Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v19n1p235

Adaptive Neural Networks for predicting Research Excellence Framework results
By Andria et al.

March 15, 2026

Electronic Journal of Applied Statistical Analysis
Vol. 19, Issue 01, March 2026, 235-254
DOI: 10.1285/i20705948v19n1p235

Adaptive Neural Networks for predicting Research Excellence Framework results

J. Andria *[a], G. di Tollo[b], S. Cruz Rambaud[c], and M.Squillante[d]

[a] *Università degli Studi di Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, Viale delle Scienze, 90128 Palermo IT*
[b] *Università Politecnica delle Marche, Dipartimento di Management , Piazzale Martelli Raffaele, 8, 60121 Ancona It*
[c] *University of Almería, Department of Economics and Business, La Cañada de San Urbano, Almería, Spain, ES*
[d] *Accademia Peloritana dei Pericolanti, Palazzo Università - Piazza Pugliatti, 1 Messina, 98122 IT*

March 15, 2026

In many countries, the task of evaluating high-quality scientific research is performed by public agencies that produce an assessment of the activities carried out by universities and other High Education Providers (HEPs). This evaluation process is crucial as it constitutes the main basis of allocating national financial resources among HEPs.
In our contribution, we are using neural networks for predicting the *next* UK Research Excellence Framework exercise (REF) outcomes, i.e., a HEP's rank and its quality-related research recurrent funding, and comparing our results with respect to other linear models in literature. We show that our approach lead to accurate results and success in predicting the outcomes of the UK's research evaluation exercise. The benefits of predicting quality research are noteworthy and involve alike national funding agencies, educational institutions and researchers.

1 Introduction

Policy makers distribute funds to universities and other High Education Providers (HEPs) on the basis of evaluations performed by national agencies. These supply comparative

*Corresponding author: joseph.andria@unipa.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

information regarding the research quality (along with other parameters) at HEPs. The evaluation task is long, time-consuming, and strongly relies on peer review processes. In this context, universities and other HEPs are more and more interested in modeling and predicting the outcome of the whole evaluation process, since an accurate prediction could lead them to understand their pro and cons, to define appropriate strategies and, eventually, to improve their performance and thus increase the amount of the granted public funds. To this extent, in literature some studies proposed quantitative approaches: Mryglod et al. (2015a) and Mryglod et al. (2015b) performed a correlation analysis between departmental h-indices and the outcome of the 2008 Research Assessment Exercise (RAE) in the UK, proposing to use the departmental h-index to predict the outcome of the next research assessment exercise, but failing to obtain good results. Basso and di Tollo proposed generalised linear models Basso and di Tollo (2017) and linear models Basso and di Tollo (2022) that use, as predictors, the previous research assessment and the departmental h-indices. Although they obtained good results, their models, as we will show in the following, are affected by over-fitting, that hinders their suitability for real life applications.

The task of effectively predicting the outcomes of performance-based research funding systems, throughout the involved complementary targeted sub-objectives (if any), is not only desirable, but even necessary in order to let an educational institution implement an impactful policy strategy without unanticipated side effects Banal-Estañol et al. (2023). Based on the previous considerations, this contribution introduces a neural network approach to predict the *next* research assessment exercise. Neural Networks are algorithms that mimic the behaviour of the human brain to detect non-linear relationships and to perform complex tasks and/or predictions, showing good *generalization* capabilities, and, if trained properly, decreasing the risk of over-fitting. Furthermore, since neural networks' performances can turn to be sensitive to their parameter values, we have resorted to a dynamic parameter setting procedure in order to select the most appropriated values of the introduced parameters.

In this framework, the UK's public research funding system, along with other countries' research-boosting policies, aims at improving national research performances in both term of quality and quantity Jonkers and Zacharewicz (2016); Geuna and Piolatto (2016), based on the idea to apply a rewarding and motivating scheme for allocating scarce funds between competing institutions. Hence, we have applied our approach to data from the UK Research Excellence Framework (REF), also comparing our results with those in Basso and di Tollo (2022) and with those obtained by a classical neural network. Compared to the state-of-the-art, our approach shows better performances, supporting the idea that neural networks can be effective in predicting the result of a research evaluation exercise when coupled with optimal parameter settings and well-defined data pre-processing routines. The optimized neural networks were then tested to forecast the results of the next research assessment, in order to obtain some insights about their generalisation effectiveness. We finally tested the reliability of our approach to predict the next exercise research-funded grants: this objective of predicting the next Research Excellence exercise is the ultimate goal of every HEP, and the use of an ineffective and unreliable evaluation model may crucially affect both the mission and the

governance of a HEP.

Our paper is structured as follows: Section 2 describes the problem at hand along with the related literature, while Section 3 outlines the problem instances; these instances will be tackled in the experimental analysis whose components will be detailed in Section 4, and whose performances will be analysed in Section 5. Last, we are using our approach to forecast the results of the next evaluation exercise and to test the estimation reliability of the predicted granted fund in Section 6, before concluding with Section 7.

2 Related Literature

There is an ongoing debate about using bibliometric indicators in the research evaluation carried out by sovereign institutions, instead of implementing costly and time-consuming peer-review procedures. Bibliometric indicators are computed based on the amount of citations received by a paper, which makes their computation deterministic and easy to understand: they represent a fair measure of a researcher (or institution) excellence and impact (Burger et al. (1985)), but they come from different databases, and often make difficult to compare researchers and institutions from different disciplines (MacRoberts and MacRoberts (1989)), on top of the fact that citations accumulate too slowly to be used in a short-term research assessment (Bornmann and Leydesdorff (2014)). For these reasons some scholars discourage their use for research assessment (Lawrance (2003)).

On the other hand, peer-review greatly relies on experts assessments (Bertocchi et al. (2011)), and although it is universally seen as the most effective tool to assess the quality of academic research (Evidence (2010)), it presents some shortcomings (i.e., the Hawthorne effect reported by Bornmann (2012)). Traditionally, research evaluation has been mainly made by peer-review; nevertheless, more and more relevance is given to bibliometric indicators: for instance in the UK, the RAE evaluation carried out in 2008 formally forbade to use citations data by panels and resorted to peer-review only, while the following REF 2014 included bibliometrics as a tool to be used by panels (Sgroi and Oswald (2013)). Even if the role of bibliometrics is still secondary with respect to peer-review, we can state that its use in research evaluation is increasing, and that in the literature one can find more and more contributions aimed at comparing the outcome of a peer-review based evaluation system with bibliometric indicators only.

Each country has its own rules and procedures to combine both peer-review and bibliometric analysis: for example, the Italian Evaluation of Research Quality (VQR) must peer-review at least 50% of the submitted research, though in the humanities panel this value raises to 100% (Bertocchi et al. (2011)); the Australian Government's Excellence in Research for Australia (ERA) uses bibliometric indicators for natural sciences and peer-review in social sciences (Bruns and Stern (2015)). Providing an annotated bibliography of this aspect is out of the scope of our contribution, and we forward the interested reader to Basso and di Tollo (2022).

In Thelwall (2025), the authors discuss on Large Language Models, LLMs, for automated scholarly paper review and to what extent they correlate with REF scores. In

Inglis et al. (2024), the authors analyse the full text of all journal articles returned to the education subpanel of the 2021 Research Excellence Framework (REF2021). They found that their predicted Grade Point Aaverages, GPAs, were strongly correlated with the scores assigned by the REF2014 subpanel. In Thelwall and Yaghi (2024), authors investigate on whether ChatGPT 4o-mini can be used to estimate the quality of journal articles across academia. They sample up to 200 articles from all 34 Units of Assessment (UoAs) in the UK's Research Excellence Framework (REF) 2021 and compare ChatGPT scores with departmental average scores.

3 Our Data

In this contribution we are tackling the case study provided by the analysis of the UK's Research Excellence Framework, in which HEFCE[1] defines (and updates) a formula for ranking universities and other HEPs on the basis of a performance-based approach across different disciplines (dedicated panels), each of which linked to a specific unit of assessment (*UoA*s: engineering, sociology, biology etc.). Each researcher is invited to submit four research products, which are then evaluated with respect to three distinct criteria, i.e., *Outputs*, which relates to the originality, significance and rigour compared to international research quality standards; *Impact*, which relates to the positive spillover of research outside of the academic context, and *Environment*, which concerns the research environment in terms of its "vitality and sustainability". For each of these criteria, all products are assigned to one out of five rating classes: the highest one is labeled 4^* and represents the world leading research; the others are labeled 3^*, 2^*, 1^* and *Unclassified*, in decreasing order of relevance. On the basis of these evaluations, for each criterion, the percent values of the products submitted belonging to the five rating classes are computed to produce O_i^* (Output), I_i^* (Impact) and E_i^* (Environment), where i^* corresponds to each of the five-point star scale, $4^*, 3^*, 2^*, 1^*$ and $Unclassified$; then, the results for all criteria are weighted to produce the global percentages p_4, p_3, p_2, p_1, $p_{Unclassified}$: each criteria accounts, respectively, for 60%, 25% and 15% of the overall percentage of research activity $p_{i^*,REF}$, which is computed as:

$$p_{i^*} = 0.60O\_i^* + 0.25I\_i^* + 0.15E\_i^*. \tag{1}$$

From Eq. (1), the quality-weighted volume (QWV) of each HEP is obtained by:

$$QWV = FTE \times (p_{4^*} \times 4 + p_{3^*}), \tag{2}$$

where, FTE stands for the Full-Time Equivalent of submitted staff and the term in parenthesis stands for the overall REF 2020 score. HEP's total granted QR mainstream is then obtained by multiplying the result from Eq.(2) by the amount of funding per unit of quality-weighted volume for the whole UoA.

[1] *Higher Education Funding Council for England*, `htpp://hefce.ac.uk/`.

Please notice that this formula may change over time: at the time of RAE 2008 the score s'_{RAE} was computed as

$$s'_{RAE} = p_{4,RAE} + \frac{3}{7}p_{3,RAE} + \frac{1}{7}p_{2,RAE} \tag{3}$$

where $p_{i,RAE}$ represents the percent value of research belonging to class i^*, $i \in [1 \dots 4]$; REF 2014, instead, defines the formula

$$s_{REF} = p_{4,REF} + \frac{1}{3}p_{3,REF} \tag{4}$$

which is, along with the REF2020 score, the dependent variable to be predicted by our approach.

Please notice that the aforementioned percent values for RAE and REF are publicly available, and that the formers are used in our study as predictors. On top of them we dispose, for every year in the period $2008-2014$, of the departmental h-index of all HEPs in the sample, that was kindly provided to us by Olesya Mryglod and will be also used in the predictors set. The goal of our prediction exercise is to assess the relationship, if any, between the set of predictors (h-indices and $p_{i,RAE}$) and the score s_{REF} used to allocate funds.

With respect to the predictor set, we want to point out a great element of novelty compared to the study in Basso and di Tollo (2022): starting from the consideration that it is not desirable using many potential predictors when only small sample sizes are available, they have not used all yearly h-indices, but rather the average yearly h-index variation over the period $2008 - 2014$, jointly with the initial h-value (h_{2008}). In our contribution instead, based on the generalization skills of neural networks, and their good performances over small samples, we are using all yearly h-indices together without any assumption of linearity nor data loss.

As for the sets of data, we have used those in Mryglod et al. (2015a) and Mryglod et al. (2015b), which has been kindly provided us by the authors, and that consist of the following *UoA*s: biology (39 Universities and other HEPs), sociology (29 Universities and other HEPs), chemistry (34 Universities and other HEPs), and physics (33 Universities and other HEPs): these UoAs have been selected because they have been included in the list of UoAs for both RAE and REF. Table 1 shows the main statistics of the sample used.

4 Our Neural Network Approach

Artificial Neural Networks (ANN) are Artificial Intelligence based algorithms that mimic and simplify the behavior of the human brain (Haykin (2009)): they are used to solve problems stemming from many fields, and they are particularly useful in situations in which there are no assumptions about the relations amongst variables. They are composed of elementary units (i.e., neurons), connected to (some of) each other by

Table 1: Summary statistics: $p_{i^*}^{'08}$ and $p_{i^*}^{'14}$ are, respectively, the five-point star scale global percentages for the years '08 and '14.

		h-index							$p_{i^*}^{'08}$					$p_{i^*}^{'14}$				
		'08	*'09*	*'10*	*'11*	*'12*	*'13*	*'14*	4*	3*	2*	1*	*Un*	4*	3*	2*	1*	*Un*
Sociology	Mean	21.90	26.55	31.61	36.71	42.06	47.55	51.90	22.26	29.84	32.90	13.55	1.45	19.14	48.20	30.84	1.58	0.25
	Std dev	6.22	7.60	8.98	10.56	11.86	13.41	15.13	9.47	5.84	4.96	4.51	2.31	4.90	6.75	6.86	2.00	0.57
	Skew	-0.53	-0.51	-0.57	-0.41	-0.45	-0.48	-0.42	0.08	0.33	0.53	-0.24	0.97	0.30	-0.55	-0.17	2.24	2.48
	Kurt	0.47	0.25	0.10	-0.15	-0.15	-0.08	-0.09	-1.01	-0.82	0.84	-0.57	-1.13	-0.43	1.27	-1.35	6.54	5.65
Physics	Mean	44.03	50.97	57.55	63.52	68.97	73.58	77.48	15.65	39.68	35.97	8.55	0.16	20.18	66.99	12.15	0.48	0.21
	Std dev	15.14	17.73	20.27	22.28	24.47	25.93	27.29	6.02	4.46	7.46	3.91	0.90	5.51	6.85	7.68	0.72	0.33
	Skew	0.88	0.86	0.95	1.02	1.07	1.06	1.05	-0.63	0.13	1.19	1.02	5.57	-0.32	-1.18	1.37	1.75	1.40
	Kurt	0.56	0.45	0.67	0.87	1.04	1.05	1.00	0.31	-0.23	2.51	0.95	31.00	0.71	1.73	2.17	2.61	0.85
Chemistry	Mean	37.93	44.71	51.64	58.32	64.86	70.61	75.86	13.57	45.54	37.50	3.39	0.00	19.75	70.34	9.62	0.06	0.23
	Std dev	10.78	13.24	15.22	17.53	20.01	22.06	23.39	9.70	9.16	12.44	5.28	0.00	11.45	7.87	8.49	0.21	0.85
	Skew	-0.15	-0.01	0.05	0.07	0.29	0.26	0.17	0.89	-0.60	0.23	2.74	–	0.67	-0.83	1.36	3.75	4.80
	Kurt	0.12	0.04	0.18	0.24	0.39	0.35	0.33	0.63	0.05	-0.70	9.96	–	0.15	0.57	1.03	14.78	24.04
Biology	Mean	58.79	68.46	78.68	88.82	98.14	107.07	115.21	13.57	39.64	35.71	9.64	1.43	27.09	52.77	18.09	0.77	1.28
	Std dev	18.53	21.30	24.16	27.39	30.44	33.07	35.35	7.80	5.92	7.90	4.89	2.30	10.55	6.62	7.61	0.87	1.41
	Skew	0.74	0.74	0.73	0.72	0.80	0.84	0.86	1.59	0.29	-0.74	0.41	1.00	0.20	0.15	0.02	1.08	1.14
	Kurt	0.87	0.81	0.74	0.58	0.85	1.03	1.10	3.60	0.69	0.06	-0.14	-1.08	-0.13	-0.49	-0.27	0.41	0.48

weighted edges (i.e., synapses): each synapse is associated to a continuous value, and each neuron receives as input the weighted sum of the values sent by its neighboring neurons. This weighted sum is then processed through a specific activation function to get the activation value which is sent, through synapses, to the connected neurons. Synapses' values are set by the learning algorithm, which can be seen as the procedure to modify the synapses' weights in order to obtain the desired outputs. Amongst the learning algorithms, Backpropagation (Werbos (1994)) is still the most used, and is based on the back-propagation of the difference between the network output and the actual output of the sample, from the output to the input neurons.

Neurons are organized in topologies, the most common of which are the *layered* (featuring neurons organised in layers, in which each neuron is connected to neurons of adjacent layers) and the *completely connected* (featuring all neurons connected to each other).

In what follows, we are describing the main components of our neural network approach: the pre-processing operations; the procedure to partition data into *training* and *test* sets; the network topology, and the learning parameter.

Data Pre-Processing A deep data analysis is important to detect eventual similarities and anomalies, and to preserve the most information as possible. We have implemented the pre-processing operations used in Angelini et al. (2008) and Corazza et al. (2021), i.e.:

- **Removal and replacement** Missing and/or wrong values are often found in sets of data coming from real-world scenarios, and some procedures have to be devised to tackle this aspect Angelini et al. (2008). Differently from Basso and di Tollo (2022), that remove all entries with missing data, we have followed the guidelines indicated in Angelini et al. (2008) and Corazza et al. (2021) and replaced the missing values by the variable's average over all HEPs: this led to preserve the most information as possible, that will be useful in the training phase.

- **Normalization** Data has to be normalised so that its values belong to the same interval over the different variables. Many mathematical formulations have been suggested to this aim (Khashman (2010)), and we are using the logarithmic normalization in di Tollo et al. (2015), di Tollo et al. (2014), and di Tollo et al. (2012), in which the relation between the post-normalisation $\overline{x}_i$ and pre-normalisation x_i values is defined as follows:

$$\overline{x}_i = \log_u (x_i + 1) \tag{5}$$

  where $u = x_{\max} + 1$, in order to have $\overline{x}_i \in [0, 1]$. Please notice that this transformation leads to have the departmental h-index always in the same rage, completely eliminating the problem arising from the fact that the h-index may increase due to the effect of time rather than of an increase of the scientific quality. This makes linear models inappropriate when dealing with actual h-index values unless an

appropriate corrective action is taken, as detailed in section 6. Actually, this is another hint towards the good generalisation power of neural networks; that is, they do not need extra-actions after the experimental setting is properly defined.

Training and Test Set As for the training-test partition, we have sampled two disjoint sets of observations out of the total number of HEPs belonging to the each specific UoAs: the training set (used for the learning phase) and the test set (used for stopping the learning phase and assessing the network performances). Since we are dealing with small-sized data, HEPs have been randomly allocated to these two sets to have the training set consisting of 70% of whole set (and the test set of 30%, accordingly). This sampling has been repeated 30 times for each UoA, each time leading to a different definition of the training and test sets, and for each time the adopted performance measures have been computed over the different training-test partitions.

Network Topology and Learning Parameters As for the network topology, we have used the classical feed-forward architecture, in which the most important parameters to be set are the number of hidden neurons and the number of hidden layers. To this extent, we have used the adaptive procedure defined in Corazza et al. (2021) to adaptively select the best network topology w.r.t. the performance measures. Furthermore, Back-Propagation requires the user to pre-define the learning and momentum parameters, but according to several authors there is no way of determining these parameters a-priori Reed and Marks (1998). Hence, we have used the generic parameter tuning procedure F-Race Birattari et al. (2010), whose joint application with the aforementioned adaptive procedure, is apt to define a robust approach that can be instantiated at every algorithm's execution.

Supervised learning has been applied on the training set, but the termination criterion has been evaluated on the test set in order to avoid overfitting. To this goal, several error measures can be used, such as the mean absolute error (MAE), the mean absolute percentage error (MAPE), the root mean square error (RMSE), etc. In our experiments, we have used the mean square error (MSE), defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i - a_i)^2, \tag{6}$$

where n is the number of observations, e is the expected output and a is the actual neural network output. From an operational point of view, neural networks have been implemented in Python, and the experiments have been performed on a cluster with AMD Opteron 2216 dual core CPUs running at 2.4 GHz with 2x1 MB L2 cache and 4 GB of RAM under Cluster Rocks distribution built on CentOS 5.3 Linux. The average execution time was between 22 and 57 seconds for each run.

4.1 The applied adaptive network algorithm

The task of designing the topology of a network is neither an automatic nor a simple operation, especially if one considers that no unique criteria exists to determine the number and the size of hidden layers. In particular, the number of hidden neurons is set by means of an adaptive procedure aimed at minimizing the following misclassification error metric calculated for every data set:

$$MIS = \frac{\sum_{i=1}^{n} I\{(\hat{\delta}s_{REF} = 0 \wedge \delta s_{REF} = 1) \vee (\hat{\delta}s_{REF} = 1 \wedge \delta s_{REF} = 0)\}}{\sum_{i=1}^{n} I\{(\delta s_{REF} = 1) \vee (\delta s_{REF} = 0)\}}, \qquad (7)$$

where $I(\cdot)$ is an indicator function that equals 1 when the argument is true, 0 otherwise while $\hat{\delta}s_{REF}$ and δs_{REF} are, respectively, the predicted and observed direction of change score over the period 2008-20014. For practical purposes, we have arbitrary associated the value 1 to an increasing change score whereas 0 stands for a negative sign change.
In detail, the network is initially constructed with one neuron in the hidden layer and, by an iterative procedure, one more neuron is added until the last K iterations error as in Eq.(7) remains unchanged (or no significant improvement is observed).
Obviously, the same adaptive procedure could be applied to choose the number of hidden layers, but here, following Corazza et al. (2021), we set the maximal number of hidden layers equal to two. For the sake of the reader, we report the flowchart of the proposed algorithm in Fig. 1.

5 Algorithm Performances

In this section we are going to describe the results of our experimental analysis. As a first analysis, Table 2 (a) reports the R^2 obtained by the linear and a log-linear models as in Basso and di Tollo (2022): for each couple of UoA and model, they have reported the R^2 along with the corresponding AIC and F-test, showing satisfactory performances over all instances except *physics*. We have performed a more in-depth analysis of their results by computing the corresponding *predicted* R^2, which conveys interesting information about the over-fitting of a model, which occurs whenever its value is significantly lower than the corresponding pure R^2. We can remark that this is the case for all instances, and this result fully justify the introduction of our neural network approach. Of course, the computed *predicted* R^2s are relative to the whole set of data for each instance, whilst in our case, as described in section 4, each dataset splits into 30 different partitions of *training* and *test* sets. Table 2 (b) reports the R^2 average over the 30 *test* sets by applying both our adaptive approach (adaptive topology and parameter tuning) and a classical neural network approach whose parameters are user-defined before the run. For each partition, the algorithm has been run 30 times, and the best run (with respect to the MSE) has been recorded. Please notice that we do not conduct the F-test for our neural network approach, since there is no guarantee that F-statistics follows an F-distribution under the null hypothesis in our experimental settings. Also, we have not computed the AIC, since the devised procedure makes the network topologies (hence, the network's

Table 2: Predicting s_{REF}: significance statistics of: (a) models by Basso and di Tollo (2022); (b) our network approach and a classical neural network approach. Table (a) refers to statistics on the whole set of data; table (b) refers to the average statistics on 30 different train-test partitions.

(a)

	Linear Model				Log-linear Model			
UoA	biology	chemistry	physics	sociology	biology	chemistry	physics	sociology
R^2	0.35	0.61	0.18	0.69	0.38	0.60	0.21	0.65
Predicted R^2	0.10	0.35	-0.14	0.61	0.12	0.36	-0.15	0.46

(b)

	Adaptive Neural Network				Classical Neural Network			
UoA	biology	chemistry	physics	sociology	biology	chemistry	physics	sociology
R^2	0.41	0.44	0.55	0.18	0.27	0.38	0.11	0.32
Predicted R^2	0.39	0.39	0.43	0.15	0.15	0.21	-0.03	0.13

parameters) vary over the different instances and executions, making difficult to penalize the log-likelihood by the complexity of the model.

It is interesting to remark that the classical neural network approach obtains results that are even worse than the ones in Basso and di Tollo (2022), even though the standard deviations are kept low. By means of our Adaptive Neural Network approach we instead obtain R^2 values that are higher than those in Basso and di Tollo (2022) over two out of four instances, at the cost of standard deviations that are comparable to the classical approach. Anyhow, it can be seen that the difference between the predicted and pure R^2 are always kept low, indicating that neural networks are less affected by over-fitting and that they can be successfully applied to predict research quality evaluations.

6 Generalization: Predicting the Results of the Next REF

Basso and di Tollo (2022) applied the implemented regressive models for REF also to forecast the outcome of the next research evaluation exercise, that was supposed to be held in 2020. This relied on the fact that one could insert in the predictor set $p_{i,REF}$ $(i \in \{1 \dots 4\})$ (instead of the formerly used $p_{i,RAE}$ $(i \in \{1 \dots 4\})$) and h-2014 (instead of the formerly used h-2008), and assign them the coefficients determined in the estimation phase in the resulting model. Unfortunately, this way of operating does not take into account the fact that the h-index is cumulative, hence it follows an increasing trend over time (Mryglod et al. (2015a)), and that a h-2014 value bigger than h-2008 could be partly due to the time-delay effects and not completely attributable to an improvement of the HEP's research quality. Hence, they have proposed to neutralise the average cumulative effect of time by subtracting the average h-index increase observed over the 6-year period 2008–2014 (the average being computed over all the HEPs of the UoA under

examination) to the value of h-2014. As stated in Section 4, our neural network approach does not need this kind of procedure, since input values are normalised according to pre-processing operations: this is a further proof of robustness of our approach. Hence, we have used the implemented neural network, whose results are outlined in table 2 (b), to forecast the 2020 s_{REF} exercise. An instance of the obtained forecasted results is shown in Tab. 3, which exhibits for *Phisics UoA* the expected direction of change of the 2020 assessments with respect to those of the 2014 REF. " $\uparrow$ " indicates that a higher (i.e., better) value is obtained, while " $\downarrow$" indicates that a lower (i.e., worse) score is obtained than that achieved in the previous REF exercise. Results reported in Table 3 show that the two approaches lead to comparable forecastings, with on average higher scores than those observed in the previous exercise. More detailed results are given in Table 4, where we reported the predicted *Physics UoA* scores by different methods: linear (L) and Log-Linear (LL) as in Basso and di Tollo (2022) and our Classical (CNN) and Adaptive (ANN) Neural Network approach. Results show that our Adaptive Neural Network leads to the best performances on 12 out of 33 $HEPs$. Interestingly, results obtained using the linear model as in Basso and di Tollo (2022) show the best performances for only 7 out of 33 $HEPs$, being comparable to those resulting from our classical neural network (Fig. 2). This result is also confirmed across all the other UoAs, for which comparable performances are observed between the linear model and the classic neural network approach, while our Adaptive-based neural network is always the best performing one.

Table 3: Forecasting REF 2020: expected s_{REF} direction of change for our Adaptive Neural Network and Classical Neural Network on *Phisics UoA*. Both methods lead to an increase of the score value on 22 out of 33 HEPs.

HEP	Adaptive Neural Network	Classical Neural Network
Cardiff University	↓	↓
Heriot-Watt University	↑	↑
Imperial College London	↑	↑
Kings College London	↑	↑
Lancaster University	↑	↑
Loughborough University	↑	↓
Queen Belfast	↑	↑
Royal Holloway, University of London	↑	↑
Swansea University	↑	↑
University College London	↑	↑
University of Bath	↑	↑
University of Birmingham	↑	↑
University of Bristol	↑	↑
University of Cambridge	↑	↑
University of Durham	↓	↓
University of Edinburgh	↓	↑
University of Exeter	↑	↑
University of Glasgow	↑	↑
University of Kent	↑	↓
University of Leeds	↑	↑
University of Leicester	↑	↑
University of Liverpool	↑	↑
University of Manchester	↓	↑
University of Nottingham	↓	↑
University of Oxford	↓	↑
University of Sheffield	↑	↑
University of Southampton	↑	↑
University of St Andrews	↓	↑
University of Strathclyde	↓	↑
University of Surrey	↑	↑
University of Sussex	↑	↑
University of Warwick	↓	↑
University of York	↑	↑

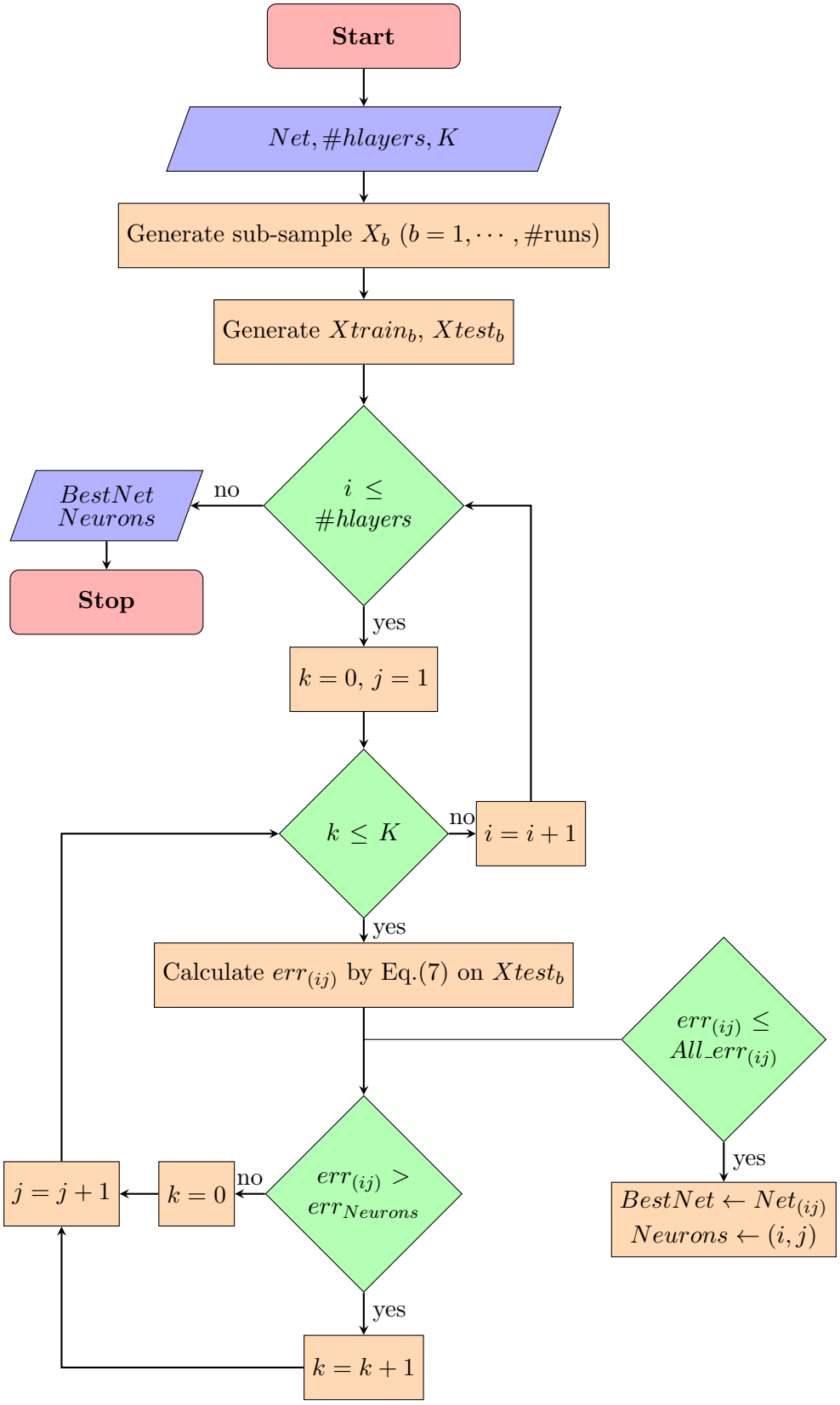

Figure 1: Adaptive Network's flowchart algorithm.

Table 4: Forecasting REF 2020: actual (A) Vs predicted s_{REF} scores (eq.4) for the linear (L) and Log-Linear (LL) models by Basso and di Tollo (2022), and for our Classical (CNN) and Adaptive (ANN) Neural Networks on the *UoA Phisics*. The method leading to the smallest difference between actual and predicted value is highlighted in bold.

HEP	A	L	LL	CNN	ANN
Cardiff University	63.67	64.59	59.06	59.29	**63.70**
Heriot-Watt University	69.00	58.36	56.22	53.91	**69.63**
Imperial College London	64.67	73.91	54.30	59.22	**60.41**
Kings College London	54.33	65.82	53.81	55.28	**53.85**
Lancaster University	62.67	**55.90**	47.44	52.18	48.25
Loughborough University	45.00	40.49	**44.41**	42.16	41.74
Queen Mary University of London	54.00	64.79	53.82	**54.00**	63.18
Queen Belfast	51.67	54.38	**51.54**	52.64	53.97
Royal Holloway, University of London	52.00	59.41	**52.04**	55.28	53.01
Swansea University	51.33	53.95	**52.88**	54.03	53.20
University College London	60.00	71.37	54.00	52.66	**60.99**
University of Bath	60.00	**58.96**	55.77	61.04	61.08
University of Birmingham	75.67	**71.09**	54.40	70.03	70.06
University of Bristol	73.67	**63.00**	50.67	52.88	52.85
University of Cambridge	75.67	**80.58**	54.34	55.18	56.56
University of Durham	61.33	62.83	56.03	62.37	**61.59**
University of Exeter	63.00	52.56	52.23	60.15	**61.26**
University of Glasgow	63.00	66.52	52.54	**65.23**	65.70
University of Kent	39.33	37.07	44.27	**40.03**	40.18
University of Leeds	68.33	58.99	55.99	**63.02**	56.25
University of Leicester	49.33	44.92	**49.24**	49.03	50.63
University of Liverpool	65.67	**68.74**	55.73	58.00	57.94
University of Manchester	76.67	69.43	53.99	55.72	**70.03**
University of Nottingham	72.00	63.48	56.53	**64.51**	57.38
University of Oxford	69.67	76.70	52.48	**70.03**	71.39
University of Southampton	57.00	67.92	55.39	**58.37**	56.02
University of Strathclyde	67.00	59.76	54.28	62.81	**64.30**
University of Surrey	50.33	**50.27**	49.69	49.01	52.88
University of Sussex	59.33	66.64	56.20	57.23	**59.86**
University of Warwick	64.00	68.92	56.98	57.38	**65.77**
University of York	66.67	54.64	51.18	52.11	**62.94**

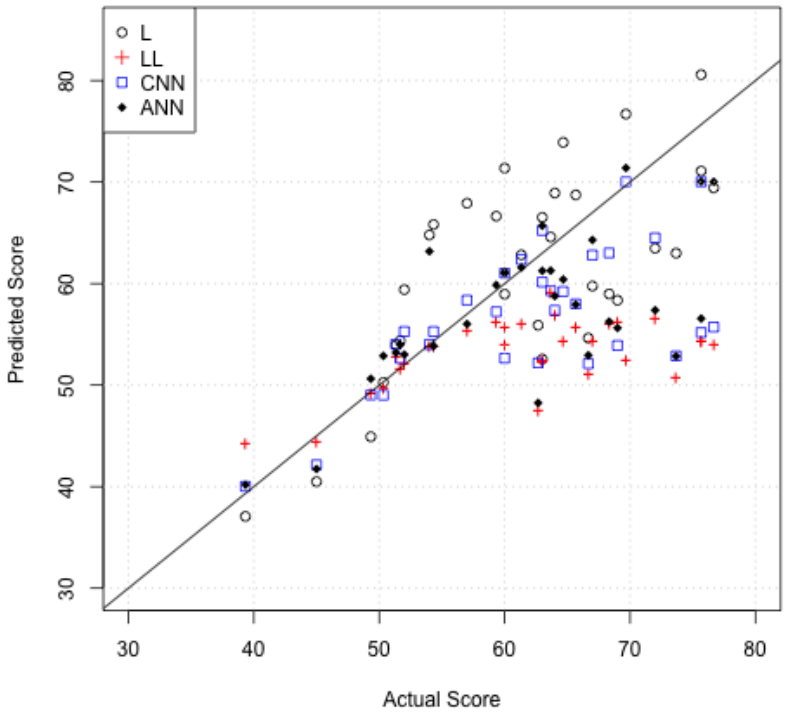

Figure 2: Scatter plot of Actual against Predicted values for the different methods taken into account: Linear (L), Log-Linear (LL), Classical Neural Network (CNN) and Adaptive Neural Network (ANN).

As a last analysis, we want to analyze the reliability of the proposed methods to predict the actual granted research funds. In light of the acknowledged role of public funding in boosting and disseminating knowledge creation Fleming et al. (2019) (which in turn generates positive spillover effects on economic and social variables Hausman (2022)), an effective predictive management model may provide an important policy tool at both national and institutional levels. Please notice that HEFCE ceased to exist as of 1 April 2018 and was replaced by the newly created *Research England* and that, for determining the overall quality-related (QR) research profile, Research England's allocated mainstream budget is split into three sub-profile pots. This is to reflect the different weights given to the three elements, as described in Sec. (3), assessed in each submission to the REF, i.e., *Output*, *Impact* and *Environment*.

Looking at Fig.3, we see that our ANN proposed approach shows a better fit in terms of both skewness and kurtosis. In fact, as Fig. 3 supports, the mean and the variance of Sub-Profiles and Total Grant predicted distributions are quite similar to the actual ones, however, the Linear approach fails to fit properly the third and the fourth moment of the

Table 5: Relative variation from the actual assigned funds by Sub-profiles/Total mainstream and by approach: (L) Linear - (ANN) Adaptive Neural Network - (Act.) Actual. The best performance is highlighted in bold.

	Output		Impact		Environment		**Total**	
	Δ(L,Act)%	*Δ(ANN,Act)%*	*Δ(L,Act)%*	*Δ(ANN,Act)%*	*Δ(L,Act)%*	*Δ(ANN,Act)%*	*Δ(L,Act)%*	*Δ(ANN,Act)%*
Cardiff University	**-1.258**	-5.872	-13.791	**+0.544**	**-5.005**	+1.772	-4.553	**-3.392**
Heriot-Watt University	2.202	**-1.461**	+0.381	**-0.346**	-5.210	**+4.832**	+1.090	**-0.567**
Imperial College London	**-0.252**	-9.883	-13.704	**+1.313**	-9.392	**+0.532**	**-4.511**	-5.903
King's College London	**+0.810**	-1.815	+34.020	**+12.809**	+7.920	**+2.629**	+4.378	**-0.099**
Lancaster University	**-0.472**	-4.736	-6.815	**+6.053**	-7.201	**+0.883**	**-2.374**	-2.423
Loughborough University	**-1.040**	-4.358	+41.095	**+11.555**	+26.870	**+3.316**	+6.078	**-1.986**
Queen Mary University of London	+3.073	**-2.319**	-12.262	**-5.391**	-6.248	**+2.588**	**-1.128**	-2.033
Queen's University Belfast	**-4.199**	-10.105	**+40.975**	+66.600	+8.879	**-1.069**	**+0.930**	-2.930
Royal Holloway University of London	-3.527	**-1.341**	+78.410	**+1.520**	-6.880	**-1.057**	**+0.172**	-1.136
Swansea University	+1.909	**-0.336**	-32.587	**-2.090**	+26,902	**+3.716**	-4.025	**-0.371**
University College London	**+0.095**	-0.909	**-2.194**	-3.882	-6.354	**+1.612**	-1.435	**-0.832**
University of Bath	**+0.144**	+1.933	-5.690	**-1.190**	**-5.144**	+8.612	-2.264	**+2.075**
University of Birmingham	+1.552	**-1.526**	-11.219	**+1.452**	**-0.384**	+1.457	-1.023	**-0.518**
University of Bristol	**+0.821**	+1.697	-3.171	**+1.389**	-6.294	**+1.243**	**-1.091**	+1.564
University of Cambridge	+4.410	**+0.410**	+35.152	**+0.016**	-9.527	**-2.223**	+10.789	**-0.040**
University of Durham	**-0.254**	+1.475	-21.575	**-0.373**	**-4.755**	-6.646	-5.711	**-0.442**
University of Exeter	+1.377	**-0.580**	**+2.246**	-2.391	-4.701	**+1.872**	+0.649	**-0.426**
University of Glasgow	+1.905	**+0.694**	**-3.878**	+5.906	+11.851	**+0.452**	+3.165	**+1.317**
University of Kent	+7.712	**+0.275**	+29.087	**+9.310**	**+0.754**	+4.674	+10.412	**+2.035**
University of Leeds	-1.882	**+1.201**	-2.918	**-2.509**	+9.123	**-0.023**	-1.159	**+0.192**
University of Leicester	-3.764	**+2.488**	+35.153	**+0.984**	**+0.275**	+1.562	**+0.583**	+2.202
University of Liverpool	+0.398	-0.458	-5.157	-2.120	-6.155	1.920	-1.528	**-0.297**
University of Manchester	-0.807	**-0.614**	+34.122	**+0.372**	+10.497	**-0.377**	+10.279	**-0.316**
University of Nottingham	-2.544	**-0.655**	+17.325	**+1.047**	-4.544	**-2.363**	+2.379	**-0.476**
University of Oxford	**-0.935**	-1.125	+4.507	**+2.935**	+7.496	**-3.998**	+2.113	**-0.692**
University of Southampton	+0.172	**+0.121**	-9.179	**+3.148**	**-2.131**	+11.346	-2.515	**+2.407**
University of Strathclyde	**-2.955**	+14.565	**+5.407**	**+1.111**	-5.274	+4.128	**-0.974**	+9.514
University of Surrey	+3.598	**+1.613**	-22.877	**+4.801**	+8.954	**+5.167**	**-2.916**	+2.933
University of Sussex	-1.099	**-0.459**	-12.997	**+2.542**	-4.695	**+1.329**	-3.635	**+0.351**
University of Warwick	**-1.255**	-50.828	**-10.228**	-11.116	-5.495	**+1.025**	**-3.580**	-34.650
University of York	-1.434	**+0.816**	**-6.496**	+8.889	**+2.564**	-9.098	-1.868	**+0.826**

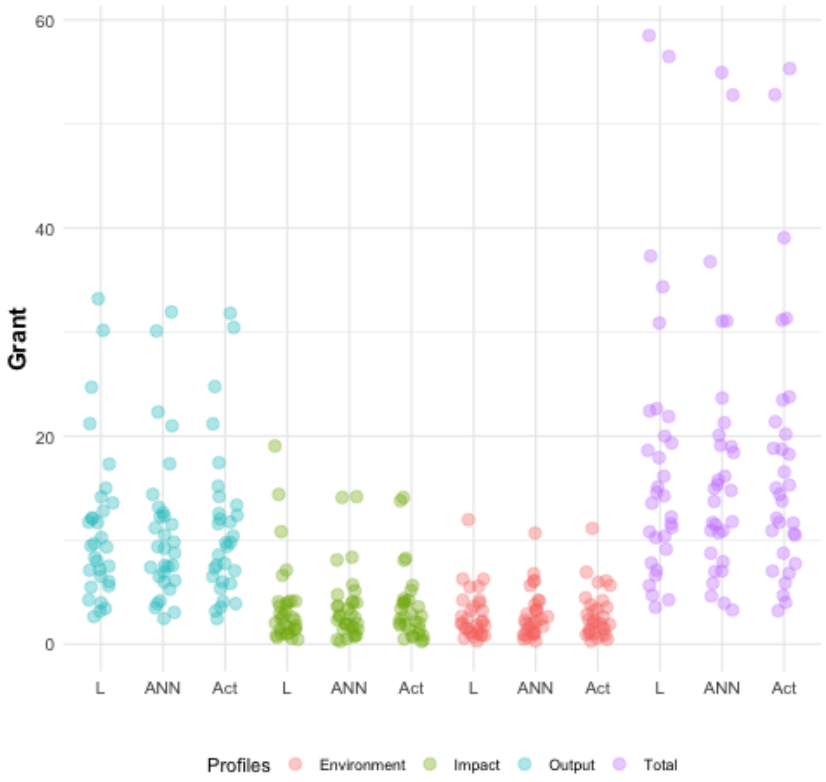

Figure 3: Linear (L), Adaptive Neural Network (ANN) and Actual (Act.) Mainstream QR funds (£). Grant values are scaled by a factor of $1e5^{-1}$.

observed QR funds distribution. Knowledge about the proper shape of the distribution is important at both local (universities) and national governance levels (national education funding bodies). At a local level, relying on proper information for an institution would imply to know its own projected position relative to the other institutions competing in the Quality-related Research Funding programme. At the upper level, more reliable predicted results might be crucial at anticipating and implementing more effective and sustaining funding policies.

7 Concluding Remarks

In this paper we have introduced an adaptive neural network approach for predicting and forecasting the results of a research quality assessment exercise. These procedures are carried out by national public bodies to allocate, among Higher Education Providers, public research funds on a competitive performance basis.

We have worked on instances from the UK's Research Excellence Framework, and compared our adaptive parameter-tuning-based approach with previous contributions

tackling the same dataset. Results show that our proposed approach does not suffer of overfitting, compares favourably with the existing literature, and, when coupled with meaningful pre-processing operations, could be successfully used to predict and forecast the outcome of quality research assessment procedures.

As for future research, we are planning to expand this approach by comparing *parameter-tuning* and *parameter-control* procedures to investigate about the optimal neural network parameters settings. Furthermore, instead of using our dynamic approach to determine the topology of the network, we plan to develop a genetic algorithm that relies on several mutation operators, and that selects, at each step, the right operator to be used according to user-defined criteria.

Last, in order to test the robustness of our approach, we aim to tackle instances from countries other than UK, such as Italy and Belgium, where the research quality assessment procedures are conducted by applying a different combination of *peer-review* and *bibliometric* criteria.

Declarations

- *Funding*: No funds, grants, or other support was received for this work;
- *Conflict of interest/Competing interests*: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Angelini, E., di Tollo, G., and Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, 48(4):733–755.

Banal-Estañol, A., Jofre-Bonet, M., Iori, G., Maynou, L., Tumminello, M., and Vassallo, P. (2023). Performance-based research funding: Evidence from the largest natural experiment worldwide. *Research Policy*, 52(6):104780.

Basso, A. and di Tollo, G. (2017). *A Generalised Linear Model Approach to Predict the Result of Research Evaluation*, pages 29–41. Springer International Publishing, Cham.

Basso, A. and di Tollo, G. (2022). Prediction of UK research excellence framework assessment by the departmental h-index. *European Journal of Operational Research*, 296(3):1036–1049.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C., and Peracchi, F. (2011). Bibliometric evaluation vs. informed peer-review: Evidence from Italy. *Research Policy*, 44(2):451–466.

Birattari, M., Yuan, Z., Balaprakash, P., and Stützle, T. (2010). *F-Race and Iterated F-Race: An Overview*, pages 311–336. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bornmann, L. (2012). The Hawthorne effect in journal peer review. *Scientometrics*, 91:857–862.

Bornmann, L. and Leydesdorff, I. (2014). Scientometrics in a changing research landscape. *EMBO reports*, 15(12):1228–1232.

Bruns, S. and Stern, D. (2015). Research assessment using early citation information. Technical report, Crawford School of Public Policy, The Australian National University, Crawford School Research Papers.

Burger, M., Frankfort, J., and van Raan, A. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy*, 14:131–149.

Corazza, M., March, D. D., and di Tollo, G. (2021). Design of adaptive elman networks for credit risk assessment. *Quantitative Finance*, 21(2):323–340.

di Tollo, G., Tanev, S., De March, D., and Zheng, M. (2012). Neural networks to model the innovativeness perception of co-creative firms. *Expert Systems with Applications*, 39(16):12719 – 12726.

di Tollo, G., Tanev, S., Liotta, G., and March, D. D. (2015). Using online textual data, principal component analysis and artificial neural networks to study business and innovation practices in technology-driven firms. *Computers in Industry*, 74:16–28.

di Tollo, G., Tanev, S., Slim, K., and De March, D. (2014). Determining the relationship between co-creation and innovation by neural networks. In Faggini, M. and Parziale, A., editors, *Complexity in Economics: Cutting Edge Research*, pages 49–62. Springer International Publishing, Cham.

Evidence (2010). The future of the UK university research base. Technical report, a Thomson Reuters business, Universities UK.

Fleming, L., Greene, H., Li, G., Marx, M., and Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141.

Geuna, A. and Piolatto, M. (2016). Research assessment in the uk and italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy*, 45(1):260–271.

Hausman, N. (2022). University Innovation and Local Economic Growth. *The Review of Economics and Statistics*, 104(4):718–735.

Haykin, S. (2009). *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition.

Inglis, M., Foster, C., Lortie-Forgues, H., and Stokoe, E. (2024). British education research and its quality: An analysis of research excellence framework submissions. *British Educational Research Journal*, 50(5):2495–2518.

Jonkers, K. and Zacharewicz, T. (2016). Research performance based funding systems: a comparative assessment. Technical Report EUR 27837, JRC101043.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9):6233–6239.

Lawrance, P. (2003). The politics of publications. *Nature*, 422:259–261.

MacRoberts, M. and MacRoberts, B. (1989). Problems of citation analysis: a critical review. *Journal of the American Society for Information Science*, 40(5):342–349.

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2015a). Predicting results

of the research excellence framework using departmental h-index. *Scientometrics*, 102(3):2165–2180.

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2015b). Predicting results of the research excellence framework using departmental h-index: revisited. *Scientometrics*, 104(3):1013–1017.

Reed, R. and Marks, R. (1998). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press, Cambridge, MA, USA.

Sgroi, D. and Oswald, A. (2013). How should peer-review behave? *Economic Journal*, 123(570):F255–F278.

Thelwall, M. (2025). Research quality evaluation by ai in the era of large language models: advantages, disadvantages, and systemic effects –an opinion paper. *Scientometrics*, 130(10):5309–5321.

Thelwall, M. and Yaghi, A. (2024). In which fields can chatgpt detect journal article quality? an evaluation of ref2021 results. *arXiv*.

Werbos, P. (1994). *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. USA: Wiley-Interscience.