On the Usage of Rank Percentile in Evaluating and Predicting Scientific Impacts

Sen Tian[1] **and Panos Ipeirotis**[1]

[1]Department of Technology, Operations, and Statistics, Stern School of Business, New York University, New York, NY, 10012, USA. Correspondence and requests for materials should be addressed to S.T. (email: st1864@stern.nyu.edu)

ABSTRACT

Bibliographic metrics are commonly utilized for evaluation purposes within academia, often in conjunction with other metrics. These metrics vary widely across fields and change with the seniority of the scholar; consequently, the only way to interpret these values is by comparison with other academics within the same field and of similar seniority. We propose a simple extension that allows us to create metrics that are easy to interpret and can make comparisons easier. Our basic idea is to create benchmarks and then utilize percentile indicators to measure the performance of a scholar or publication over time. These percentile-based metrics allow for comparison of people and publications of different seniority and are easily interpretable. Furthermore, we demonstrate that the rank percentile indicators have reasonable predictive power. The publication indicator is highly stable over time, while the scholar indicator exhibits short-term stability and can be predicted via a simple linear regression model. While more advanced models offer slightly superior performance, the simplicity and interpretability of the simple model impose significant advantages over the additional complexity of other models.

Introduction

Comparing the scientific impact of scholars or publications often occurs when making academic decisions. For instance, academic committees evaluate a candidate scholar relative to other cohorts in the same department to award tenure promotions. Directly comparing the number of citations can be biased, since the citations change with the seniority of the scholar. Another example is assigning research funding in which the scholar's portfolio is compared to other candidates from various facilities and disciplines. The magnitude of the citations that a publication receives varies drastically across disciplines. Evaluating whether to utilize citations favors scholars from more active fields and hence is not an appropriate measure.

To make the comparison feasible, we propose utilizing the rank percentile indicator, whose four fundamental elements are entity, benchmark, evaluation metric, and age. The entity can be either a publication (P) or a scholar (S). The benchmark (b) characterizes the reference set to which the entity is compared, and it is specified by the problem of interest. In the example of tenure promotions, the benchmark can comprise all the cohorts in the same department, while in the research funding allocation example, the benchmark contains all the candidates in competition. The cohorts in the benchmark are evaluated utilizing some metric (m), such as the number of citations, and the age t specifies the time at which the evaluation is executed. The publication j is ranked based on its evaluation metric at t years since published, and the rank is further transformed into the rank percentile indicator, denoted as $\mathrm{P}_m^{jb}(t)$. The rank percentile indicator for scholar i can be calculated following the same procedure and is denoted as $\mathrm{S}_m^{ib}(t)$.

The rank percentile indicator is significantly interpretable. It describes the performance of a publication or a scholar at certain age relative to the cohorts in the benchmark. Additionally, the rank percentile indicator is flexible in the choice of evaluation metric. For scholars, the h-index [1] is probably the most popular metric and is defined as the maximum number h for which the scholar has h publications, each with at least h citations. The h-index removes some of the bias introduced by utilizing the number of citations. A scholar participating in a small number of frequently cited works or a large number of low-profiled projects can have a high citation count but a low h-index, since h-index rewards a consistent stream of impactful efforts. The problem with the h-index is that the actual number of citations is irrelevant once it exceeds h, and hence, the h-index does not reward astonishing works differently than publications that attract barely sufficient citations to boost the h-index. Numerous indices have been proposed to improve the h-index, such as the g-index [2], the m-index [1], and the i-10 index [3]. All these metrics treat citations equally and do not distinguish between a citation from a highly regarded journal and a citation from a workshop panel. PageRank index [4–6] utilizes the citation network and evaluates a publication by assigning different weights to its citations. It can further be aggregated to measure the impact of a scholar [7]. However, there does not exist a single ideal metric, and all of the aforementioned metrics are biased in certain ways.

The rank percentile normalizes the citations by their ranks relative to the citations of other publications in the benchmark. The usage of a normalized indicator to compare the performance of publications has been studied in the literature. A mean-based indicator normalizes the citations of publications in the benchmark with respect to the expected citation impact of the benchmark, which can be estimated by the arithmetic mean of citations for all publications in the benchmark [8]. Since the citation distribution is skewed and heavy-tailed, the arithmetic mean is not a reasonable representation of the expected citation impact, and therefore mean-based indicators can be largely influenced by a small number of frequently cited publications. These drawbacks can be largely avoided by utilizing the rank percentile indicator [9–11].

The projection of the future performance for a publication or a scholar is often of great interest for evaluation purposes. There have been extensive discussions regarding prediction of the number of citations and the resultant h-index score. The mechanism model unveils the factors driving the citation dynamic of publications, in which the three main factors are the scaling-law distribution of citations [12–15], aging [13, 16–18], and perceived novelty [19]. The mechanism model can be applied to predict the future evolution of citations [19], but it relies on a long citation history [20, 21]. Each publication must be addressed individually, and hence, it is not appropriate for large-scale analysis. Another type of predictive model formulates the task as a supervised learning problem. By utilizing sophisticated machine learning algorithms and an extensive list of features, these models can be utilized to predict citations [22–27] and h-index scores [27–30], and they can be scaled to account for large-scale datasets.

To the best of our knowledge, little is known about the evolution of the rank percentile indicator over time and its predictive power. In this paper, we discuss the framework for calculating the rank percentile indicator. Additionally, we propose and justify a novel rank percentile indicator for scholars, and we demonstrate its advantage over rank percentiles based on the existing evaluation metrics. Furthermore, we study the predictability of the rank percentile indicator, illustrating that the publication percentile is highly stable over time, while the scholar percentile offers short-term stability and can be predicted via a simple linear regression model.

Calculation of the rank percentile indicator

In this section, we start by discussing the framework for calculating the rank percentile indicator. For publications, the indicator is based on the number of citations. We further propose utilizing an aggregation of rank percentile indicators for publications as the evaluation metric, based on which we then construct the indicator for scholars. We discuss the advantage of the proposed indicator compared to indicator based on existing evaluation metrics, such as the number of citations or h-index.

Dataset

The dataset from Google Scholar includes scholars in multiple disciplines from the top 10 universities in the United States, which totals 14358 scholars. It includes the citation history through 2016 for each publication from these scholars; they contributed to more than $800,000$ publications total, which received approximately 100 million citations collectively. The discipline for a scholar is determined by the area of interest specified on the Google Scholar page. For instance, a scholar and his/her publications are labeled as in the area of biology if the area of interest contains any of the following keywords: biology, genetic, neuroscience, or cell. An exploratory description of the dataset can be found in Supplemental Material (Figure S5 and Table S4). The dataset and the code to reproduce the results in this paper are available online at `https://github.com/sentian/SciImpactRanking`.

Framework for calculating the rank percentile indicator

Recall that the four fundamental elements for the rank percentile indicator are as follows.

- Entity: publication (P) or scholar (S).
- Benchmark (b): the reference set to which the entity is compared.
- Metric (m): the measure that evaluates the performance of the entity.
- Age (t): the time when the evaluation is executed.

The rank percentile for publication j, $\mathrm{P}_m^{jb}(t)$, is calculated in the following way.

(1) Evaluate the publications in the benchmark by their performance at age t (publications with life shorter than t are not included). Utilize the evaluation metric to calculate the rank $\mathrm{r}_m^{jb}(t)$ of publication j against the publications in the benchmark. An average rank is assigned to $\mathrm{r}_m^{jb}(t)$ if there exist other publications that have the same value of the metric.

(2) The rank percentile is indicated by $\mathrm{P}_m^{jb}(t) = \left(\mathrm{r}_m^{jb}(t) - 0.5\right)/N$.

With the compromise $0.5/N$ in the final step, the median paper is assigned 50% percentile, and the tails of the citation distribution are treated symmetrically [31]. The above framework can be easily adapted to compute the rank percentile indicators for scholar i: $S_m^{ib}(t)$.

The rank percentile indicator for scholars

For publication j, we utilize the number of citations by age t as the evaluation metric and denote the rank percentile indicator as $P_c^{jb}(t)$. We further utilize the publication indicators to construct the rank percentile for scholars. For scholar i, the performance is determined by the qualities of the scholar's publications, in which each publication is evaluated via $P_c^{jb}(5)$, meaning the rank percentile for the publication at the fifth year since published. The evaluation metric for scholar i is determined by aggregating the performance of all $N(t)$ papers that the scholar publishes by age t, that is $\sum_{j=1}^{N(t)} P_c^{jb}(5)$. We denote the resulting rank percentile indicator as $S_{P5}^{ib}(t)$, where P5 indicates the evaluation metric based on rank percentile indicator of publications at age 5. In the discussion that follows, we utilize the simplified notations P_c and S_{P5} to refer to the publication and scholar rank percentiles in general, respectively.

Figure 1 presents an example of S_{P5} for a random scholar in our dataset in which the benchmark includes all tenured professors in the top 10 universities, as ranked by the U.S. News. The scholar's career started in 2004, and our dataset tracks the citation information until 2016. The indicator S_{P5} ranks the scholar to be in the top 40% throughout the majority of their career. The figure indicates two other types of rank percentile indicators, S_c and S_h, that utilize the number of citations and h-index score as evaluation metrics, respectively. We see that S_h largely agrees with S_{P5}, and S_c ranks the scholar lower than the other two indicators.

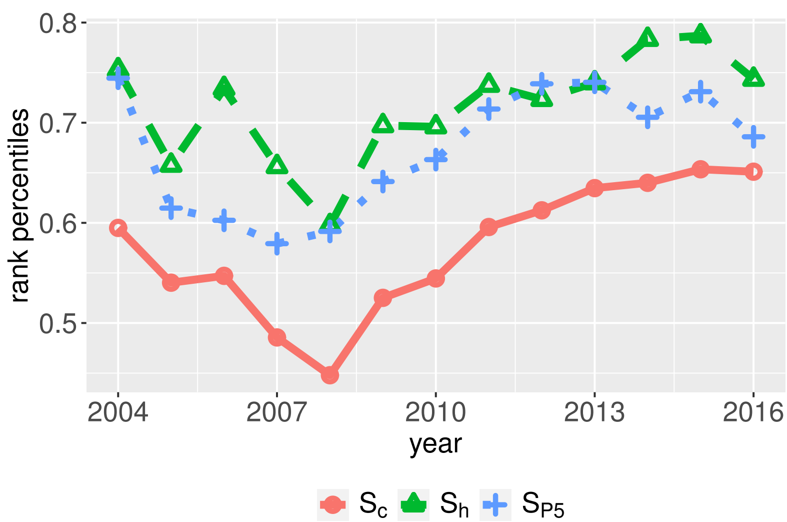


Figure 1. The rank percentile indicators for a random scholar. The benchmark contains all tenured professors in the top-10 universities, as ranked by the U.S. News.

The indicator S_{P5} improves some major drawbacks of S_c and S_h. First, it removes the seniority effect of publications. The evaluation metric for $S_c^{ib}(t)$ represents the citations that scholar i receives by age t, which is the sum of citations for the scholar's publications by t. Compared to newly published works, publications with longer histories are more likely to attract citations and therefore contribute more to formulating S_c. A similar argument can be made for S_h. However, S_{P5} treats the publications equally and evaluates them based on their performances at the publications' age of five. Additionally, a scholar who publishes a considerable number of low-impact works or participates in only a small number of high-impact projects can have a high value of S_c, since the absolute number of citations can be unlimited and is significantly influenced by extreme values. However, the S_{P5} and S_h of these scholars are not necessarily large, since these indicators limit the contribution of a single publication to be, at most, 1 by definition of rank percentile and h-index score. Furthermore, compared to S_h, S_{P5} penalizes scholars who are not truly innovative but carefully massage their h-index scores by publishing a number of papers that attract barely sufficient amounts of citations to increase their h-index scores. As long as a paper is among the top h papers, the actual number of citations is irrelevant for h-index and S_h, but it can still impact S_{P5}. Finally, S_{P5} requires less data compared to S_c and S_h, since it only relies on the 5-year citation history of each publication. Hence, S_{P5} is better suited to large-scale analysis.

We demonstrate the advantages of S_{P5} by examining some extreme cases. We created three synthetic academic careers. Scholar A publishes a substantial number of publications throughout his/her career (more than 90% of his/her cohorts in the benchmark), while all of the publications have little impact. Scholars B and C only publish one paper each at the beginning of their careers; B's paper is astonishing, while C's paper is average. Both scholars have an h-index equal to 1 throughout their careers. Figure 2 illustrates the rank percentile indicators for these three artificial scholars. We see that flooding low-impact publications can increase S_c at the beginning of Scholar A's career. We also see that a single high-impact work improves the value S_c throughout Scholar B's career; the author remains in the top 50% at age 12, as indicated by S_c. Both S_{P5} and S_h better characterize the performances of these authors. Finally, S_h remains the same for Scholars B and C since they each have an h-index of 1 throughout their careers. However, S_{P5} considers that Scholar B's publication has a greater impact and therefore ranks higher than Scholar C.

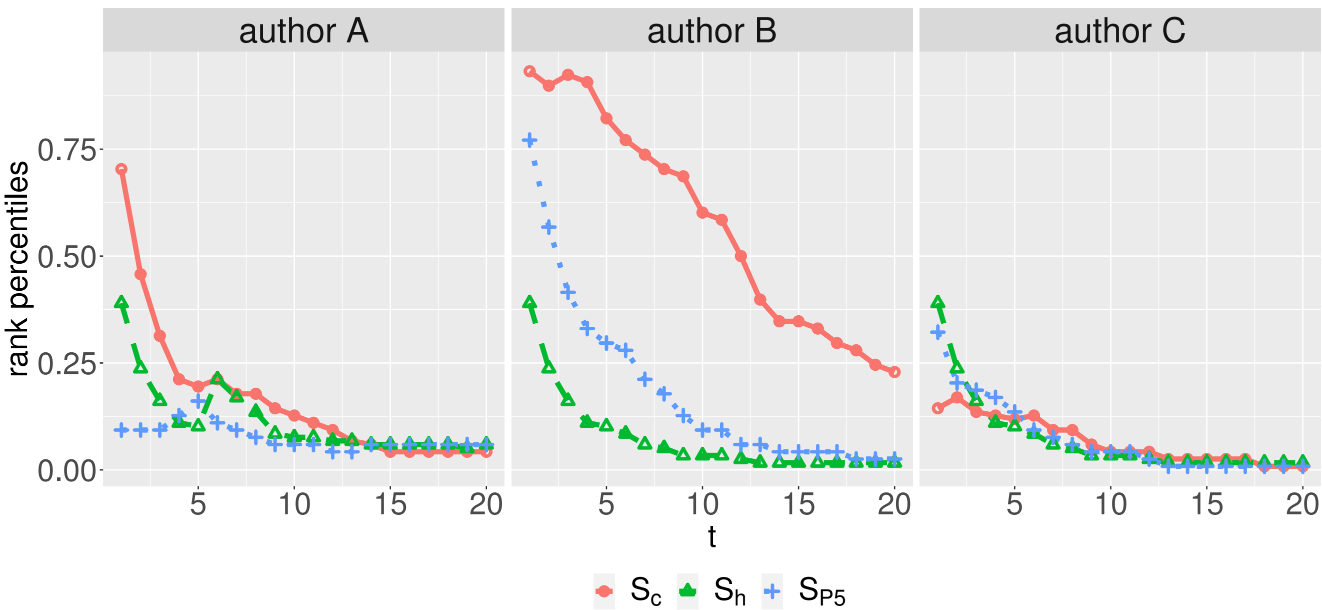


Figure 2. Three artificial scholars illustrate the difference between various types of rank percentile indicators for scholars. Scholar A is highly productive at all ages, but all of the works have little impact. Scholars B and C only publish one paper each at age 1; the paper of Scholar B has substantial impact, while the paper of Scholar C has medium impact. The h-index scores of Scholars B and C remain at 1 throughout their careers. The benchmark includes scholars in biology who started their careers in 1990.

With the exception of the above-mentioned discrepancies, we present in Supplemental Material Section A that for the majority of scholars in our dataset, S_{P5} largely agrees with S_c and S_h. Furthermore, we consider utilizing various metrics to evaluate the publication, for example $P_c^{jb}(10)$ which is based on 10-year citation history. As we reveal Supplemental Material Section B, the differences between the resultant rank percentile indicators and S_{P5} are not statistically significant, thus indicating the robustness of S_{P5}. Intuitively, the publication percentile $P_c^{jb}(t)$ is highly stable over t, and therefore $P_c^{jb}(5)$ can be a reasonable indicator of the performance for the publication. The details will be discussed in the next section.

The predictability of rank percentile indicator

In this section, we study the predictability of the rank percentile indicator. In general, we find the indicator to exhibit stability over time, and the indicator can be predicted via some simple linear regression models.

The stationarity of the rank percentile indicator for scholars

In the example of the tenure promotion, we utilized the rank percentile to compare the candidate with senior cohorts. The comparison is not valid if the candidate is more likely to attract citations than senior colleagues who started their careers years earlier. In such a case, it may well be the academic environment that results in a better performance of the candidate rather than the internal factors, such as creativity and productivity.

Figure 3 portrays S_{P5} at age 10, grouped by the starting year of academic careers. We see that S_{P5} does not exhibit an obvious trend and is approximately stationary over the starting year of careers, thus providing empirical evidence for the validity of the rank percentile.

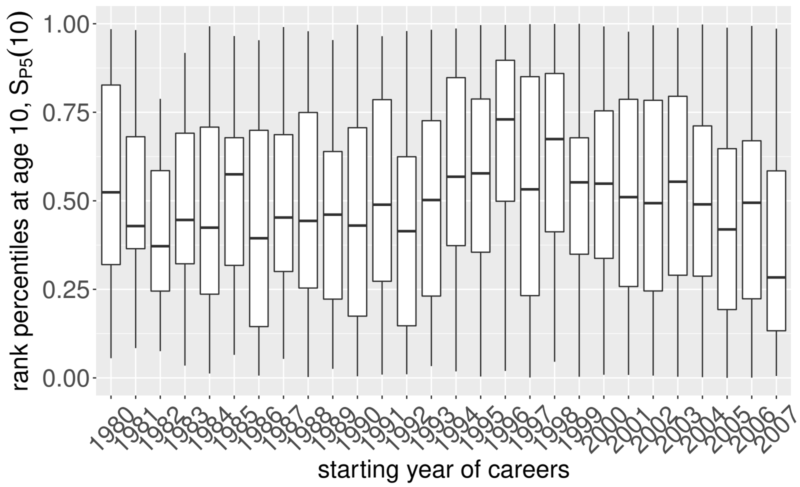


Figure 3. $S_{P5}(10)$ grouped by the starting years of academic careers. The benchmark contains all tenured professors in the top-10 universities, as ranked by the U.S. News.

The predictability of rank percentile indicators

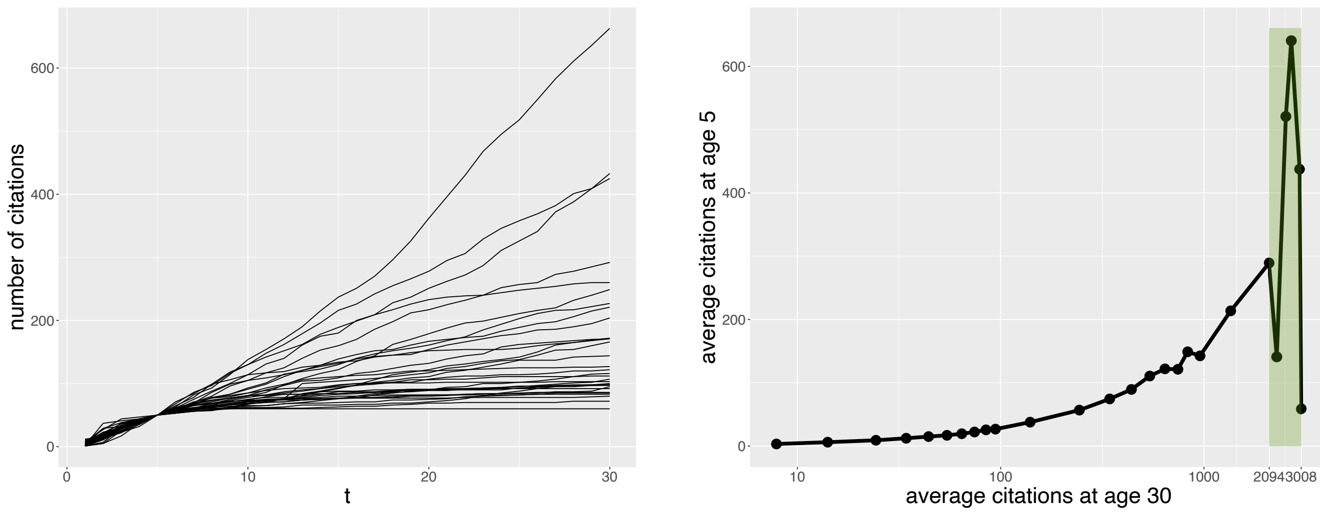
Citations have been proven to lack long-term predictive power [19]. Figure 4a illustrates that papers with the same citations by the fifth year since published can have noticeably different citation paths and long-term effects. Additionally, exceptional and creative ideas typically require a lengthy period to be appreciated by the scientific community. As presented in Figure 4b, the correlation between short- and long-term citations breaks down for the most highly-cited publications (the shaded rectangle). These problems can be largely avoided by utilizing rank percentile indicators, as evidenced in Figures 5a and 5b. The considerable variation in the long-term effect of citations is restricted by utilizing rank percentiles. For publications with high impact, the correlation between short- and long-term effects persists when utilizing rank percentiles.

We further characterize the predictability of rank percentile indicators. Figure 6a presents the correlation between rank percentiles at two ages, $P_c^{jb}(t_1)$ and $P_c^{jb}(t_2)$ where $t_1 < t_2$. We noticed overall large correlations for both benchmarks. The correlation diminishes as the forecast horizon $(t_2 - t_1)$ increases, which simply reflects the difficulty of long-term forecasting. Additionally, the correlation increases as t_1 increases while holding the forecast horizon fix. This indicates that the performance of a senior publication is easier to predict, since the longer history removes more uncertainties regarding its performance. We further noticed a slightly higher predictive power when we restricted the benchmark to be the area of biology.

Figure 6b illustrates that the patterns discussed above generally hold for rank percentiles of scholars. The magnitude of correlations is smaller than those for publications, especially for long-term forecasts. This results from the fact that forecasting the future impact of future works is considerably more difficult than forecasting the future impact of existing works. Intuitively, predicting $S_{P5}^{ib}(t_2)$ involves predicting the performance of papers published before t_1 and predicting the performance of those published between t_1 and t_2. The former is predicting the future impact of existing works, while the latter is predicting the future impact of future works. However, predicting the publication indicator $P_c^{jb}(t_2)$ only involves predicting the future impact of publication j, which is a considerably easier task. Furthermore, when the forecast horizon increases while fixing t_1, additional future works are involved in predicting $S_{P5}^{ib}(t_2)$; therefore, we see that the correlation decreases more quickly than when we predict $P_c^{jb}(t_2)$.

The strong linear relationship between $P_c^{jb}(t_1)$ and $P_c^{jb}(t_2)$ is further characterized in Figure 7. The data are along the 45° line, and the linear regression coefficient of $P_c^{jb}(t_2)$ on $P_c^{jb}(t_1)$ is close to 1 with small standard errors, thus indicating the high stability of P_c. A similar figure for S_{P5} is displayed in Supplemental Material Figure S6, in which we find that S_{P5} exhibits short-term stability.

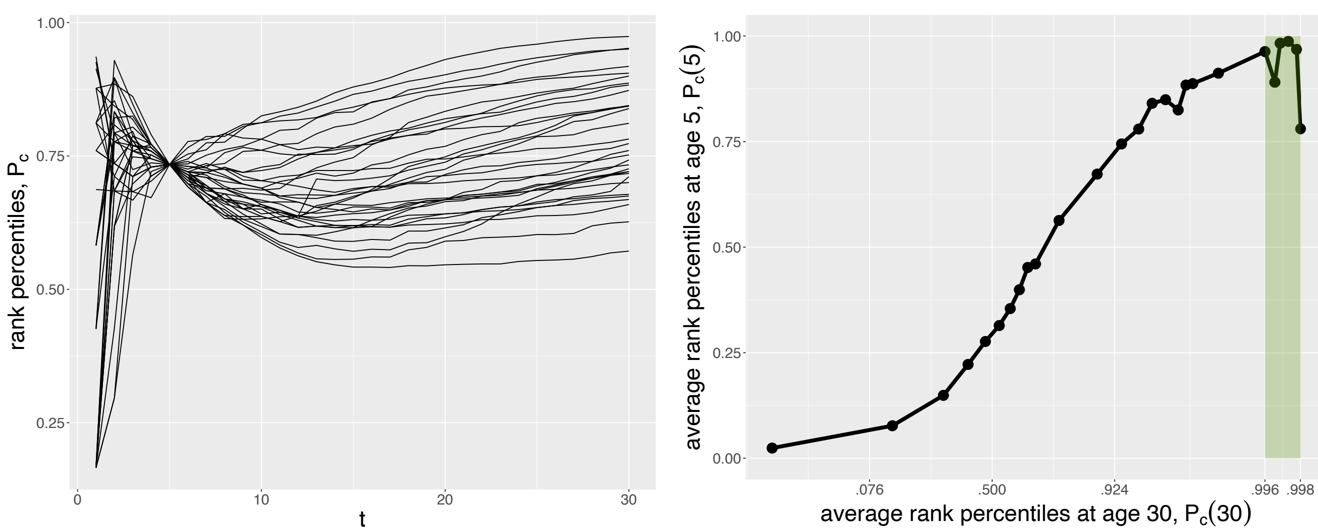
Predicting $S_{P5}^{ib}(t_2)$ can assist in decision making for faculty positions or granting tenure, since the committee would like to examine the cumulative scientific impact of the scholar. In assigning research funding or allocating research resources regarding planned studies and potential future publications, the future impact of future works is often of interest. We utilize



(a) Number of citations versus age

(b) Average citations at age 5 versus those at age 30

Figure 4. The predictability of citations. The benchmark contains the publications in biology. Figure 4a portrays the cumulative citations for publications that have 50 citations by the fifth year since published. Figure 4b displays the average citations by age 5 versus the average citations by age 30. The averages are calculated over groups of publications, which are prespecified by dividing the range of citations by age 30 into equal intervals on the log scale. Note that we do not claim the originality of the figures, which have been illustrated via a different dataset [19].



(a) Rank percentiles P_c versus age

(b) Average P_c at age 5 versus those at age 30

Figure 5. The predictability of rank percentiles. Figure 5a demonstrates the rank percentiles for the publications considered in Figure 4a. Figure 5b presents the average of $P_c(5)$ versus the average of $P_c(30)$ for the same groups of publications as in Figure 4b.

$S^{ib}_{P5}(t_2|t_1)$ to denote the rank percentile indicator calculated based on papers published between t_1 and t_2. Figure 6c illustrates the correlation between $S^{ib}_{P5}(t_1)$ and $S^{ib}_{P5}(t_2|t_1)$. The magnitudes of correlation are moderately high, indicating an approximately linear relationship, although the strength is not as strong as it was in predicting the cumulative impact, that is, predicting $S^{ib}_{P5}(t_2)$, which is consistent with our expectation.

Predictive models

We now formulate the prediction tasks as supervised learning problems, and we illustrate that the rank percentile indicators can be predicted via simple linear models. We consider the following fitting procedures; these models are ordered by increasing complexity:

- Baseline: simple linear regression model.
- Simple Markov model (sm).
- Penalized linear regression models, including the ridge [32], lasso [33], elastic net (enet) [34] and the Gamma lasso (gamlr) [35].
- Ensemble methods of regression trees, including the random forest (rf) [36] and extreme gradient boosting trees (xgbtree) [37].
- Neural networks (nnet).

The baseline model fits a simple linear regression of the target variable on the autoregressive feature, e.g. $\mathrm{P}_c^{jb}(t_1)$ for predicting the publication impact. The simple Markov model further considers the change of the autoregressive feature in the past two ages, e.g. $\mathrm{P}_c^{jb}(t_1)$-$\mathrm{P}_c^{jb}(t_1-2)$, in addition to the autoregressive feature and fits a linear regression model.

Features and model fitting

For the rest of the methods, we created an extensive list of features based on the citation histories. The features are characterized as either scholar- or publication-based features. For example, to predict the scholar indicator $\mathrm{S}_{P5}^{ib}(t_2)$, a scholar-based feature is the number of papers that scholar i publishes by age t_1, and a publication-based feature is the average number of citations for these papers. We established a total of 30 features for predicting the publication indicator and 42 features for predicting the scholar indicator, which can be found in the Supplemental Material Tables S2 and S3, respectively. Note that many of the features have been utilized when formulating the prediction task for number of citations and h-index scores [27, 29].

The features were created utilizing the citation information available by t_1, and the dependent variable was specified at t_2. The data was split into the training and testing set based on a 9:1 ratio. We considered five stages of a publication or a scholar, that is, $t_1 \in \{5,10,15,20,25\}$, and we forecasted up to 30 years of age, that is, $t_2 = t_1+1,\cdots,30$; this resulted in 75 pairs of (t_1,t_2) in total. The models were independently trained 75 times.

We also noted (in the Supplemental Material Section C) that both S_{P5} and P_c are non-stationary time series, as evidenced by the Dicky-Fuller test [38] and the KPSS test [39]. The differenced series are stationary (also presented in the Supplemental Material) and are utilized as the response variable, that is $\Delta\mathrm{P}_c^{jb}(t_2) = \mathrm{P}_c^{jb}(t_2) - \mathrm{P}_c^{jb}(t_1)$, $\Delta\mathrm{S}_{P5}^{ib}(t_2) = \mathrm{S}_{P5}^{ib}(t_2) - \mathrm{S}_{P5}^{ib}(t_1)$, and $\Delta\mathrm{S}_{P5}^{ib}(t_2|t_1) = \mathrm{S}_{P5}^{ib}(t_2|t_1) - \mathrm{S}_{P5}^{ib}(t_1)$. Note that the stationarity discussed here characterizes the property of S_{P5} and P_c as time series. It is different from the stationarity as discussed in Figure 3, where we fix $t = 10$ and examine the stationarity of $\mathrm{S}_{P5}(10)$ over the starting year of the scholars' careers.

The machine learning methods were trained using *R* [40] with the package *mlr* [41], which provides a pipeline of training, validating, and testing for the model. The lasso, ridge, elastic net, random forest, and xgbtree are inbuilt learners of the package. The Gamma lasso and neural network were trained utilizing *R* packages *gamlr* [35] and *keras* [42], respectively.

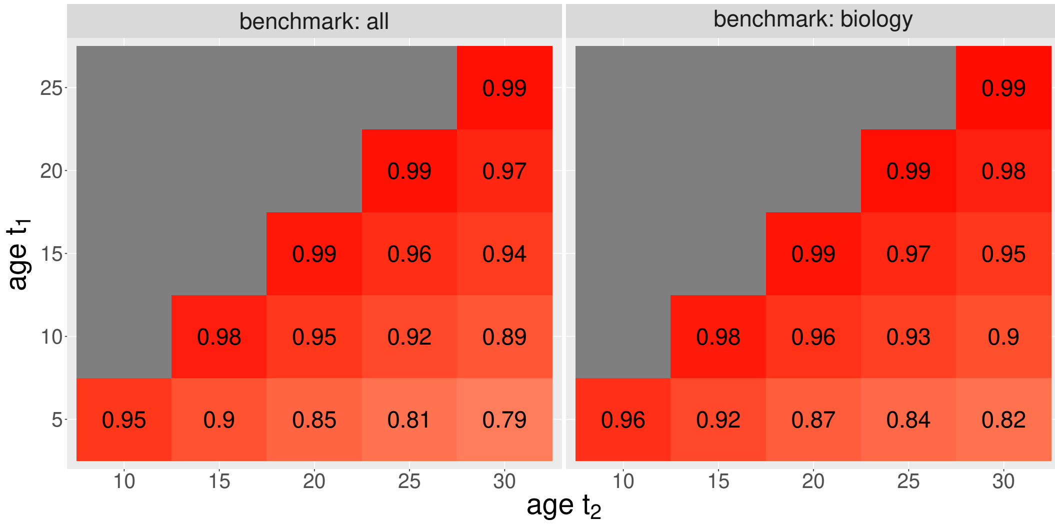
The machine learning methods require hyperparameter tuning, which involves deciding the search space of parameters and evaluating the sets of parameters utilizing the validation data. The optimal model is the one that minimizes the validation error. The hyperparameters for each machine learning model considered in this paper are presented in Supplemental Material Table S1. The parameter space can be substantial for methods such as xgbtree, which utilizes an extensive list of tunable parameters. We applied Bayesian optimization, which searches over the parameter space based on the performance gain.

Results

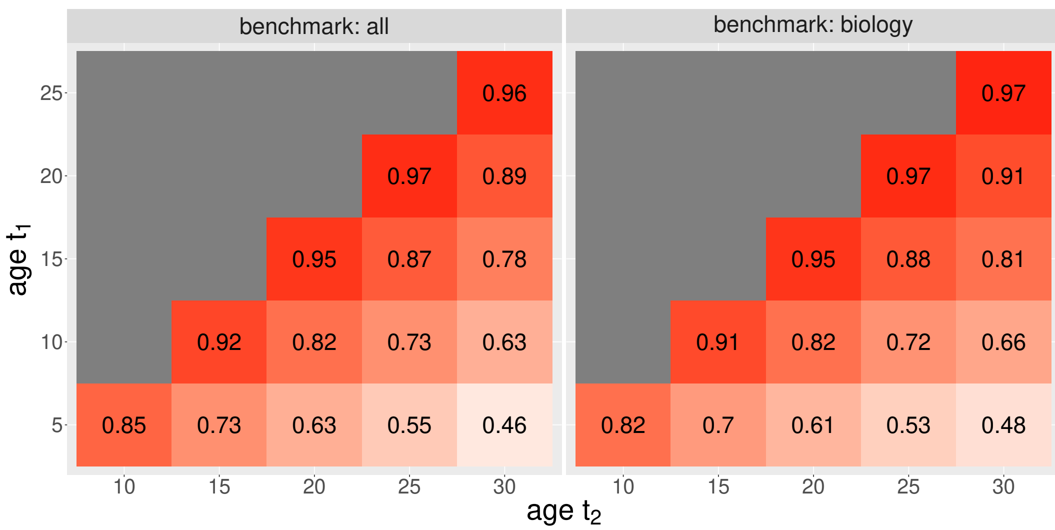
The prediction accuracy is presented in Figure 8. We see that the baseline model predicts the cumulative impacts, $\mathrm{P}_c^{jb}(t_2)$ and $\mathrm{S}_{P5}^{ib}(t_2)$, well, and the usage of a large number of features and complex machine learning models offers little improvement. Predicting the future impact of scholars, $\mathrm{S}_{P5}^{ib}(t_2|t_1)$, is more difficult. The baseline model can still provide reasonable predictions, but the performance is not as satisfactory as others, especially when t_1 is large. By simply adding the difference $\mathrm{S}_{P5}^{ib}(t_1) - \mathrm{S}_{P5}^{ib}(t_1-2)$ as an extra feature, the simple Markov model achieves similar performance compared to the complex machine learning models, which rely on an extensive list of features and exhibit non-linear relationships. Other types of evaluation metrics for the predictive models, including the root mean squared error, root median squared error, and mean absolute error, can be found in the Supplemental Material (Figures S7, S8, and S9, respectively).

Discussion

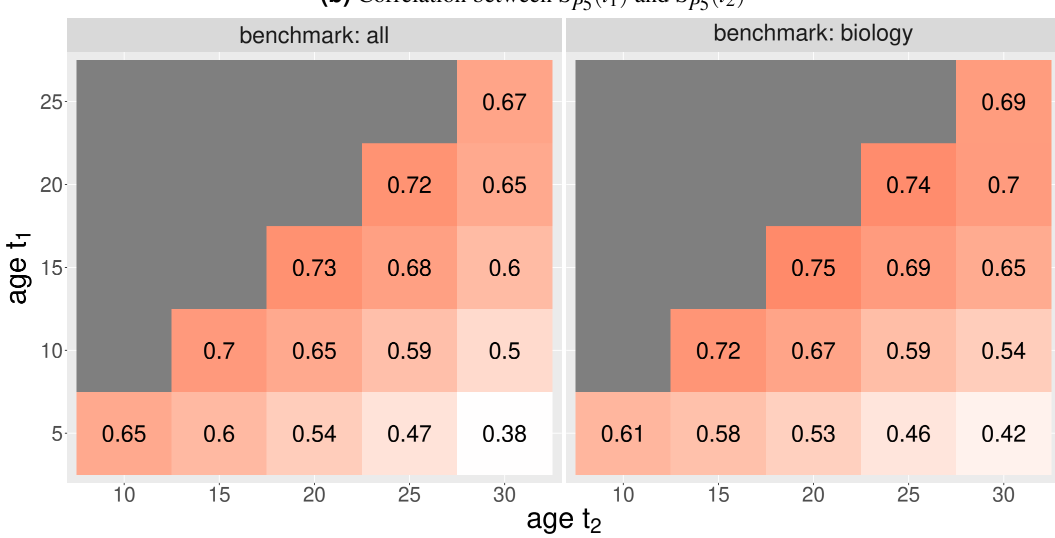
Rank percentile indicators compare the performance of a publication or a scholar relative to cohorts in the benchmark. They have the advantage of being interpretable and flexible in the choice of benchmark and evaluation metrics. We illustrate that the publication percentile is highly stable, while the scholar percentile exhibits short-term stability and can be predicted utilizing a simple linear regression model. In practice, the highly predictable rank percentiles can be utilized in combination with other metrics to picture the trajectory of a scholar or a publication and assist in academic decision making.



(a) Correlation between $P_c^{jb}(t_1)$ and $P_c^{jb}(t_2)$



(b) Correlation between $S_{P5}^{ib}(t_1)$ and $S_{P5}^{ib}(t_2)$



(c) Correlation between $S_{P5}^{ib}(t_1)$ and $S_{P5}^{ib}(t_2|t_1)$

Figure 6. The Pearson's correlation between rank percentile indicators at two different ages.

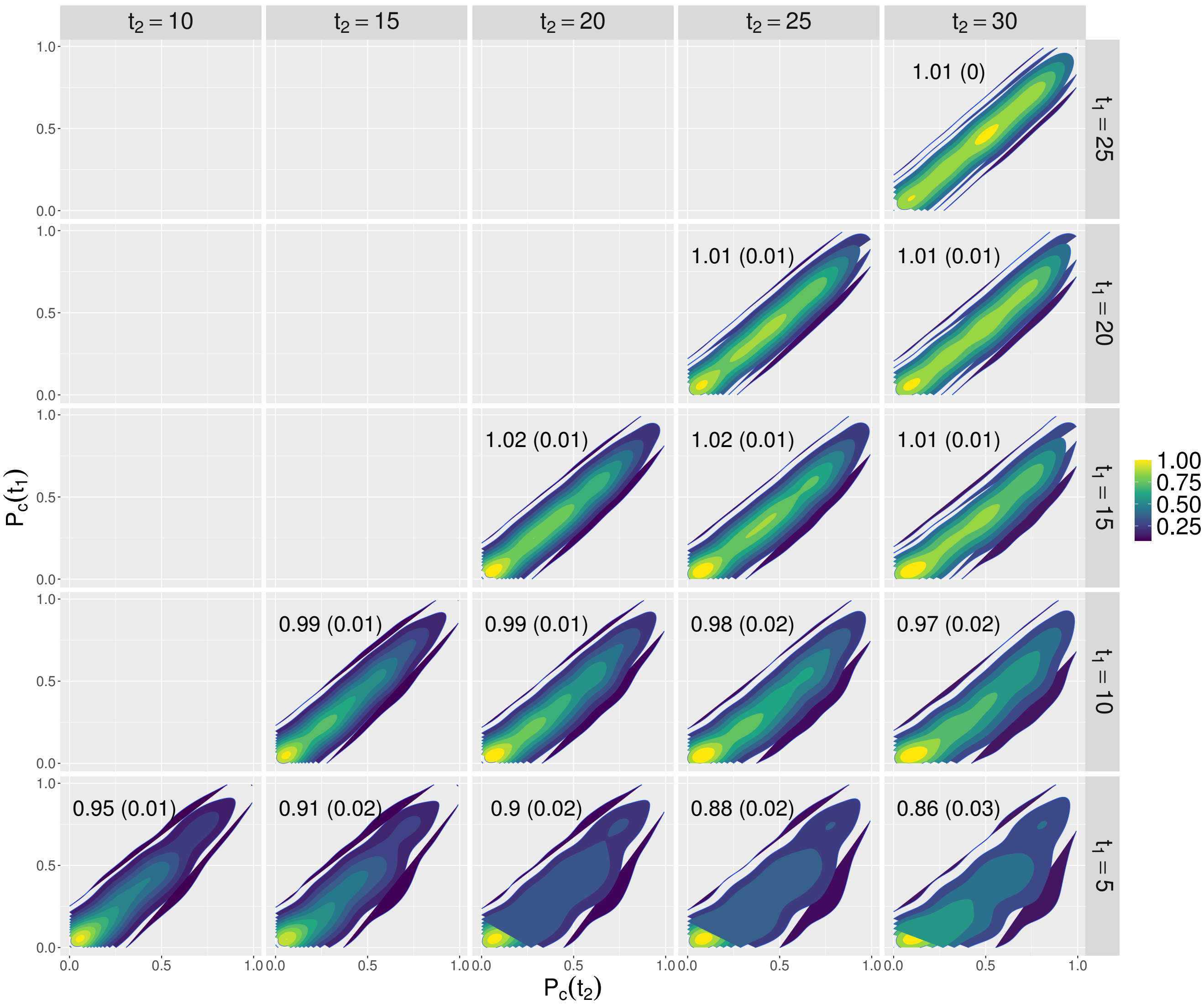


Figure 7. Kernel density estimation for the scatter points of $\mathrm{P}_c^{jb}(t_1)$ and $\mathrm{P}_c^{jb}(t_2)$. We also fitted a simple linear regression of $\mathrm{P}_c^{jb}(t_2)$ on $\mathrm{P}_c^{jb}(t_1)$. The estimated coefficient and the corresponding standard error (in parentheses) are displayed in each plot. The benchmark here includes publications in biology that were published in 1980.

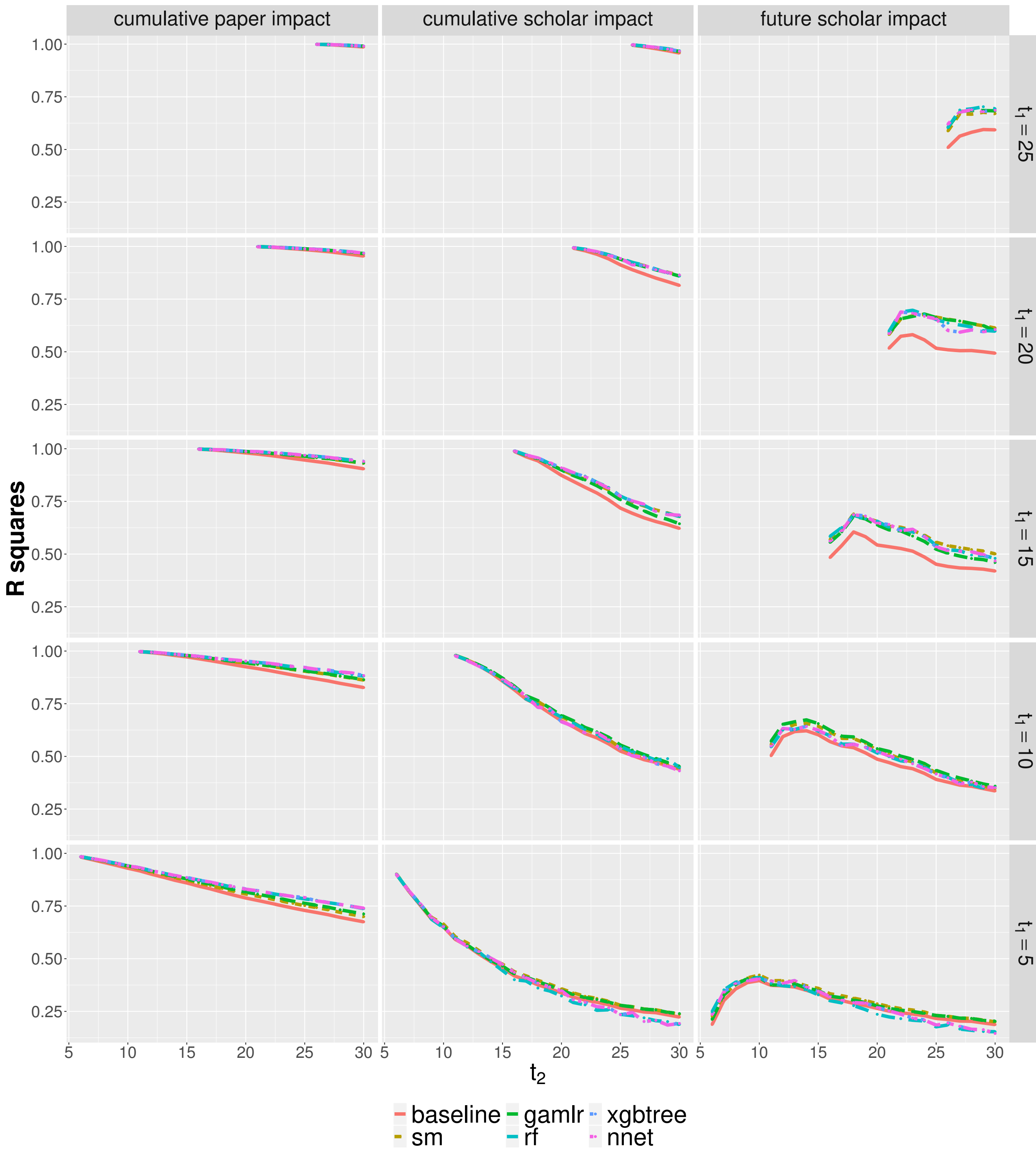


Figure 8. Testing R^2 of the predictive models. The lasso, ridge, and elastic net are outperformed by the Gamma lasso and hence are ignored for a better visualization.

References

1. Hirsch, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* **102,** 16569–16572 (2005).
2. Egghe, L. Theory and practise of the g-index. *Scientometrics* **69,** 131–152 (2006).
3. Connor, J. Google Scholar citations open to all. *Google Scholar Blog* (2011).
4. Chen, P., Xie, H., Maslov, S. & Redner, S. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* **1,** 8–15 (2007).
5. Walker, D., Xie, H., Yan, K.-K. & Maslov, S. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment* **2007,** P06010 (2007).
6. Ma, N., Guan, J. & Zhao, Y. Bringing PageRank to the citation analysis. *Information Processing & Management* **44,** 800–810 (2008).
7. Senanayake, U., Piraveenan, M. & Zomaya, A. The PageRank-index: Going beyond citation counts in quantifying scientific impact of researchers. *PLoS One* **10,** e0134794 (2015).
8. Schubert, A. & Braun, T. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics* **9,** 281–291 (1986).
9. Bornmann, L., Leydesdorff, L. & Mutz, R. The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics* **7,** 158–165 (2013).
10. Mingers, J. & Leydesdorff, L. A review of theory and practice in scientometrics. *European Journal of Operational Research* **246,** 1–19 (2015).
11. Bornmann, L., Tekles, A. & Leydesdorff, L. How well does I3 perform for impact measurement compared to other bibliometric indicators? The convergent validity of several (field-normalized) indicators. *Scientometrics* **119,** 1187–1205 (2019).
12. Price, D. d. S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27,** 292–306 (1976).
13. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
14. Peterson, G. J., Pressé, S. & Dill, K. A. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences* **107,** 16023–16027 (2010).
15. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105,** 17268–17272 (2008).
16. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74,** 47 (2002).
17. Hajra, K. B. & Sen, P. Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and its Applications* **368,** 575–582 (2006).
18. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Physical Review E* **62,** 1842 (2000).
19. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342,** 127–132 (2013).
20. Wang, J., Mei, Y. & Hicks, D. Science communication. Comment on "Quantifying long-term scientific impact". *Science* **345,** 149–149 (2014).
21. Wang, D., Song, C., Shen, H.-W. & Barabási, A.-L. Response to Comment on "Quantifying long-term scientific impact". *Science* **345,** 149–149 (2014).
22. Fu, L. D. & Aliferis, C. *Models for predicting and explaining citation count of biomedical articles* in *AMIA Annual symposium proceedings* **2008** (2008), 222.
23. Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L. & Haynes, R. B. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ* **336,** 655–657 (2008).
24. Ibáñez, A., Larrañaga, P. & Bielza, C. Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics* **25,** 3303–3309 (2009).
25. Mazloumian, A. Predicting scholars' scientific impact. *PLoS One* **7,** e49246 (2012).

26. Stern, D. I. High-ranked social science journal articles can be identified from early citation information. *PLoS One* **9,** e112520 (2014).

27. Weihs, L. & Etzioni, O. *Learning to predict citation-based impact measures* in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (2017), 49–58.

28. Hirsch, J. E. Does the h index have predictive power? *Proceedings of the National Academy of Sciences* **104,** 19193–19198 (2007).

29. Acuna, D. E., Allesina, S. & Kording, K. P. Future impact: Predicting scientific success. *Nature* **489,** 201 (2012).

30. Penner, O., Pan, R. K., Petersen, A. M., Kaski, K. & Fortunato, S. On the predictability of future impact in science. *Scientific Reports* **3,** 3052 (2013).

31. Allen, H. The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers* **77,** 1539–1669 (1914).

32. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67 (1970).

33. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 267–288 (1996).

34. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320 (2005).

35. Taddy, M. One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics* **26,** 525–536 (2017).

36. Liaw, A., Wiener, M., *et al.* Classification and regression by randomForest. *R news* **2,** 18–22 (2002).

37. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785–794.

38. Dickey, D. A. & Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74,** 427–431 (1979).

39. Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* **54,** 159–178 (1992).

40. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2019). `http://www.R-project.org/`.

41. Bischl, B. *et al.* mlr: Machine Learning in R. *Journal of Machine Learning Research* **17,** 1–5. `http://jmlr.org/papers/v17/15-066.html` (2016).

42. Allaire, J. & Chollet, F. *keras: R Interface to 'Keras'* R package version 2.2.4.1 (2019). `https://CRAN.R-project.org/package=keras`.

Supplemental Material:
On the usage of rank percentile in evaluating and predicting the scientific impact

Sen Tian, Panos Ipeirotis

A Rank percentile indicators for scholars

We consider the benchmark being the area of biology. In order to study the agreement of various indicators, for each indicator, we classify the scholars into four classes, class 1: $0 \leq \mathrm{S}_m^{ib}(t) < 0.25$, class 2: $0.25 \leq \mathrm{S}_m^{ib}(t) < 0.5$, class 3: $0.5 \leq \mathrm{S}_m^{ib}(t) < 0.75$ and class 4: $0.75 \leq \mathrm{S}_m^{ib}(t) \leq 1$. An agreement is when two (or three) different indicators belong to the same class. The overall agreement for all three indicators, S_{P5}, S_c, and S_h, is 51% at age 5 and 68% at age 30; that is, for about half of the scholars the three indicators agree with each other at age 5, while that number becomes around two third at age 30. Figure S1 displays pairwise agreement of the three indicators. We see that the agreement increases with the age. Furthermore, S_{P5} has large agreement with both S_c and S_h, which are 69% and 67%, respectively, at age 5, and 71% and 81%, respectively, at age 30.

Figure S2 shows the correlation between $\mathrm{S}_c^{ib}(t_1)$ and $\mathrm{S}_c^{ib}(t_2)$, and the correlation between $\mathrm{S}_h^{ib}(t_1)$ and $\mathrm{S}_h^{ib}(t_1)$. The magnitudes of correlations are similar to those for S_{P5} as shown in Figure 6b.

B Robustness of S_{P5}

Recall that $\mathrm{S}_{P5}^{ib}(t)$ is calculated based on an aggregation of the performances of publications that scholar i publishes by age t. Denote $\mathrm{m}_{P5}^{ib}(t)$ as the evaluation metric of scholar i at age t based on $\mathrm{P}_c^{jb}(5)$, that is $\mathrm{m}_{P5}^{ib}(t) = \sum_{j=1}^{N(t)} \mathrm{P}_c^{jb}(5)$, where $N(t)$ is the total number of publications of scholar i by age t. We illustrated in the paper that P_c exhibits high stability, and hence $\mathrm{P}_c^{jb}(5)$ can be applied to represent the performance of the publication.

We further demonstrate the robustness of S_{P5} by considering a longer citation history for each publication. Figure S3 illustrates that m_{P5} is highly correlated with m_{P10} at age $t = 1, \cdots, 30$, where $\mathrm{m}_{P10}^{ib}\mathrm{t}= \sum_{j=1}^{N(t)} \mathrm{P}_c^{jb}(10)$. We also consider the maximum, mean and median values of $\mathrm{P}_c^{jb}(t)$, e.g. $\mathrm{m}_{Pmax}^{ib}\mathrm{t}= \sum_{j=1}^{N(t)} \max_{t'} \mathrm{P}_c^{jb}(t')$. We see from the figure that these metrics also exhibit high correlations with m_{P5}. Furthermore, we perform a Wilcoxon paired signed-rank test to compare the differences between S_{P5} and S_{P10} at each age $t = 1, \cdots, 30$, and the p-values are close to 1; indicating that the differences are not statistically significant. Similar conclusions can be drawn for other indicators being considered.

C Stationarity test

Two commonly used statistical tests for stationarity are the Dicky-Fuller test [1] and KPSS test [2]. These two tests formulate the hypothesis testing problems differently. Dicky-Fuller test assumes a unit root presented in the series. A unit root means that the series is $I(1)$, i.e. integrated order 1 and the first differenced series is stationary. The more negative the test statistic is, the stronger the rejection of the null. On the other hand, KPSS test assumes the null as the series being stationary, i.e. $I(0)$. KPSS test is slightly more general since it allows testing a series being non-stationary but does not present a unit root. The more positive the test statistic is, the stronger the rejection of the null. Both tests include the drift in the test equations but exclude the trend, since we do not observe significant trends in the series.

The test statistics are shown in Figure S4. The dashed lines indicate the critical values at 5% level. KPSS test indicates that P_c and S_{P5} are non-stationary series, and we do not have enough evidence to reject them being $I(1)$ according to the Dicky-Fuller test. Furthermore, the differenced series are stationary based on both tests.

Method	Tuning parameters
Lasso	Penalty strength parameter
Ridge	Penalty strength parameter
Elastic net	Penalty strength parameter Penalty gap parameter
Gamma lasso	Penalty strength parameter Convexity parameter
Random forest	Number of trees to grow Number of variables used at each split Minimum number of observations in a node
xgbtree	Maximum number of iterations Learning rate Regularization parameter Maximum depth of the tree Minimum number of observations in each child leaf Number of observations supplied to a tree Number of features supplied to a tree Regularization parameter for ridge penalty Regularization parameter for LASSO penalty
Deep neural network	Number of layers Learning rate Number of hidden units at each layer Dropout rate Regularization parameter

Table S1. Hyperparameter(s) of the machine learning models.

Feature	Description
pub_cit_cumulative	total citations of publication j
pub_cit_yearly	yearly citations of publication j received in t_1
pub_cit_peryear	average citations of publication j over age
pub_rp_cumulative	rank percentile indicator calculated based on total citations, i.e. $\mathrm{P}_c^{jb}(t_1)$
pub_rp_yearly	rank percentile indicator calculated based on yearly citations at t_1
aut_cit_cumulative	total citations of author i
aut_cit_yearly	yearly citations of author i at t_1
aut_npub_cumulative	total number of publications of author i
aut_npub_yearly	yearly number of publications of author i at t_1
aut_cit_perpaper	average citations per paper for author i
aut_h_index	h-index of author i
aut_g_index	g-index of author i
aut_maxcit_pub	largest citation that a single paper of author i has received
aut_rprp5_cumulative	rank percentile calculated based on all papers, i.e. $\mathrm{S}_{P5}^{ib}(t_1)$
aut_rprp5_yearly	rank percentile calculated based on just papers written at t_1
*_delta	the difference over the last two ages for each of the above features

Table S2. Features for predicting the impact of publication j of scholar i. The features are created at t_1.

Feature	Description
aut_cit_cumulative	total citations of author i
aut_cit_yearly	yearly citations of author i at age t_1
aut_npub_cumulative	number of publications of author i
aut_npub_yearly	yearly number of publications of author i at age t_1
aut_h_index	h-index of author i
aut_g_index	g-index of author i
aut_cit_peryear	average citations per age of author i
aut_rprp5_cumulative	rank percentile calculated using all publications, i.e. $S^{ib}_{P5}(t_1)$
aut_rprp5_yearly	rank percentile calculated using just publications written in age t_1
pub_cit_cumulative_{min,mean,max}	citations received by each of the publications
pub_cit_yearly_{min,mean,max}	citations received by each of the publications written at age t_1
pub_rp_cumulative_{min,mean,max}	publication rank percentiles calculated based on total citations
pub_rp_yearly_{min,mean,max}	publication rank percentiles calculated based on citations at age t_1
*_delta	the difference over the last two ages for each of the above features

Table S3. Features for predicting the impact of scholar i. The features are created at t_1.

Figure S1. Rank percentile indicators are classified into four groups, that are class 1: $0 \leq S_m^{ib}(t) < 0.25$, class 2: $0.25 \leq S_m^{ib}(t) < 0.5$, class 3: $0.5 \leq S_m^{ib}(t) < 0.75$ and class 4: $0.75 \leq S_m^{ib}(t) \leq 1$. The agreement of classes (sum of the anti-diagonal elements) is displayed in the title of each panel. The benchmark is biology. The agreement for all three indicators is 51% at age 5 and is 68% at age 30.

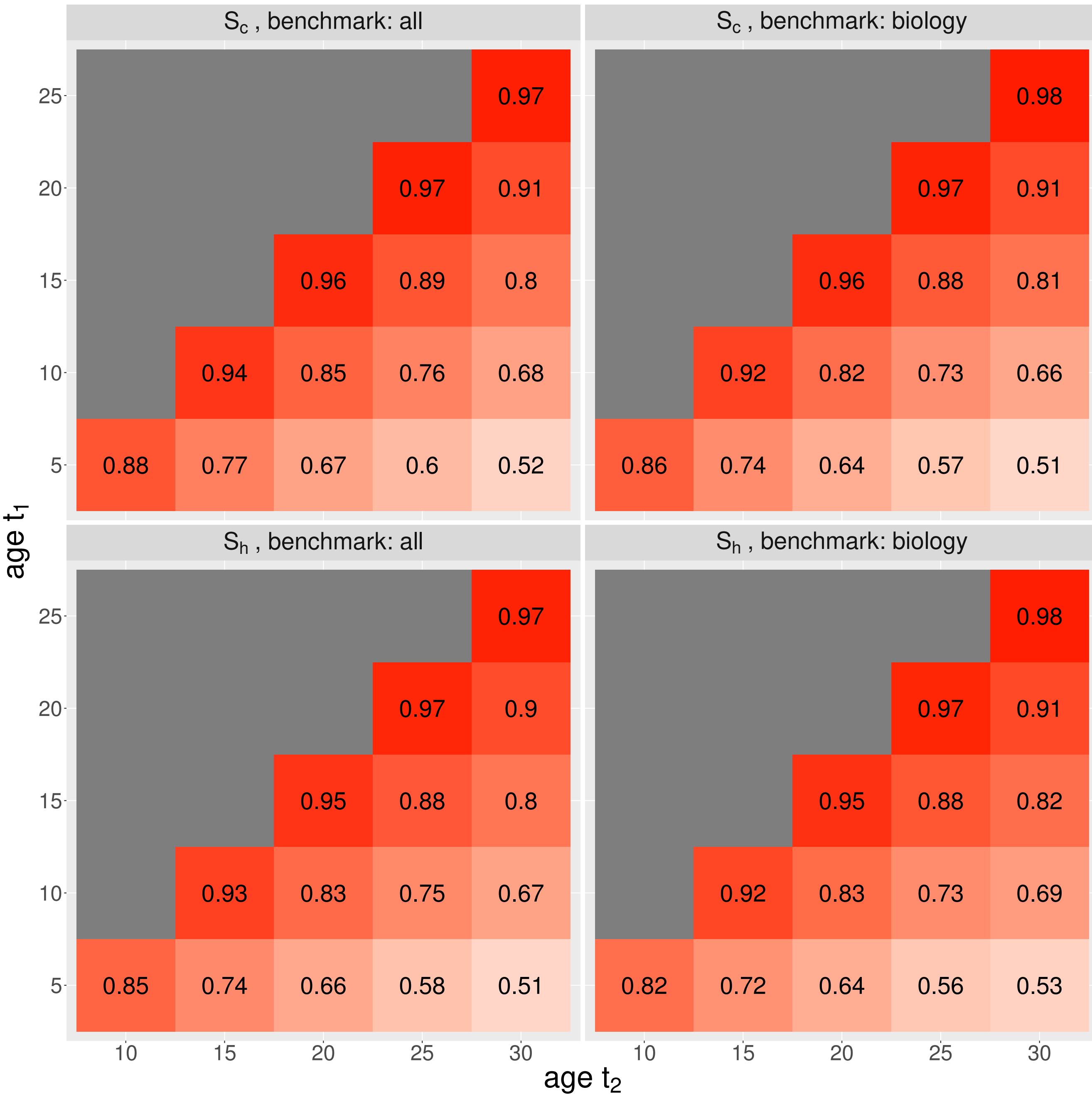


Figure S2. The Pearson's correlation between rank percentile indicators at two different ages.

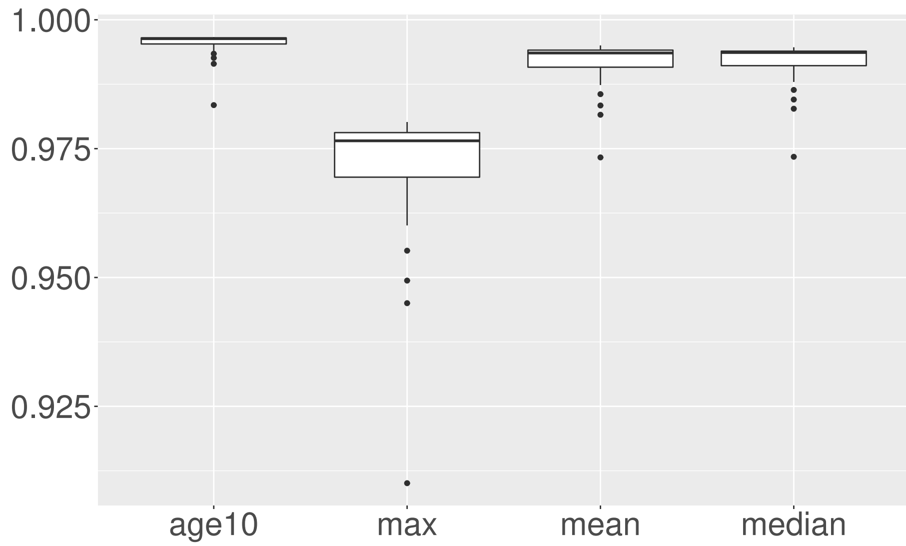


Figure S3. Pearson's correlation between m_{P5} and other choices of evaluation metric. Age 10, max, mean and median correspond to m_{P10}, m_{Pmax}, m_{Pmean} and $m_{Pmedian}$, respectively. The correlation is calculated at each age $t = 1, \cdots, 30$.

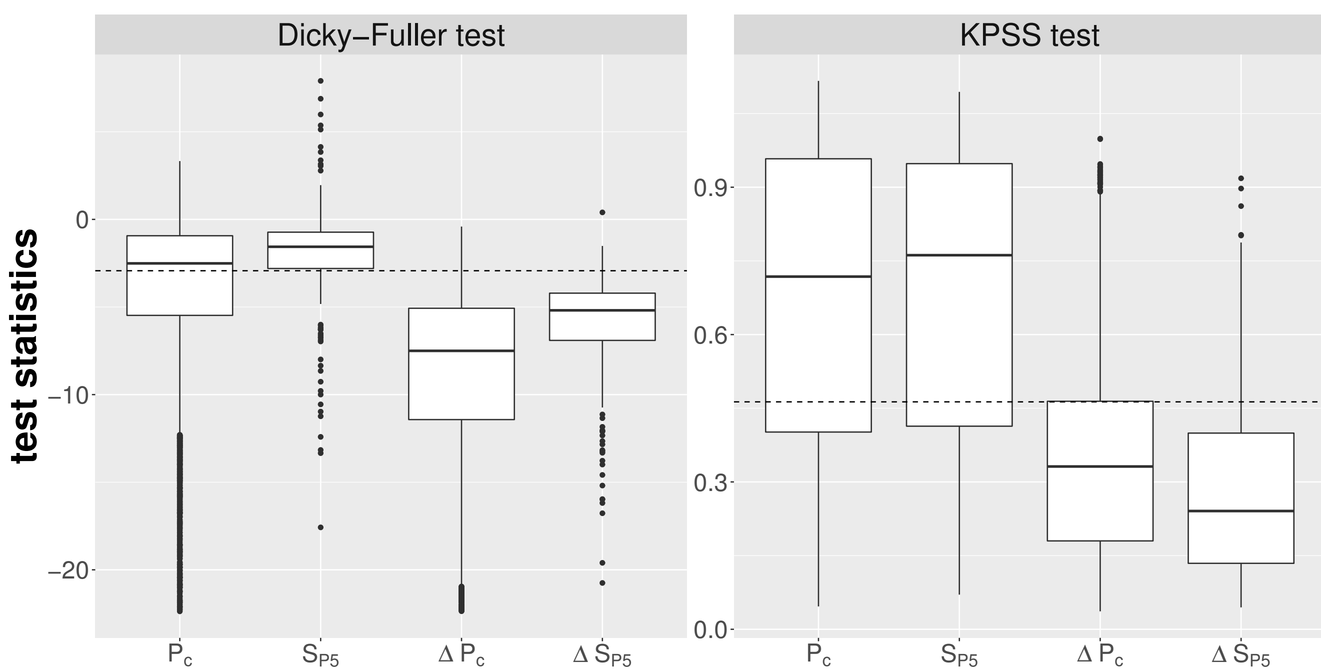


Figure S4. Statistical tests for the stationarity of rank percentile series. Both tests are applied on every individual series, and the test statistics are presented. The 5% critical value for each test is illustrated by the dashed horizontal line. Both tests suggest that publication indicator P_c and scholar indicator S_{P5} are non-stationary, while their differenced series are stationary.

D Other tables and figures

benchmark	all	biology	tenured
# publications	801239	194713	176404
# scholars	14358	3410	2706
# citations per publication by age 5	45	56	49
# citations per scholar by age 5	172	209	332

Table S4. Summary statistics of the dataset.

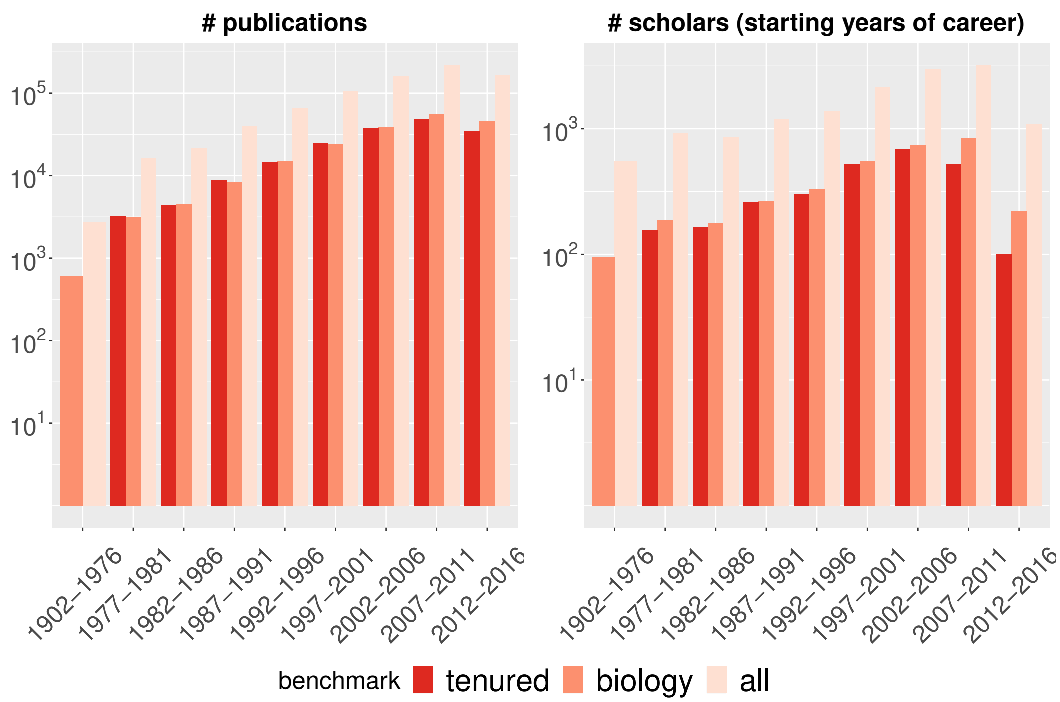


Figure S5. Left panel: number of papers published in a certain period; **right panel**: number of authors who start their careers in a certain period.

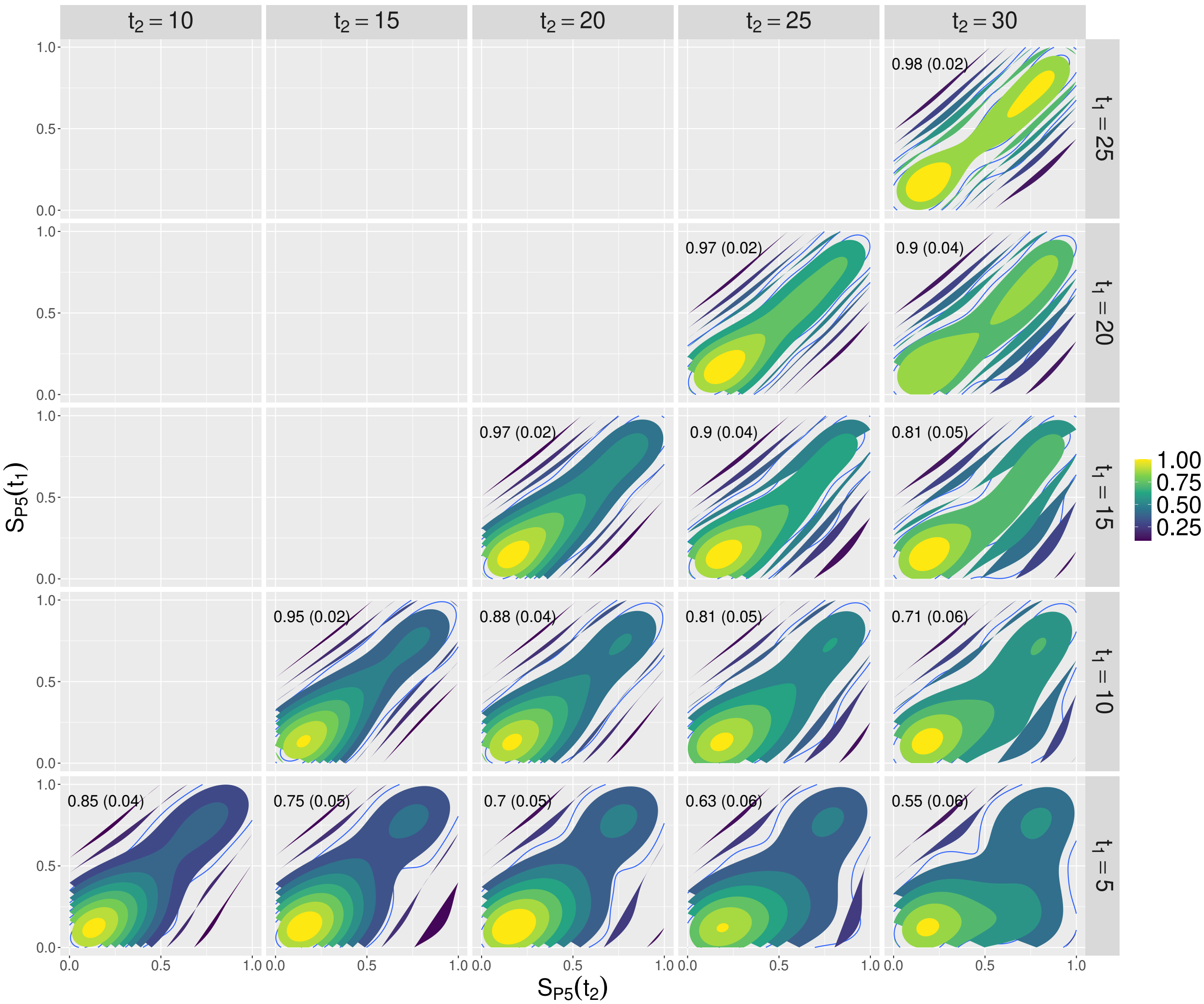


Figure S6. Kernel density estimation for the scatter points of $S^{ib}_{P5}(t_1)$ and $S^{ib}_{P5}(t_2)$. We also fit a simple linear regression of $S^{ib}_{P5}(t_2)$ on $S^{ib}_{P5}(t_1)$. The estimated coefficient and the corresponding standard error (in the parentheses) are displayed in each plot.

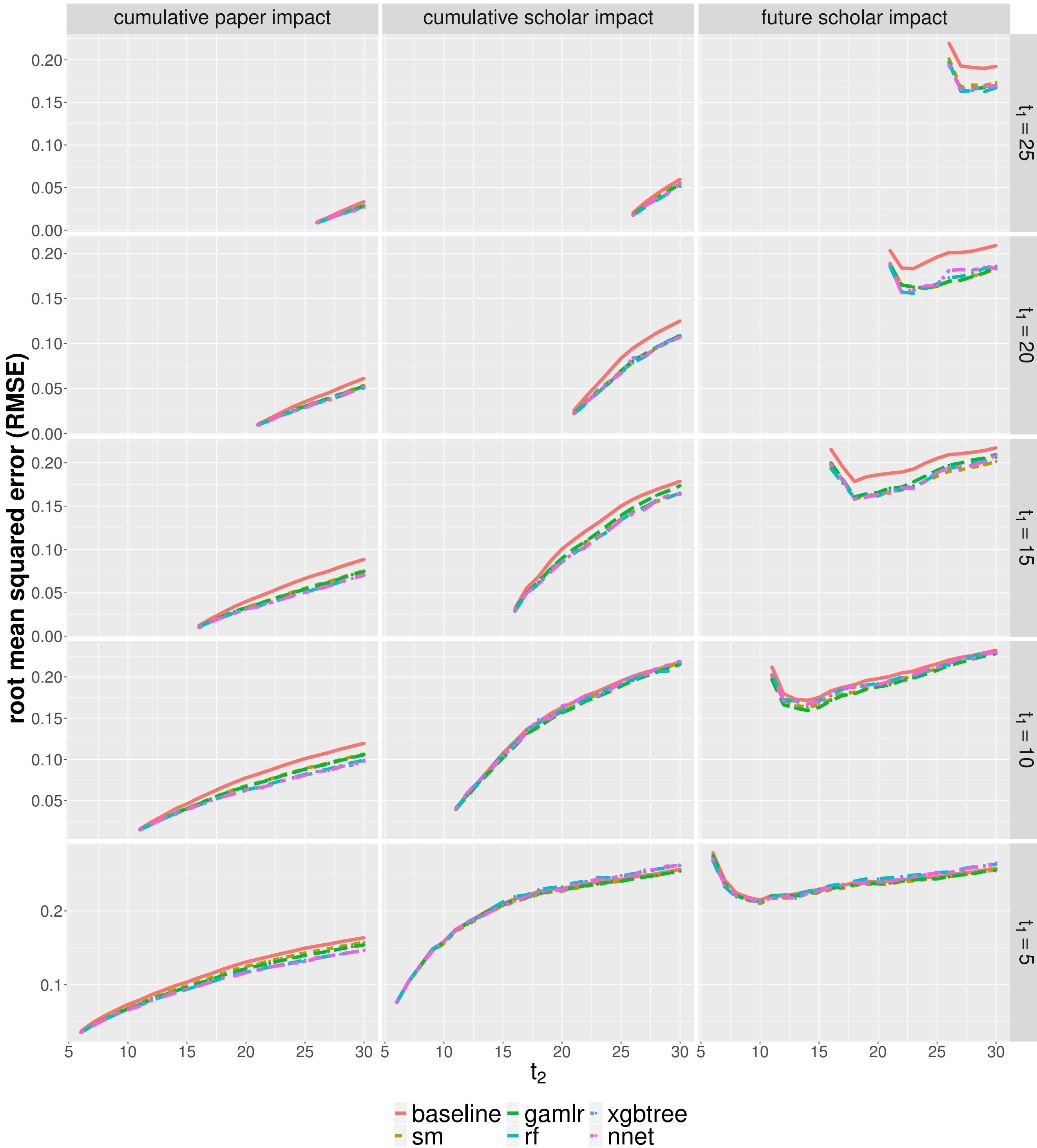


Figure S7. Root mean squared error of the predictive models. The lasso, ridge and elastic net are outperformed by Gamma lasso, and hence are ignored for a better visualization.

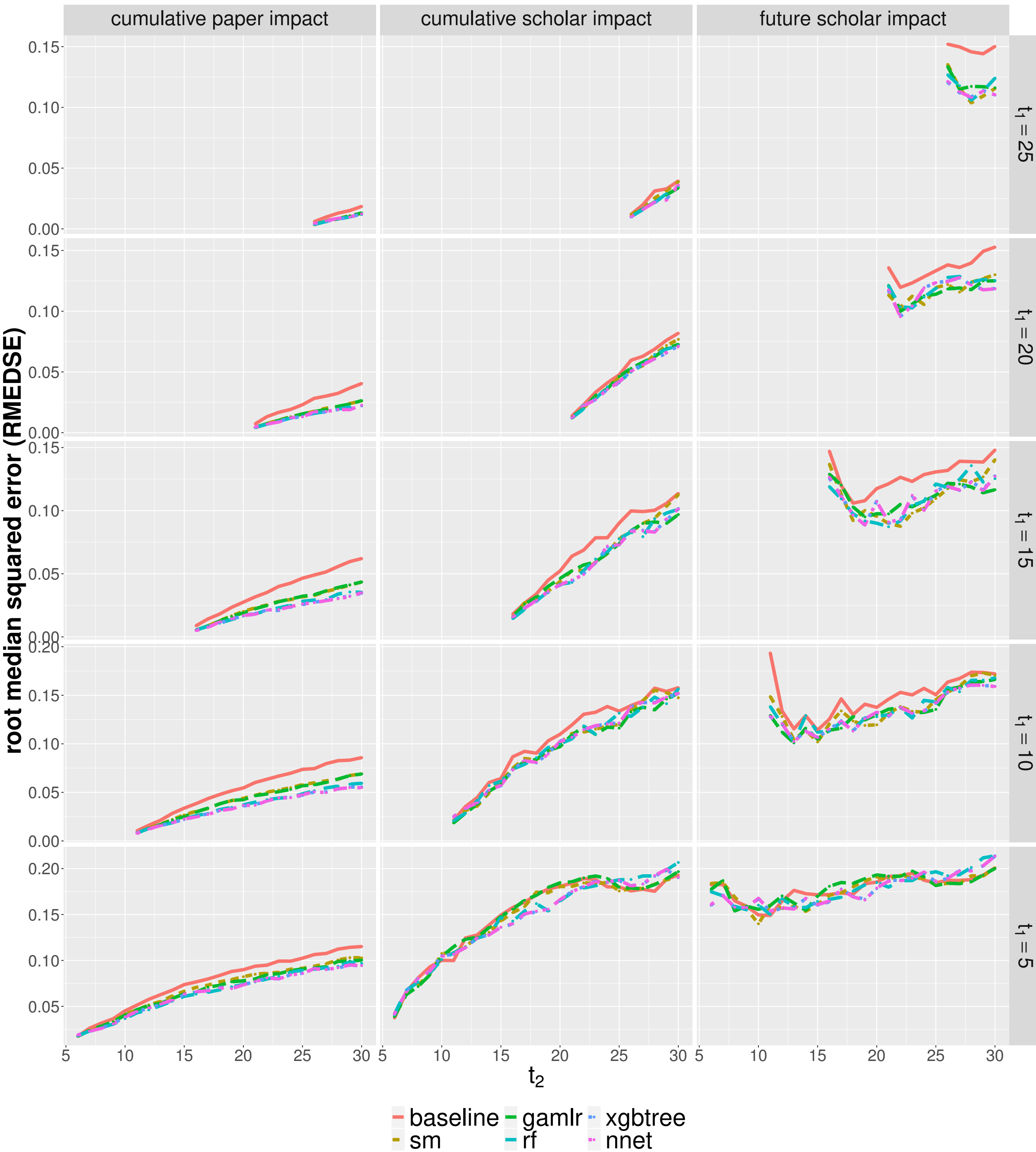


Figure S8. Root median squared error of the predictive models.

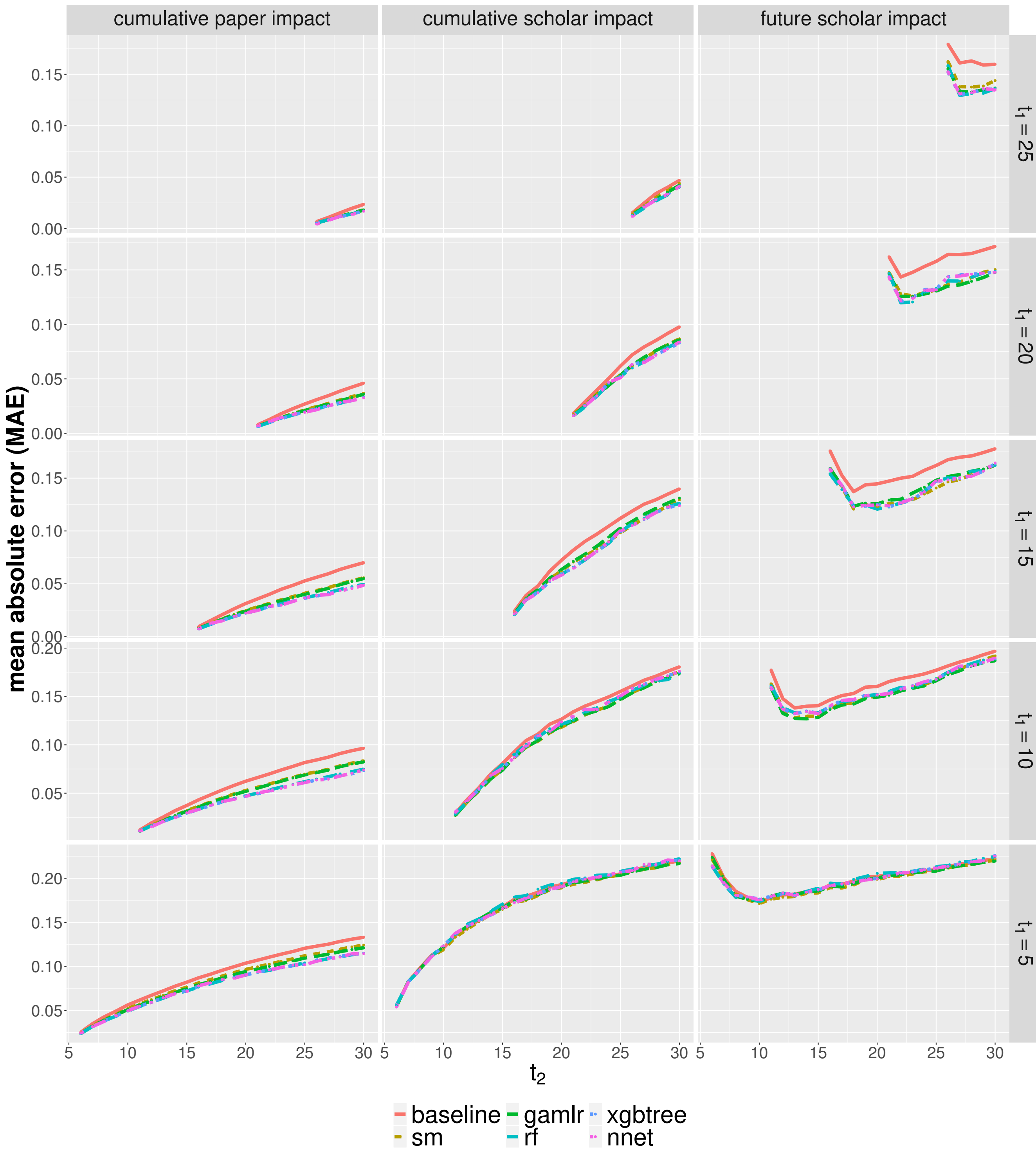


Figure S9. Mean absolute error of the predictive models.