of the higher education institutions in Australia [Australian Research Council, 2014]. When ERA was first conducted in 2010, as part of the process, a normalised spreadsheet was produced which ranked international journals based on inputs from expert, academic and public. The spreadsheet was then used as part of the assessment to evaluate the work of the higher institutions in Australia. The analysis on the results further on proceeded with every caution, since there is a time gap between the 2010 publication of the journal rankings and the 2014 REF process.

4.2.2 Higher Education Statistics Agency (HESA)

HESA was legally formed in 1993 as a not-for-profit private entity and has a vision of becoming HEIs related analytical powerhouse in UK. This organization works with HEIs in supporting their reporting of data under statutory requirement by UK Government. During REF 2014, HESA has provided contextual data related to estimates of total Full-Time Equivalent (FTE) expected from HEIs as part of REF 2014 analysis requirement [HESA, 2014].

4.2.3 SCimago Journal and Country Rank

The SCImago Journal & Country Rank is a publicly available portal that includes the journals and country scientific indicators developed from the information contained in the Scopus database [Scimago Journal and Country Rank, 2017]. The 2014 dataset containing international journals with the SJR score that SCImago uses to rank journals, which is an indicator based on citation data from Scopus, was used for this analysis.

4.3 Statistical Tests

4.3.1 Pearson & Spearman Correlation

The Pearson correlation coefficient measures the linear relationship between two datasets. This varies between -1 and +1 with 0 implying no correlation. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so [The Scipy community, 2014]. correlation test is useful provided the data tested comes from normal distribution. If data do not have normal distribution, it is advisable to use the non-parametric correlation test of Spearman [Lund Research Ltd, 2013] Table 1 aims to show how the results for correlation were interpreted for both correlation tests.

Table 1: The boundaries for the strength of correlations [Chowdhury et al., 2015]

Range of correleation coefficients	Degree of correlations	Range of correleation coefficients	Degree of correlations
0.80 - 1.00	very strong positive	(-0.19) - 0.00	very weak negative
0.60 - 0.79	strong positive	(-0.39) - (-0.20)	weak negative
0.40 - 0.59	moderate positive	(-0.59) - (-0.40)	moderate negative
0.20 - 0.39	weak positive	(-0.79) - (-0.60)	strong negative
0.00 - 0.19	very weak positive	(-1.00) - (-0.80)	very strong negative

4.4 Data Overview

From the original 36 Unit of Assessments (UOAs) present in the REF results spreadsheet, 10 UOAs were selected in random from each main panel as seen in Table 2. The number of UOAs selected was based on the number of UOAs present in each main panel, the larger the number of UOAs a panel covers, the more UOAs were selected.

Table 2: UOA Selection

Main Panel	Selected UOAs
A	2 - Public Health, Health Services and Primary Care 5 - Biological Sciences
B	9 - Physics 11 - Computer Sciences & Informatics 12 - Aeronautical, Mechanical, Chemical and Manufacturing Engineering
C	18 - Economics and Econometrics 24 - Anthropology and Development Studies 25 - Education
D	32 - Philosophy 36 - Communication

4.5 Feature 1 - Unusual words in “Additional Information” Text

4.5.1 Additional Information

Additional information constitutes part of the REF output submission where a paragraph of text is given for each submission to describe how significant the output is, how earlier work (before 2008) has been reviewed to incorporate new findings (if any), and how much co-authors have been contributing to the work [REF, 2012b]. There are different requirements for additional information across various main panels and even various UOAs. Nevertheless, additional information provides textual descriptions that most closely reflect on the actual content of the submissions, at least for some of the UOAs. It is therefore hypothesized that the more well-written additional information is, the higher the score would be. A clear definition for “well-written” additional information is therefore required.

4.5.2 Text analysis in REF

According to HEFCE [2015], text mining techniques were deployed in the assessment of impact case studies, which included mainly large bodies of descriptive texts. Given the time constraint of the current team project, it might be difficult to replicate the text mining analysis for the 6679 impact case studies. It was also noticed that the published case studies had certain words or phrases removed from the original submissions for publication purposes, which might be the keywords that lead to higher scores.

Nonetheless, text processing at a smaller scale could still be performed to analyse the texts involved in the tabulated REF submission data - 100-word additional information.

4.5.3 Attempt on Levenshtein distance

Levenshtein distance measures how similar two strings are in a character-by-character manner. In this context, a string refers to one paragraph of additional information. The distance is equivalent to the number of edits (e.g. deletion, insertion or substitution) required to transform one string to another [Gilleland, 2006]. Each paragraph of additional information is stored as a string in Python. Given that Levenshtein distance is one of the easiest-to-compute string similarity measurements, it was used in an attempt to measure the distance from the strings in top-ranking institution to the strings in every other institutions, by treating the top institution as a gold standard.

The "gold standard" institution was chosen through ranking all institutions in a certain way. The institutions were ranked based using all 4*, 3*, 2*, 1* and UC scores - they were firstly sorted by 4* scores, and if two or more institutions had the same 4* score, those were further sorted by 3* scores and so on. The "*" scores were sorted in descending order, while UC scores were sorted in ascending order.

When computing the Levenshtein distance between two institutions, all the strings in both institutions were compared against each other before an average was calculated as the final distance.

This approach was used to test the correlation between the averaged Levenshtein distance and the overall/output 4* for UOA 12, which contained the highest percentage of submissions with additional information available. It was noted that, for UOA 12, the top university ranked based on both overall and output scores was the same, thus the same "gold standard" was compared against. The results showed no correlation for both overall and output scores. It was therefore questionable whether the additional information written by the top university could be served as gold standard; and even if it could, whether the edit distance between two strings of roughly 100-word long would be a suitable measure. For example, since the characters in strings were compared in order, reversing the order of two words would give rise to a large Levenshtein difference. Althernative approaches needed to be explored.

4.5.4 TFIDF approach

TFIDF, short for term frequency-inverse document frequency, provides a numerical measure to how important a term is in a collection of documents. In this case, a term corresponds to a word and a document refers to a paragraph of additional information. The idea of TFIDF is to find a term that appears commonly in a particular document (high TF) but not often mentioned in other documents of the collection (high IDF), by assigning a weighting factor to each word as a product of TF and IDF [Lekovec et al., 2014]. "Important" words found with high TFIDF values are considered to characterize the topic of the document to which they belong. It is therefore hypothesized that the more "important" terms a document has, the more "well-written" the document is.

Given that the documents in this case are paragraphs of 100-words, to cover the broad range of topics in which a research work is involved, the number of occurrences of "important" words in a document is expected to be small. The "important" words are thus referred to "unusual" words. The number of "unusual" words in a document is counted as the number of words with TFIDF values above a threshold. Detailed explanation in threshold determination would be included later on in this report.

4.5.5 Computing TFIDF

One typical question for TFIDF is how to calculate TF and IDF. It is easy to find standard formula for such calculations in text books, however, adjustments are necessary for this specific application.

According to Lekovec et al. [2014], TF and IDF are calculated as follows:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \tag{1}$$

$$IDF_i = \log \frac{N}{n_i} \tag{2}$$

where the term frequency of a word i in document TF_{ij} is the number of occurrences of a word i in document f_{ij} normalised by the maximum number of occurrences of any word k in the document $max_k f_{kj}$, and IDF_i gives a measure for a word i appearing in n_i documents out of total number of documents N. TFIDF is therefore simply the product of TF_{ij} and IDF_i.

The first adjustment was associated with the computation of TF_{ij}. It was noticed that not every submission was provided with a solid paragraph of description as its additional information, some only had one generic sentence which was not considered very descriptive. In such cases, the value of $max_k f_{kj}$ would be relatively small, giving rise to a relatively higher TF_{ij} values comparing to more descriptive paragraphs which simply contained more words. These falsely large TF_{ij} values failed to reflect on the importance of words, thus the normalisation factor $max_k f_{kj}$ was removed from the calculation leaving only $TF_{ij} = f_{ij}$.

For the computation of IDF_i, one had to raise the question whether the collection of documents was defined per institution (N represents the number of submissions per institution) or as an overall pool of documents (N represents the total number of submissions in the UOA). To answer this question, both normalisation scales were tested for UOA11. It was proved that normalising by institution gives a much better correlation with overall 4* scores than that by normalising across all submissions. This could be explained as each institution has its writing style or a template provided for additional information, a more informative document should stand out from its peers – i.e containing more “unusual” words comparing to documents of the same style (in the same institution). Thus, when normalised by institution, a better correlation is shown as higher scores are awarded to the institution with distinctively written additional information for each of its submissions. The weak correlation shown for normalisation across all submissions probably infers that even within a particular UOA, there seems not to be any absolutely powerful word that guarantees high scores when mentioned in additional information. Consequently, IDF_i was computed by normalising documents per institution.

4.5.6 Determining threshold

A threshold defined such that words with a TFIDF value above the threshold are considered unusual. Based on the fact the requirements for additional information differ between main panels and even between UOAs, it is difficult to set a uniform threshold across all sample UOAs. An alternative approach was to use the threshold that gave the best correlation between unusual word count and 4* scores inside each UOA, and then to compare the best correlations across different UOAs or panels.

A document threshold was given as a certain percentile of the TFIDF values in a document, and the final threshold would be an average of document thresholds across all submissions in a UOA regardless of institutions. The final threshold was applied to count the number of unusual words per document, and the counted numbers were then averaged across each institution to give a unusual word count for every institution.

To obtain the threshold that gave the best correlation, a range of document thresholds were initially set from 90 to 99.5 percentile at a 0.5 percentile increment. Pearson correlations were computed between unusual word count and overall/output 4* scores at each threshold level, and the best correlation was chosen to represent, when the optimal threshold was chosen, how well the unusual word count could give indication to the overall/output 4* scores awarded to each institution.

4.6 Feature 2 - Citations Count

Alan Dix's [Dix, 2015a] and Morris Sloman's [Sloman, 2014] work focused only on UOA 11 for the citation analysis, and not any of the other UOAs. Additionally the outcome from reading through assessment criteria specified by the Main Panels[REF, 2012c], is that the citations should have some effect on the output rating but should not be a deciding factor. Main Panel D is the exception though as they clearly specified that they would not request or use any citation data to assess submitted work [REF, 2012c].

This analysis therefore is focused on identifying if there is any correlation between citations and the output ratings by the REF for all UOAs. The results will then be used to assess the difference in the extent of use of the citations as an indicator for the performance of work in different sub-panels and UOAs.

From the UOA's selected the following had citation data supplied by Scopus after request by the REF:

- 2 - Public Health, Health Services and Primary Care (Panel A)
- 5 - Biological Sciences (Panel A)
- 9 - Physics (Panel B)
- 11 - Computer Science & Informatics (Panel B)
- 18 - Economics and Econometrics (Panel C)

The analysis was therefore performed only on the UOAs listed above, due to the confirmed source of the citation data [REF, 2014a], using the data provided by the REF [REF, 2014b]. The original data contained the citation counts as integers in almost all UOAs, including 0, for each submission and in certain cases data was not available and the citation count was given as blank. To account for the unavailable data, instead of comparing the output score against the total citations per university, the total citations were divided by the number of submissions in each university to obtain a normalised value. UOA 11 was the only UOA were the citation data was supplied as string numbers and where therefore converted to integers.

The citations per submission for each university were then compared against the percentage of output ranked 4* and the combination of 2* and 1* papers in each university. The Pearsons method was used to return the correlation of the two variables. The reason why the combination 2* and 1* output was used, is to confirm that the inverse relationship that is expected from the 4* output is true.

4.7 Feature 3 - Research Income

In Morris Sloman's work [Sloman, 2014], income by staff is used to check whether there was any correlation between that and the REF outputs, in UOA 11. This analysis aims at examining whether there is any correlation between income per submission for each university in each UOA and the REF output. Through the results it is expected to observe if generally higher research income, regardless of the income source, implies higher quality research and thus a higher REF output.

The original data obtained from the REF website for each UOA, contained the research income by source for each university for each year during the period of the process. To obtain the income per submission per university, the total income for each university, for all sources and years was calculated. The incomes were supplied as integers. This number was then divided by the total number of submissions in each university. The income per submission for each university was then compared against the percentage of 4* and a combination of the percentages of 2* and 1* outputs. It is expected that universities with higher income per submission, regardless of the income source, will have a higher percentage of 4* output and a lower percentage of 2* and 1* outputs.

4.8 Feature 4 - Journal Ranking

Morris Sloman's [Sloman, 2014] work did not show any correlation between journal ranking and the REF output. The main aim of this assessment was to examine the hypothesis is that the journals and their reputation, or ranking, in which the research work was submitted to affects the 4* output score of the paper. The ERA 2010 journal ranking list [Australian Research Council, 2014], was used as the base dataset for the analysis on all the UOAs and the SCImago Journal Ranking list for the year 2014 for a complementary analysis on the results from the base dataset [Scimago Journal and Country Rank, 2017]. The SCImago Journal Ranking List was discovered later on in the analysis of the results, and therefore due to time limitations it was used as a supplementary dataset for confirming the results of the base dataset.

4.8.1 2010 ERA Journal Ranking List

The ERA 2010 journal ranking list ranks journals by A*, A, B, C and Not Ranked, with A* being the highest rated and C the lowest rating. The original data from the REF contains (per UOA) contains the submission, volume title and the ISSN code, which is a unique number for each journal.

Since there is a four year gap between the publication of the REF and the 2010 ERA, it is assumed that the journal ranks remained unaffected. Additionally due to the previous fact and the fact that not all work submitted in the REF was published in journals, it was expected that there would be some data loss when matching the two data sets.

The 2010 ERA journal ranking list and the REF data for each UOA were initially matched by volume titles, with most UOAs having half of the data matched. To match by volume titles, the text was normalised by lower-casing, and punctuation and any symbols were removed. This raised the issue though that there might be a mismatch between similarly named journals. To avoid this it was decided that it would be better if the datasets were matched by the ISSN numbers.

This returned a matched dataframe which contained the university, submission, volume title and the 2010 ERA rank. To be able to quantify the effect the journal ranks have on the output, the journal rank scores were given a representative score. The scores were converted as follows:

- A* : 4
- A : 3
- B : 2
- C : 1
- Not Ranked : 0

The idea behind the scores is that a paper published in an A* journal will have a larger gravity in affecting the output 4* percentage of the university, C will have the least and Not Ranked will have none. Different weighting schemes were trialled on several UOAs before settling with the scores above. Different scores produced better correlation in some cases, for example in the Physics UOA, the scores A*: 1, A : 1, B : -2.5, C: -5 and Not Ranked:0, returned a correlation of 0.61. This was better but the pattern was the same and the overall outcome was not largely affected. After converting the 2010 ERA ranks to their scores, the scores were summed up for each university and divided by the number of submissions to find the average journal rank per university. The average journal rank per university was then compared against the 4* output.

4.8.2 SCImago Journal Ranking List

The SCImago Journal Ranking List ranks journals by an indicator, the SJR score, which is based on data from the Scopus data created by SCImago. This dataset was used only in strongy correlated UOAs from the 2010 ERA Journal Ranking List. Unlike with the 2010 ERA dataset, the SCImago dataset and the REF dataset were matched by the name of the journal title rather than the ISSN. Matching by ISSN, returned very low matched percentages, whilst matching by journal name after normalisation resulted in higher matching percentages.

Following the data matching by name the new dataframe for each UOA contained the university, submission, journal title and the SJR . The SJR scores where then summed up for each university and divided by the number of submissions to find the average SJR score per submission per university in each UOA. This value was then compared against the output 4* of each university.

4.9 Feature 5 - Distance of Institutions from London

This feature aims to examine if the distance from London affects the grades the REF gives an institution, as in England London, and the area near and around London, is known to have the most Prestigious universities, e.g. Cambridge, Oxford and Imperial College London are all within 85km of London.

This was tested in two ways, firstly with a correlation test between the distance in kilometres from London and the mean percentage of 4* grades per institution. The second method tested whether if there was any significant difference in the mean % of 4* grades if the institutions were grouped by distance, for this a KruskalWallis test was ran.

To test for a correlation between the distance institution is from London and the percentage of 4* grades given by the REF, the distance from London for each institution was first of all found out using an online distance calculator, it was then placed in a dataset with the mean overall 4* % per institution. Finally a Spearmans rank correlation test [Lund Research Ltd, 2013] was ran on these newly generated distances against the mean 4* % of the corresponding institution. A further test was also ran to see if there was any significant difference between the 4* grades given if the data was grouped by distance, to this end the data was grouped into seven groups containing twenty two institutions.

4.10 Feature 6 - Russell Group vs Non-Russell Groups

Some of the universities on the REF's list of institutions are classified as "Russell Groups", which are considered to be the best institutions in the UK. This feature is being tested for as a possible cause of bias amongst the judging bodies, as they may be inclined to award a higher score to a Russell group university even if the submitted output is the same as a non-Russell group institutes. To this end a Mann-Whitney U test was ran between the various REF grades of the Russell and non-Russell institutions.

4.11 Feature 7 - Selective Submission of Output

4.11.1 Higher Number of Outputs Per Institution - Initial Hypothesis

Feature 7 was selected based on the outcome of the initial checking of linear relationship between the number of outputs submitted per institution versus percentage of outputs with four stars ranking for UOA 11 (Computer Science and Informatics). The initial analysis intent was to see if there was any impact of higher submissions by institutions to the outcome of the ranking.

However, the original plot done in R Studio revealed an institution which perceived to be good ranking was having low rating of four stars (between 0 to 10%), although the number of submissions was high (more than 400 outputs), whereas there were another two institutions with lower submissions count (between 200 to 300 outputs), but have higher four stars rating (between 40 to 50%).

Though the initial analysis was off, this triggered a hypothesis of institutions practice of being selective on the outputs, and whether this could lead to a higher output ranking upon weird observation of the plot.

4.11.2 Selective Submission of Output - Current Hypothesis

Checking for this hypothesis required additional dataset outside of existing REF 2014 data as the current submitted Full-Time Equivalent (FTE) data for Category A was not sufficient to prove this hypothesis. Instead, the data used was from HESA, which was given to REF committee during REF 2014 assessment exercise. The calculated FTE data by HESA using internal records approximated the eligible number of staff that were supposed to be returned by HEIs. The calculation was further scaled due to the differences of definitions between HESA and REF. (HESA, 2014).

The analysis was first done by checking the distributions of the submitted FTE of Total Category A by HEIs and Scaled FTE provided by HESA, using Shapiro-Wilk normality test in R. Upon checking the distribution, since there was not enough evidence that both were from normal distribution, then the test proceeded with Spearman correlation method, instead of Pearson. Since three variables were involved, a 3D plot was attempted between Eligible FTE, Four Star Output, and Number of Submissions Per University to see how these two variables could have affected four stars ranking. Then, a multiple correlation coefficient was calculated.

The same processes were replicated for all UOAs, with minor differences on pre-processing part for certain UOAs, where the outlier was removed to check if it will improve the correlation. However, five UOA's were not able to be pre-processed due the some of the output submitted by HEIs were marked as two submissions, instead of one. Therefore, this introduce two different ranking instead of one.

Prior to this current hypothesis analysis, the initial hypothesis was again checked and written properly in R Notebook with different scripts produced for data wrangling and data analysis. This was again confirmed by having similar checking on three stars ranking and average scores of three and four stars ranking.

4.11.3 Higher Number of Outputs Per Institution - Future Hypothesis

However, upon re-checking the result of data produced by higher number of outputs versus four stars ranking produced by R, and compared against similar data produced by Python code written in Jupyter Notebook, the output was different although the methodology was similar.

5 Results

5.1 Feature 1 - Unusual words in "Additional Information" text

5.1.1 Correlations

Table 3: Summary of best correlations between unusual word count and overall/output 4* score in sample UOAs.

Main panel	UOA	Submissions with additional information (%)	Overall 4*		Output 4*	
			Threshold (percentile)	Best correlation	Threshold (percentile)	Best correlation
A	2	50.75	98	-0.54*	99.5	-0.39*
	5	25.35	99	0.22	99	0.21
B	9	32.57	91	0.53*	95	0.51*
	11	93.74	98	0.46*	97	0.41*
	12	95.19	97.5	0.62**	95.5	0.55*
C	18	6.66	90.5	0.08	94	0.06
	24	7.97	91	0.4	99	0.34
	25	8.49	96	0.19	91.5	0.15
D	32	5.79	99	0.06	99	0.06
	36	15.60	99.5	-0.15	99.5	-0.25

As seen in Table 3, correlations of moderate level were indicated using * while strong correlations were indicated using **. It was noticed that strong or moderate positive correlations were found in UOAs that belong to panel B. All UOAs in other panels showed weak or no correlations between unusual word count and 4* scores, except for UOA2 in panel A where moderate negative correlation was discovered. The results could be easily visualised in Figures 2 – 3.

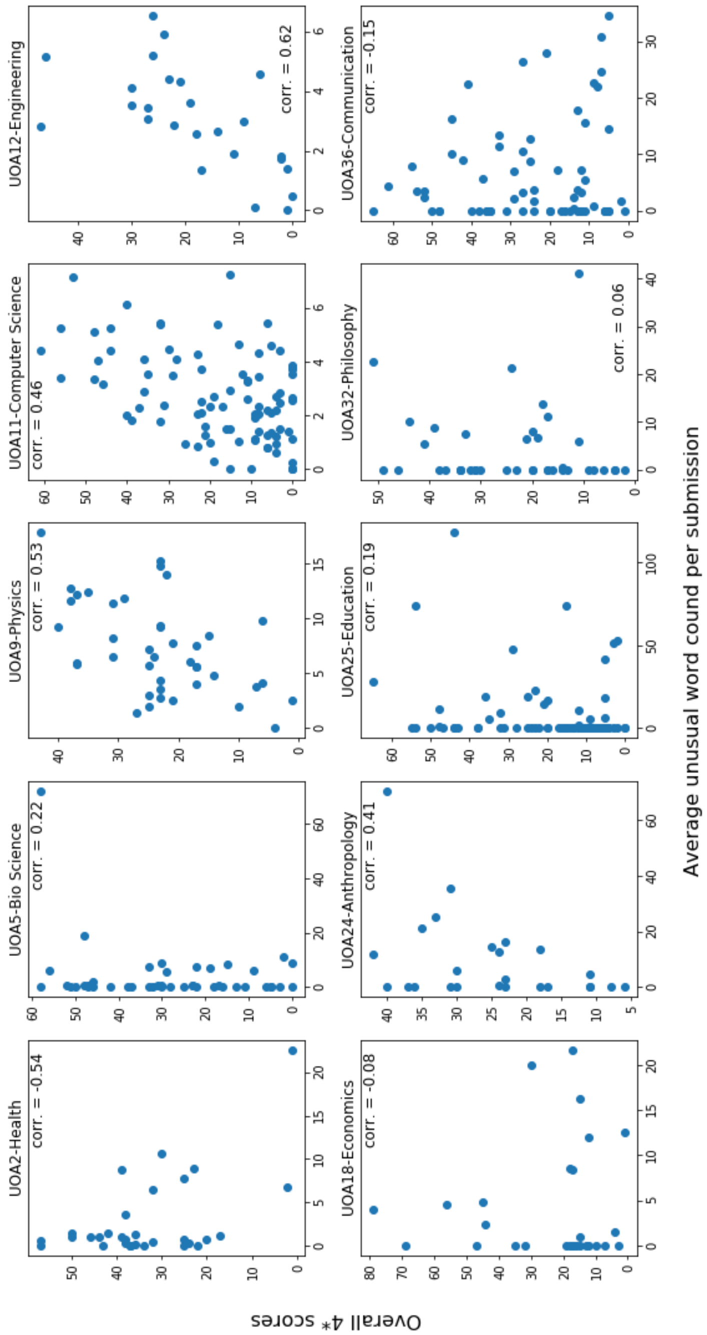


Figure 2: Average unusual word count per submission vs overall 4* scores plots for every sample UOA.

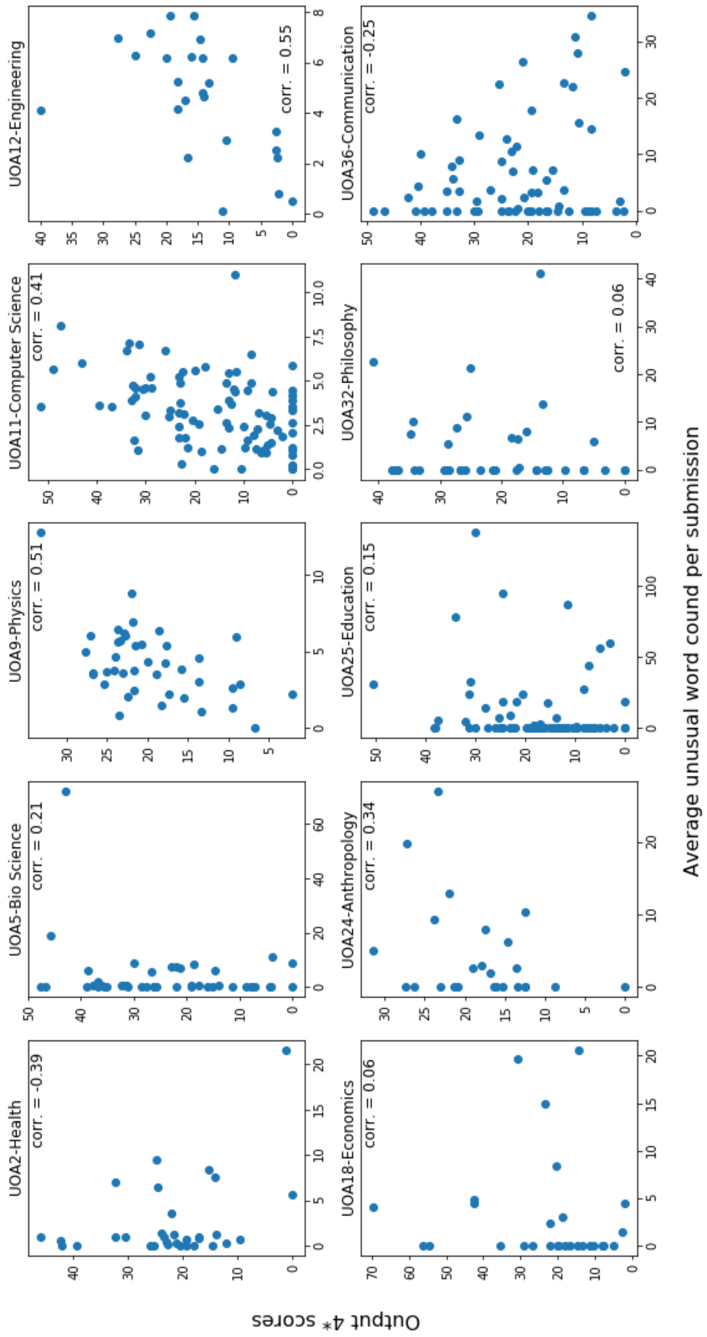


Figure 3: Average unusual word count per submission vs output 4* scores plots for every sample UOA.

5.1.2 Difference in correlations

Based on the results above, different levels of correlations were obtained for various UOAs. This section attempts to discuss all the possible reasons behind such difference thus to provide insights into the level of influence additional information has on the scores awarded.

Percentage of submissions with additional information
In the first instance, one would suspect the difference in correlations is largely associated with the amount of data available - the more data there is, the more likely a correlation is going to be found. This could be illustrated in Figure 4, which seems to show a trend that Pearson correlation coefficient becomes more positive as more data is given. However, there are two distinctive cases where negative and near-strong negative correlations are seen.

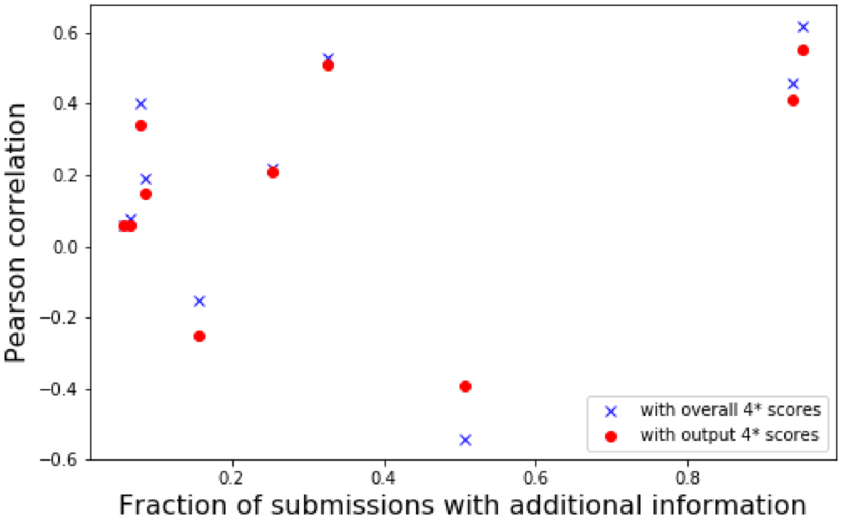


Figure 4: Relation between correlation and amount of additional information available.

Main panels
Table 3 shows that the highest positive correlations are mainly found in main panel B, while negative correlations are found in panels A and D respectively. This reflects on the fact that the UOAs in panel B generally provide the largest amount data in the form of additional information, while panel A being the second largest. Strong correlations (both positive and negative) are found in these panels.

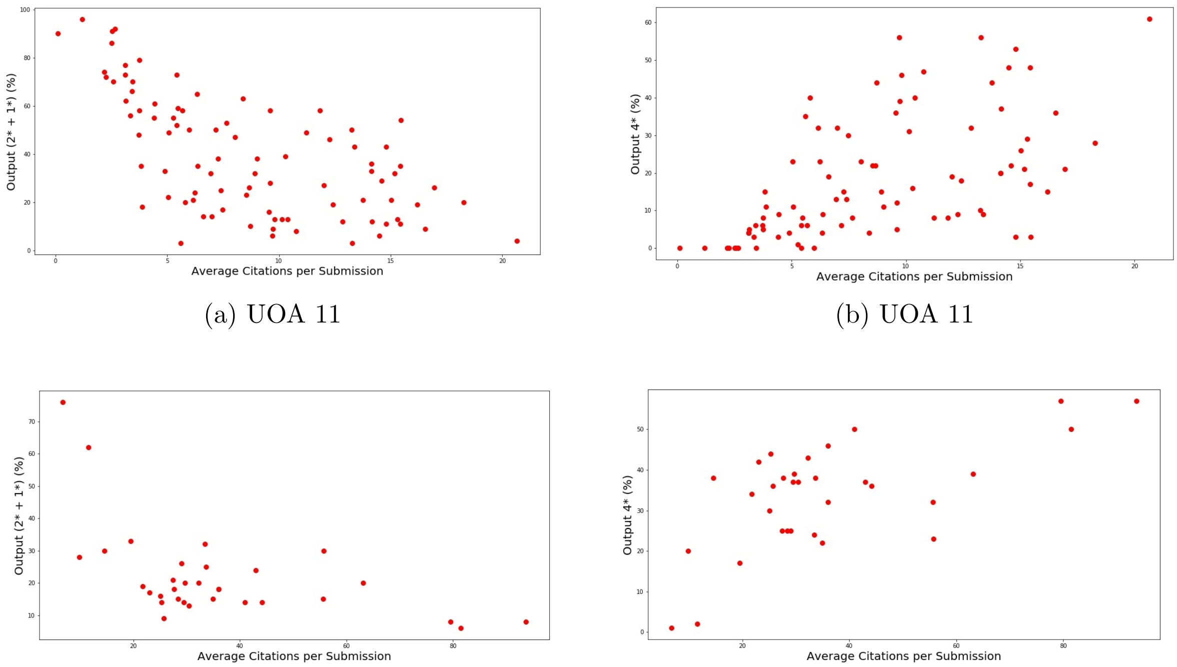
5.2 Feature 2 - Citations Count

The R values from the Pearson test performed on the data can be seen in Table 4. The R-values returned indicate the correlation between average citations per submission for each university in each UOA and the REF output listed in the table.

Table 4: Pearson correlation R-values between average citations per submission vs. Output 4* / Output 2* + 1*. ** refers to strong Pearson correlation; * refers to moderate correlation.

Main Panel	UOA	Output 4*	Output 2* + 1*
A	2	0.63**	-0.52*
	5	0.40*	-0.35
B	9	0.49*	-0.61**
	11	0.57*	-0.63**
C	18	0.73**	-0.63**

To illustrate the two best correlated UOAs, Figure 5 can be seen below. From the figures it appears that there is a linear relationship between the two data and this is supported by the R-values returned. When comparing against the combined 2* and 1* outputs the relationship in the Economics and Econometrics UOA appears to follow a linear-to-quadratic pattern.



(a) UOA 11

(b) UOA 11

(c) UOA 2

(d) UOA 2

Figure 5: Average citations per submission vs. REF output percentages 4* and (2* +1*) for UOAs 2 and 11.

The results suggest that there is large variance between the effect of the citation numbers and output scores in each UOA. Even comparing the correlation results between UOAs of the same panel, we can see some variances, with the obvious example being Main Panel A. What this can suggest is that although there were guidelines on the assessment criteria, when considering citations, published by each Main Panel [REF, 2012c], several sub-panels used the citation data more than others.

In UOA 18, which is the UOA with the strongest correlations on both outputs compared, the results can imply that the sub-panel responsible for assessing the research work did not adhere to the degree described by Main Panel C in the assessment criteria [REF, 2012c].

5.3 Feature 3 - Research Income

The R values from the Pearson test performed on the data can be seen in Table 5. The R-values returned indicate the correlation between average income per submission for each university in each UOA and the REF output listed in the table.

Table 5: Pearson correlation R-values between average income per submission vs. Output 4* / Output 2* + 1*. ** refers to very strong Pearson correlation; * refers to strong correlation.

Main Panel	UOA	Output 4*	Output 2* + 1*
A	2	0.45	-0.39
	5	0.82**	-0.70*
B	9	0.45	-0.48
	11	0.74*	-0.67*
	12	0.65*	-0.52
C	18	0.73*	-0.60*
	24	0.38	-0.47
	25	0.41	-0.36
D	32	0.63*	-0.55
	36	0.34	-0.22

The results indicate that there is a large variance on the effect of income on the REF output result between UOAs. This variance is also observed between UOAs of the same Main Panel, with the example of Main Panel A, similarly as in the citations. The larger number of strongly correlated UOA's can suggest that the research quality is largely dependent, in these areas, and heavily influenced by the research income per submission.

5.4 Feature 4 - Journal Rankings

5.4.1 Results using the ERA 2010 Journal Ranking List

The R values from the Pearson test performed on the data can be seen in Table 6, along with the percentage of matched data in each UOA. The R-values returned indicate the correlation between average journal ranking for each university in each UOA and the REF output listed in the table.

Table 6: Pearson correlation R-values between average income per submission vs. Output 4*. Table also includes matched data in percentage for each UOA. ** refers to very strong Pearson correlation; * refers to strong correlation.

Main Panel	UOA	Matched Data (%)	Output 4*
A	2	90	0.63*
	5	95	0.80**
B	9	92	0.53
	11	62	0.73*
	12	92	0.55
C	18	89	0.82**
	24	60	0.57
	25	53	0.43
D	32	54	0.60*
	36	No results could be extracted	

Whilst the python code was implemented successfully for this analysis most UOAs, this was not the case for UOA 36. During the analysis it was attempted to identify the issue and troubleshoot the code, but with no result.

The results from several UOAs returned a very strong positive correlation, therefore possibly implying that the journal rank does affect the output score of a paper. In an attempt to further analyse the results the Computer Science and Informatics (UOA 11) and the Economics and Econometrics (UOA 18) were selected for further analysis. The figures below show the plotted results for both of the UOAs, with the analysis on each graph below them.

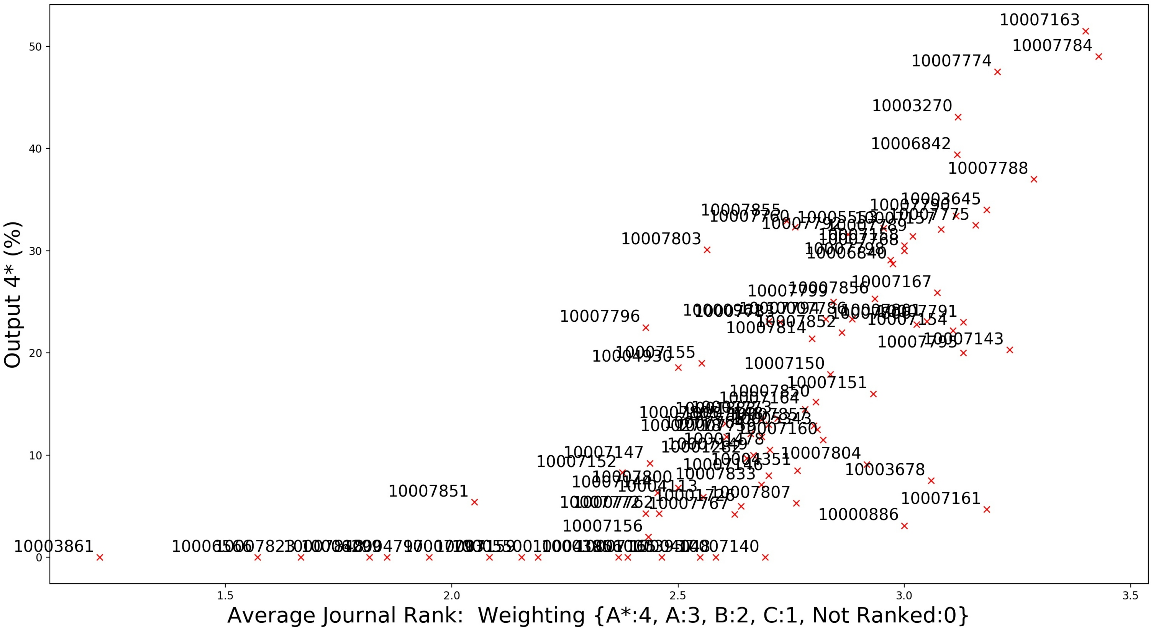


Figure 6: UOA 11 - Average journal rank vs. Percentage of 4* Output. The annotations represent the UKPRN code for the universities. Each data point is a university.

What is visible from Figure 6 is that almost all of the universities with up to approximately 2.3 in average journal rank have 0% of 4* output rated papers. After that threshold the data appears to follow a linear pattern, with universities having a higher average journal rank having also a higher percentage of 4* output. What is interesting are the universities that appear not to follow the pattern, or are no close to the rest of the data. Specifically universities with UKPRN 10000886 and 10007161 have a high average journal rank but a very low percentage of 4* output papers, when comparing with universities with similar average journal rank. What was observed is that these universities have significantly less submissions than the universities with an average rank close to these two.

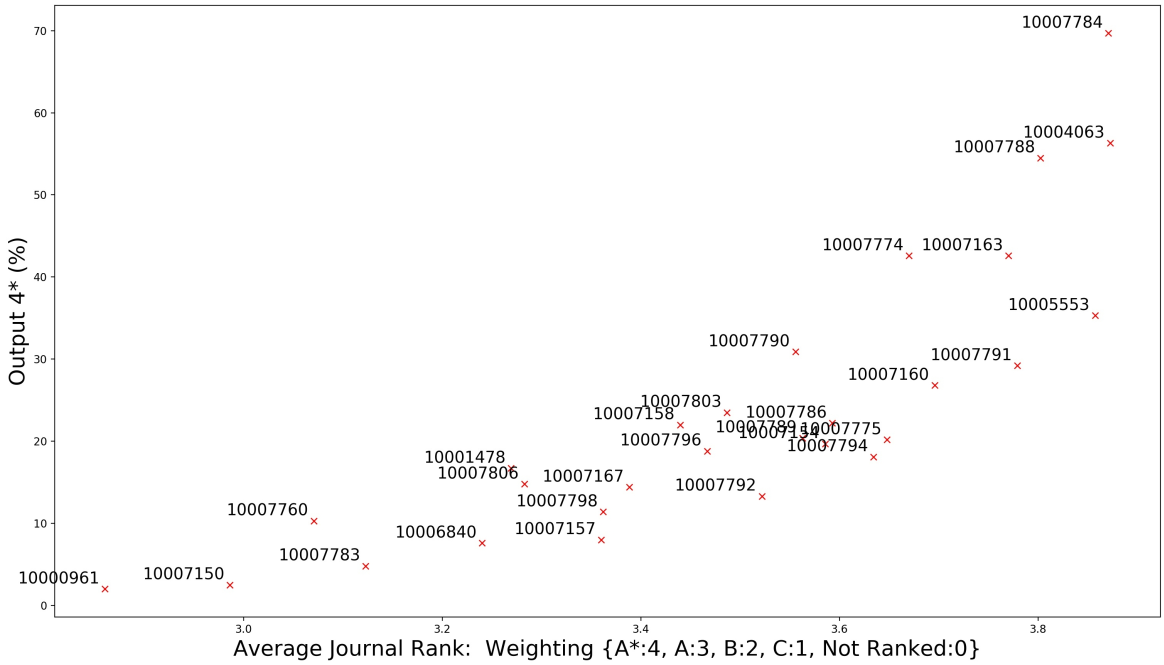


Figure 7: UOA 18 - Average journal rank vs. Percentage of 4* Output. The annotations represent the UKPRN code for the universities. Each data point is a university.

By selecting the university with the UKPRN 10005553 from Figure 7 and comparing to the universities with a higher 4* output but similar journal rank, the same pattern is observed. University 10005553 has significantly less submissions than the universities with a close average journal rank.

5.4.2 Further analysis using the SCImago 2014 Journal Ranking List

To further assess the most strongly correlated UOAs from Panels A, B and C to check see if the hypothesis that journal rankings could possibly affect the REF output results still holds valid. Journal ranking scores from the SCImago Journal & Country Rank [Scimago Journal and Country Rank, 2017] for the year 2014 were used. Table 7 summarises the R-values returned with the Pearson test, between the average SJR score per institution and the output 4* percentage.

Table 7: R-values of correlation between average SJR score per institution vs. Output 4* for UOAs 2, 11 and 18.

Main Panel	UOA	Matched Data(%)	Output 4*
A	5	73	0.83
B	11	59	0.63
C	18	73	0.95

The results produced with the SCImago dataset agree with the results produced with the 2010 ERA Journal Ranking List. The correlation results of the three UOAs still indicate that there is a strong correlation between the journal ranking in any sort of form against the 4* output of the REF. In this case the results even show that in UOA 18 the two variables are very strongly correlated. The plot for UOA 18 can be seen below in Figure 8.

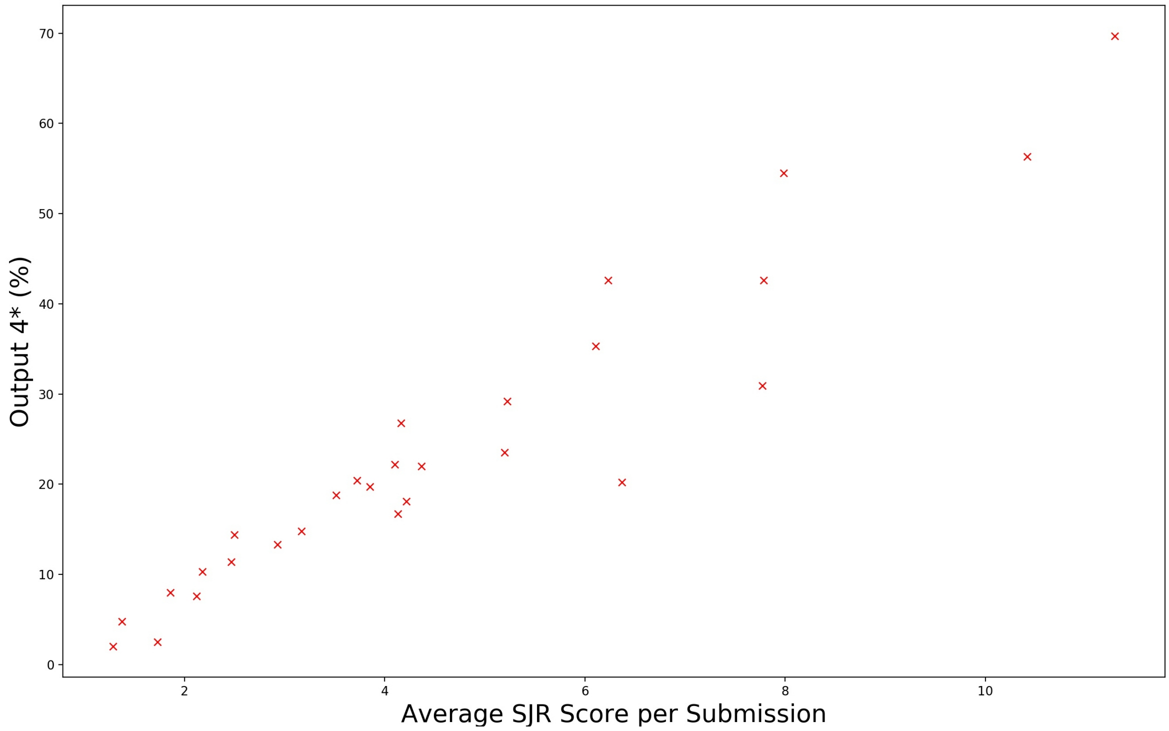


Figure 8: UOA 18 - Average SJR score per submission vs. Percentage of 4* Output per institution. Each data point is an institution

5.5 Feature 5 - Distance of Institutions from London

The Spearman's Rank Correlation test that was ran between the institution's distance from London and it's mean % of 4* Grades had a result of $r_s = -0.2088194, p = 0.00935.$, giving evidence for a weak negative correlation between the grade the REF gave an institution and its distance from London. The scatterplot of this correlation can be seen in figure 9

The KruskalWallis test ran between the location groups gave a result of $H = 19.384, p = 0.003562$, giving strong evidence to support the idea of there being a difference between the 4* grades of the grouped distances. This can also be seen in figure 10.

5.6 Feature 6 - Russell Group vs Non-Russell Groups

All of the Mann-Whitney tests gave strong evidence towards there being a significant difference in the mean % of each grade, with the Russell groups showing a higher mean % of 4* grades, and a lower mean % of all the other grades, when compared to the non-Russell groups. The mean scores, the Mann-Whitney U score and the p value for each grade can be seen in Table 8, and a visual difference can be seen in the boxplot of these results, as shown in Figure 11.

Table 8: The results from the Mann-Whitney U test between the REF grades of Russell and Non-Russell grouped Institutions

Grade	W Score	P value	Russell Mean	Non-Russell Mean
4*	137860	<2.2e-16	34.24288	16.66318
3*	331630	4.41E-13	47.62969	42.23875
2*	687050	<2.2e-16	16.17391	31.67685
1*	622240	<2.2e-16	1.661169	8.135048
Unclassified	481990	1.77E-13	0.2923538	1.286174

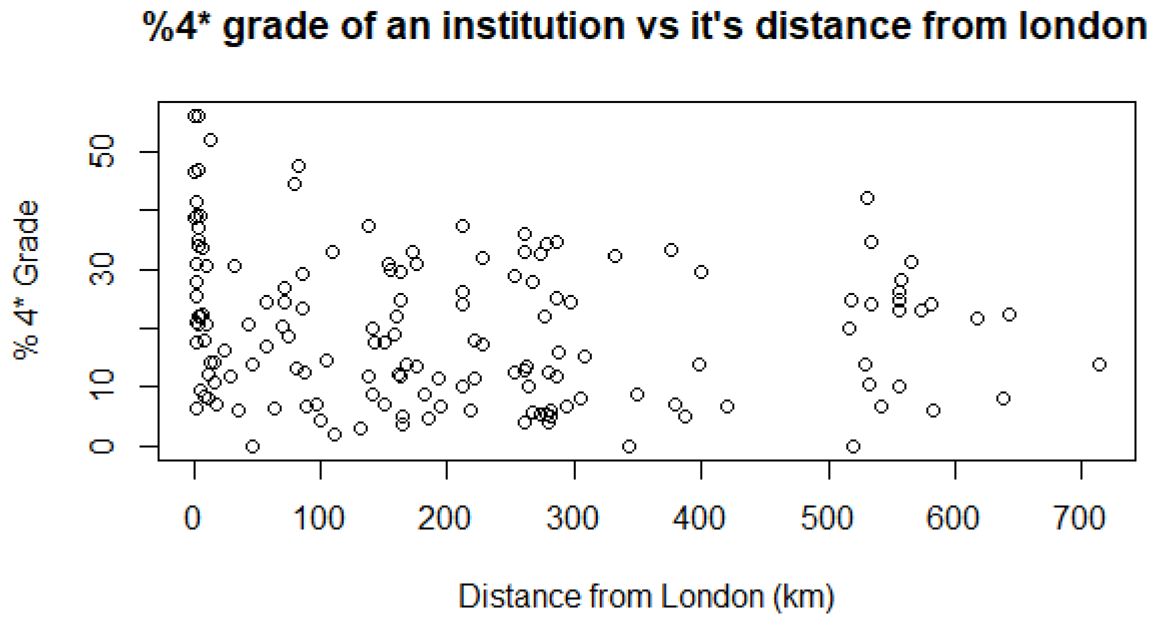


Figure 9: An Institutions distance from London plotted against it's mean % grade from the REF

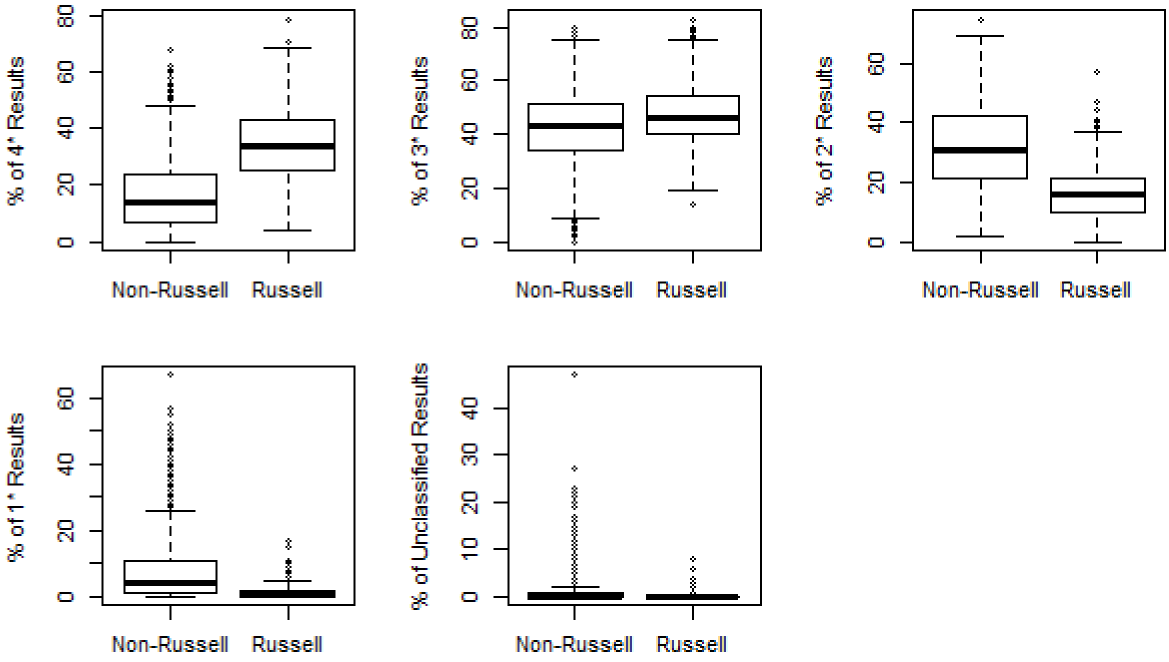


Figure 11: The difference between the REF grades between Russell and non-Russell groups

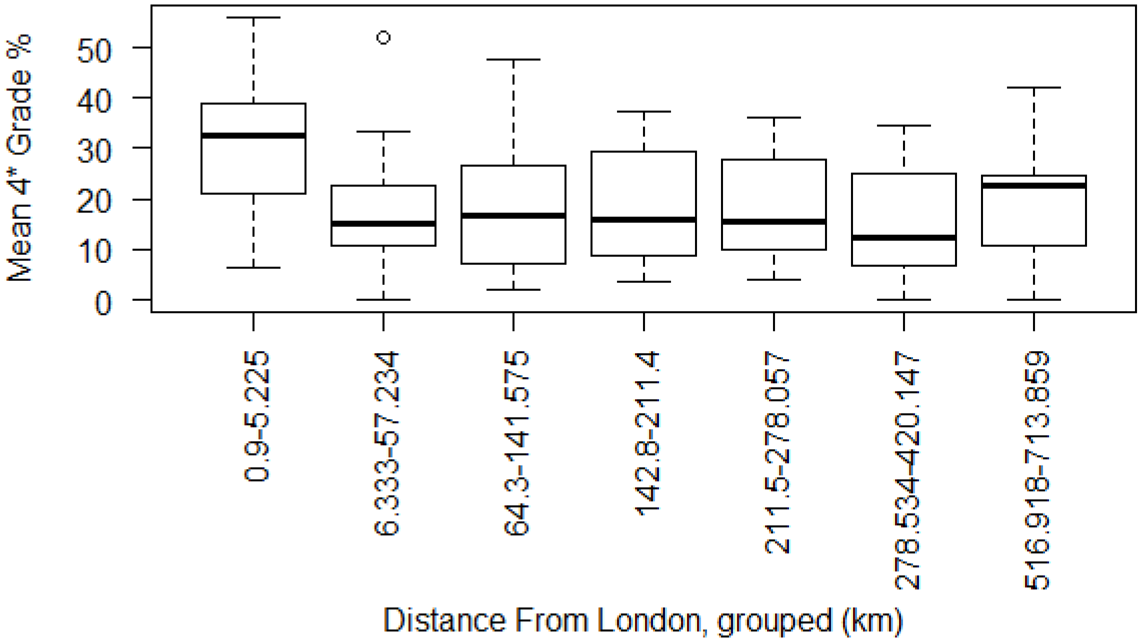


Figure 10: Distance from London as groups plotted against their mean % 4* grade from the REF

5.7 Feature 7 - Selective Submission of Output

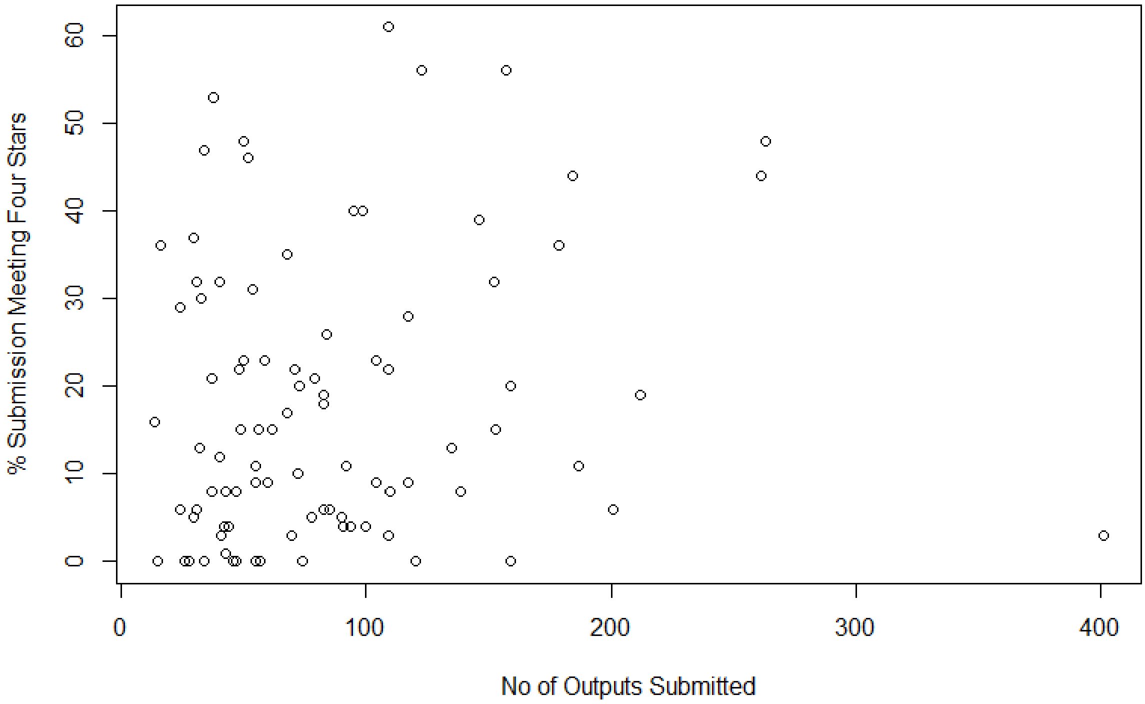


Figure 12: The original plot of number of outputs submitted per university versus percentage of submissions meeting four stars ranking of output for UOA 11

Based on the original plot using R Studio in Figure 12, there was no apparent pattern of the populations, and the highest point of the outputs submitted (400) showed that it was having a ranking between 0 to 10%, but the second and third highest point of the submission counts showed that both have good ranking of four stars, which were between 40 to 50%. These initial observations triggered the analysis of selective staff submission to be performed. At this point of time, no correlation test have been performed yet based on the plot result achieved.

After the pre-processing of the UOA 11 dataset was re-created step-by-step in R Notebook to ensure the correct methodology was in-place, Figure 13 was achieved. Plots in Figure 13 showed that there were still no apparent distributions for the amended plot. However, between Figure 12 and 13, two of the plots have shifted to a different ranking of 0 to 10% compared to earlier ones of 40 to 50%.

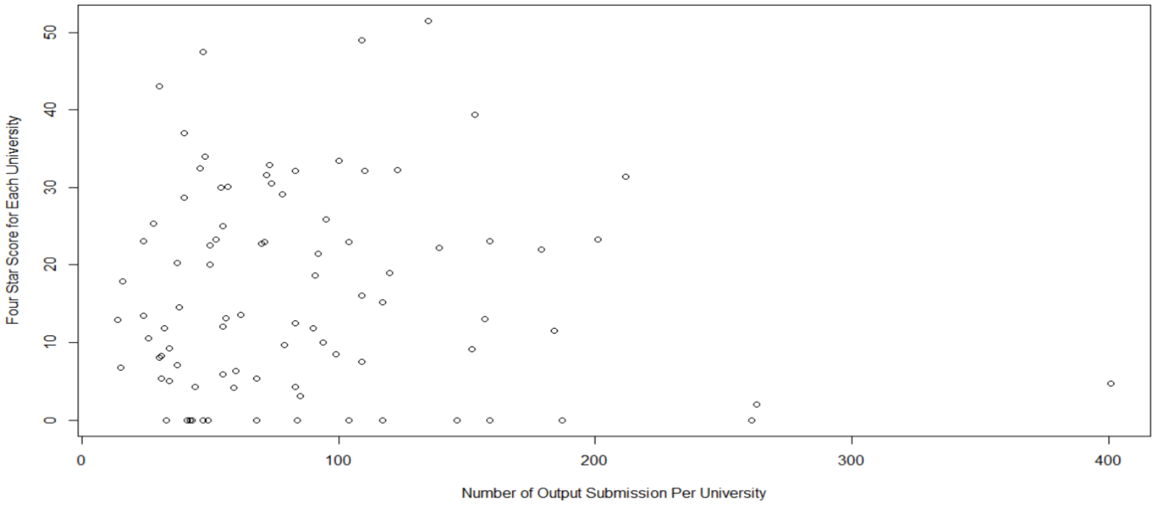


Figure 13: The amended plot of number of outputs submitted per university versus percentage of submissions meeting four stars ranking of output for UOA 11

Upon checking the actual data table, these last three plots were plotted wrongly, therefore, there was a possibility that the all the distribution of the points were wrong. Because of this, the code was reproduced using Python and Figure 14 was achieved which has different distribution compared to the earlier plots drawn using R. With this new plot, the correlation achieved for UOA 11 is 0.698. So, this feature should be re-looked and re-attempted for future research work.

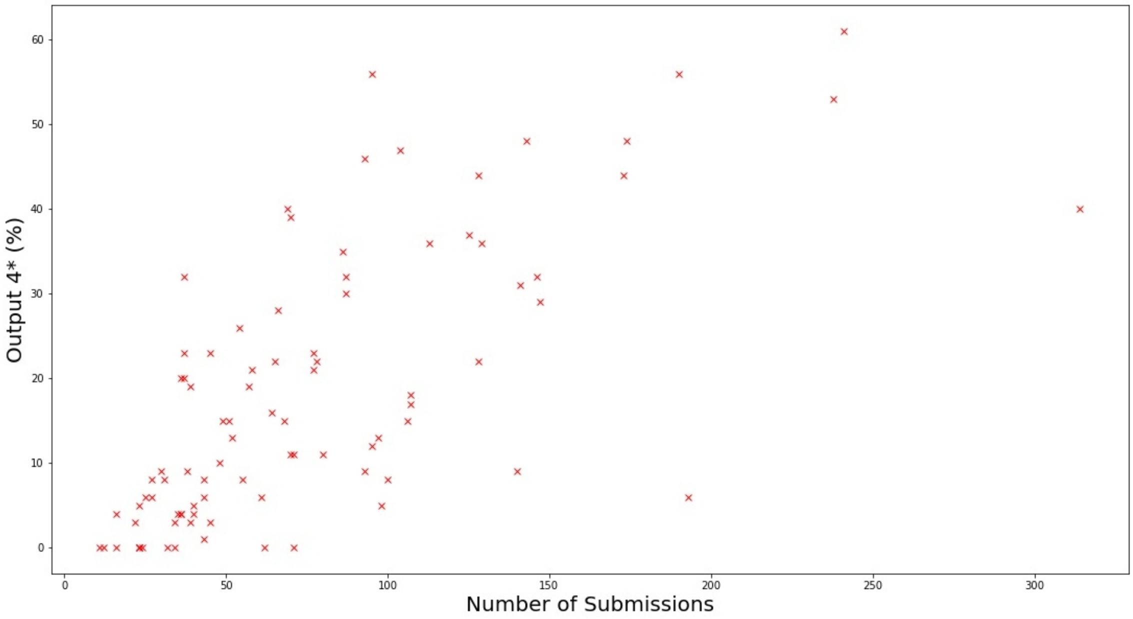


Figure 14: Re-created plot using Python in Jupyter Notebook for UOA 11

However, to continue further on the existing HESA data obtained to check on the selective submissions of output feature, Figure 15 was attempted to see if there were any correlation between these three variables.

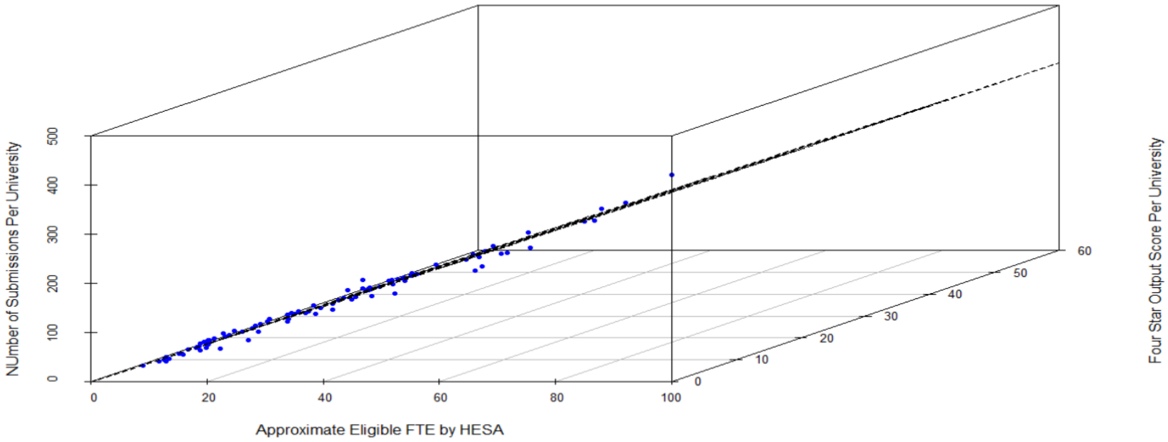


Figure 15: 3D Plot of Output Submissions versus Four Stars Ranking versus Approximate Eligible FTE by HESA

Based on the multiple correlation coefficient in Table 9, there are strong correlation suggested between "Eligible FTE by HESA" and "Number of Submissions Per University", which is 0.98. However, the coefficient returned between "Four Stars Ranking Per University" and both the other variables were not strong. But, because the earlier plots of the distribution was wrongly calculated for number of output submission versus four stars ranking, it was possible that the correlation coefficient was not returned properly. For other UOAs, the results achieved were similar, where there are strong correlations between "Eligible FTE by HESA" and "Number of

Submissions Per University".

Table 9: Multiple Correlation Coefficient of X - Approximate Eligible FTE by HESA, Y - Four Stars Ranking Per University, and Z - Number of Submissions Output Per University

	X	Y	Z
X	1.0	0.064	0.98
Y	0.064	1.0	0.042
Z	0.98	0.042	1.0

6 Discussions

6.1 Feature 1 - Unusual words in “Additional Information” Text

6.1.1 Correlations and amount of data available

In general, the more data available, the stronger the correlation is between unusual word count and 4* scores.

6.1.2 Negative correlations

Since 2 out of 10 chosen UOAs are showing negative correlations, it is therefore hard to conclude that unusual word count correlates positively with 4* scores. Further clarification should be made by considering more UOAs. However, given the chosen sample UOAs for this project, it is suggested to look at the main panels in which a UOA belongs on efforts to identify any pattern that gives rise to different directions of correlations.

6.1.3 Correlations and main panels

The strongest positive correlations were found mainly in panel B, and negative correlations were also possible in panel A and D.

This could be explained by REF submission requirements - for additional information, panel A emphasizes the contributions of co-authors and co-researchers, panel B focuses on abstract-like description for the actual work, while panels C and D do not have any strict emphasis [REF, 2012b]. This casts a doubt on the previous hypothesis whether a more “well-written” and descriptive additional information would give rise to a better score in all main panels, given that the additional information in some panels does not summarise the submissions. Thus, this hypothesis could be modified such that more “well-written” additional information only contributes positively to 4* scores for UOAs in panel B. Since mainly co-author/co-researcher information is involved in additional information in panel A, it is reasonable to question to what extent this information contributes to the scores. Therefore, more investigations in panel A are required to clarify whether or not the negative correlation is a coincidence.

6.2 Feature 2 - Citations

The results have indicated differences in correlation values between the UOA’s when comparing over the entire sample but also when comparing the UOAs within the same Main Panel. How this can be interpreted is that the sub-panels responsible for the UOAs with higher correlation values, may have possibly used to a larger extent the citations as an indicator of the quality of submitted work.

6.3 Feature 3 - Research Income

The strong correlation results in several UOAs can suggest that in those UOAs, the quality of work is largely affected by the research income, irrespective of the source. The variance between correlation results in the UOAs is still apparent.

In UOAs where the results indicate strong correlations, it is possible that higher income per submission is an indicator of higher quality work. There is also the possibility that although the assessing members were trying to be objective, there was some institutional bias in the strongly correlated UOAs. The assumption is that papers from higher research income institutions were assessed more leniently than lower research income institutions.

6.4 Feature 4 - Journal Ranking

By using the ERA 2010 Journal Ranking List [Australian Research Council, 2014], the results show strong correlations in several UOAs. In particular there is a repetitive pattern with the Research Income feature. As it appears the three UOAs with the strongest correlation results in both features are UOAs 5, 11, 18 and the UOA with the weakest correlation in both features is UOA 25. The strong correlation results of the three UOAs mentioned previously, have been verified by using a different dataset and a different methodology for Journal Ranking. This was achieved by using the SCImago SJR score [Scimago Journal and Country Rank, 2017] the strong correlations were replicated, with the cases of UOA 5 and 18 returning even stronger correlation results.

This can hardly be a coincidence, and is a strong indication that both Research Income and the Journal Rankings were good indicators in predicting the output for the REF 2014 process in the UOAs that returned strong correlation results.

6.5 Feature 5 - Distance of Institutions from London

The weak negative correlation results as shown in figure 9 suggest there is no real relationship between the distance from London and the 4* grade given by the REF, however the results in figure 10 would suggest that there is some difference between the distance and 4* grades, at least when the distances are grouped together. These two conflicting results would imply that although the distance by itself doesn't affect the grading of the REF that much, it could be a latent variable that occurs when the distances are grouped that could cause the difference in grades.

6.6 Feature 6 - Russell Group

The results observed in figure 11 could be a result of bias amongst researchers, however it could also just represent the quality of work Russell groups produce. Since we know the REF is not blind marked the possibility of bias is ever present, and unless the next REF is conducted without the panel members knowing which university submits which paper, there is no way to know if the grades given are influenced by the Russell Group standing or not. What is clear is that there is a obvious difference between the results of the Russell and non-Russell groups.

6.7 Feature 7 - Selective Submission of Output

Given more time and uncovering the error in the analysis earlier on in the analysis could lead to more reliable and accurate results. Therefore, no real conclusion can be drawn yet from this feature.

7 Future Research Work

7.1 More on "unusual words"

7.1.1 UOA sampling

In the current study, UOAs were sampled randomly across all main panels in the aim of obtaining an overview of how academic disciplines differ from one another in terms of how much each feature contribute differently to the scores in various UOAs. More UOAs should be looked into in the future to focus on the characteristics of one or each of the main panels. For instance, more UOAs in panel A should be examined to determine whether the negative correlation seen in UOA2 is an anomaly or a representative of the entire panel.

7.1.2 "Powerful words"

In UOAs where positive correlations are found, it is worth extracting the "unusual words" that give rise to high TFIDF values. Analysis could be done in identifying the use of any words that are more likely to result in higher 4* scores – the "powerful words".

7.1.3 Overall vs output scores

Theoretically, stronger correlations should be seen between unusual word count and output scores since additional information is directly associated with output submissions. However, the correlations for output are usually less than those for overall for every sampled UOA. Given that the actual content of additional information has no relation with impact or environmental case studies, which contribute to the overall scores, the only possible relation would be the writing style. Knowing that text mining played a part in assessing impact case studies, one might be able to guess that writing style does have an effect on the scores, and this effect is greater on overall results than output since overall rating is associated with the assessment of more text materials. However, this needs to be proved through looking into more UOAs and performing analysis on impact and environmental case studies.

7.2 More on the Citations

To get a better understanding of the results it would be ideal if citation data is extracted from multiple sources, such as Google Scholar and Scopus. This would provide a richer data set for all UOAs and more reliable results.

7.3 More on the Journal Rankings

The SCImago Journal & Country Ranking List could be used to assess how the journal ranking influences the REF results in more UOAs. By fully utilising this dataset together with a more up-to-date ERA Journal Ranking List could provide more reliable results. Additionally given more time it would be better for the analysis if all the journals in the REF datasets were matched and given a score.

What would also be interesting to explore, would be to sample the REF datasets and form training and validation datasets, and thus build a predictive model to whether journal rankings could possibly be used in the next REF process as indicators. Ideally in this case data from previous processes, such as the RAE could be included.

7.4 More on the Selective Submission of Output

Based on the investigation study conducted by HEFCE, there are quantitative biases by selection of staff processes, but the scope did not cover the qualitative portion of the process, which is related to the ranking outcome. Therefore, it is worth to re-attempt this feature in future research work, especially when there is a possibility of strong correlation between higher submissions of output by HEIs to the four star ranking, and eligible FTE data to the submissions of output.

What could be explored is finding the correct and more accurate variables representation to represent the feature being tested, for example, the differences between the eligible FTE data to the submitted FTE data against the REF output rating for the institutions.

8 Conclusion

Analysis was performed on several features, in an attempt to identify whether they have any effect on the output score produced by the REF.

Allocating the institutions to the Russell group and the non-Russel group, indicated that there is a large variance between the REF outputs between the two. The Russell group universities had a better average of 4* percentage papers produced. This can either suggest that there is institutional bias in the REF process or that simply the Russell Group universities are better.

When citations are considered, the results showed that there were variances in the correlation results between UOAs, with some UOAs having stronger correlation than others. This can suggest that there was a difference in the extent of how the citations were used as an indicator of the academic performance of work in the assessment process.

The strong correlation results in several UOAs between the average income per submission and the REF output, can provide an indication that higher research income institutions produce higher quality work, or that possibly the assessment process was biased. Biased in the sense that work submitted by higher research income institutions was assessed more leniently than the other institutions.

For the journal ranking several UOAs returned very strong correlation results, with UOA 18 having the strongest correlation of R-value 0.95 with the SCImago Journal Ranking list. What the results can imply in these UOAs is that the higher the quality of the paper the higher the journal it is submitted to. Alternatively there is the possibility that the sub-panels responsible for the strongy correlated UOA's were biased, and papers published in higher ranked journals were more leniently assessed in comparison to the papers published in the lower ranked journals.

The counting of unusual words in additional information suggests a possibility, when descriptive texts are part of the requirements of output submission, that a piece of better written additional information might lead to higher scores for the submission, or that a department with diverse research interests (thus less repetitive words) is considered better.

Bibliography

Australian Research Council. Journal list relating to the 2010 excellence in research for australia round. Available at `https://www.righttoknow.org.au/request/journal_list_relating_to_the_201` (12/12/2017), 2014.

Chowdhury et al. Novel methods for assessing urban air quality: Combined air and noise pollution approach. *Journal of Atmospheric Pollution*, 3(1):1–8, 2015. doi: 10.12691/jap-3-1-1.

Alan Dix. Ref2014 citation analysis. Available at `http://alandix.com/ref2014/` (12/12/2017), 2015a.

Alan Dix. Ref redux 3 plain citations. Available at `http://alandix.com/blog/2015/08/20/ref-redux-3-plain-citations/` (12/12/2017), 2015b.

Professor David Eastwood. Future framework for research assessment and funding. Available at `http://webarchive.nationalarchives.gov.uk/20100303171159/http://www.hefce.ac.uk/pubs/circlets/2007/cl06_07/` (12/12/2017), 2007.

Michael Gilleland. Levenshtein distance, in three flavors. Available at `http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm` (12/12/2017), 2006.

HEFCE. The nature, scale and beneficiaries of research impact: An initial analysis of research excellence framework (ref) 2014 impact case studies. Available at `http://www.hefce.ac.uk/pubs/rereports/Year/2015/analysisREFimpact/` (12/12/2017), 2015.

HEFCE. Guide to funding 2017-18. Available at `http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2017/201704/HEFCE_Funding_Guide_2017-18_.pdf` (12/12/2017), 2017.

HESA. Contextual data for the research excellence framework 2014. Available at `https://www.hesa.ac.uk/news/18-12-2014/research-excellence-framework-data` (12/12/2017), 2014.

Lekovec et al. Data mining. In *Mining of Massive Datasets*, pages 7–9. Cambridge University Press, 2014.

Lund Research Ltd. Spearman's rank-order correlation. Available at `https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php` (12/12/2017), 2013.

NIST SEMATECH. What is eda? Available at `http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm` (12/12/2017), 2013.

REF. Assessment framework and guidance on submissions. Available at `http://www.ref.ac.uk/2014/pubs/2011-02/` (12/12/2017), 2011.

REF. Main panel b criteria. Available at `http://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12_2B.pdf` (12/12/2017), 2012a.

REF. Summary of additional information about outputs. Available at `http://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12a.pdf` (12/12/2017), 2012b.

REF. Panel criteria and working methods. Available at `http://www.ref.ac.uk/2014/pubs/2012-01/` (12/12/2017), 2012c.

REF. Research excellence framework. Available at http://www.ref.ac.uk/2014/ (12/12/2017), 2014a.

REF. Units of assessment. Available at http://www.ref.ac.uk/2014/panels/unitsofassessment/ (12/12/2017), 2014b.

REF. Citation data. Available at http://www.ref.ac.uk/2014/about/guidance/citationdata/ (12/12/2017), 2014a.

REF. Download submission data. Available at http://results.ref.ac.uk/DownloadSubmissions/SelectUoa (12/12/2017), 2014b.

Scimago Journal and Country Rank. About us. Available at http://www.scimagojr.com/aboutus.php (12/12/2017), 2017.

Morris Sloman. Computer science and informatics ref analysis. Available at http://alandix.com/docs/ref2014/SP11-REF-analysis.pdf (12/12/2017), 2014.

The Scipy community. Statistical functions (scipy.stats). Available at https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html (12/12/2017), 2014.