Judging a book by its cover: how much of REF 'research quality' is really 'journal prestige'?

David Antony Selby   David Firth

July 2020

Abstract

The Research Excellence Framework (REF) is a periodic UK-wide assessment of the quality of published research in universities. The most recent REF was in 2014, and the next will be in 2021. The published results of REF2014 include a categorical 'quality profile' for each unit of assessment (typically a university department), reporting what percentage of the unit's REF-submitted research outputs were assessed as being at each of four quality levels (labelled 4*, 3*, 2* and 1*). Also in the public domain are the original submissions made to REF2014, which include — for each unit of assessment — publication details of the REF-submitted research outputs.

In this chapter, we address the question: to what extent can a REF quality profile for research outputs be attributed to the journals in which (most of) those outputs were published?

The data are the published submissions and results from REF2014. The main statistical challenge comes from the fact that REF quality profiles are available only at the aggregated level of whole units of assessment: the REF panel's assessment of each individual research output is not made public. Our research question is thus an 'ecological inference' problem, which demands special care in model formulation and methodology. The analysis is based on logit models in which journal-specific parameters are regularized via prior 'pseudo-data'. We develop a lack-of-fit measure for the extent to which REF scores appear to depend on publication venues rather than research quality or institution-level differences. Results are presented for several research fields.

Contents

1   Introduction

The Research Excellence Framework (REF; successor to the Research Assessment Exercise, or RAE) is the method used by UK funding bodies to evaluate the quality of research. The last REF took place in 2014 and the next one is currently scheduled for 2021. Panels of experts rate universities and research institutions in three categories: impact outside academia, research environment and quality of outputs, based on written submissions. In the sciences and some other fields, submissions are more likely to comprise academic journal articles than books or reports (Wilsdon et al. 2015; Marques et al. 2017).

Expert panels can judge a submission to be 'world-leading' (4*) 'internationally excellent' (3*), 'recognised internationally' (2*), 'recognised nationally' (1*) or unclassified. Results published online describe, for each subject area ('unit of assessment') the proportion of each institution's outputs that were assigned to each of these categories.

Though it is publicly known which works were submitted for assessment, the ratings are only published in aggregate, by institution and subject: it is not disclosed which rating was assigned to which paper. Thus, it is not obvious what constitutes a '4* paper' or which authors wrote them. However, rumours have long circulated about lists of '4* journals' that peer review panels might use to help them determine the quality of articles (Oswald 2007).

Given that the purposes of the REF are explicitly '[to] establish reputational yardsticks' (i.e. rank academic departments) and 'to inform the selective allocation of funding for

research' (REF web site 2019), it is not surprising that it has had an effect on institutional behaviour, allegedly increasing the number of staff hired on short-term contracts that coincide with the assessment period (Jump 2013), changing the way departments submit members of staff and publications for evaluation (Marques et al. 2017) and increasing productivity just before the deadline (Groen-Xu et al. 2017).

The popularity of journal-level citation metrics such as the impact factor raises the question: might some expert panels be influenced (consciously or otherwise) by a journal's reputation or citation count when judging an individual paper?

In this chapter, we investigate the extent to which research institutions' REF ratings (for outputs) might be attributed to the journals in which their outputs were published. Paper-level ratings are missing, but the margins—institutions' REF profiles and the numbers of articles they submitted from each journal—are known, so the research question becomes an 'ecological inference' problem. Using both frequentist and Bayesian approaches, we will estimate latent 'quality' scores for journals, and then quantify the variation in REF results that is explained by these scores. We also compare these scores with published journal citation metrics.

Initially, we demonstrate the methodology on the field of economics, a relatively small and well-defined discipline, which mostly publishes its outputs in academic journals and has a well-established 'Top Five' journals that act as a baseline. Results will then be compared with several other, larger academic fields whose REF-submitted outputs are also mostly in the form of journal articles.

2   Background

2.1   Modelling research assessments

Koya and Chowdhury (2017) suggested that there is, for some subject areas, a correlation between journal rankings and REF performance. Their approach computed a 'monetary value' (funding allocation) for each research output as rated in the REF, using a similar method to that described in an earlier blog post by Reed and Kerridge (2017).

Let F be the total amount of funding awarded to an institution based on the REF, let n_3 and n_4 be the number of 3* and 4* outputs and let x_3 and x_4 be the respective monetary value of an output with each rating. According to the then Higher Education Funding Council for England (HEFCE), a 4* output is worth four times as much as a 3* output (Else 2015), so $x_4 = 4x_3$. The numbers of outputs are known. Then $F = n_3 x_3 + n_4 x_4 = n_3 x_3 + 4n_4 x_3$, from which we obtain

$$x_3 = \frac{F}{n_3 + 4n_4},$$

for a given institution and subject area.

Consider the example of general engineering at the University of Cambridge. It was awarded $F = £5,328,295$ in 2015–16 as a result of its outputs submitted to the 2014 REF[1]. Of the submitted outputs, 37.4% were rated 4* and 55.8% at 3*, for 177.2 full-time equivalent staff. Each staff member was allowed up to four submissions, and

---

[1] according to the HEFCE 2015–16 funding allocation tables for research

funding was allocated assuming that staff submitted this maximum, even if they did not. So the theoretical (not actual) number of outputs was $4 \times 177.2 = 708.8$. Thus $n_3 = 708.8 \times 55.8\% = 395.5104$ and $n_4 = 708.8 \times 37.4\% = 265.0912$, from which we obtain $x_3 = \pounds 5328295/1455.875 = \pounds 3{,}659.86$ and $x_4 = \pounds 14{,}639.43$.

From here, Koya and Chowdhury (2017) studied the relationship between the distribution of an institution's REF scores with the venues in which the outputs were published. REF results do not reveal which article/submission received each rating; the data are only published in aggregate. Koya and Chowdhury (2017) "identified how many of the submitted articles were in top quartile [sic] journals" based on impact factors published in the 2013 edition of Thomson Reuters'[2] *Journal Citation Reports*, and compared this proportion with the percentages of articles awarded 4* and 3* ratings[3]. Positive correlations, where found, were weak and only present in some subject areas. Surprisingly, Koya and Chowdhury (2017) did not directly compare the computed 'monetary value' of research outputs with the corresponding bibliometric indicators for each institution.

Wilsdon et al. (2015 Section 9.1) commissioned HEFCE to perform a more detailed study of the relationship between bibliometric indicators and REF scores, with privileged access to ratings at the level of the individual outputs. That analysis found low (< 0.5) positive correlations between citation metrics and 4* outputs, but with stronger relationships for some fields such as medicine, biology, chemistry, physics and economics. The strongest predictors were full-text clicks (on Scopus), number of authors, citation count (according to Google Scholar), SJR (a Scopus-published journal metric based on PageRank score), source-normalized impact per paper (a another Scopus metric, similar to a weighted impact factor), tweets, and downloads from the web site *Science Direct*.

In the field of Art and Design, Mansfield (2016) ranked journals according to their popularity in REF submissions, but did not attempt to infer star ratings for the publications.

Stockhammer, Dammerer, and Kapur (2017) investigated the 'grade point average'—the average star rating—of each institution in the 2014 REF, modelling it as a linear function of either the SCImago Journal Rank citation score (SJR) or of journal ratings assigned by the Chartered Association of Business Schools. That analysis, applied to the fields of economics, found a coefficient of determination of up to $R^2 = 89\%$, with the 2014 log-SJR score having a statistically significant effect under their model.

Italy's research assessment exercise, the *Valutazione Triennale della Ricerca* (triennial research evaluation) began in 2003 with a similar remit to the UK's RAE/REF and was initially 'fully based on peer review'. Franceschet and Costantini (2011) found positive correlations between the peer review assessments and citation metrics, but the strength of the correlation varied between fields, and was particularly weak for journal impact factor. The then-recently proposed h-index (Hirsch 2005) provided a better approximation.

From 2004, the *Valutazione della Qualità della Ricerca* (research quality evaluation; VQR) introduced a 'dual system of evaluation' using a combination of peer review and bibliometrics. The Italian National Agency for the Evaluation of the University and Research Systems (ANVUR) compared the results from each approach and found a 'more than adequate concordance', apparently justifying the decision to use bibliometrics.

---

[2] Now operated by Clarivate Analytics

[3] As pointed out by Hill (2017), this means only a subset of journal articles are being compared with the full range of outputs submitted to the REF—not a like-for-like comparison.

However, this conclusion has been strongly challenged by Baccini and Nicolao (2016), who insist the methodology was 'fatally flawed' and undermines the results for the field of economics and statistics in particular.

Between the RAE2008 and REF2014, Mryglod et al. (2015a) compared departmental h-indices with performance in the RAE, finding a correlation between h-index and certain grade-point averages of RAE results. Using this relationship, they made predictions for the upcoming REF2014 for several institutions and fields. However, in a follow-up after the REF2014 results were published, Mryglod et al. (2015b) reported the predictions "failed to anticipate with any accuracy either overall REF outcomes or movements of individual institutions in the rankings relative to their positions in the previous Research Assessment Exercise". Thus care should be taken in trying to predict one research assessment from the results of another that took place years before.

Our research is not the first attempt at producing a journal ranking from REF results for economics, let alone for academic fields in general. Hole (2017) used a greedy iterative algorithm to assign star ratings to individual papers (assuming these were entirely dependent on the journals in they appeared) minimizing the squared error in predicted ratings for institutions,

$$Q = \sum_{i=1}^{I} \sum_{r=1}^{4} N_i (p_{ir} - \hat{p}_{ir})^2,$$

where N_i is the number of submissions from each institution, p_{ir} is the observed proportion of r-star submissions from that institution and $\hat{p}_{ir}$ is the predicted proportion, based on the imputed ratings. The algorithm first assigns an arbitrary star-rating r to each journal, calculates the objective function Q, then iterates over the list of journals, changing each journal's star rating to that which would decrease Q the most, terminating when a full pass over all journals produces a change in Q smaller than a pre-specified threshold. The analysis of Hole (2017) excluded journals with fewer than five submissions in the REF. Since this would result in the number of submissions no longer adding up to the total number of ratings, they assigned arbitrary ranks to these left-out journals. The results had a correlation of approximately $\rho = 0.5$ with previously-published economics journal rankings.

More recently, Balbuena (2018) adopted a machine learning approach, using a Bayesian additive regression tree model to predict grade point average from a range of institutional covariates, including the number of attributed documents indexed in the *Web of Science* and the proportional intake of students from state schools. However this analysis focussed more on possible inequities in distribution of funding, rather than investigating an explicit journal identity effect.

Yan (2017) used a Metropolis-within-Gibbs sampling regime to fit an ordinal response model to Economics & Econometrics outputs for REF2014. Whilst broadly similar to our approach, their framework is based on a proportional-odds cumulative probit model, which assumes a common set of thresholds between star ratings for all journals. In other words, the increase in difficulty of attaining a 4* rating over a 3* one is the same for every journal. Our analysis fits models for several different subjects and finds that this assumption does not hold, even for Economics & Econometrics.

2.2 Ecological inference

The previous section provided examples of limited analyses comparing some citation indices and other journal- or institution-level covariates with REF results, and of approaches to produce journal rankings from institutional scores. However, to our knowledge, no principled *statistical* analyses (that is, with quantified uncertainty) of the relationship between journal identities and UK research assessment have been published. Moreover, modelling REF ratings as a function of citation metrics is problematic; criticisms abound of certain indicators under inspection—impact factor and its variants, as well as 'alternative metrics' such as tweets and download counts (e.g. Colquhoun and Plested 2014; MacRoberts and MacRoberts 2017). Instead of using a flawed and imprecise proxy such as a citation metric to analyse the relationship between publications and research assessment, one might consider modelling published REF results against the actual journal identities instead. The problem with this approach is that HEFCE (or since April 2018 its successor, Research England) will never publish the individual ratings given to submissions in the REF; indeed they were destroyed upon completion of the research assessment (REF 2015).

We are therefore left in a quandary: how do we model the effect of journals on star ratings, if we don't know which journal articles received which ratings? What if we wanted to try to infer these publication-level ratings? This would allow us to construct a ranking of *journals*, not just institutions, from the REF results, similar to the work by Hole (2017). Moreover we might attempt to answer the question: is an institution's REF rating simply a function of the journals in which it published? Were that to be the case, it would suggest that the REF is directly measuring prestige rather than quality—a common criticism of citation indices. On the other hand, if an institution's REF score is *more* than the sum of its output journals then it might be used as evidence against using journal-level metrics to assess research quality.

However, as already mentioned, the REF scores are aggregated by institution, not by journal. Journal-level scores must therefore be imputed rather than observed. Such a task—inferring individual-level properties from aggregate data—is known as *ecological inference* or *ecological regression* (Goodman 1953), typically applied to estimate voting behaviour in a secret ballot, when exit polls are infeasible or unreliable. Examples include modelling voter transitions between parties (Brown and Payne 1986) and estimating who voted for the Nazi Party in Weimar Germany (Rosen et al. 2001). A detailed review of the topic is provided by Wakefield (2005). The following is a brief summary.

Sociologists and political scientists often use the term 'ecological inference' to refer to inference on voting populations—for example, voter transitions between elections in a two-party system, or turnout for two demographics.

Consider an election where, to comply with civil rights legislation, authorities in the US desire to compare turnout amongst black and white voters. Suppose for a given electoral district (constituency) i, the demographic makeup is known with proportion X_i of the population black and the remainder white. Overall voter turnout, T_i, is observed for a particular election but the ballot is secret, so turnout among blacks and whites, respectively β_i^b and β_i^w, are unknown. These data yield the following 2×2 table of proportions.

Table 1: Observed and unobserved proportions for a two-dimensional voter turnout model

	Vote	Not vote	
Black	β_i^b	$1 - \beta_i^b$	X_i
White	β_i^w	$1 - \beta_i^w$	$1 - X_i$
	T_i	$1 - T_i$	

At first glance, it may not appear that one can really glean any information about individuals only from the margins. Via the *method of bounds* however, we can obtain deterministic bounds on (at least one of) the parameters: black turnout β_i^b must be greater than $\frac{T_i - (1 - X_i)}{X_i}$ and smaller than $\frac{T_i}{X_i}$, whilst white turnout β_i^w must be between $\frac{T_i - X_i}{1 - X_i}$ and $\frac{T_i}{1 - X_i}$, to ensure they are valid proportions that add up to one (Duncan and Davis 1953). For example, if a district's population were 70% black and overall turnout were 40%, then black turnout must be in the range (14%, 57%), but white turnout could still be anywhere in (0, 100%).

Unlike this limited deterministic approach to the ecological inference problem, *ecological regression* or *Goodman regression* (1953, 1959) is one of the first *statistical* solutions. Using the identity

$$T_i = X_i \beta_i^b + (1 - X_i)\beta_i^w,$$

one may construct a simple linear regression model of turnout on racial proportions:

$$\mathbb{E}[T_i | X_i] = \alpha + \beta X_i,$$

where $\alpha = \beta_i^w$ and $\beta = \beta_i^b - \beta_w$. A notable criticism is that these voting propensities are assumed to be homogeneous over districts, regardless of the racial mix in each area. Moreover, least squares does not constrain these parameters to lie within the bounds described above, or even between zero and one (Wakefield 2005).

Brown and Payne (1986) proposed modelling voter turnout using a convolution of Dirichlet–multinomial distributions, with the response approximated by a multivariate normal distribution. However, this model is sensitive to the choice of prior (Wakefield 2005). More recently, King (2013) combined the method of bounds with a pseudo-'likelihood' function—equivalent to an asymptotic form of the binomial distribution—and imposed a truncated bivariate normal distribution to tighten the bounds. This approach describes itself as 'a solution to the ecological inference problem', however this claim was criticized as overly optimistic (Cho 1998; Freedman et al. 1999).

Since then, King, Rosen, and Tanner (1999) proposed a different solution in the form of an hierarchical Dirichlet–multinomial model where the unobserved probabilities (the voter turnouts by ethnicity) are beta-distributed latent random variables. For a constituency/district i with total voting-age population N_i and observed voter turnout count $Y_i = N_i T_i$ the hierarchical model takes the form

$$Y_i \sim \text{Binomial}(N_i, T_i),$$

where the marginal probability of voter turnout in constituency i is

$$T_i = \sum_{j=1}^{J} x_{ij}\beta_i^j,$$

with x_{ij} denoting the proportion of people of ethnicity j in constituency i, and where the prior constituency-level probabilities of voter turnout, by ethnic group, are

$$\beta_i^j | a_j, b_j \overset{\text{iid}_i}{\sim} \text{Beta}(a_j, b_j)$$

with hyper priors

$$a_j, b_j \overset{\text{iid}_i}{\sim} \text{Exp}(\lambda)$$

and the default hyper-parameter setting $\lambda = 0.5$.

In the two-dimensional (black–white voter turnout) case described above, $J = 2$ and the middle level is

$$T_i = X_i \beta_i^b + (1 - X_i) \beta_i^w.$$

The model generalizes to $J > 2$ ethnic groups (or journals, in our case) and can be further extended to multiple outcomes (beyond binary 'vote or not') by replacing the beta–binomial distribution pair with a Dirichlet–multinomial (Rosen et al. 2001).

In our view, the top level of King's hierarchical model possibly adds an unnecessary random component, for the total turnout should simply be a deterministic, weighted sum of the turnout among each ethnic group. The election result is not an approximation of the counted votes: it *is* the counted votes. All that is necessary is for the βs to be constrained so that the sum over ethnic groups of voters adds up to the observed overall turnout. This is perhaps more easily said than done, however.

More recent approaches to ecological inference make use of *distribution regression*, by treating the makeup of each electoral district as a probability distribution (Flaxman, Wang, and Smola 2015; Szabó et al. 2014). The basic idea is to project the distributions into a feature space, then fit a regularized regression model, such as kernel ridge regression, using this embedding. Flaxman, Wang, and Smola (2015) used this technique to combine demographic and spatial information and infer the groups who voted for Barack Obama in the 2012 US presidential elections, and again for the 2016 elections (Flaxman et al. 2016).

Rosenman and Viswanathan (2018) derived a 'heteroscedastic Gaussian' approximation to the Poisson binomial log-likelihood, via a central limit theorem, and later applied this to a large voter transition model (Rosenman 2019), which they term a *Poisson binomial generalized linear model*. Unlike the presidential election studies by Flaxman et al., which used Bayesian techniques, the Poisson binomial GLM is 'purely frequentist'. This offers the advantages of 'simpler fitting procedures, straightforward estimation of individual-level probabilities, and greater model interpretability' at the expense of reduced flexibility (Rosenman 2019).

3 Model

The REF ratings received by institutions on their outputs could be assumed to be drawn from a Poisson binomial distribution (Poisson 1837), which describes the probability of obtaining K successes in n independent but non-identically distributed Bernoulli trials, with probability mass function

$$\Pr(K = k) = \sum_{A \in F_k} \prod_{i \in A} \pi_i \prod_{j \in A^c} (1 - \pi_j),$$

8

where F_k is the set of all subsets of k integers that can be selected from $\{1, 2, \ldots, n\}$, for n the number of Bernoulli trials and π_i the success probability of the ith trial (Wang 1993). For our purposes, π_i represents the probability that an output i was awarded a particular rating, for example 4*.

The Poisson binomial is a special case of the aggregated compound multinomial model used by Brown and Payne (1986). That paper describes a Dirichlet-multinomial ('compound multinomial') model for the unobserved numbers of voters who switched between each of the major parties from one election to another. In their notation, each election featured the same set of political parties. The model estimates the probability, p_{ijk}, that a voter for party i in constituency k becomes a voter for party j.

Our analogy is rather different: there are J parties at the first election, representing the journals in which the articles are published, but only two parties at the next election: '4*' and 'not 4*'. Voters are articles, and constituencies are academic institutions. We model the probability that an article published in a particular journal is awarded a 4* rating, or not.

Let x_{ij} denote the (known) number of articles published by institution i in journal j. Let y_{ij} denote the (unknown) number of such articles that attained a 4* rating in the REF, with $0 \leq y_{ij} \leq x_{ij}$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let $y_i = \sum_j y_{ij}$ denote the published number of 4* ratings awarded to each institution and let $x_j = \sum_i x_{ij}$ denote the total number of articles submitted from each journal. Then the marginal totals, $\mathbf{Y} = (Y_1, \ldots, Y_I)$, are aggregated compound multinomial (Dirichlet-Poisson-binomial) random variables with expectation

$$\mathbf{E}(\mathbf{Y}) = \mathbf{P}^T \mathbf{x}$$

and covariance

$$\text{cov}(\mathbf{Y}) = \text{diag}(\mathbf{P}^T \mathbf{w}) - \mathbf{P}^T \text{diag}(\mathbf{w}) \mathbf{P},$$

where $\mathbf{P}$ is the J-vector of journal success probabilities[4], $\mathbf{x} = (x_1, \ldots, x_J)$ is a vector of the number of articles in each journal and $\mathbf{w}$ is a J-vector of weights $w_j = x_j(x_j + \alpha_j)/(1 + \alpha_j)$. Brown and Payne (1986) note that 'election data involve more variability than a multinomial would suggest' and add the α vector of J precision/dispersion parameters to account for this.

The variance of a Poisson binomial-distributed random variable is

$$\text{Var}(Y_i) = \sum_j (1 - \pi_j)\pi_j = \sum_j (\pi_j - \pi_j^2),$$

which differs from the variance of the aggregate compound multinomial model only by the w_j term. We notice that as α_j grows large, then (dropping the subscripts for the moment)

$$\lim_{\alpha \to \infty} w = \lim_{\alpha \to \infty} x \left(\frac{x}{1 + \alpha} + \frac{1}{\frac{1}{\alpha} + 1} \right) = x,$$

and since $\alpha = \infty$ corresponds to the (non-compound) aggregate multinomial distribution (Brown and Payne 1986), we can see the Poisson binomial and aggregated multinomial models are equivalent.

---

[4]More generally, $\mathbf{P}$ is a *matrix* of multinomial probabilities

If we consider every paper grading to constitute an independent trial, with success probability dependent on the journal in which it is published (but not the institution or any paper-level characteristic), then for each institution i, the number of 4* ratings received is distributed

$$Y_i \sim \text{Poisson-Binomial}(\boldsymbol{\pi}), \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is the vector of journal probabilities[5]. Here the success probabilities are not identical for every publication, but they do coincide wherever two submissions are published in the same journal; if an institution submits two or more articles from the same journal then each of these articles is regarded as a separate independent trial. Of course, independence might be an heroic assumption here, but *ideally* one would hope the REF panels consider each article on its own merits rather than ranking them against one another.

We can fit the model twice: firstly with 'success' defined as 4* ratings, and secondly with success defined as 3* *or* 4* ratings, i.e. 3* or better. As the star ratings are ordinal responses (4* is better than 3*, which is better than 2* and so on), it seems reasonable to assume cumulative odds, and infer the probability of 3* from the estimated probabilities of 4* and of 3*-or-better. Thus a journal's probability of obtaining a 3* rating is assumed to be

$$\pi_j^3 = \pi_j^{34} - \pi_j^4,$$

where we introduce superscript notation: π_j^3, π_j^4 and π_j^{34} respectively denote journal j's probability of accruing 3*, 4* and $\geq 3*$ ratings.

This separate fitting of the model for the 4* and 3*-or-4* is a key difference from the work of Yan (2017), which used a cumulative odds model with common thresholds for every journal. That is, under Yan's model, $\text{probit}(\pi_j^{34}) - \text{probit}(\pi_j^4) = c$, a constant offset that does not depend on the journal j. Our approach replaces c with c_j, a difference that can might be distinct for every journal.

On top of the likelihood (1) we impose a prior on the journal success parameters,

$$\pi_j \sim \text{Beta}\big(\gamma\mu, \gamma(1 - \mu)\big) \tag{2}$$

for each journal j, such that the mean probability of success is μ, and γ is a regularizing concentration parameter. On top of these we impose hyper-priors

$$\begin{aligned} \mu &\sim \text{Uniform}(0, 1) \\ \gamma &\sim \text{Gamma}(\tfrac{1}{10}, \tfrac{1}{20}), \end{aligned} \tag{3}$$

where the given hyper-parameters of the gamma distribution are the shape and rate, respectively—corresponding to a mean of 2 and variance 40. In principle one could set these manually, for instance setting μ equal to the empirical mean institutional profile, but we shall try to learn them from the data.

There are more differences between institutions than just the journals in which they publish, so to check for aggregation bias, we extend the model (1) such that an article success depends not just on the journal parameter, but on an institutional covariate linked to the REF Environment profiles. In this way we might hope to detect any institutions that

---

[5]In practice, each element π_j is repeated x_{ij} times, representing repeated trials for the number of articles in journal j that were submitted by institution i to the REF.

perform better or worse in output scores due to the quality of their research environment rather than on the journals in which they publish. Thus the success probability of an article from institution i in journal j is

$$\text{log-odds}(4*|i,j) = \text{logit}\,\pi_j + \alpha\,\text{envir}_i \tag{4}$$

where α is a parameter to be estimated and envir_i is the proportion of 'Environment' in institution i rated 4* (centred by subtracting the mean). If the α is near zero, then we might conclude that output profiles depend more on the journals than on the unique characteristics of each research institution.

4   Methods

We will employ two different methods to estimate the parameters of the model. Firstly, a Bayesian Monte Carlo method, and secondly a maximum likelihood approach using an expectation–maximization algorithm. This section describes the details behind each technique.

4.1   Hamiltonian Monte Carlo

Owing to the limited computational power available at the time, Brown and Payne (1986) employed a normal approximation to the Poisson binomial model to estimate the unknown coefficients. The Poisson binomial distribution can also be approximated by a Poisson distribution, though the performance of this approximation is poor when the number of trials is large (Hong 2013).

Advances in computation capacity allow us to consider a couple of different approaches of fitting a Poisson binomial model. The first would be to employ the probabilistic programming language Stan (Carpenter et al. 2017; Stan Development Team 2018) to sample from the posterior distribution via Hamiltonian Monte Carlo (also known as hybrid Monte Carlo or HMC; Duane et al. 1987). Because enumerating all possible sets of integers F_k is computationally infeasible, instead one can program a routine to compute the mass by enumerating with a recursive formula (Shah 1973)

$$\Pr(K=k) = \begin{cases} \prod_{j=1}^{n}(1-p_j) & k=0 \\ \frac{1}{k}\sum_{j=1}^{k}(-1)^{j-1}\Pr(K=k-j)\sum_{l=1}^{n}\left(\frac{p_l}{1-p_l}\right)^j & k>0, \end{cases}$$

however this may not be numerically stable for large n (Hong 2013) unless computed on the logarithmic scale.

Using such a dynamic programming algorithm on the logarithmic scale, we fit the model in Stan and report the results in Section 6. As a robustness check, we also consider a maximum likelihood approach, described in the next subsection.

4.2   Expectation–maximization algorithm

The expectation–maximization (EM) algorithm can provide alternative maximum likelihood point estimates, albeit without any covariance estimate as a measure of uncertainty.

The EM algorithm makes use of the *extended multivariate hypergeometric distribution*. Recall the more familiar hypergeometric distribution describes the probability that, given an urn of N balls, K of them white and $N - K$ black, that if we draw n balls at random then k of them are white. The *extended* hypergeometric distribution, also known as Fisher's *noncentral* hypergeometric distribution, extends this scenario to non-uniform sampling—where the white balls are more likely to be drawn than black ones because of differences in size or weight.

A multivariate hypergeometric distribution generalizes to a situation where there are more than two colours of balls and describes the probability of picking a particular mixture of colours. Hence, an extended multivariate hypergeometric distribution describes the probability of picking a certain mix from an urn of balls whose weights are not all equal (McCullagh and Nelder 1989, 260–61). For dimension d different colours, the probability mass function for drawing a mixture $\mathbf{x} = (x_1, \ldots, x_d)$ of n balls is

$$f(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \frac{1}{P_0} \prod_{i=1}^{d} \binom{m_i}{x_i} \omega_i^{x_i}$$

where $\mathbf{m} = (m_1, \ldots, m_d)$ is the number of each colour of balls in the urn and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ are their respective weights. The denominator P_0 is

$$P_0 = \sum_{\mathbf{y} \in \mathcal{S}} \prod_{i=1}^{d} \binom{m_i}{y_i} \omega_i^{y_i}$$

with $\mathcal{S}$ denoting the set of all possible non-negative integer d-vectors $\mathbf{y} = (y_1, \ldots, y_n)$ where $\sum_{i=1}^{d} y_i = n$.

If the article-level ratings were known, we could fit a Rasch-like generalized linear model of the following form to estimate journal effects:

$$\text{logit } \mathbb{E}[\Pr(4^*|i,j)] = \mu + \alpha z_i + \beta_j, \tag{5}$$

for a paper by institution i in journal j, where parameter μ, analogously to (1), acts as a 'grand mean' intercept term, here in the logit space, and where z_i is an indicator variable for a 'pseudo-institution'. The latter submits an equal number of 4* and not-4* papers from every journal, acting as a regularizing prior on the strength of the journal effects to avoid overfitting. The more pseudo-papers submitted, the stronger the effect of the regularization. The maximum likelihood estimate for such an artificially augmented dataset is equivalent to the posterior mode with a conjugate Bayes model (Jannarone, Yu, and Laughlin 1990). The optimum strength of regularization (i.e. the number of pseudo-articles augmenting the data) is determined via cross-validation, described at the end of this section.

We adopt an expectation–maximization procedure as follows.

1. Initialize the weights of the noncentral multivariate hypergeometric distribution. That is, randomly generate a probability for each journal–institution that corresponding outputs will be 4*-rated in the REF. (We use a logit-normal distribution for this.)

2. Compute the (approximate) expectation of the noncentral multivariate hypergeometric distribution with these odds (for this, we use R package `BiasedUrn` by Fog 2015). This vector forms an imputation of the latent individual-level ratings.

3. Fit the model described in Equation (5). Extract coefficients from this model to get new odds for the noncentral multivariate hypergeometric distribution.

4. Repeat steps 2–3 until convergence.

To obtain new odds form the logistic regression model for the noncentral hypergeometric distribution, we simply use the relation

$$\text{odds}(4^*|j) = \exp(\hat{\mu} + \hat{\beta}_j)$$

for all journals j, where $\hat{\mu}$ and $\hat{\beta}_j$ are the estimated parameters from the previous EM step.

Our chosen prior for this model is essentially uninformative on the expected journal ranking. In principle, one could attempt to elicit distributions for the relative strengths of the journals, or (by asking someone who might have served on REF/RAE expert panels in the past) the probability that papers in a certain journal might accrue 4* ratings. However, such an approach is not very scalable to large numbers of journals or fields, so we do not adopt it here.

The cross-validation procedure works as follows.

1. Randomly divide the institutions into (say) 10 groups.
2. For each group:
   a. Run the above expectation–maximization algorithm on data from the other 9 groups.
   b. Use the estimated journal parameters to predict the institutional results for the held-out group.
   c. Compute the index of dissimilarity between the predicted and actual institutional results.
3. Repeat steps 1–2 for different levels of regularization.

We seek the parameter that minimizes the index of dissimilarity (described in the next section) between the predicted institutional scores and the actual scores of the held-out institutions.

4.3   Diagnostics and summary statistics

To obtain a 'prediction' or fitted value from the Poisson binomial model, we take the posterior median of the journal probability estimates p_j and take them to be the proportion of the time that articles in those journals were awarded 4*.

That is, we compute

$$\hat{y}_i^4 = \sum_{j=1}^{J} n_{ij} \hat{\pi}_j^4$$

from the model fitted to 4* outputs and

$$\hat{y}_i^{34} = \sum_{j=1}^{J} n_{ij} \hat{\pi}_j^{34}$$

from the same model fitted to a dataset of 3* *or* 4* outputs, where y_i^{34} denotes the number of an institution's outputs rated 3* or better (i.e. 3* *or* 4*), π_j^{34} represents the probability

that articles in journal j are awarded 3* or better, and n_{ij} denotes the number of articles from institution i in journal j. Hence we can compute the predicted number of 3* outputs,

$$\hat{y}_i^3 = \hat{y}_i^{34} - \hat{y}_i^4,$$

for each institution $i = 1, \ldots, I$.

Recall that our main aim is to answer the question: to what extent are REF output profiles a function of journal identities? In other words: given the journals in which an institution published its submissions, can we predict that institution's REF score?

To determine the quality of fit of the Poisson binomial model we adopt the index of dissimilarity (Duncan and Duncan 1955; Kuha and Firth 2011), which here represents the proportion of an institution's articles predicted a different rating to that observed in the REF. It is computed using the formula

$$\Delta = \frac{1}{2N} \sum_i \left(|y_i^4 - \hat{y}_i^4| + |y_i^3 - \hat{y}_i^3| + |y_i^4 + y_i^3 - \hat{y}_i^4 - \hat{y}_i^3| \right),$$

where $N = \sum_i \sum_j n_{ij}$, the total number of submitted outputs.

From the index of dissimilarity we propose another metric, the *redistribution of monetary reward*, based on the notion that a 4* output is worth four times as much in research funding as a 3* output, and outputs rated 2* or lower accrue no direct funding at all (see e.g. Koya and Chowdhury 2017). This metric describes the fraction of total monetary reward that would move between institutions if the estimated REF profiles $(\hat{y}_i^4, \hat{y}_i^3)$ were used instead of the observed profiles (y_i^4, y_i^3), and is measured by

$$\Delta_{\pounds} = \frac{\frac{1}{2} \sum_i m_i |r_4(p_i^4 - \hat{p}_i^4) + r_3(p_i^3 - \hat{p}_i^3)|}{\sum_i m_i (r_4 p_i^4 + r_3 p_i^3)},$$

where m_i is the number of full-time equivalent (FTE) staff submitted by institution i in the unit of assessment, $\hat{p}_i^4 = \hat{y}_i^4 / \sum_j n_{ij}$, $\hat{p}_i^3 = \hat{y}_i^3 / \sum_j n_{ij}$ and r_4 and r_3 are the respective monetary reward per FTE for the 4* and 3* components of output profiles, in arbitrary units with $r_4 = 4r_3$. (Implicitly, terms for 2*, 1* and unclassified outputs can appear in the above formula, but we take $r_2 = r_1 = r_u = 0$.)

The monetary index might even be combined with calculations of the kind by Koya and Chowdhury (2017) to compute an absolute sterling figure for the amount of funding that would move institutions in a switch from the actual REF profiles to those estimated our model.

5 Data

5.1 Units of assessment

To demonstrate the method, we will first consider the 'Economics and Econometrics' unit of assessment. We choose this particular subject area because it is small, relatively self-contained, and a high proportion of output submissions (92%) are in the form of journal articles (rather than books, conference proceedings or other works). One might expect (this being a statistics PhD thesis) to look at statistical science submissions first,

however these fall under the umbrella of Mathematical Sciences—along with research in probability, pure and applied mathematics and mathematical physics—which is a larger and more hetereogeneous field.

From Table 2 it is easy to see that the hard sciences (REF panels A and B) mostly submitted outputs in the form of journal articles; the arts and humanities (panel D) used other formats, and social sciences (panel C) were somewhere in between. A notable exception to this rule is the field of Computer Science and Informatics, where the role of academic journals is often supplanted by conference proceedings.

After the initial analysis of the Economics and Econometrics sub-panel, we also examine three other REF subpanels, all from the Physical Sciences main panel, to see how they compare. The arts (main panel D) publish too few of their outputs as journal articles for this model to be of practical relevance.

5.2 Wrangling REF2014 data

In the published REF2014 submissions data, outputs are explicitly categorised into types, such as journal articles, book chapters, conference proceedings, working papers and so on. However, there is no sure-fire way to group together articles published in the same journal or book, as the titles are unstandardised, ISSNs, if provided, can vary between print and online editions and DOIs, where present, can be difficult to parse. Labour-intensive manual tagging of the data has rather little appeal, not least because it is error-prone and does not scale well to larger future data sets. But there is a network science solution to the problem.

We coerced the output submissions data into a long-format table comprising just a journal identifier—the ISSN, ISBN, DOI or standardised journal title (coerced to lower case, with punctuation, diacritics, spaces and leading "the"s removed)—and a unique identifier for each output, then constructed an undirected bipartite graph between the journal identifiers and individual output identifiers. Each connected component in this graph represents a unique journal, containing outputs with a common journal title, DOI, ISSN and/or ISBN. Each is assigned a unique journal ID, as well as a human-readable title, the latter sampled from one of the journal title variants found in the component.

Unfortunately, this methodology on the published REF data alone assumes integrity of the published data, which was later found to be lacking. Some administrators entered article metadata by hand, rather than retrieving it programmatically via CrossRef, as perhaps they should have done. This inevitably introduced human error; for example one entry that should have been from the *Annals of Mathematics* had the correct DOI, article and journal title, but the ISSN was that for the separate *Advances in Mathematics* journal, which causes the above mini-algorithm to merge the works in *Annals* and *Advances* as if they came from the same journal. In turn, the *Advances in Mathematics* journal was grouped with *Advances in Applied Mathematics* due to similarly careless data entry. Further issues were caused by journal titles that were ambiguous if not completely erroneous, for example various articles published in *Physical Review Letters*, *Physical Review A*, *Physical Review B* and so on all being given the unhelpful abbreviation *PHYS REV*.

Evidently, the metadata in the published REF outputs data set cannot be trusted, except possibly the DOIs. To remedy this, we used the R package **rcrossref** (Chamberlain et al. 2019) to access CrossRef application programming interface (API), allowing retrieval
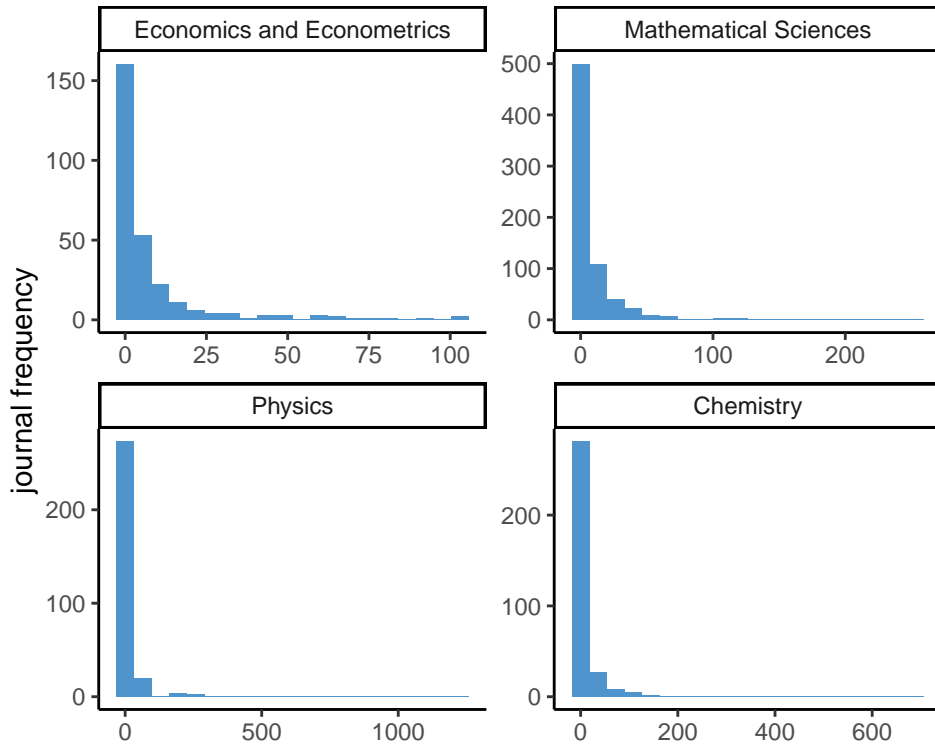
(a) number of submitted articles
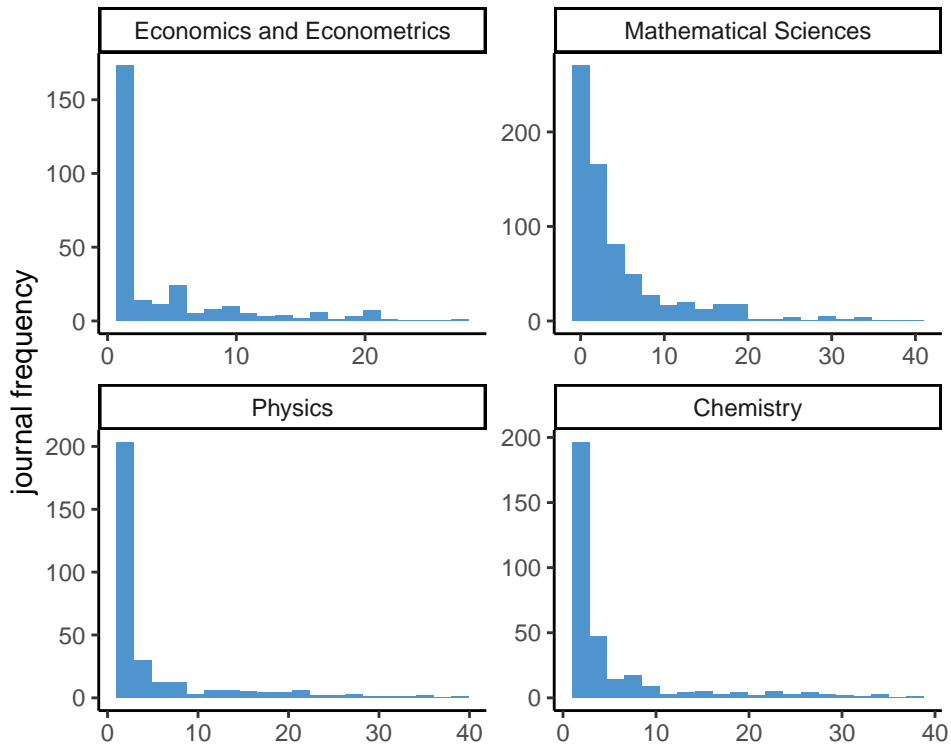
(b) number of submitting institutions

Figure 1: Distribution of journal articles across journals and institutions, by unit of assessment. Some journals are much more popular than others, and not all institutions publish in the same journals

of metadata associated with the 25,000 unique DOIs for the Economics & Econometrics, Mathematical Sciences, Chemistry and Physics submissions. All except 22 returned results. Of these few 'invalid' DOIs, manual inspection showed the same broken DOIs to be published on publishers' own web pages (and this was reported to CrossRef) so these were not a problem with the REF data itself. For the remaining (vast majority) of DOIs, the CrossREF API returned the titles of the articles and the names and ISSNs of the containing periodicals.

A small amount of data wrangling remained, however. Though no single DOI yielded multiple entries in the CrossREF database, our mini clustering algorithm was still required to merge journals which have multiple titles appearing in CrossRef, for example *The Review of Economic Studies* and *Review of Economic Studies*. These were able to be clustered by shared ISSNs (and we assume that CrossRef, at least, gets these correct).

This approach can easily be applied to every field with no manual or *ad hoc* data processing necessary (except those articles with missing or invalid DOIs). The distribution of outputs to journals and to institutions is illustrated in Figure 1, where we can see that it is quite skewed. An uneven spread of journals between institutions is desirable for an ecological inference model; if every institution published in the same profile of journals then it would be impossible to learn any journal-level effect.

Nevertheless, estimating several hundred journal parameters from just a few dozen institution-level observations is particularly ambitious, especially when it is evident that many journals accounted for very few submitted outputs.

Ordinarily in high-dimensional data analysis, one can apply some level of regularization to the model, the exact level of regularization to be determined by, say, empirical Bayes estimation. However, standard techniques of 'soft' regularization do not seem to work very well for aggregated data like those found in our ecological inference problem. Instead we adopt a fairly pragmatic approach: any journal containing fewer than some threshold number of articles is aggregated into a single *super-journal* entitled 'Other journals'. We choose the threshold such that *most* (i.e. $\geq 50\%$) of the articles in the data fall into a named journal rather than an anonymous 'other' journal, while hopefully also keeping the number of parameters low enough to be practical for reporting and visualization. Conference proceedings and other non-journal outputs cannot be ignored, as the Poisson binomial model requires we account for all submitted outputs, so these publications are aggregated into their own respective 'other' categories.

We apply our methodology to the Economics and Econometrics sub-panel as well as three other fields: Mathematical Sciences, Physics and Chemistry, representing three units of assessment from REF2014 main panel B. Biological Sciences (main panel A) was also considered, but modelling this field proved too computationally intensive, possibly due to the large number of submitted outputs (8,608) and institutions (44) or the distributions thereof. (This unit of assessment could still be analysed in future with a more efficient model fitting implementation.)

Compared to Economics and Econometrics, several times more outputs were submitted to each of these sub-panels (see Table 2), the vast majority of them ($\geq 96\%$) in the form of journal articles.

5.3   Economics & Econometrics

Our first REF sub-panel of interest, the 'Economics and Econometrics' unit of assessment, received 2600 publications from 28 institutions for the outputs submission. Of these, 2388 were journal articles, distributed in various publications as shown in Table 3.

Using a combination of CrossRef data and the clustering algorithm described in the previous section, eventually we were able automatically to assign the 2388 economics outputs into 277 unique journals.

Setting the threshold at all Economics and Econometrics journals containing ≥ 20 submitted articles, we obtain the distribution shown in Table 3. There are 29 named journals, representing over half of the total submissions.

5.4   Mathematical Sciences

After the field of economics, we study the Mathematical Sciences unit of assessment, which encompasses pure and applied mathematics, statistics and probability—though no distinction was made between these sub-fields, so the REF panel perhaps had the dubious honour of trying to assess subfields as diverse as pure mathematics and applied statistics together on the same measurement scale.

Mathematical Sciences was larger than Economics & Econometrics, with 53 submitting research institutions. The 6994 Mathematical Sciences outputs, of which 6731 were classified as journal articles, span some 696 unique scholarly journals. In this case, and for the remaining three sub-panels, the larger number of articles per journal necessitates a higher threshold for named journals: we increase the minimum number of submitted articles to 30 for Mathematical Sciences, Physics and Chemistry. This ensures that 'named' journals still provide a good representation ($\geq 50\%$ coverage) of outputs in the data, while keeping model complexity reasonably low.

Figure 1a suggests a similarly skewed distribution of articles across journals: many journals represented just one or two article submissions each, but a small number of mainly physical science journals had article counts in triple figures, including *Journal of Fluid Mechanics* with 254 articles and *Physical Review Letters* with 209. Some sub-fields appear to have published (or at least been submitted) more prolifically than others: the biggest statistical journal submission number was from *Biometrika* with 57 articles. See Table 5 for a full breakdown.

Across institutions, the journal submissions data for Mathematical Sciences are skewed: most journals were published in by only a handful of unique institutions, but there were a small number of journals popular with nearly all of the institutions assessed by the sub-panel. See Figure 1b.

5.5   Physics

We now turn to Physics, with 6446 REF2014 outputs, of which 6376 were journal articles in 304 unique journals, which we might expect to have some overlap with the Mathematical Sciences. Indeed, some 100 journal titles appear in both submissions.

There were 41 different institutions who submitted to the Physics sub-panel for REF2014.

As with Mathematical Sciences, in Physics we used a cut-off of 30 articles for a publication to be 'named' in the model, rather than aggregated under 'Other journals'.

The distribution of journals by article count and across institutions, shown in Figure 1, appear largely similar to the aforementioned subjects, but the breakdown of article counts by journal in Table 4 reveals that two journals—*Physical Review Letters* and *Monthly Notices of the Royal Astronomical Society*—were extremely strongly represented, constituting nearly 30% of all outputs.

5.6   Chemistry

Our fourth unit of assessment to model is Chemistry. The data for this field comprise 37 institutions, who submitted 4698 outputs, of which 4688 were journal articles in 326 unique journals.

The distributions of submissions, shown in Figure 1, once again look similar to the other fields. As in Physics, a couple of journals stand out for containing a very high proportion of outputs: the *Journal of the American Chemical Society* and *Angewandte Chemie* together represent nearly 25% of all submitted works.

6   Results

Figure 18 represents the posterior marginal density for the parameter α, defined in (4) as the effect of institutions' research environments—rather than journal submissions—on the probability of their outputs attaining 4* ratings in the REF.

For Economics and Econometrics, Figure 18a suggests there is little evidence for the environmental effect being distinct from zero, either when estimating the probabilities of journals attaining 4* or ≥ 3* ratings in the REF. The same was found also for Mathematical Sciences, Physics and Chemistry. This simple diagnostic check—for sensitivity of the results when controlling for an institution-level covariate—provides some, albeit limited, reassurance: the results appear robust to potential effects of aggregation bias, and there is no indication from this check of anything like Simpson's paradox.

Trace plots for the Hamiltonian Monte Carlo runs are given in Figure 19, and suggest good mixing of the chains for each of the parameters.

To catch any glaring errors in the results, and for a more informed interpretation of the findings (especially the implied journal rankings in each field) the initial results were presented to several senior University of Warwick academics with expertise in their respective disciplines. This was invaluable, for example, in spotting the conspicuous absence of *Annals of Mathematics* from the rankings, due to the aforementioned coding error in the REF2014 data. With such anomalies fixed, our informal panel of experts provided useful context for the final results, presented in the following sections.

6.1   Economics & Econometrics

Figure 2 provides a 'league table', in the form of a series of box plots of the marginal posterior distributions, of the estimated Economics and Econometrics journal probabilities of attaining 4* and 3* or 4* ratings in the REF. Strikingly, the five journals considered

among economists to be the 'Top Five' in their field (Heckman and Moktan 2018) are near the top of this ranking as well: namely, the *American Economic Review*, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and even the *Journal of the Political Economy*, despite the latter only representing a handful of outputs, at 22 articles. Looking at the probability of achieving 3* or greater (Figure 2b), the top probabilities are all so close to 1 that little can be inferred from the ordering of the journals.

The 95% posterior intervals are quite wide, especially for publications with fewer articles submitted in the REF, which is to be expected given the inherent uncertainty associated with estimating a large number of parameters from a small number of incomplete observations.

Our journal ranking has several notable omissions: the *Journal of Labor Economics* and the *RAND Journal of Economics* are highly respected (Sgroi 2019; Oswald 2019), as are general science journals such as *Science*, *Nature* and *PNAS*. However none of these journals met the minimum threshold of 20 articles submitted to the REF2014 Economics & Econometrics sub-panel, so they do not appear as 'named journals' in our results.

As a robustness check, Figure 3 compares Hamiltonian Monte Carlo estimated journal probabilities of attaining 4* with the respective maximum likelihood estimates computed via the expectation–maximization algorithm (with the level of pseudo-data set (arbitrarily) at one article per journal). The maximum likelihood estimates come without any uncertainty quantification, but we can see a strong correlation in point estimates between the β_j estimates of Equation (5) and the (logit) success probabilities corresponding to the same journals, so the general approach seems sound.

Figure 4a shows the predicted versus actual 4* output profiles for each of the institutions in Economics and Econometrics. With the predicted 4* and 3* profiles converted into funding allocations, Figure 4b shows the resulting discrepancies between the predicted institutional funding versus that actually allocated (based on the methodology of Koya and Chowdhury 2017) based on HEFCE data. Not all institutions are based in England, of course, so the 'actual' funding figures for other nations in the UK assume that the respective research councils used similar formulae to allocate funding based on REF2014 outputs.

The quality of prediction appears reasonably good, with most points falling close to the line of $y = x$. Some institutions appear to have received more 4* ratings than predicted from their journal choices, notably Cambridge and UCL, and Queen Mary University London appears to have received fewer 4* ratings than suggested by the model. Otherwise there are no noticeable outliers.

When it comes to funding, Figure 4b shows how much funding would be allocated, when combining the estimated 4* profiles with 3* ratings and the FTE headcount for each department. The only institution with a significant discrepancy is Brunel, and this can be accounted for by the fact that most of that university's Economics and Econometrics outputs were published in less popular journals not named in Table 3, so the model has less information available to predict this institution's results.

Across the four fields, we compute the index of dissimilarity, Δ, and the index of redistribution of monetary reward, $\Delta_{\pounds}$. The distributions of these metrics are plotted in Figure 5. Lower numbers are better.

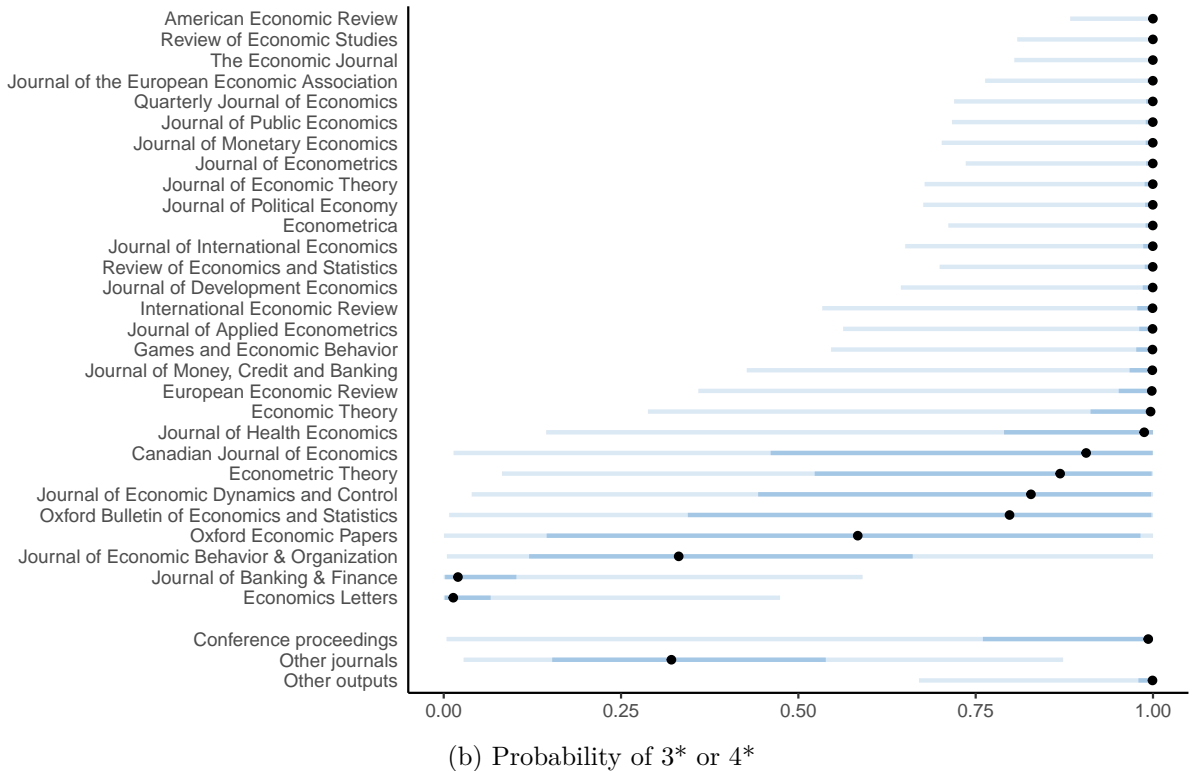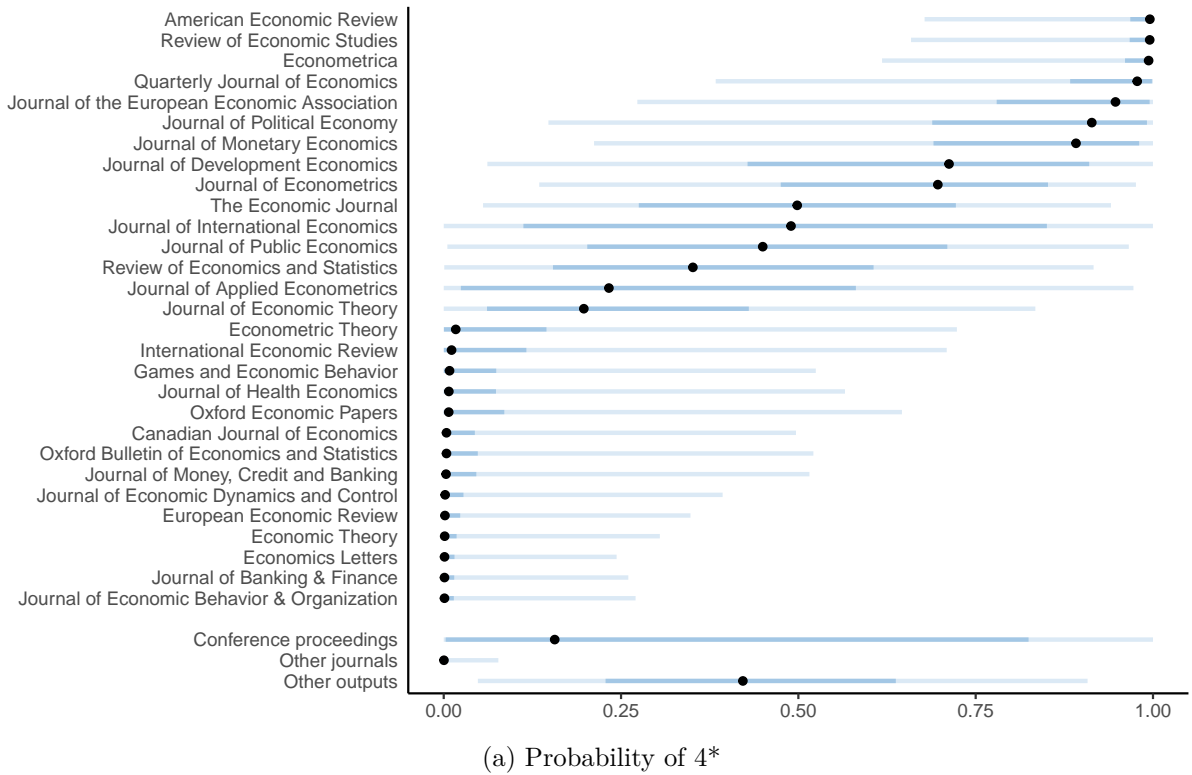(a) Probability of 4*

(b) Probability of 3* or 4*

Figure 2: Median estimated journal success probabilities in Economics and Econometrics. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 20 or more articles submitted in REF2014
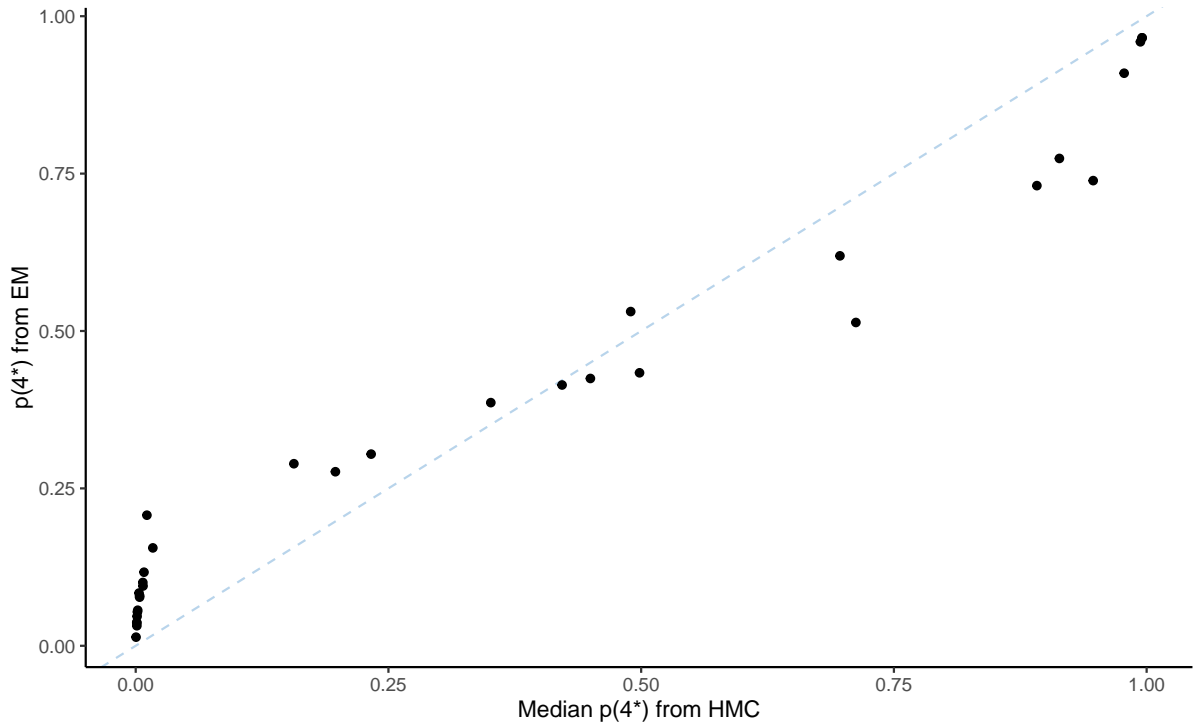
Figure 3: Maximum likelihood estimates of journal effects, $\hat{\beta}_j$, versus Hamiltonian Monte Carlo estimates of journal success probabilities (on a logit scale), for Economics and Econometrics, with line of best fit

The median value for Economics and Econometrics is $\Delta = 17.9\%$, that is, this proportion of articles would need to be reclassified for the estimated institutional profiles to match exactly those published in the REF. As a metric, 82.1% accuracy sounds like it might be quite good, but we should be careful not to draw too many conclusions from a single number. In funding terms, that translates to $\Delta_\pounds = 8.7\%$ of funding in Economics and Econometrics needing to be reallocated if an initial allocation was made based on the Poisson binomial model alone. Across dozens of institutions, that represents a substantial amount of money, though.

Figure 6 provides evidence against the model of Yan (2017), which assumed a constant cumulative probit difference between the probability of getting 4* and the probability of getting 3* or better. It is clear that 'better' journals (those more likely to attain 4*) have a smaller cumulative probit difference, suggesting that it is not much harder for an output in such an apparently high-achieving journal to get a 4* than a 3* rating, whereas for 'weaker' journals, it is harder to improve from 3* to 4*.

6.2   Mathematical Sciences

In Mathematical Sciences, we face the problem of several partly disjointed sub-fields, such as pure mathematics, statistics, mathematical physics and mathematical biology, all falling under the same umbrella. As a result it is harder to gauge what might be considered a group of 'top' mathematical sciences journals—mathematicians might declare that statistics, for example, is merely applied mathematics and that a pure mathematics journal should lead the field (Monroe 2008) whilst statisticians might counter that statistical journals should
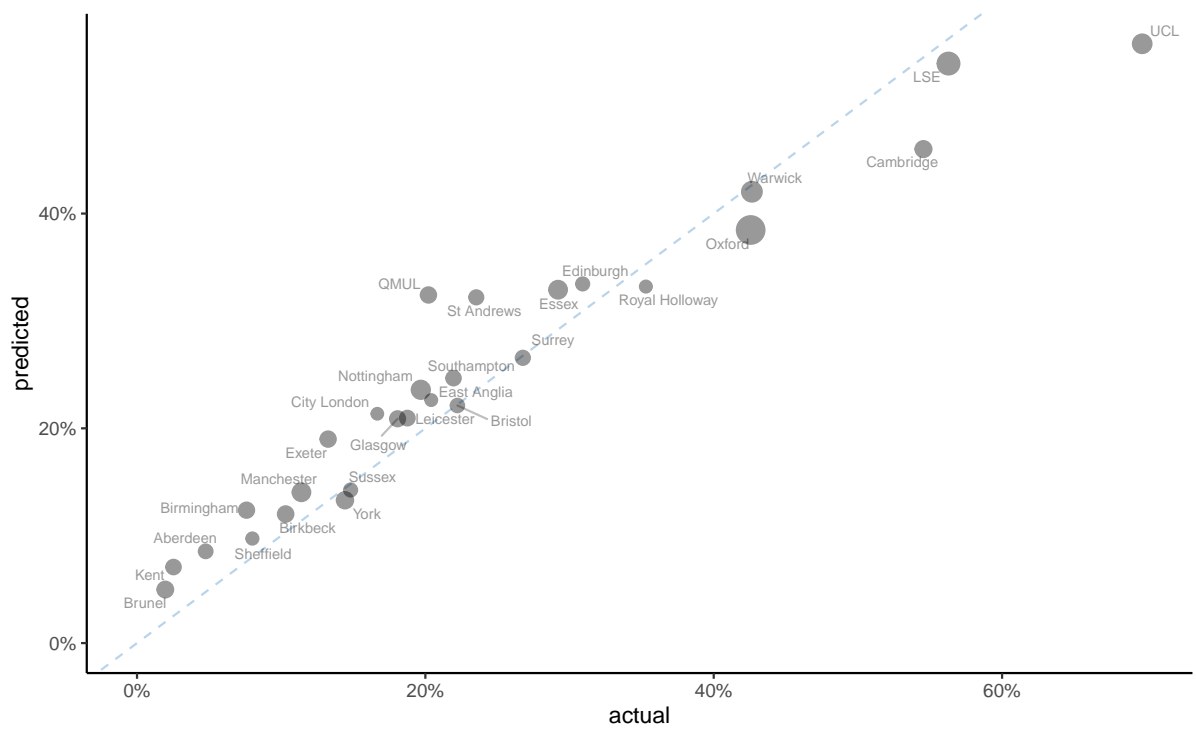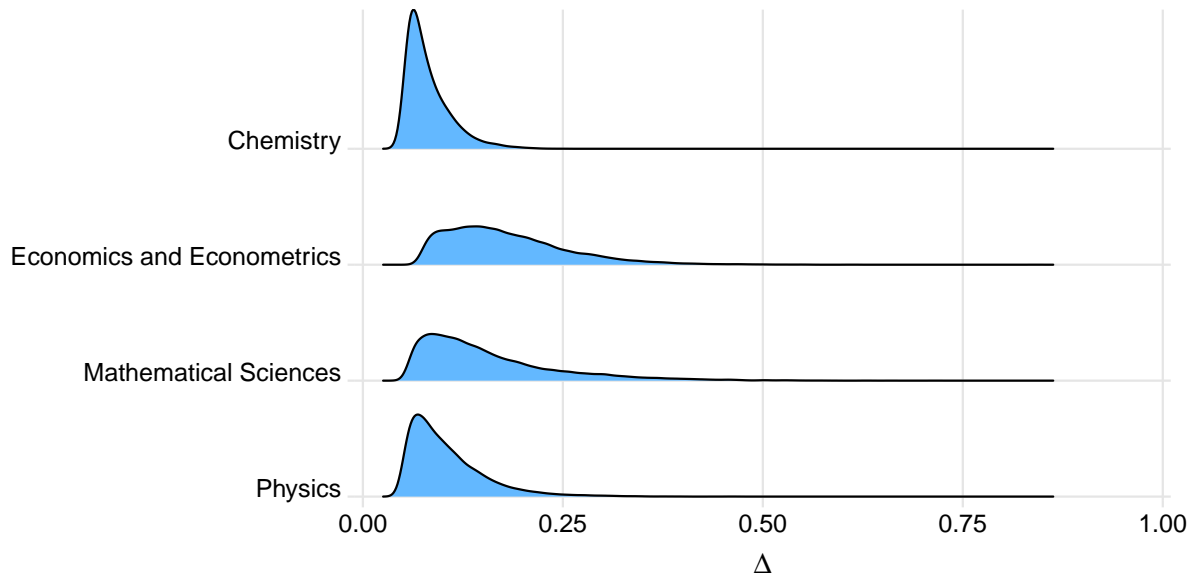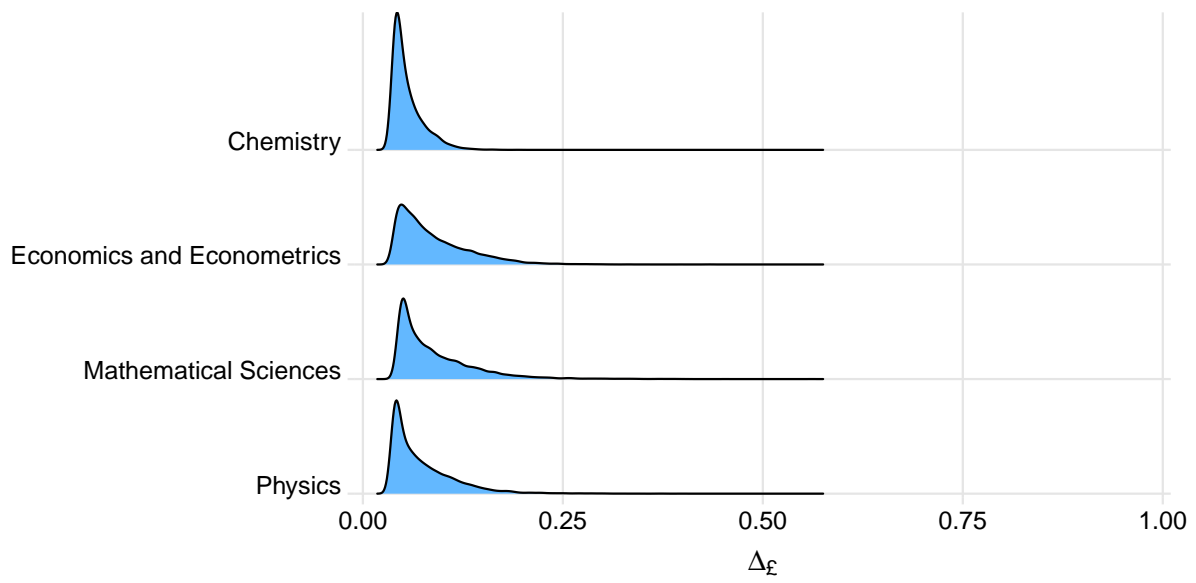
(a) % articles at 4*
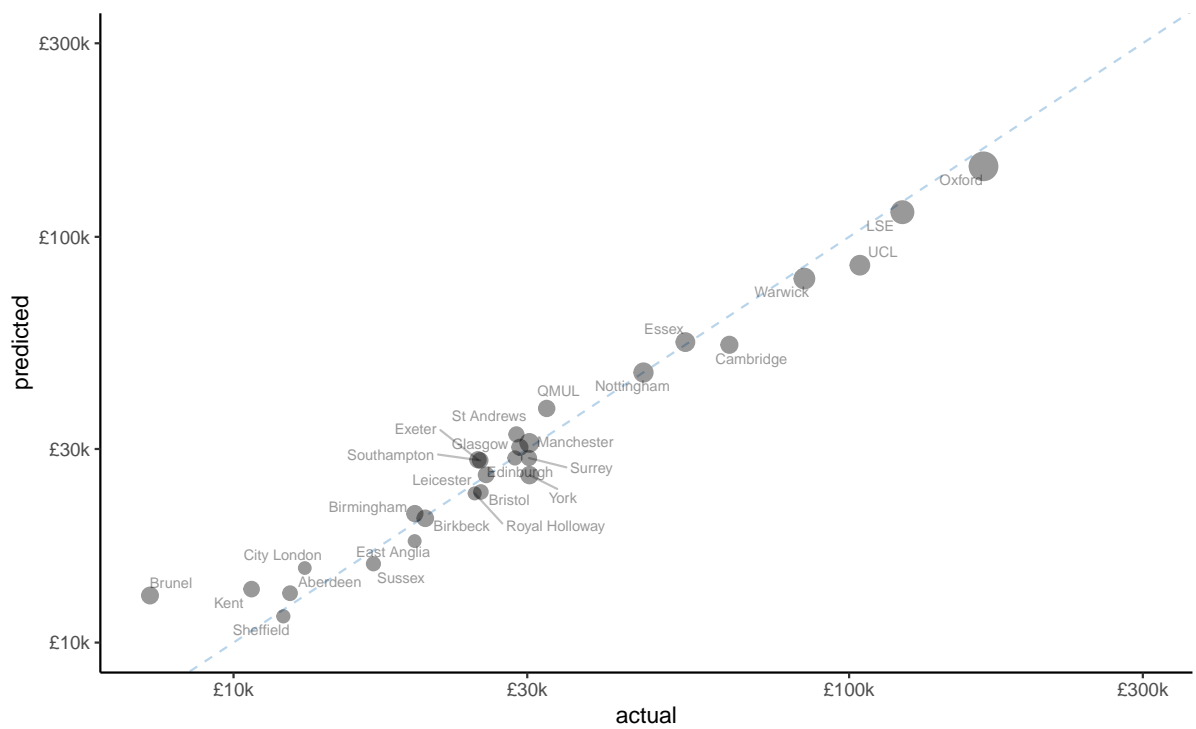
(b) funding allocation

Figure 4: Predictions versus observed REF2014 results for institutions submitting outputs to the Economics & Econometrics sub-panel, with point sizes proportional to number of FTE staff

23

(a) Index of dissimilarity

(b) Index of redistribution of monetary reward

Figure 5: Density plots of indices of dissimilarity and of redistribution of monetary reward, by unit of assessment
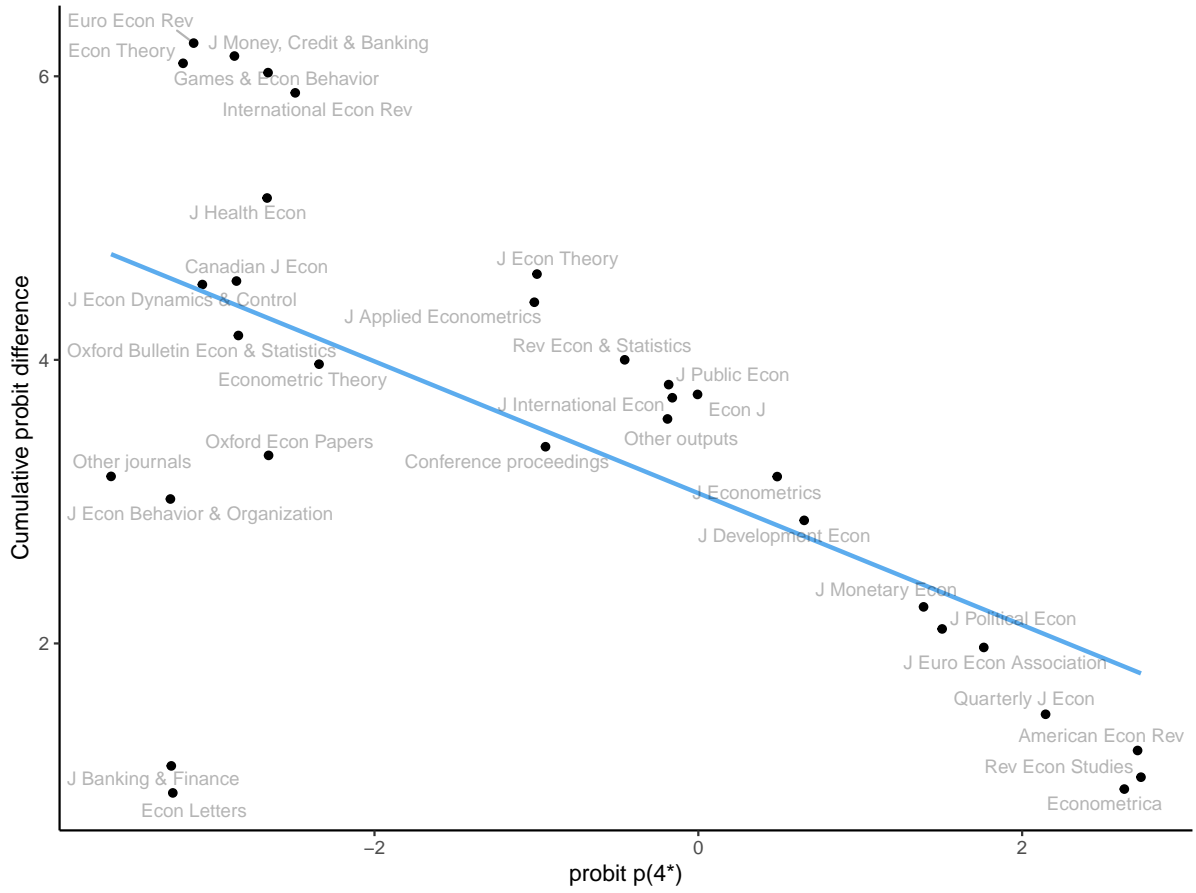
Figure 6: Comparison of cumulative probit differences, $c_j = \text{probit}(p_j^{34}) - \text{probit}(p_j^4)$, versus estimated probit probability of attaining 4*, by journal in Economics and Econometrics in REF2014, with line of best fit. A non-zero slope implies $c_j \neq c$, that the cumulative probit difference is not constant across journals

come top because of the widespread application of statistics. It is perhaps surprising, then, that some of the reputed top journals in statistics, *Annals of Statistics*, *Biometrika* and the *Journal of the Royal Statistical Society: Series B* (Varin, Cattelan, and Firth 2016) are still ranked highly based on the model for attaining 4* ratings in the REF. See Figure 7. However, the *Journal of the American Statistical Association*, also a highly-regarded statistics journal, has a low estimated probability of obtaining 4* ratings.

For the mathematicians, *Inventiones Mathematicae* and *Annals of Mathematics* are both highly reputed and have the highest estimated probabilities of yielding 4* ratings in the REF. The *Journal of the American Mathematical Society* and *Publications Mathémathiques de l'IHÉS* are also highly regarded (Loeffler 2019), but do not appear in the results as named journals because fewer than 30 of their respective articles were submitted to the REF.

In Mathematical Sciences, the predicted versus allocated 4* ratings and funding allocations, by institution, are presented in Figure 9. Apparent outliers (such as Coventry University or the University of Greenwich in the 4* plot) are among the smallest institutions by number of full-time equivalent (FTE) research staff. There appears to be a pattern, however: weaker institutions are expected to do better, and stronger institutions are expected to do worse, than their actual published performance in the REF.

This shrinkage effect implies that some variation in assessed quality of outputs is not explained by journal identities alone. It indicates that there is variation in quality within at least some journals, and that high-ranked institutions tend to publish more of the high-quality papers in such journals.

In terms of summary measures, the median index of dissimilarity for Mathematical Sciences is 15.5% and the median required redistribution of monetary reward is 8.9%; the posterior distributions of these statistics are plotted in Figure 5.

6.3 Physics

Posterior probabilities for the Physics sub-panel are presented in Figure 10. Journals from Nature Publishing Group have the highest estimated probabilities of attaining 4*, though no probabilities are near 100%, perhaps owing to the relatively small number of 4* ratings awarded in this field generally. The appearance of *Physics Review Letters* above the *Proceedings of the National Academy of Sciences (PNAS)* in the ranking might imply a preference by physicists in the review panel for physics-specific journals over general science ones. In the international astrophysics community, *Astrophysical Journal* might be considered more prestigious than *Monthly Notices of the Royal Astronomical Society*, but the latter has a slightly higher estimated probability of 4*, which might be interpreted as a UK-centric bias (Ball 2019). Relatively low success probabilities for *Physics Review B* and *C* could be attributed to an inter-journal dependence: namely, some works published in these journals also being announced in the highly-ranked *Physical Review Letters*.

As in Mathematical Sciences, a comparison of the predicted versus actual institutional REF results in Physics, shown in Figure 11, reveals a linear relationship, but an apparent shrinkage effect, implying some variation in assessed quality not explained by journal identities. The performance of the University of Oxford, in particular, appears to be under-estimated by the model, suggesting that where there is variation of assessed quality
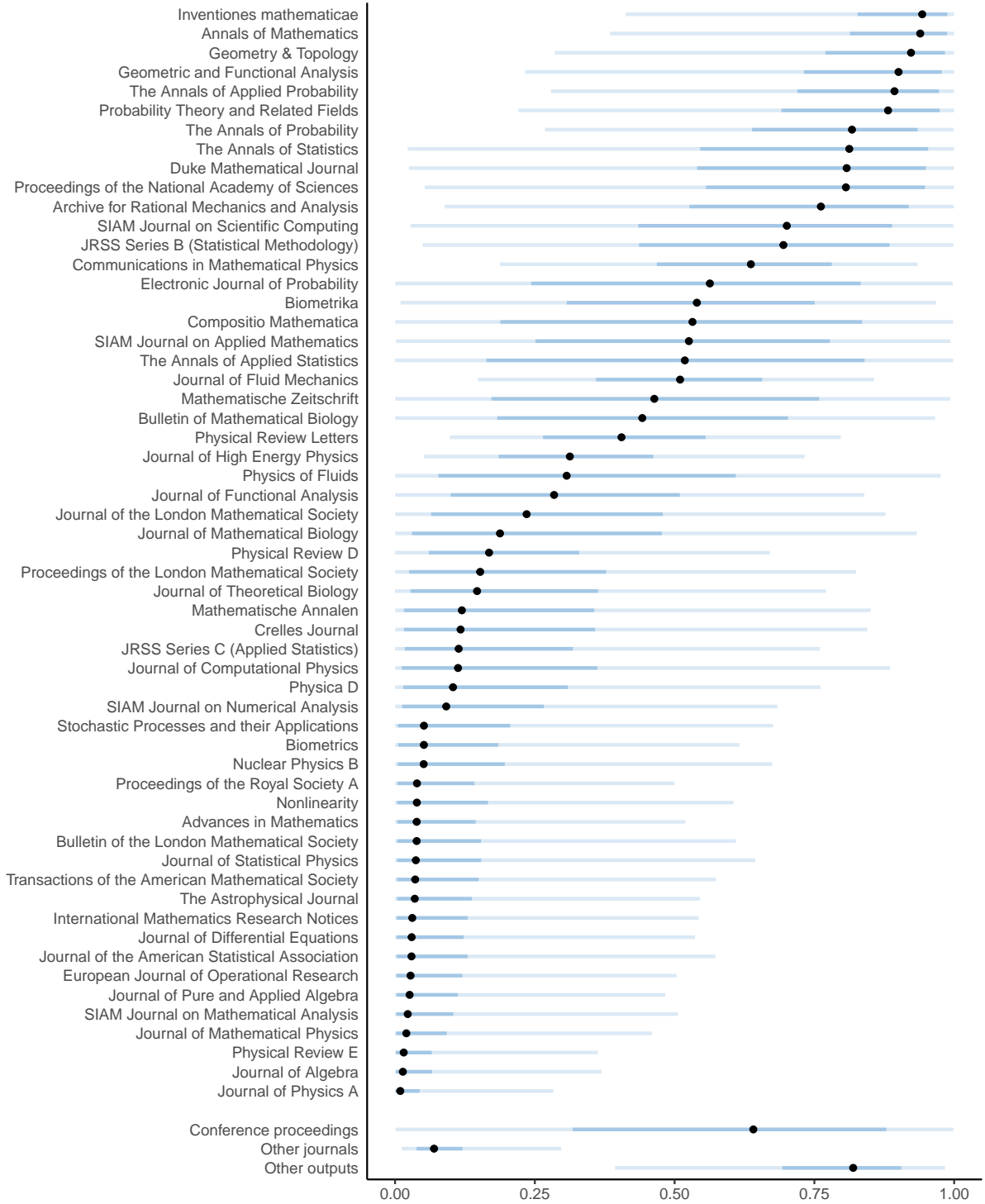
Figure 7: Median estimated journal success probabilities of 4* ratings in Mathematical Sciences. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014
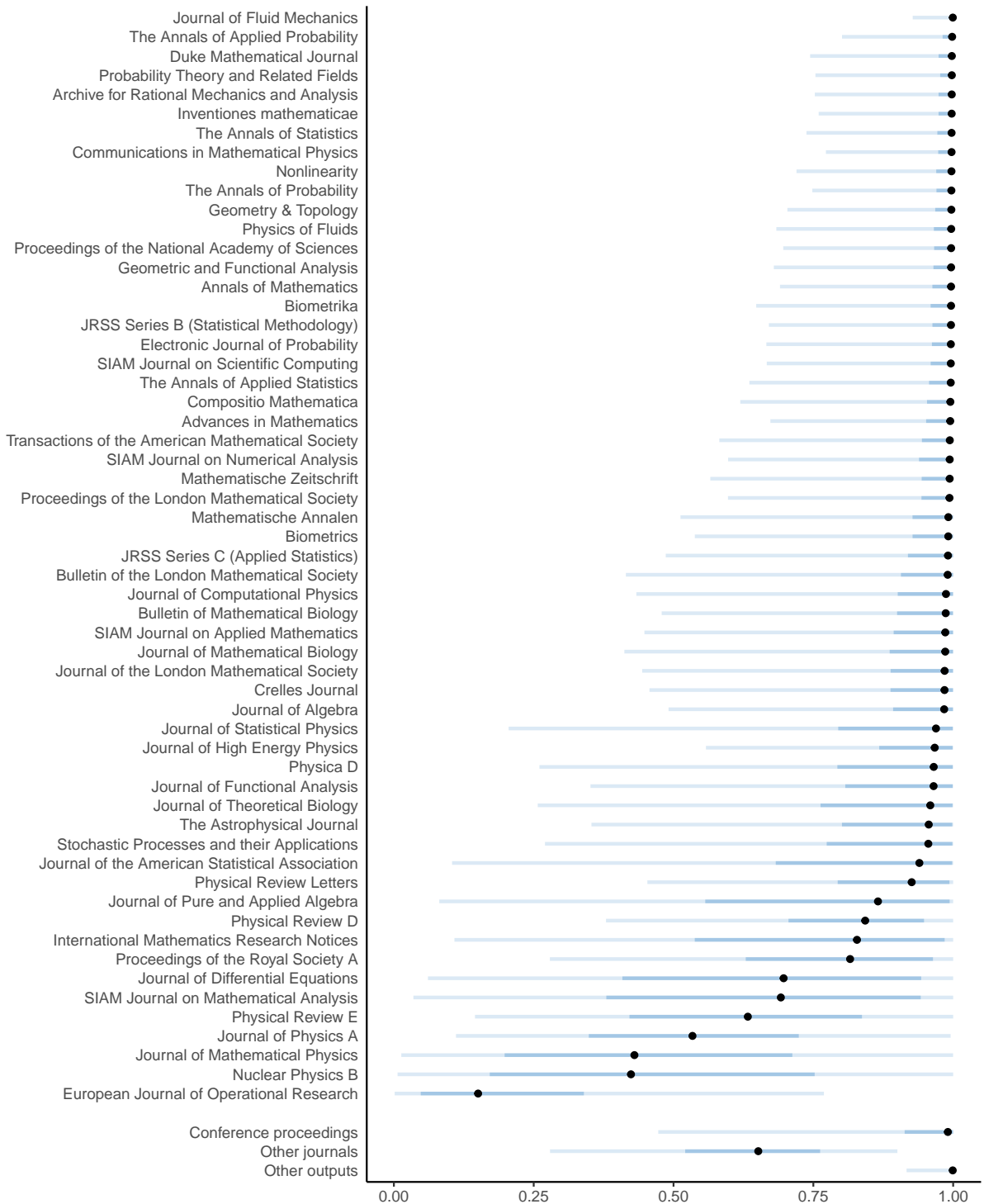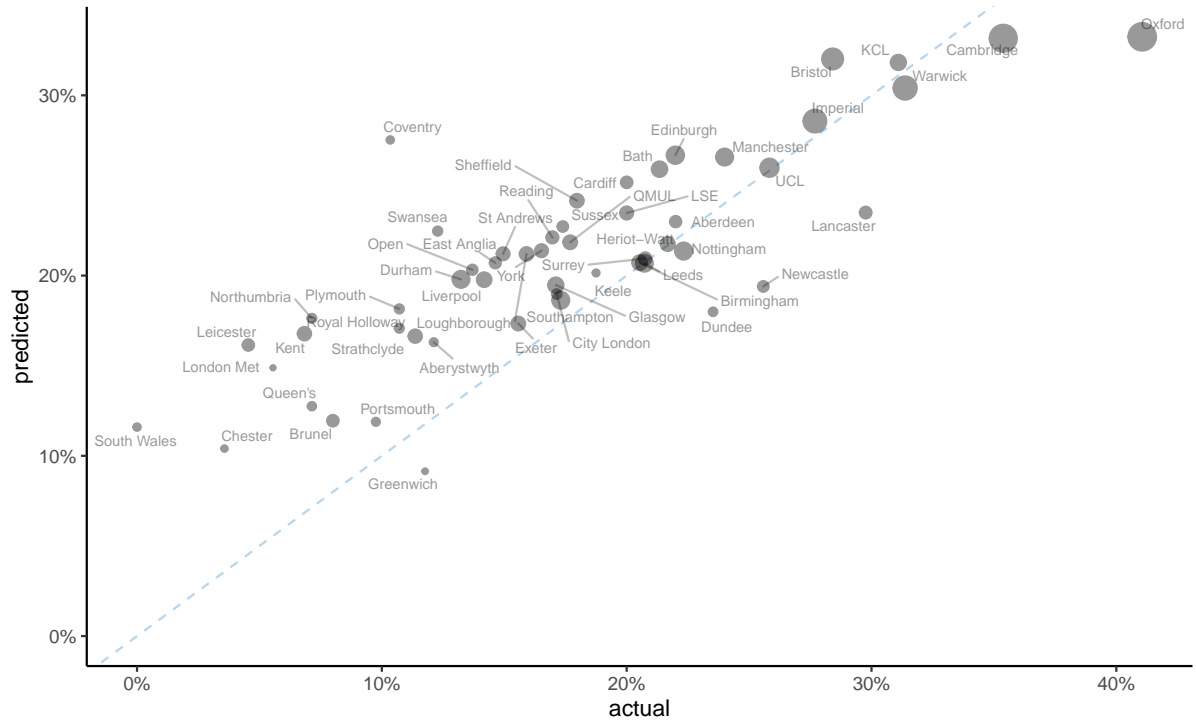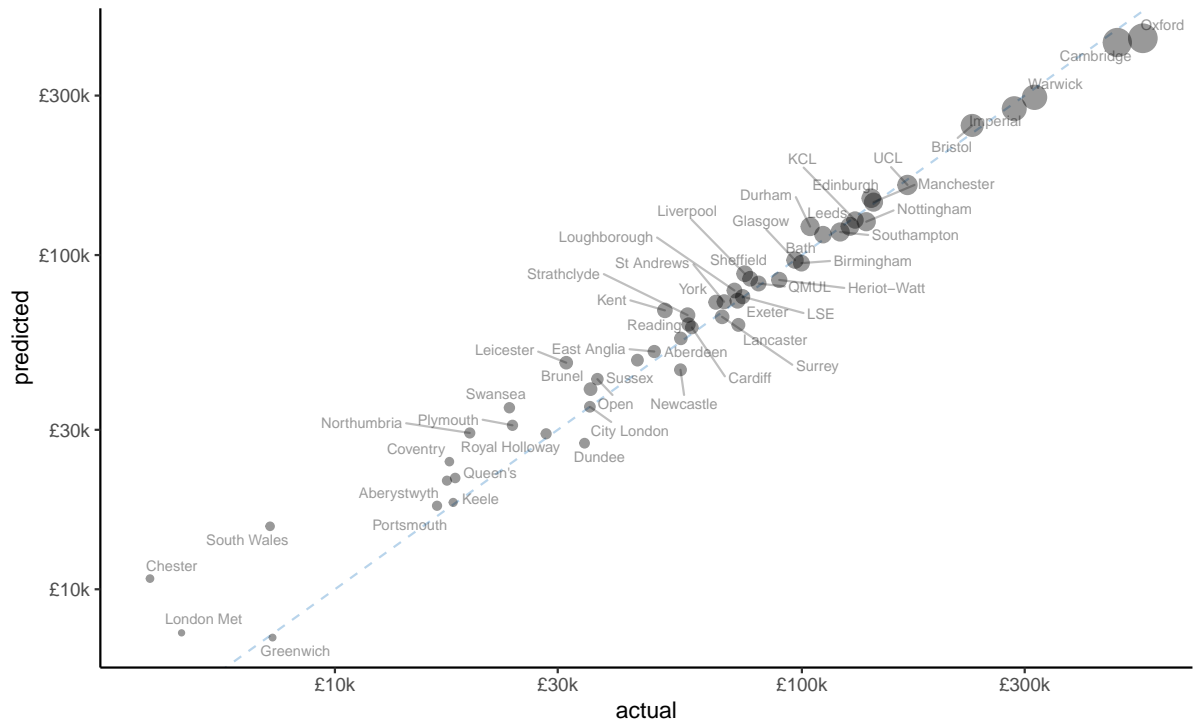
Figure 8: Median estimated journal success probabilities of 3* or 4* ratings in Mathematical Sciences. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014
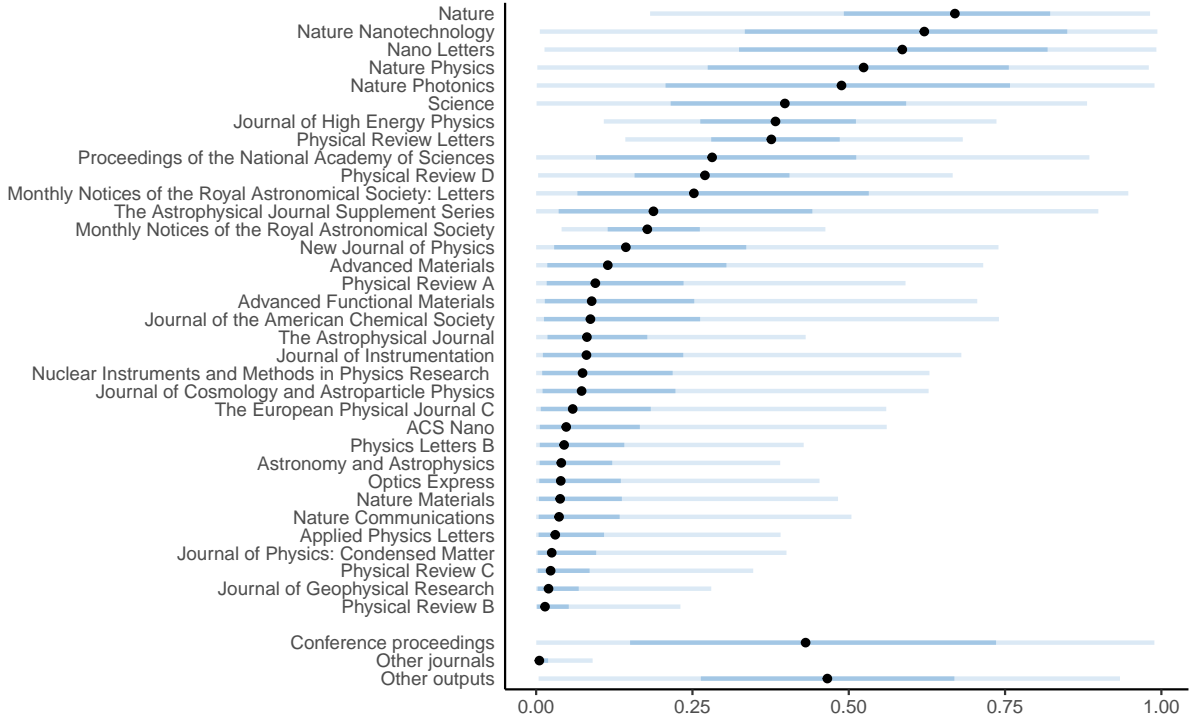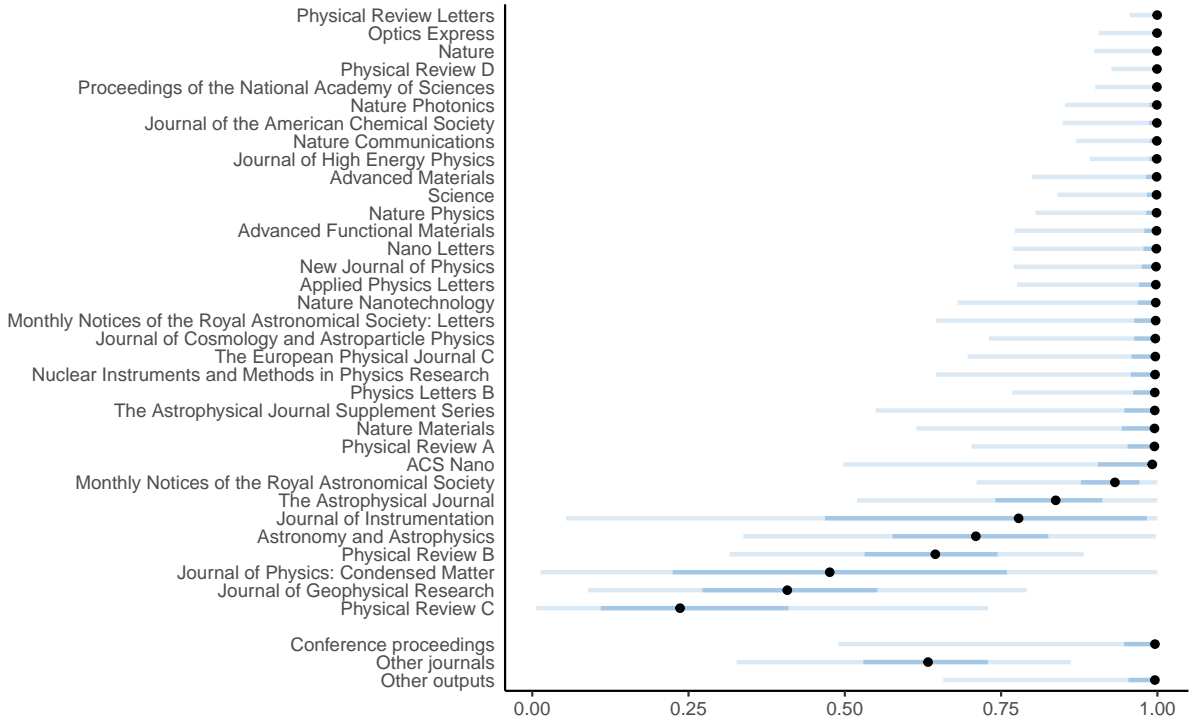
(a) % articles at 4*

(b) funding allocation

Figure 9: Predictions versus observed REF2014 results for institutions submitting outputs to the Mathematical Sciences sub-panel, with point sizes proportional to number of FTE staff

(a) Probability of 4*

(b) Probability of 3* or 4*

Figure 10: Median estimated journal success probabilities of 4* ratings in Physics. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014

within journals, the higher-quality outputs may be more likely to have been from Oxford researchers.

By summary measures, the median index of similarity in Physics is 10.5% and the median proportion of reallocated research funding is 7.3%.

6.4   Chemistry

Figure 12 provides league tables of estimated journal REF success probabilities in Chemistry. This field, unlike the others studied here, seems to be dominated by popular general science outlets, in *PNAS*, *Nature* and *Science*, rather than dedicated chemistry journals. There may be some dependence on types of articles published: some periodicals print different mixtures of 'full' research papers and communications (letters). *Nature Chemistry* and *Nature Communications* fall lower in the ranking than might be expected (Bugg 2019; Scott 2019). This appears to be simply a result of work published in journals being submitted by several low-scoring institutions that were not awarded many 4* ratings in the REF.

The expected versus actual institutional results are plotted in Figure 13. The pattern around the line of $y = x$ is similar to that in the other sub-panels: broadly a linear relationship, but with lower-scoring institutions having higher predicted than actual results, and the converse for stronger institutions. There are no noticeable outliers.

By summary measures, the median index of similarity in Chemistry is 8.2% and the median proportion of research funding that would need to be reallocated would be 5.6%. The posterior distributions are plotted in Figure 5. Performance, according to these metrics, appears similar to for other fields.

6.5   Comparison with Journal Impact Factors

Using data from Clarivate Analytics' *Journal Citation Reports*, we can compare the latent journal REF effects with journal impact factors for the respective year. For this article, we use the 2014 edition of *Journal Citation Reports*, as this is based on citation data from the preceding two years. (One could also consider the 2013 edition, though the results should not be too different.)

It may also be possible to compare with rival metrics, such as the CiteScore and Scimago Journal Rank (SJR), Scopus's versions of the impact factor and the Eigenfactor, respectively, however we do not make those comparisons here.

Economics & Econometrics

Comparisons are plotted in Figure 14. Note the logarithmic scale for the Eigenfactor score. Broadly speaking, there is a (weak) positive correlation between both citation metrics and the probability of attaining 4* in the REF. Evidence for the supposed dominance of the 'top 5' economics journals is mixed. Whilst these periodicals are indeed highly ranked by journal impact factor, Eigenfactor and apparent REF effect, they do not completely dominate the top five spots, so their reputation must depend on other factors or perhaps be undeserved. Moreover, as economists have explicitly known of the 'top 5' designation for years, it may be a self-fulfilling prophecy.
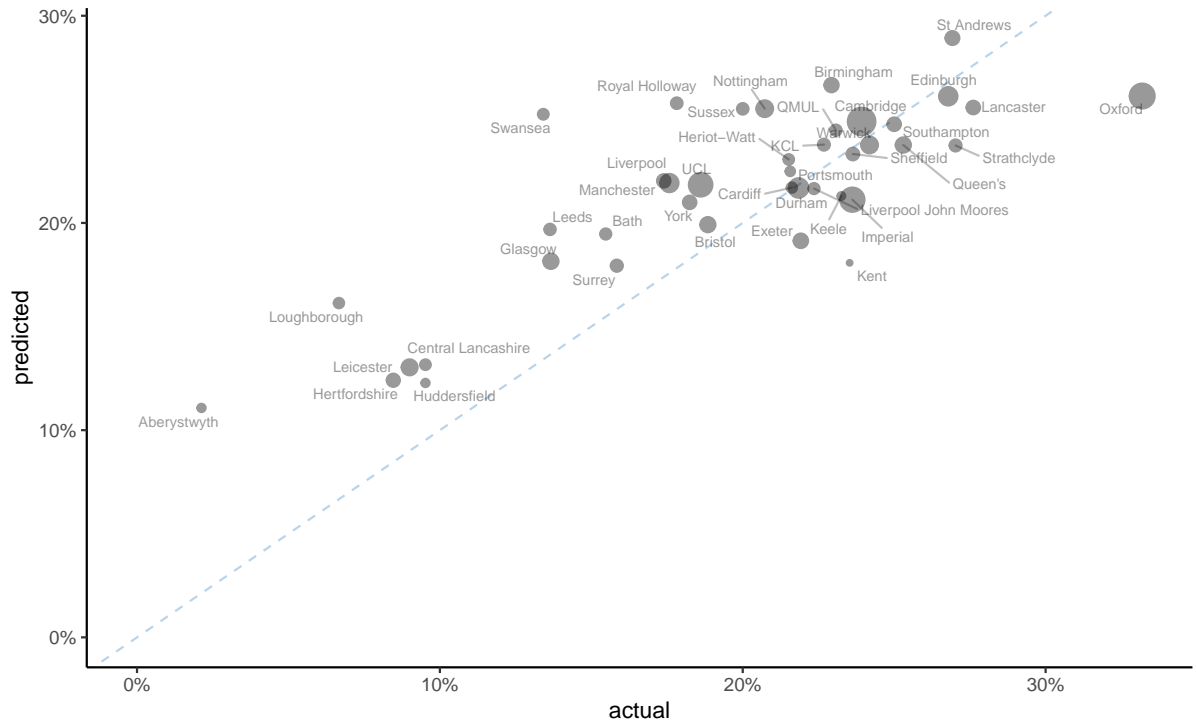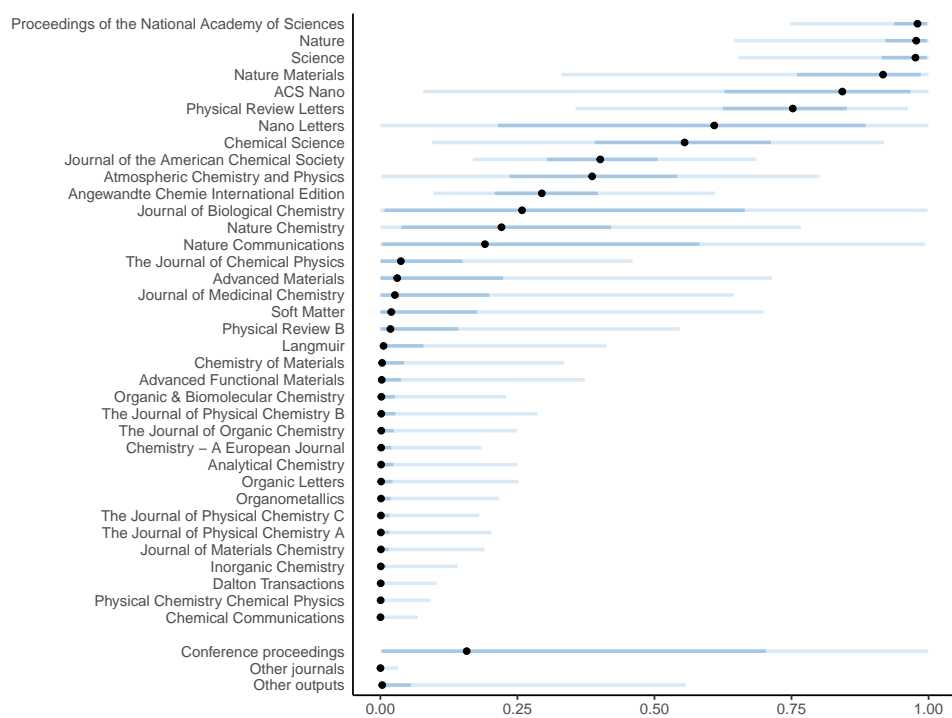
(a) % articles at 4*
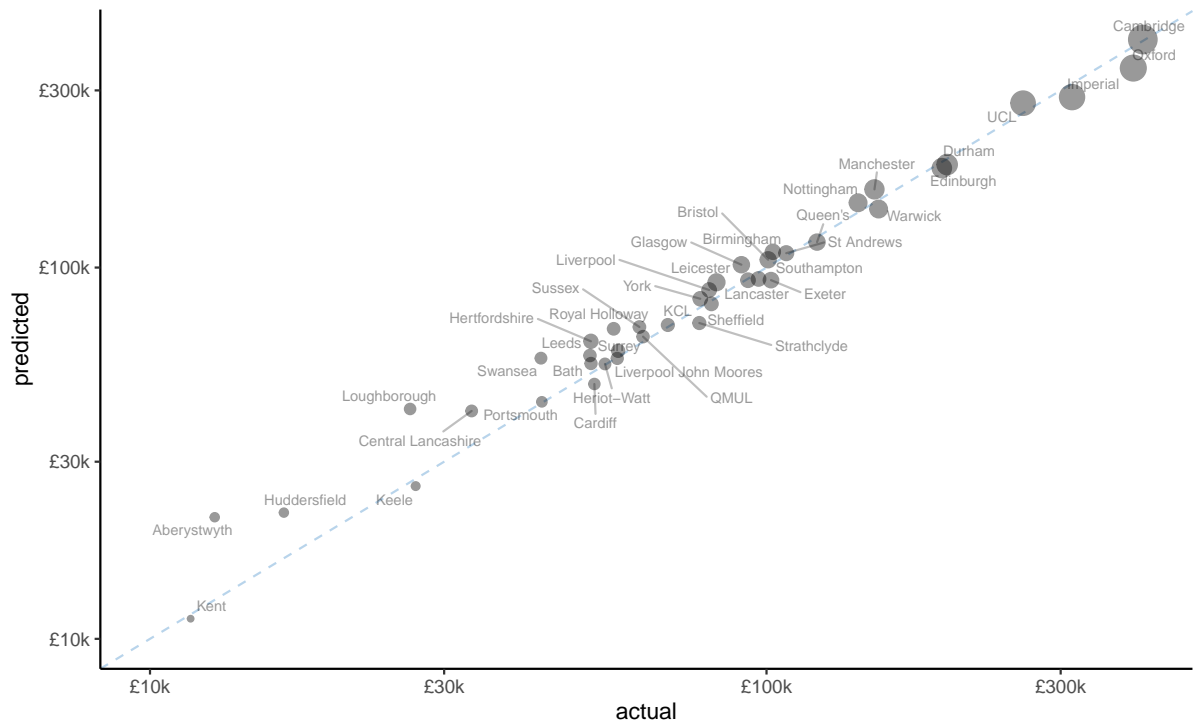
(b) funding allocation

Figure 11: Predictions versus observed REF2014 results for institutions submitting outputs to the Physics sub-panel, with point sizes proportional to number of FTE staff

(a) Probability of 4*

(b) Probability of 3* or 4*

Figure 12: Median estimated journal success probabilities of 4* ratings in Chemistry. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014
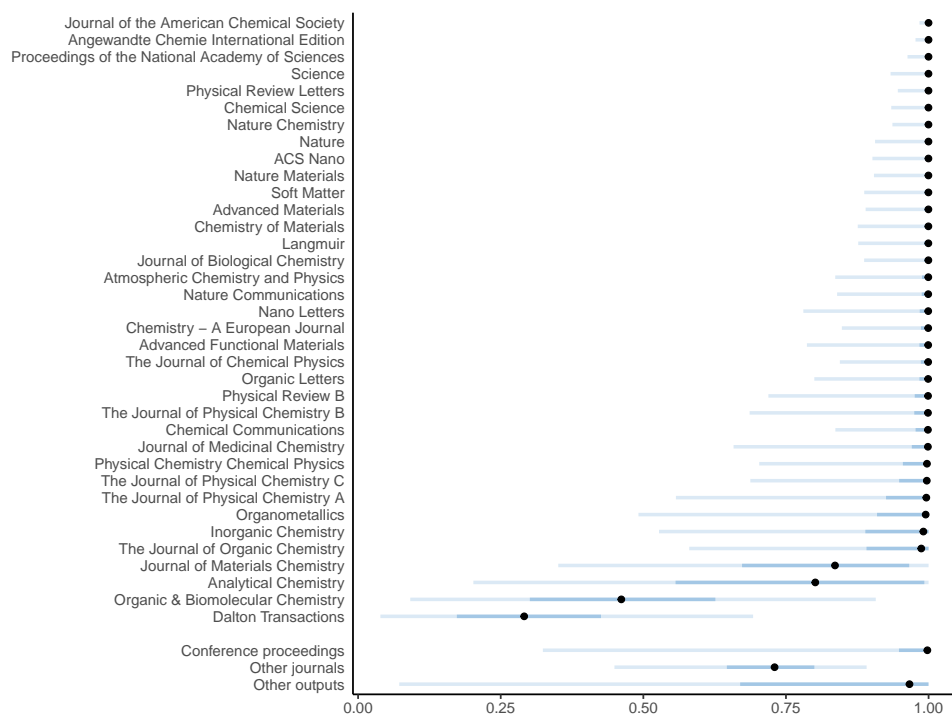
(a) % articles at 4*

(b) funding allocation

Figure 13: Predictions versus observed REF2014 results for institutions submitting outputs to the Chemistry sub-panel, with point sizes proportional to number of FTE staff
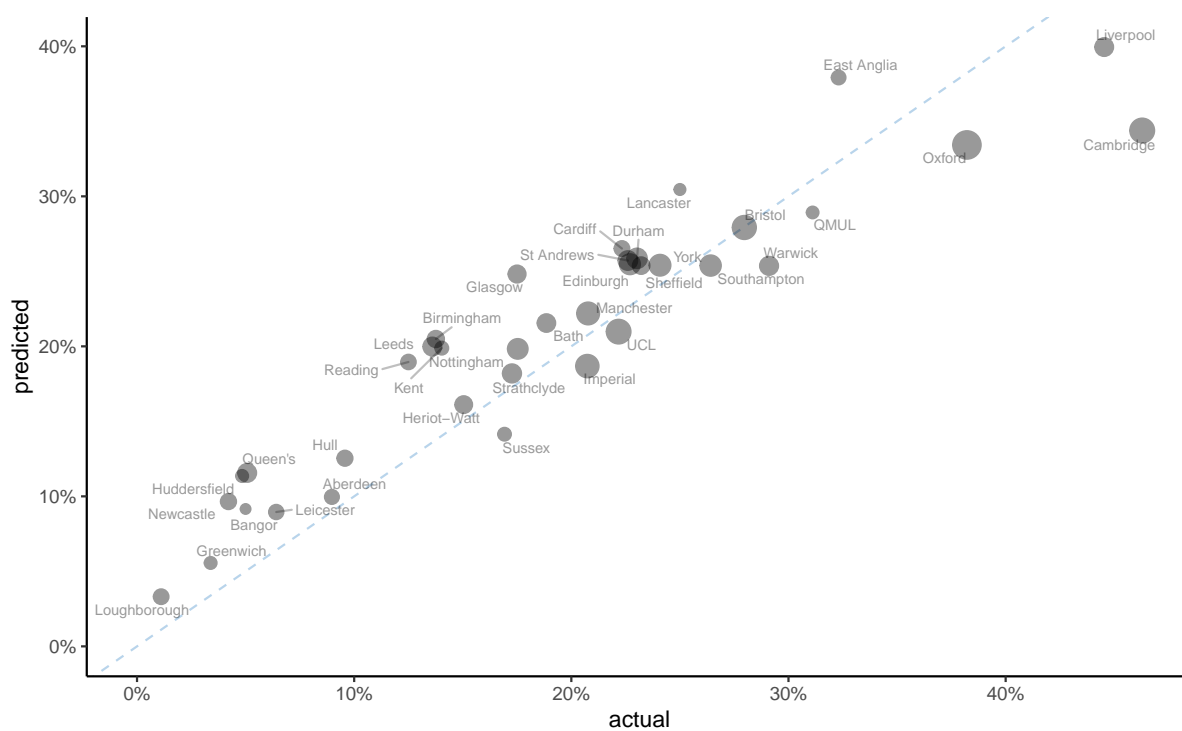
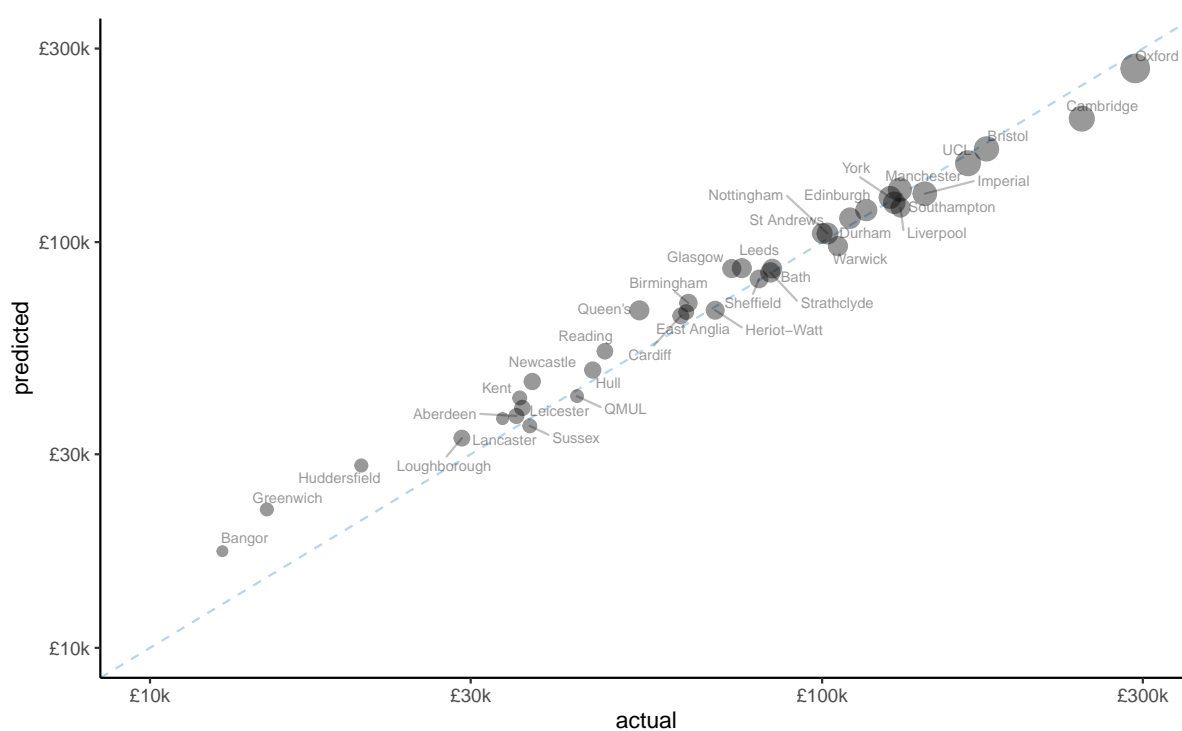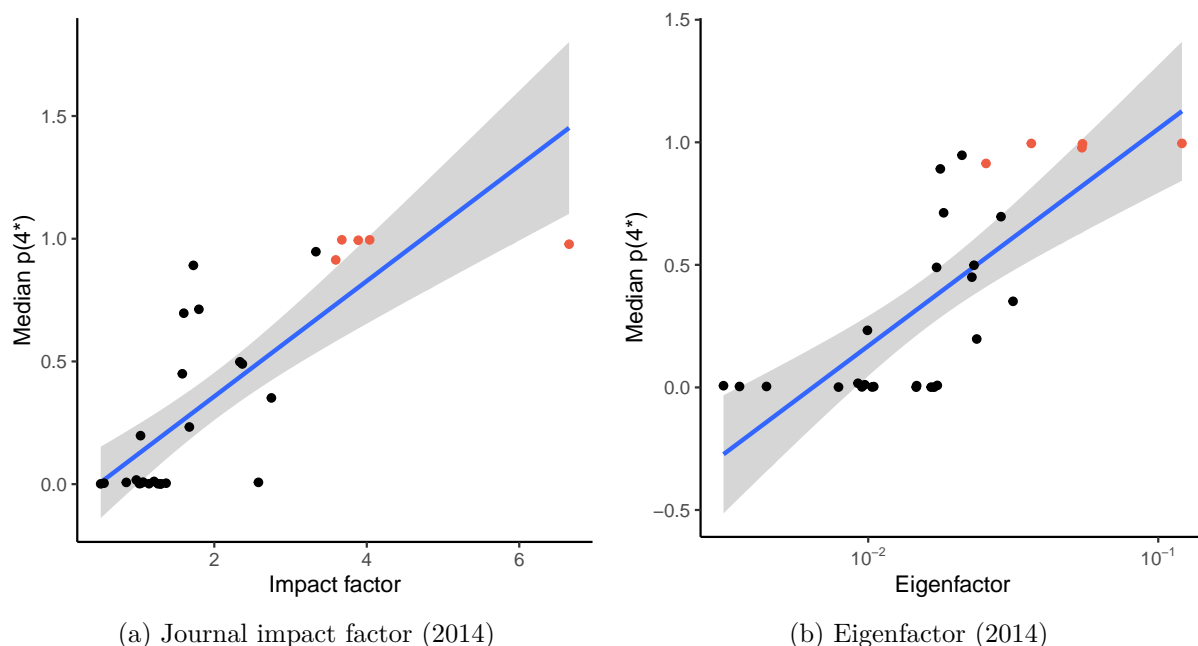(a) Journal impact factor (2014) (b) Eigenfactor (2014)

Figure 14: Comparison of Economics and Econometrics journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit. So-called 'top 5' journals are highlighted in red

Mathematical Sciences

In Mathematical Sciences, however, there is almost no correlation between journal impact factor and the estimated probability of 4* in the REF; see Figure 15. This phenomenon could partly be explained by mathematical journals generally receiving lower impact factors; mathematics papers tend to have short reference lists and take longer to be noticed, when compared with publications in microbiology and other applied disciplines, so the journal impact factor (roughly speaking, counting citations over two years) is an especially poor metric for mathematics work. Most mathematics journals here had an impact factor of around 1 or 2, so most of the variation between those scores might be attributed to random noise—see the left hand side of Figure 15a.

Physics

A positive correlation is present between Clarivate citation metrics and estimated probability of 4* for Physics as illustrated in Figure 16.

Chemistry

The story for Chemistry is hard to interpret because so many journals have median estimated 4* probabilities close to zero. But the top three journals by impact factor were also estimated to have the highest chances of their articles attaining 4* in the REF. See Figure 17.

(a) Journal impact factor (2014)

(b) Eigenfactor (2014)
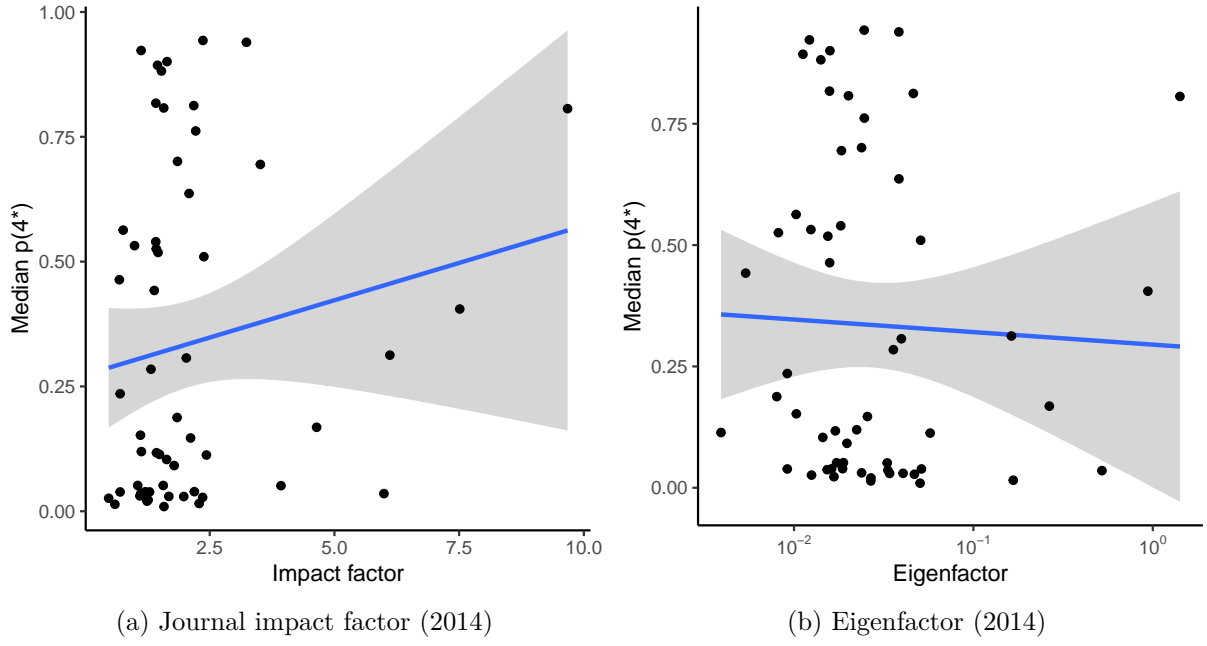
Figure 15: Comparison of Mathematical Sciences journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

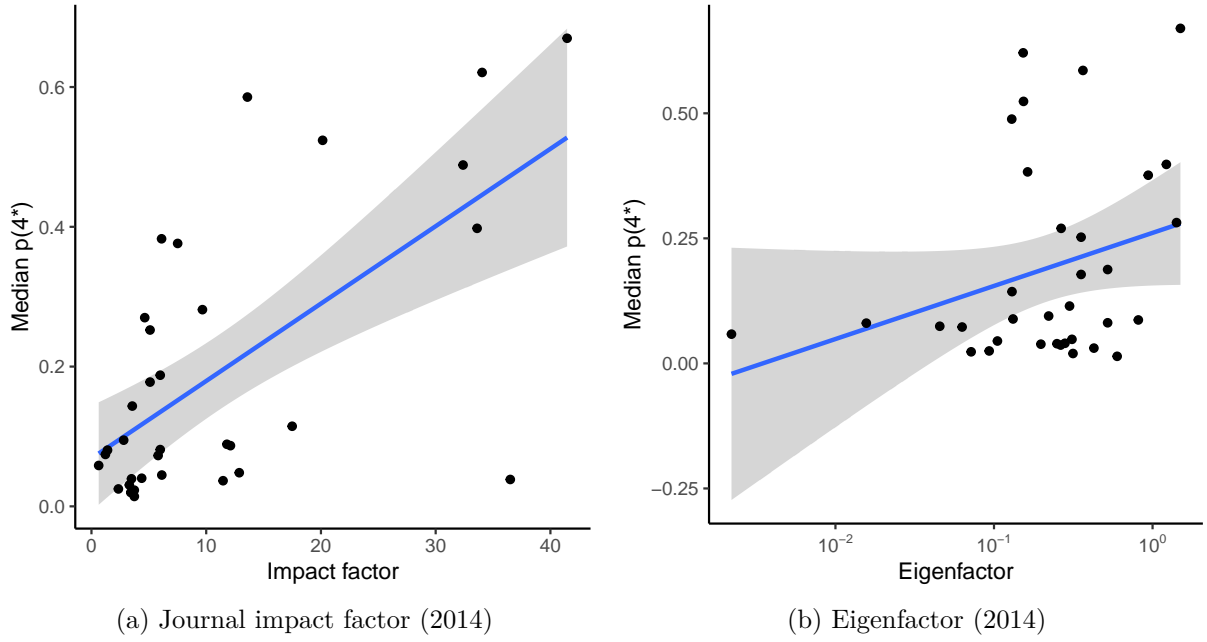(a) Journal impact factor (2014)

(b) Eigenfactor (2014)

Figure 16: Comparison of Physics journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

(a) Journal impact factor (2014)
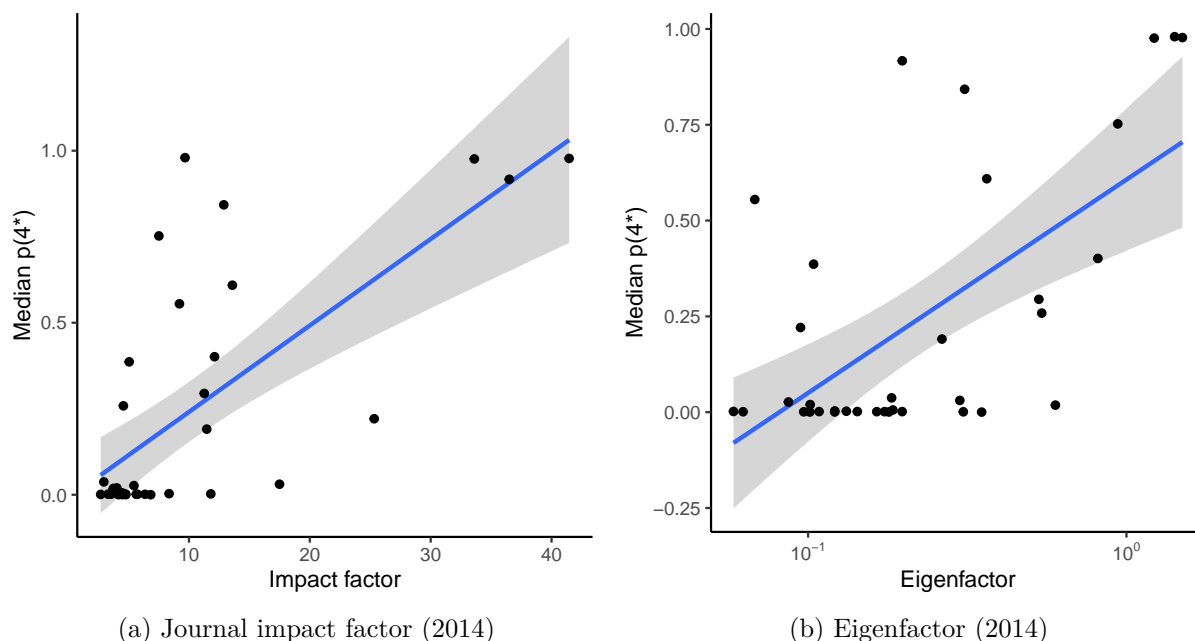
(b) Eigenfactor (2014)

Figure 17: Comparison of Chemistry journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

7 Discussion

This chapter has explored the relationship between published REF2014 results and the journals in which institutions published research outputs submitted for REF2014 assessment. The results are informative in various ways, including:

- implied rankings of the main journals from which work was submitted in each REF2014 sub-panel, together with measures of uncertainty on such rankings; and

- for each REF2014 sub-panel studied, measurement of the *maximum* extent to which REF2014 outcomes can be explained (retrospectively) by the identities of the journals from which work was submitted by each institution.

One reassuring aspect of the journal rankings derived for the four disciplines studied here is that they broadly agree with the informed opinions (informally elicited) of senior Warwick academics in those disciplines. That is to say, for the main 'named' journals in each field, the estimates and uncertainty intervals for the journals' probabilities of attaining 4* ratings in REF2014 made sense in the minds of the experts who were consulted. Had the opposite been found, it would have been a strong reason to distrust the statistical methodology used here.

Our analysis of four disciplines found that in each of them there is—as expected—a strong or very strong relationship between the composition of journals seen in an institution's REF2014 submission and its published REF2014 Outputs profile results.

Naïvely, one might infer that the REF could therefore be replaced—at least for some disciplines—by a more automated 'algorithmic' assessment that assigns quality ratings based on the journal in which each piece of research is published, rather than on an expert panel's reading of the work itself. However, such an interpretation would not be justified.

As well as the potential for such an algorithmic approach to produce undesirable changes in behaviour, it is important to emphasise two aspects of our analysis. Specifically:

1. The analysis performed here is *retrospective*, not predictive. The question asked, in each discipline, was effectively: if we imagine that the REF2014 panel based its assessments on journal identities alone, then what set of 'journal quality' scores would yield the best match with the actual published REF2014 results? How good would such a 'best match' be? The implied 'journal quality' scores in our analysis came directly from the REF2014 *results*; they were not known in advance by the REF panel, nor were they based on any explanatory covariates other than the journal identities themselves.

2. Although strong correlation was found between REF outcomes and aggregated 'journal quality' scores (see Figures 4, 9, 11 and 13), there is a clear *pattern* of deviation from that relationship, for each of the disciplines studied here. The 'top' institutions are seen typically to do better in REF2014 than their aggregated 'journal quality' scores would suggest; and conversely institutions at the other end of the scale tend to do worse, relative to purely journal-based scores. This indicates that REF assessment panels are in fact doing more than simply using journal identity to determine research quality. This finding is unsurprising: the published remit of REF panels is to *read* the submitted research and evaluate its quality against clearly stated criteria. With that in mind, it is fully to be expected that a diligent REF panel will distinguish the 'best' papers in each journal from those papers that are more ordinary.

It could perhaps still be argued that the relationship between journal-based scores and REF outcomes is sufficiently strong that deviations from it could be ignored, in the interest of reducing the overall cost of the REF exercise. But the clear *pattern* of deviation described in point 2 implies that the resulting redistribution (of research funding, but also prestige) would systematically disadvantage those institutions where predominantly top-quality research is done. While such redistribution of funds might represent a fairly modest fraction of the national funding total, its effects would systematically be concentrated in a few institutions at opposite ends of the scale.

Furthermore, the notion of judging work based on the container in which it is published, rather than on its own merits, seems to miss the point of research assessment entirely. As Traag and Waltman (2019) points out, by relying on metrics, even those which correlate strongly with peer review results, 'the goal of fostering "high quality" science may become displaced by the goal of obtaining a high metric' and have unintended consequences such as 'favouring problematic research methods'.

More pragmatically, there is nothing to say that the esteem of academic journals in 2014 will remain constant until 2021. Editors and authors change and publications can go defunct or start anew in such a long period. Mryglod et al. (2015a, 2015b) already showed that one research assessment exercise cannot necessarily be used to predict the next.

Perhaps the most interesting avenue for future research would be to apply these methods to *all* subject areas in the REF and determine which fields are most beholden to the effect of journals on institutional rankings. Data for all 36 units of assessment in the REF are readily available, and it should be straightforward to apply the methods developed here to those other fields. With the 2021 REF approaching, this could be a topic of interest

to many in academia, publishing and research assessment. As seen in Table 2, however, subjects in the hard sciences tend to submit to scholarly journals more than other fields, such as the arts, who may produce books or artefacts, so the methodology would need to be adapted carefully for such areas, if indeed it can be applied at all.

A   Appendix

Panel	Unit of assessment	Outputs	Jo
A	Clinical Medicine	13400	
	Biological Sciences	8608	
	Public Health, Health Services and Primary Care	4881	
	Psychology, Psychiatry and Neuroscience	9126	
	Agriculture, Veterinary and Food Science	3919	
	Allied Health Professions, Dentistry, Nursing and Pharmacy	10358	
B	Chemistry	4698	
	Earth Systems and Environmental Sciences	5249	
	Aeronautical, Mechanical, Chemical and Manufacturing Engineering	4143	
	Electrical and Electronic Engineering, Metallurgy and Materials	4025	
	Physics	6446	
	General Engineering	8679	
	Civil and Construction Engineering	1384	
	Mathematical Sciences	6994	
	Computer Science and Informatics	7651	
C	Sport and Exercise Sciences, Leisure and Tourism	2757	
	Business and Management Studies	12202	
	Economics and Econometrics	2600	
	Geography, Environmental Studies and Archaeology	6017	
	Education	5519	
	Architecture, Built Environment and Planning	3781	
	Social Work and Social Policy	4784	
	Sociology	2630	
	Politics and International Studies	4365	
	Anthropology and Development Studies	2013	
	Law	5522	
D	Philosophy	2173	
	Area Studies	1724	
	Communication, Cultural and Media Studies, Library and Information Management	3517	
	Modern Languages and Linguistics	4932	
	History	6431	
	Theology and Religious Studies	1558	
	English Language and Literature	6923	
	Music, Drama, Dance and Performing Arts	4246	
	Classics	1386	
	Art and Design: History, Practice and Theory	6321	

Table 2: Units of assessment in REF2014, the number of outputs submitted and the percentage of which that were classified as journal articles

Table 3: Distribution of Economics and Econometrics REF2014 submissions by containing journal (named titles contained $\geq$ 20 submissions)

Volume title	Outputs	%
American Economic Review	104	4.0
The Economic Journal	103	4.0
Journal of Econometrics	93	3.6
Journal of Economic Theory	81	3.1
Games and Economic Behavior	78	3.0
Econometrica	68	2.6
Journal of the European Economic Association	65	2.5
Review of Economic Studies	63	2.4
Economics Letters	62	2.4
Review of Economics and Statistics	58	2.2
Journal of Public Economics	57	2.2
European Economic Review	51	2.0
Economic Theory	48	1.8
Journal of Development Economics	47	1.8
Journal of Economic Dynamics and Control	44	1.7
Journal of Economic Behavior & Organization	42	1.6
Journal of Monetary Economics	42	1.6
Econometric Theory	35	1.3
Journal of International Economics	35	1.3
Journal of Health Economics	33	1.3
Journal of Money, Credit and Banking	32	1.2
Quarterly Journal of Economics	29	1.1
International Economic Review	28	1.1
Oxford Bulletin of Economics and Statistics	28	1.1
Canadian Journal of Economics	25	1.0
Journal of Applied Econometrics	24	0.9
Oxford Economic Papers	24	0.9
Journal of Banking & Finance	23	0.9
Journal of Political Economy	22	0.8
Conference proceedings	2	0.1
Other journals	944	36.3
Other outputs	210	8.1

Table 4: Distribution of Physics REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

| Volume title	Outputs	%
Physical Review Letters	1227	19.0
Monthly Notices of the Royal Astronomical Society	678	10.5
The Astrophysical Journal	393	6.1
Physical Review D	281	4.4
Physical Review B	242	3.8
Journal of High Energy Physics	226	3.5
Nature	207	3.2
Astronomy and Astrophysics	196	3.0
Physics Letters B	189	2.9
Science	175	2.7
Applied Physics Letters	123	1.9
Nature Physics	96	1.5
Physical Review A	92	1.4
The European Physical Journal C	87	1.3
Nature Communications	85	1.3
Journal of Geophysical Research	81	1.3
Optics Express	81	1.3
Proceedings of the National Academy of Sciences	81	1.3
New Journal of Physics	76	1.2
Nano Letters	70	1.1
Nature Materials	65	1.0
Advanced Materials	58	0.9
Physical Review C	55	0.9
Journal of the American Chemical Society	52	0.8
Journal of Cosmology and Astroparticle Physics	46	0.7
Nature Photonics	46	0.7
Monthly Notices of the Royal Astronomical Society: Letters	45	0.7
Nuclear Instruments and Methods in Physics Research	45	0.7
Journal of Instrumentation	41	0.6
Nature Nanotechnology	37	0.6
Advanced Functional Materials	35	0.5
ACS Nano	30	0.5
Journal of Physics: Condensed Matter	30	0.5
The Astrophysical Journal Supplement Series	30	0.5
Conference proceedings	18	0.3
Other journals	1075	16.7
Other outputs	52	0.8

Table 5: Distribution of Mathematical Sciences REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

| Volume title	Outputs	%
Journal of Fluid Mechanics	254	3.6
Physical Review Letters	209	3.0
Journal of High Energy Physics	159	2.3
Communications in Mathematical Physics	140	2.0
Proceedings of the Royal Society A	126	1.8
Advances in Mathematics	116	1.7
Journal of Physics A	112	1.6
Physical Review E	110	1.6
Physical Review D	107	1.5
Journal of Algebra	83	1.2
Proceedings of the London Mathematical Society	70	1.0
The Annals of Probability	70	1.0
Journal of Functional Analysis	66	0.9
Nonlinearity	66	0.9
Transactions of the American Mathematical Society	66	0.9
The Annals of Applied Probability	61	0.9
Biometrika	57	0.8
SIAM Journal on Numerical Analysis	55	0.8
Journal of the London Mathematical Society	54	0.8
International Mathematics Research Notices	52	0.7
Mathematische Annalen	50	0.7
Archive for Rational Mechanics and Analysis	49	0.7
JRSS Series B (Statistical Methodology)	48	0.7
Crelles Journal	47	0.7
JRSS Series C (Applied Statistics)	47	0.7
Physics of Fluids	45	0.6
Journal of Mathematical Physics	44	0.6
Annals of Mathematics	42	0.6
SIAM Journal on Applied Mathematics	42	0.6
Probability Theory and Related Fields	41	0.6
SIAM Journal on Scientific Computing	41	0.6
Stochastic Processes and their Applications	41	0.6
Journal of Differential Equations	40	0.6
Proceedings of the National Academy of Sciences	40	0.6
Bulletin of the London Mathematical Society	39	0.6
Duke Mathematical Journal	38	0.5
Journal of Pure and Applied Algebra	38	0.5
SIAM Journal on Mathematical Analysis	38	0.5
Geometric and Functional Analysis	37	0.5
Journal of Mathematical Biology	37	0.5

Journal of the American Statistical Association	37	0.5
The Astrophysical Journal	37	0.5
Bulletin of Mathematical Biology	36	0.5
Inventiones mathematicae	36	0.5
The Annals of Statistics	35	0.5
Mathematische Zeitschrift	34	0.5
Physica D	34	0.5
European Journal of Operational Research	33	0.5
Nuclear Physics B	33	0.5
Biometrics	32	0.5
Journal of Computational Physics	32	0.5
Journal of Theoretical Biology	32	0.5
Compositio Mathematica	31	0.4
Journal of Statistical Physics	31	0.4
Electronic Journal of Probability	30	0.4
Geometry & Topology	30	0.4
The Annals of Applied Statistics	30	0.4
Conference proceedings	17	0.2
Other journals	3291	47.1
Other outputs	246	3.5

Table 6: Distribution of Chemistry REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

Volume title	Outputs	%
Journal of the American Chemical Society	690	14.7
Angewandte Chemie International Edition	472	10.0
Chemical Communications	258	5.5
Proceedings of the National Academy of Sciences	142	3.0
Chemistry - A European Journal	138	2.9
Physical Chemistry Chemical Physics	126	2.7
Nature Chemistry	119	2.5
The Journal of Chemical Physics	116	2.5
Physical Review Letters	112	2.4
Chemical Science	94	2.0
Science	90	1.9
The Journal of Physical Chemistry C	80	1.7
Dalton Transactions	76	1.6
Inorganic Chemistry	71	1.5
The Journal of Organic Chemistry	70	1.5
Organic Letters	68	1.4
Chemistry of Materials	57	1.2
Organic & Biomolecular Chemistry	55	1.2
Advanced Materials	54	1.1
Analytical Chemistry	53	1.1
Nature	53	1.1
Langmuir	51	1.1
Journal of Materials Chemistry	50	1.1
ACS Nano	49	1.0
The Journal of Physical Chemistry A	49	1.0
Nature Materials	47	1.0
Atmospheric Chemistry and Physics	44	0.9
Soft Matter	43	0.9
The Journal of Physical Chemistry B	43	0.9
Organometallics	41	0.9
Physical Review B	41	0.9
Advanced Functional Materials	38	0.8
Journal of Medicinal Chemistry	37	0.8
Journal of Biological Chemistry	34	0.7
Nano Letters	32	0.7
Nature Communications	31	0.7
Conference proceedings	2	0.0
Other journals	1064	22.6
Other outputs	8	0.2
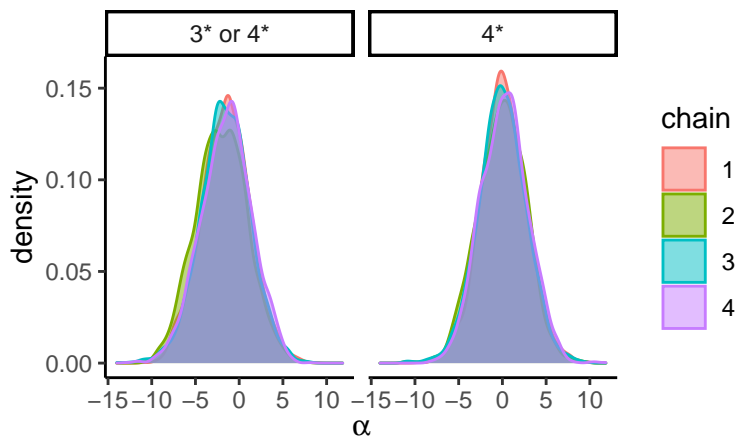
(a) Economics and Econometrics
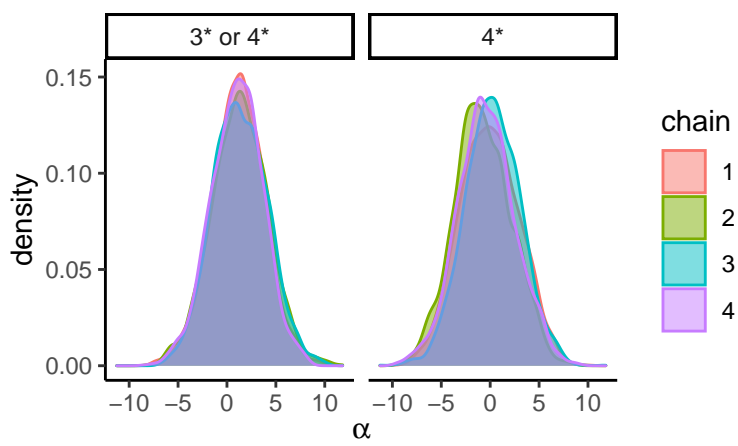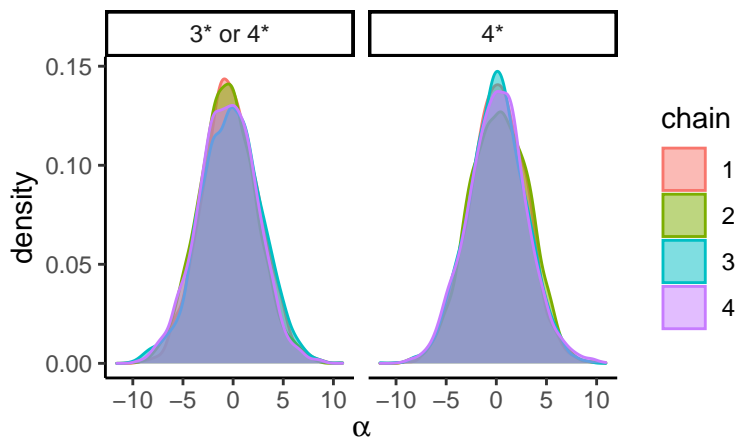
(b) Mathematical Sciences

(c) Physics

46
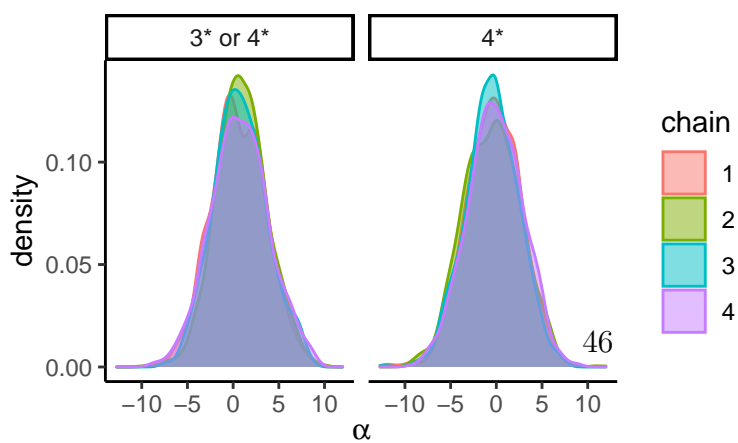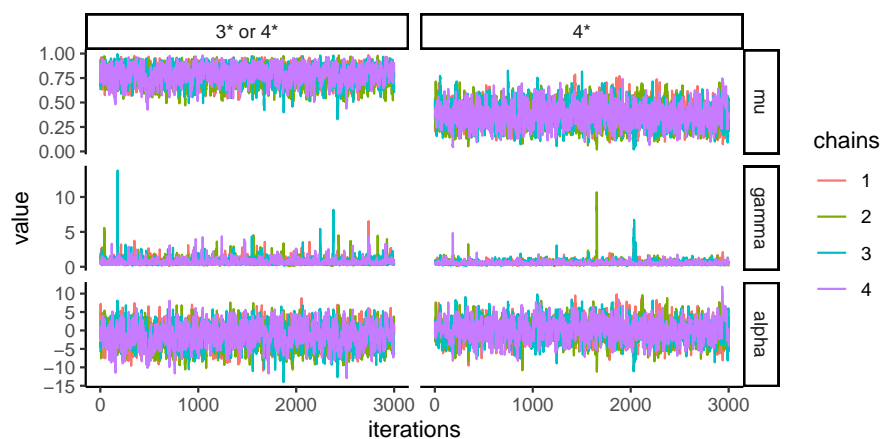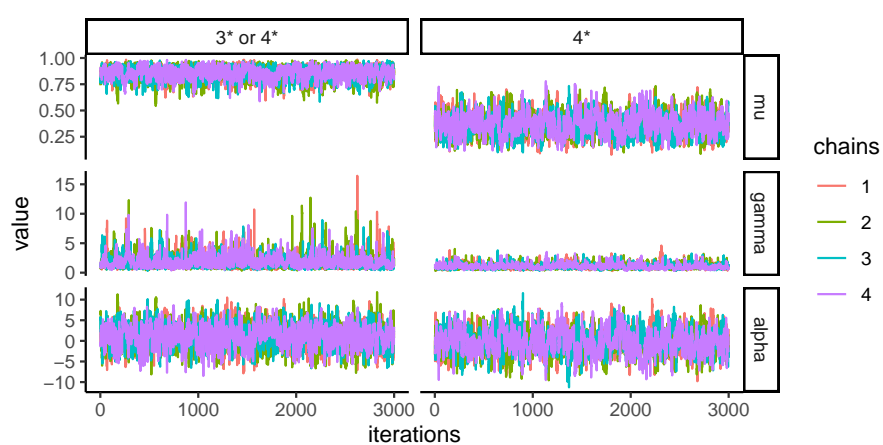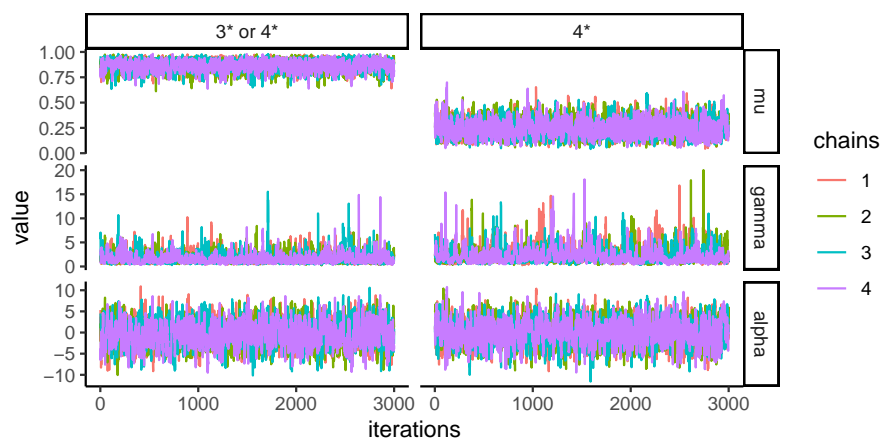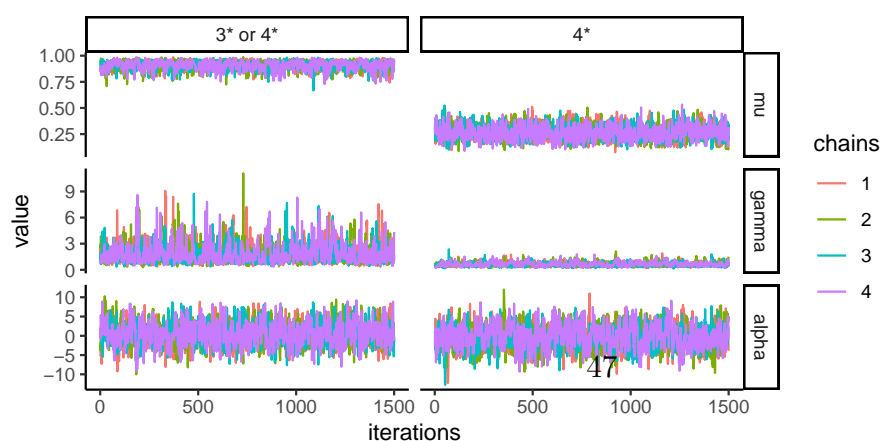
(a) Economics and Econometrics

(b) Mathematical Sciences

(c) Physics

47

(d) Chemistry

Baccini, Alberto, and Giuseppe De Nicolao. 2016. "Do They Agree? Bibliometric Evaluation Versus Informed Peer Review in the Italian Research Assessment Exercise." *Scientometrics* 108 (3): 1651–71. https://doi.org/10.1007/s11192-016-1929-y.

Balbuena, Lloyd D. 2018. "The UK Research Excellence Framework and the Matthew Effect: Insights from Machine Learning." Edited by Lutz Bornmann. *PLOS ONE* 13 (11): e0207919. https://doi.org/10.1371/journal.pone.0207919.

Ball, Robin C. 2019. Personal communication.

Brown, Philip J., and Clive D. Payne. 1986. "Aggregate Data, Ecological Regression, and Voting Transitions." *Journal of the American Statistical Association* 81 (394): 452–60. https://doi.org/10.1080/01621459.1986.10478290.

Bugg, Tim. 2019. Personal communication.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1). https://doi.org/10.18637/jss.v076.i01.

Chamberlain, Scott, Hao Zhu, Najko Jahn, Carl Boettiger, and Karthik Ram. 2019. *rcrossref: Client for Various 'CrossRef' APIs.* https://CRAN.R-project.org/package=rcrossref.

Cho, Wendy K. Tam. 1998. "Iff the Assumption Fits. . .: A Comment on the King Ecological Inference Solution." *Political Analysis* 7: 143–63. https://doi.org/10.1093/pan/7.1.143.

Colquhoun, David, and Andrew Plested. 2014. "Why You Should Ignore Altmetrics and Other Bibliometric Nightmares." *DC's Improbable Science.* http://www.dcscience.net/2014/01/16/why-you-should-ignore-altmetrics-and-other-bibliometric-nightmares/.

Duane, Simon, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22. https://doi.org/10.1016/0370-2693(87)91197-x.

Duncan, Otis Dudley, and Beverly Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18 (6): 665. https://doi.org/10.2307/2088122.

Duncan, Otis Dudley, and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indexes." *American Sociological Review* 20 (2): 210. https://doi.org/10.2307/2088328.

Else, Holly. 2015. "Research Funding Formula Tweaked After REF 2014 Results." *Times Higher Education.* https://www.timeshighereducation.com/news/research-funding-formula-tweaked-after-ref-2014-results/2018685.article.

Flaxman, Seth R., Yu-Xiang Wang, and Alexander J. Smola. 2015. "Who Supported Obama in 2012?" In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15.* ACM Press. https://doi.org/10.1145/2783258.2783300.

Flaxman, Seth, Dougal Sutherland, Yu-Xiang Wang, and Yee Whye Teh. 2016. "Understanding the 2016 US Presidential Election Using Ecological Inference and Distribution

Regression with Census Microdata," November. http://arxiv.org/abs/1611.03787v1.

Fog, Agner. 2015. *BiasedUrn: Biased Urn Model Distributions.* https://CRAN.R-project.org/package=BiasedUrn.

Franceschet, Massimo, and Antonio Costantini. 2011. "The First Italian Research Assessment Exercise: A Bibliometric Perspective." *Journal of Informetrics* 5 (2): 275–91. https://doi.org/10.1016/j.joi.2010.12.002.

Freedman, D. A., M. Ostland, M. R. Roberts, and S. P. Klein. 1999. "Reply to G. King." *Journal of the American Statistical Association* 94 (445): 355–257. https://doi.org/10.1080/01621459.1999.10473849.

Goodman, Leo A. 1953. "Ecological Regressions and Behavior of Individuals." *American Sociological Review* 18 (6): 663. https://doi.org/10.2307/2088121.

———. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64 (6): 610–25. https://doi.org/10.1086/222597.

Groen-Xu, Moqi, Pedro Teixeira, Thomas Voigt, and Bernhard Knapp. 2017. "Short-Termism in Science: Evidence from the UK Research Excellence Framework." *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3083692.

Heckman, James, and Sidharth Moktan. 2018. "Publishing and Promotion in Economics: The Tyranny of the Top Five." Technical report. National Bureau of Economic Research; National Bureau of Economic Research. https://doi.org/10.3386/w25093.

Hill, Steven. 2017. "A Paper Exploring REF, Funding and Journal Metrics." http://stevenhill.org.uk/a-paper-exploring-REF-funding-and-journal-metrics/.

Hirsch, J. E. 2005. "An Index to Quantify an Individuals Scientific Research Output." *Proceedings of the National Academy of Sciences* 102 (46): 16569–72. https://doi.org/10.1073/pnas.0507655102.

Hole, Arne Risa. 2017. "Ranking Economics Journals Using Data from a National Research Evaluation Exercise." *Oxford Bulletin of Economics and Statistics* 79 (5): 621–36. https://doi.org/10.1111/obes.12185.

Hong, Yili. 2013. "On Computing the Distribution Function for the Poisson Binomial Distribution." *Computational Statistics & Data Analysis* 59 (March): 41–51. https://doi.org/10.1016/j.csda.2012.10.006.

Jannarone, Robert J., Kai F. Yu, and James E. Laughlin. 1990. "Easy Bayes Estimation for Rasch-Type Models." *Psychometrika* 55 (3): 449–60. https://doi.org/10.1007/bf02294760.

Jump, Paul. 2013. "'Game' of One-Fifth? Part-Time Contracts Rise in Run-up to REF." *Times Higher Education*, no. 2120 (September): 6. https://www.timeshighereducation.com/news/twenty-per-cent-contracts-rise-in-run-up-to-ref/2007670.article.

King, Gary. 2013. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press.

King, Gary, Ori Rosen, and Martin A. Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research* 28 (1): 61–90.

https://doi.org/10.1177/0049124199028001004.

Koya, Kushwanth, and Gobinda Chowdhury. 2017. "Metric-Based Vs Peer-Reviewed Evaluation of a Research Output: Lesson Learnt from UK's National Research Assessment Exercise." Edited by Pablo Dorta-González. *PLOS ONE* 12 (7): e0179722. https://doi.org/10.1371/journal.pone.0179722.

Kuha, Jouni, and David Firth. 2011. "On the Index of Dissimilarity for Lack of Fit in Loglinear and Log-Multiplicative Models." *Computational Statistics & Data Analysis* 55 (1): 375–88.

Loeffler, David. 2019. Personal communication.

MacRoberts, Michael H., and Barbara R. MacRoberts. 2017. "The Mismeasure of Science: Citation Analysis." *Journal of the Association for Information Science and Technology* 69 (3): 474–82. https://doi.org/10.1002/asi.23970.

Mansfield, Neil J. 2016. "Ranking of Design Journals Based on Results of the UK Research Excellence Framework: Using REF as Referee." *The Design Journal* 19 (6): 903–19. https://doi.org/10.1080/14606925.2016.1216212.

Marques, Marcelo, Justin JW Powell, Mike Zapp, and Gert Biesta. 2017. "How Does Research Evaluation Impact Educational Research? Exploring Intended and Unintended Consequences of Research Assessment in the United Kingdom, 1986–2014." *European Educational Research Journal* 16 (6): 820–42. https://doi.org/10.1177/147490411773 0159.

McCullagh, P., and John A. Nelder. 1989. *Generalized Linear Models*. Second Edition. Taylor & Francis Ltd.

Monroe, Randall. 2008. "xkcd: Purity." https://xkcd.com/435/. https://xkcd.com/435/.

Mryglod, Olesya, Ralph Kenna, Yurij Holovatch, and Bertrand Berche. 2015a. "Predicting Results of the Research Excellence Framework Using Departmental h-Index." *Scientometrics* 102 (3): 2165–80. https://doi.org/10.1007/s11192-014-1512-3.

———. 2015b. "Predicting Results of the Research Excellence Framework Using Departmental H-Index: Revisited." *Scientometrics* 104 (3): 1013–7. https://doi.org/10.1007/ s11192-015-1567-9.

Oswald, Andrew J. 2019. Personal communication.

———. 2007. "An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers." *Economica* 74 (293): 21–31. https: //doi.org/10.1111/j.1468-0335.2006.00575.x.

Poisson, S. D. 1837. *Recherches Sur La Probabilité Des Jugements En Matière Criminelle et En Matière Civile*. Paris: Bachelier.

Reed, Mark, and Simon Kerridge. 2017. "How Much Was an Impact Case Study Worth in the UK Research Excellence Framework?" *Fast Track Impact* 1 (January): 47–49. https://www.fasttrackimpact.com/single-post/2017/02/01/How-much-was-an-impact-case-study-worth-in-the-UK-Research-Excellence-Framework.

REF. 2019. "What Is the REF?" https://www.ref.ac.uk/about/what-is-the-ref/. https://www.ref.ac.uk/about/what-is-the-ref/.

———. 2015. "REF Frequently Asked Questions." https://www.ref.ac.uk/2014/faq/.

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The R×C Case." *Statistica Neerlandica* 55 (2): 134–56. https://doi.org/10.1111/1467-9574.00162.

Rosenman, Evan. 2019. "Some New Results for Poisson Binomial Models," July. https://arxiv.org/abs/1907.09053.

Rosenman, Evan, and Nitin Viswanathan. 2018. "Using Poisson Binomial GLMs to Reveal Voter Preferences," February. https://arxiv.org/abs/1802.01053v1.

Scott, Peter. 2019. Personal communication.

Sgroi, Daniel. 2019. Personal communication.

Shah, B. K. 1973. "On the Distribution of the Sum of Independent Integer Valued Random Variables." *The American Statistician* 27 (3): 123–24. https://www.jstor.org/stable/2683639.

Stan Development Team. 2018. "RStan: The R Interface to Stan." http://mc-stan.org/.

Stockhammer, Engelbert, Quirin Dammerer, and Sukriti Kapur. 2017. "The Research Excellence Framework 2014, Journal Ratings and the Marginalization of Heterodox Economics." Research report 1715. Post-Keynesian Economics Society. http://postkeynesian.net/working-papers/1715/.

Szabó, Zoltan, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. 2014. "Learning Theory for Distribution Regression." *Journal of Machine Learning Research* 17 (152): 1–40. http://arxiv.org/abs/http://arxiv.org/abs/1411.2066v4.

Traag, V. A., and L. Waltman. 2019. "Systematic Analysis of Agreement Between Metrics and Peer Review in the UK REF." *Palgrave Communications* 5 (1). https://doi.org/10.1057/s41599-019-0233-x.

Varin, Cristiano, Manuela Cattelan, and David Firth. 2016. "Statistical Modelling of Citation Exchange Between Statistics Journals." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179 (1): 1–63.

Wakefield, Jon. 2005. "Ecological Inference for 2×2 Tables (with Discussion)." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167 (3): 385–445. https://doi.org/10.1111/j.1467-985x.2004.02046.x.

Wang, Yuan H. 1993. "On the Number of Successes in Independent Trials." *Statistica Sinica* 3 (2): 295–312.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, et al. 2015. "The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management." https://responsiblemetrics.org/the-metric-tide/; Unpublished. https://doi.org/10.13140/rg.2.1.4929.1363.

Yan, Zizhong. 2017. "How Does the REF Panel Perceive Journals? A New Approach to Estimating Ordinal Response Model with Censored Outcomes." PhD thesis, Department

of Economics, University of Warwick. http://wrap.warwick.ac.uk/91064.