On the use of rank percentile in evaluating scientific impacts, and its predictability

Panos G. Ipeirotis[1] **and Sen Tian**[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 P.G.I. (email: panos@stern.nyu.edu)

ABSTRACT

Comparing the impact of a scholar to her senior colleagues or other young scholars in a different field, is often considered in the recruitment or tenureship decisions. Bibliometric indicators such as the number of citations or h-index can not be used directly, since they are highly influenced by the subject area, seniority of the author, etc. Here, we discuss a simple framework to construct rank percentile indicators based on the bibliometric indicators, allowing us to assess the relative performance of a publication or a scholar. Furthermore, we show that the rank percentile indicators have high predictive powers, and can be modeled using 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 complexity added.

Introduction

Quantifying the scientific impact is a long-standing challenge. The most frequently used metric is the number of citations[1–4]. Despite its simplicity, citation counts have been criticized to be biased serving as the proxy for scientific importance, and various alternatives have been proposed in reliance on the citations. The most well-known metric is the h-index[1] that is defined as the maximum number h for which the scholar has h publications each with at least h citations. A scholar only participating in a small number of highly cited works, or a large number of low-profiled papers, will have a high citation counts but a low h-index, since h-index rewards a consistent stream of impactful efforts. Since then, numerous indices have been proposed to remedy the drawbacks of h-index, e.g. g-index[5], m-index[1] and i-10 index[6]. Unfortunately, these alternatives solve existing problems by creating new ones. For instance, the actual citations are irrelevant to the h-index once they exceeds h. On the contrary, g-index allows citations from highly cited works to boost the lower cited ones, but it can be saturated once the average number of citations per paper is larger than the number of papers. Furthermore, 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[7–9] 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[10]. However, there have been discussions about the dis-similarity between citation network and WWW where the PageRank algorithm is originally designed for[7].

Predicting the scientific impact is the second challenge. To potentially assign fundings or tenures, the university committees not only evaluate a scholar's past achievements, but more importantly attempt to project the future success. Much attention have focused on understanding the mechanisms that drive the citation dynamics of publications. Three main factors include the citation distributions following scaling laws[2,11–13], aging[12,14–16], and perceived novelty or importance[17]. The mechanistic model can be applied to predict the future citation evolutions of publications[17]. Unfortunately, the model estimation potentially relies on long citation history[18,19], and each publication needs to be dealt with individually, hence it is not scalable for large-scale analysis. Another type of the predictive models formulates the task as a supervised learning problem. Indicators that can potentially drive the future impact are summarized as features in order to make predictions for metrics such as citations[20–25] and h-index[25–28]. With the help of sophisticated machine learning algorithms, these models can be easily scaled to account for large-scale dataset.

Comparing the scientific impact is yet another challenge. How do we compare a publication from a less-liquid field say mathematics, with a publication in biology? Reference set or benchmark has been introduced to make such comparison feasible[29]. The benchmark characterizes a specific field, a certain publishing year or an explicit document type. The citations of publications within each benchmark are normalized with respect to the benchmark. Mean-based indicators normalize the citations by the expected citation impact of the benchmark, that can be estimated by the arithmetic mean of citations for all publications in that benchmark[30]. As 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 highly cited publications. Fortunately, these drawbacks can be largely avoided by using the rank percentile indicators[31–33], which normalize the citations by their rank relative to the citations of other publications in the benchmark.

However, little is known about the evolution of the rank percentile indicator over time and its predictive power. In this paper, we connect the third challenge with the other two. We start from discussing a general framework of formulating a rank percentile indicator, that is flexible in terms of the evaluation metric for scientific impacts, the choice of benchmark, age and the entity of interest. We then study the pros and cons of different rank percentile indicators formulated using different evaluation metrics. Furthermore, we factor the age into rank percentile indicators and demonstrate that they have high predictive powers. The dataset that we use is a large-scale Google Scholar dataset containing scholars across all disciplines in the top US universities.

Results

Fundamental elements of formulating rank percentile indicators

Four fundamental elements to calculate the rank percentile indicators are (1) entity; (2) benchmark; (3) metric; and (4) age. The entity specifies the target of interest. For example, it can be a publication (P) or a scholar (S). The benchmark (b) characterizes the individuals that the entity is compared against. For instance, the benchmark can be all the publications in computer science. It can also be the faculty members in a department competing for tenure decisions. The metric (m) specifies the way that the entity is evaluated. For example, it can be the number of citations or h-index. Finally, the age (t) is the number of years that the entity has been alive. The age of a publication is the number of years since it was published, while the age of a scholar is the number of years since his/her first paper was published. By combining these four elements, we use $\mathrm{P}^{m}_{jt}(b)$ and $\mathrm{S}^{m}_{it}(b)$ to denote the rank percentile indicators for publication j and scholar i, respectively.

The framework of calculating rank percentile indicators

Suppose the benchmark b contains N publications across T years. For each publication $j \in \{1, \cdots, N\}$, we have a history of some evaluation metric m_{jt} ($\mathrm{m}_{jt} \geq 0$) at each age $t = 1, \cdots, T_j$ where T_j denotes the total number of years since it's published. The rank percentile indicator of publication j at age $t^\star$, $\mathrm{P}^{m}_{jt^\star}(b)$, is computed as follows:

(1) Consider all the $N_{t^\star}$ publications that have $T_j \geq t^\star$, and calculate the rank $\mathrm{r}_{jt^\star}$ of publication j based on its metric $\mathrm{m}_{jt^\star}$. An average rank is assigned to $r_{jt^\star}$ if there exist other publications that have the same metric value as $\mathrm{m}_{jt^\star}$.

(2) The rank percentile is given by

$$\mathrm{P}^m_{jt^\star}(b) = \begin{cases} 0, & \text{if } \mathrm{m}_{jt^\star} = 0, \\ (\mathrm{r}_{jt^\star} - 0.5)/N_{t^\star}, & \text{otherwise.} \end{cases} \tag{1}$$

With the compromise $0.5/N_{t^\star}$ in (1), the median paper will be assigned 50% percentile, and the tails of the citation distribution are treated symmetrically[34].

The above framework can be easily adapted to compute the rank percentile indicators for scholar i at age $t^\star$, $\mathrm{S}^m_{jt^\star}(b)$. Furthermore, it is flexible in terms of the choice of the evaluation metric and benchmark. In this paper, we study two benchmarks. The first contains all the scholars in our dataset that come from various disciplines, different institutes and have different starting points of careers. The second benchmark only includes scholars in the area of biological science. Meanwhile, we consider two common choices of the evaluation metric, that are the number of citations (c) and h-index (h). The following are the indicators considered in this paper.

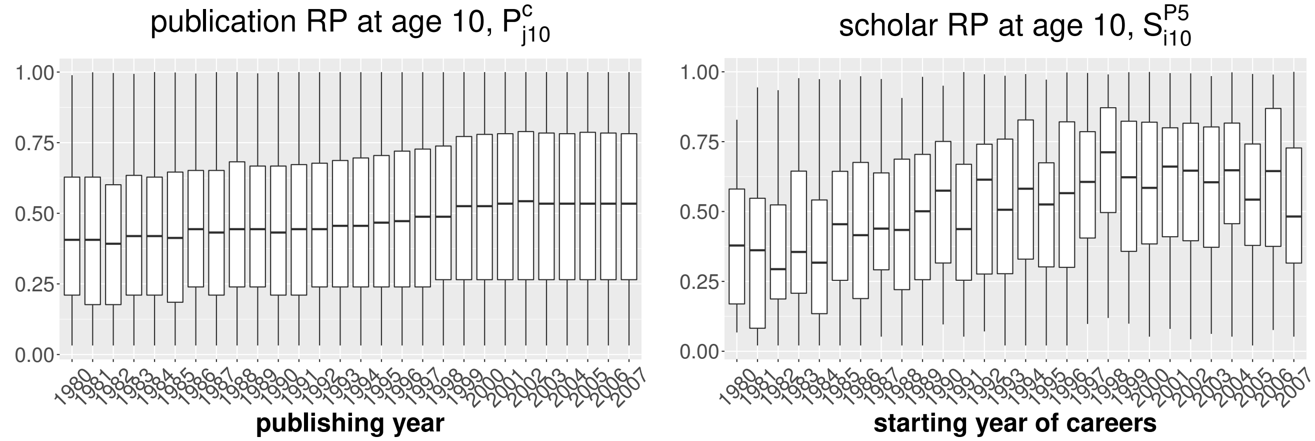
- rank percentile indicator for publication j:
  - $\mathbf{P}^c_{jt}(b)$: m_{jt} is the total citations c_{jt} that the paper receives by age t,
- rank percentile indicator for scholar i:
  - $\mathbf{S}^c_{it}(b)$: m_{it} is the total citations c_{it} that the scholar receives by age t.
  - $\mathbf{S}^h_{it}(b)$: m_{it} is the h-index h_{it} of the scholar at age t.
  - $\mathbf{S}^{P5}_{it}(b)$: $m_{it} = \sum_{j=1}^{N_t^{(i)}} \mathrm{P}^c_{j5}(b)$, where $N_t^{(i)}$ denotes the total number of papers that the scholar publishes by age t. For paper with ages less than 5, i.e. $T_j < 5$, we take the most recent performance $\mathrm{P}^c_{jT_j}(b)$ instead.

The rank percentile indicators are highly interpretable. For instance, $\mathrm{P}^c_{jt}(b) = 0.9$, immediately tells us that paper j performs better than 90% of the papers in the benchmark at age t. Furthermore, two papers from different benchmarks, such as different fields, are directly comparable using the rank percentile indicators.

The stableness of the rank percentile indicators

The benchmarks that we study in this paper do not restrict the publications to have the same publishing year or the scholars to have the same starting year of their careers. In many situations, we would like to examine a scholar's performance relative to cohorts that are more senior to her. For instance, to make a promotion decision about a candidate who is in the sixth year of her academic career, the committee may compare her with all other faculty members of the department in the sixth year of their academic careers. The comparison will not be valid, if the candidate is more likely to attract citations than her senior colleagues who start their careers many years before. It may well be the academic environment that results in a better performance of the candidate, rather than the internal factors such as the creativity and productivity.

Figure 1 shows P^{c}_{j10} grouped by the publishing years and S^{P5}_{i10} grouped by the starting year of academic careers, where the benchmark contains all the tenured scholars in our dataset. We see that both indicators are reasonably stable over the physical years. We do not see the pattern that papers or scholars attract more citations than those years before, supporting the validity of rank percentile indicators.


publication RP at age 10, P^{c}_{j10}

scholar RP at age 10, S^{P5}_{i10}

Figure 1. P^{c}_{j10} grouped by the publishing years and and S^{P5}_{i10} grouped by the starting years of academic careers. The benchmark contains professors from various disciplines who receive their tenureships no later than year 2016.

The predictability of the rank percentile indicators

A discussion on various types of rank percentile indicators for scholars

$\mathrm{S}^{c,b}_{it}$, $\mathrm{S}^{h,b}_{it}$ and $\mathrm{S}^{P5,b}_{it}$ are built upon an aggregation of the performances of the papers that the scholar publish by age τ. $\mathrm{S}^{c,b}_{it}$ and $\mathrm{S}^{h,b}_{it}$ use the citations c_{jt} to evaluate publication j while $\mathrm{S}^{P5,b}_{it}$ uses $\mathrm{P}^{c,b}_{j5}$, i.e. the rank percentile of publication j at age 5. Meanwhile, the aggregation function for $\mathrm{S}^{c,b}_{it}$ and $\mathrm{S}^{P5,b}_{it}$ is sum, and it's a threshold function for $\mathrm{S}^{h,b}_{it}$.

$\mathrm{S}^{P5,b}_{it}$ improves the major drawbacks of the other two indicators. First, it removes the seniority effect of publications, by evaluating the publications by their performances at age 5. However, both the citations c_{it} and h-index h_{it} are dominated by 'senior' publications and newly published works can hardly contribute. Second, it improves over $\mathrm{S}^{c,b}_{it}$ since it hardly rewards authors who publish a large number of low-impact works or only participate in a small number of high-impact projects. Similar to the h-index, the evaluation metric m_{it} of rp.rp5 limits the contribution of a single publication to be at most 1 due to the definition of rank percentile. However, the absolute number of citations is unlimited and is highly influenced by extreme values. Last, by comparing with rp.h, it penalizes authors who are not truly innovative, but carefully massage their h-indices by publishing a number of papers that are not top-class but attract just enough amount of citations to boost the h-index. As long as a paper is among the top h papers, the actual number of citations is irrelevant for rp.h, but it still affects rp.rp5.

We demonstrate these advantages of rp.rp5 by examining some extreme cases. We create three synthetic academic careers. Author A publishes a substantial number of publications at each age (more than 90% of other authors in the benchmark), while all of the publications have little impacts. Meanwhile, author B and author C only publish 1 paper at the beginning of their careers, where author B's paper

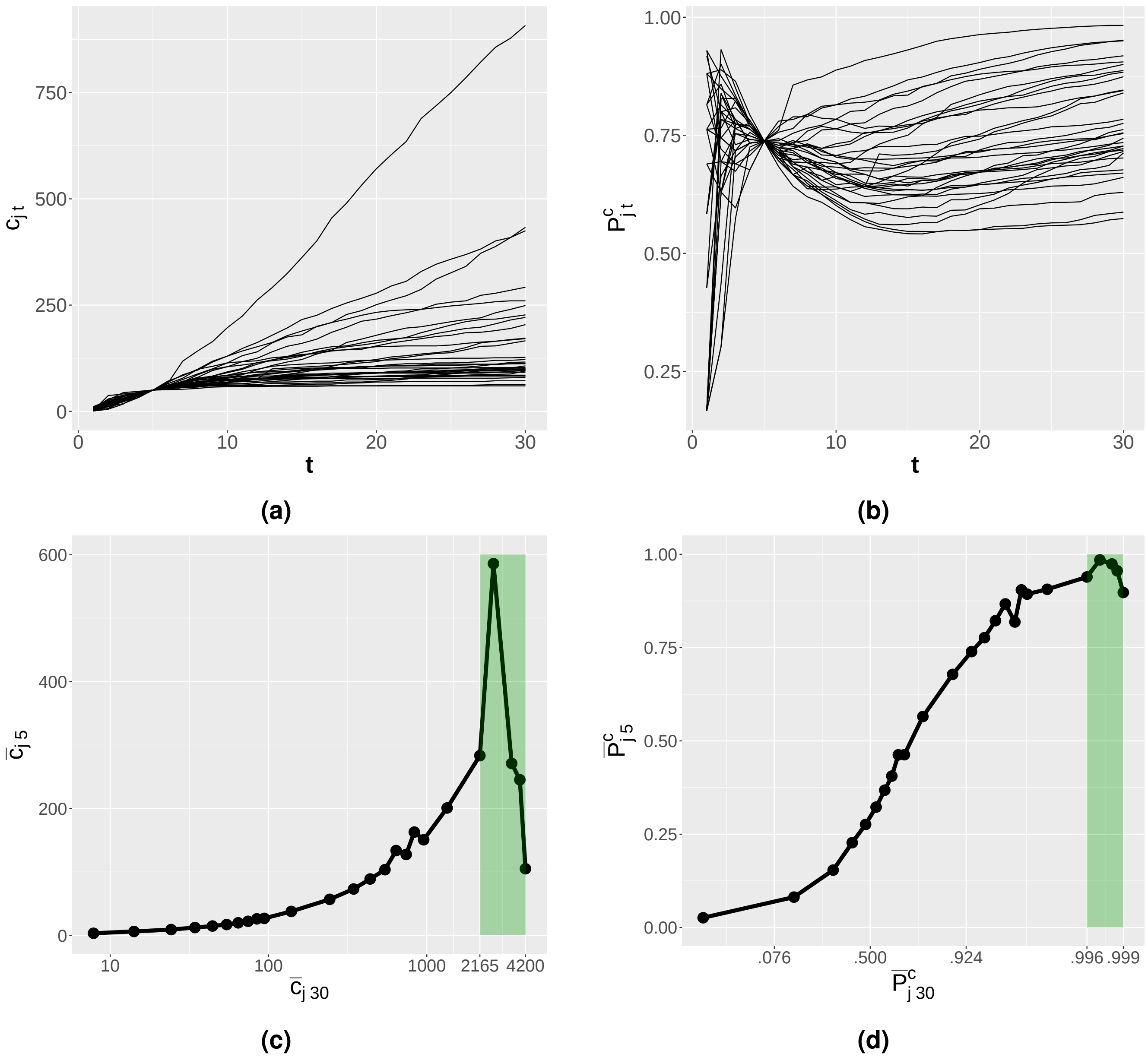


Figure 2. The predictability of citations and rank percentiles. The benchmark here is biology. Figure 2a and 2b show the citations and the corresponding rank percentiles for papers that have 50 citations by age 5. Figure 2c displays the average citations by age 5 over the average citations by age 30, for different sets of publications, which are pre-specified by dividing the range of $\bar{c}_{j30}$ into equal intervals in the log scale. Figure 2d shows the corresponding average rank percentiles for the same sets of publications. Note that we do not claim originality for these plots, as figures 2a and 2c have been illustrated via a different dataset[17].

is astonishing while author C's is average. All three authors have h-index as 1 throughout their careers. Figure **??** shows the author rp for these three artificial authors. We see that for author A and B, rp.c is substantially larger than the other two, and it dies down slowly. Even though author B only publishes one ground-breaking paper, the author remains in the top 50% even by age 12. Both rp.rp5 and rp.h are better characterizing the performances of these authors. Meanwhile, author C has the same rp.h as author B although his/her publication has much less impact. In this case, rp.rp5 and rp.c are better especially at the beginning of the careers.

Besides these special cases, for the majority of authors in our dataset, we see a large agreement between rp.rp5 and the other two types of rp. Consider the benchmark being biology. In order to study the agreement, for each indicator, we classify the authors into four classes, class 1: $0 \leq \text{rp} < 0.25$, class 2: $0.25 \leq \text{rp} < 0.5$, class 3: $0.5 \leq \text{rp} < 0.75$ and class 4: $0.75 \leq \text{rp} \leq 1$. An agreement in the classification of an author is where two (or three) different rp belong to the same class. The overall agreement for all three rp is 51% at age 5 and 68% at age 30, i.e. about half of the authors result in the same classifications of the three rp at age 5, while that number becomes around two third at age 30. Figure **??** shows the detailed classifications for every pair of the three rp. As we can see that, the agreement increases with the age. Furthermore, rp.rp5 has large agreements with both rp.c and rp.h, that are 69% and 67% respectively at age 5, and 71% and 81% respectively at age 30.

We've shown the advantage of using rp.rp5 over indicators like rp.c and rp.h. A remaining question is why we use rp.c_{j5} rather than say rp.c_{j10} to represent the quality of the paper. It turns out that $\text{rp.c}_{j\tau}$ is highly stable over ages, and rp.c_{j5} is very close to rp.c_{j10} for the majority of papers. This will be discussed in the section below. Therefore, we are using less citation information and get similar accuracies. Other choices involve taking the summary statistic of the entire history of $\text{rp.c}_{j\tau}$ $(\tau = 1, \cdots, T_j)$. For example, we can take the best performance along the history, i.e. $\max\limits_{t=1,\cdots,T_j} \text{rp.c}_{jt}$. We demonstrate that the evaluation metric $\text{m}_{i\tau}$ formulated based on these alternatives is highly correlated with $\text{m}_{i\tau}$ formulated rp.c_{j5}. Furthermore, the rank percentile indicators based on these alternatives are not statistically different from rp.rp5. The details are discussed in the supplemental material section A.

Predictive power of rank percentile indicators

Citations have been shown to lack long-term predictive power[17]. Figure 2a shows that publications with the same citations by age 5, can have noticeably different citation paths and long-term effects. Meanwhile, exceptional and creative ideas can normally take longer to be appreciated by the community. As shown in figure 2c, the correlation between short-term citations and long-term citations breaks down for most highly-cited papers (the shaded area). Such problems affect less for rank percentiles, as evidenced by figure 2b and 2d. For the papers having $\text{rp.c}_{j5} \approx 0.75$, their rp.c_{j30} are all above 0.5. Meanwhile, the correlation between short-term and long-term rp is still very strong for the most highly-cited papers.

Publication $\text{rp.c}_{j\tau}$ and scholar $\text{rp.rp5}_{i\tau}$ are highly predictive and stable over age. Figure 3a shows the ability of the indicators to predict their own values. We see an extremely high predictive power for rp.c, which indicates that a publication with high rp.c after τ_1 ages is very likely to have high rp.c after τ_2 ages where $\tau_2 > \tau_1$. Meanwhile, the predictive power dies down as the forecast horizon gets larger, which simply reflects the difficulty of long-term forecast. Also, the correlation becomes higher when we have a longer history of the publication, e.g. it increases from 0.79 to 0.86 as τ_1 changes from 5 to 10 while keeping the forecasting horizon fixed at $\tau_2 - \tau_1 = 5$. Finally, we see a slightly higher predictive power when we restrict the benchmark to only include publications in biology.

Similar patterns can be observed for rp.rp5 according to figure 3a, although it has lower predictive power than rp.c. To be more specific, it starts with a high predictive power but dies down fast as the forecast horizon becomes larger. For instance, the correlation drops from 0.99 to 0.94 for rp.c while it decreases from 0.94 to 0.77 for rp.rp5, by fixing $\tau_1 = 15$. This is result from the fact that forecasting future impact of future works is considerably harder than forecasting the future impact of existing works. Predicting $\text{rp.c}_{j\tau_2}$ belongs to the latter. On the other hand, $\text{rp.rp5}_{i\tau_2}$ is based on the set of papers that scholar i publishes by age τ_2, which contains papers published by τ_1 and papers published in the period of $(\tau_1, \tau_2]$. Hence, predicting $\text{rp.rp5}_{i\tau_2}$ involves predicting the future impact of existing works as well as the future impact of future works. As the forecast horizon $\tau_2 - \tau_1$ becomes larger, we potentially have more 'predicting the future of future' to deal with that makes the task tougher. Meanwhile, other types of scholar rank percentile indicators, rp.c and rp.h, have very similar predictive powers with rp.rp5, which is illustrated in figure S2.

It turns out that not only do publication rp.c has high predictive power, it's stable over age, i.e. rp.c$_{j\tau_1}$ $\approx$rp.c$_{j\tau_2}$. Figure 4 shows the kernel densities of the scatters of rp.c$_{j\tau_1}$ vs rp.c$_{j\tau_2}$. We see that the scatters are mostly along the 45 degree line. Meanwhile, we perform a simple linear regression by regressing rp.c$_{j\tau_2}$ upon rp.c$_{j\tau_1}$. The regression coefficient of variable rp.c$_{j\tau_1}$ together with the standard error is displayed besides each of the scatters in the figure. We see that the coefficients are very close to 1 with very small standard errors, which gives further evidence that rp.c is stable over age. Meanwhile, figure S3 shows a similar study for the stableness of rp.rp5. We do not see strong evidence of rp.rp5 being stable over age.

We know that predicting rp.rp5 involves forecasting the future impact of both existing and future works. Such question can assist making decisions for a faculty position or tenureship, since the committee would like to examine the cumulative scientific impact of the scholar. A more difficult question is to predict the future impact of the future works. Such question is of interest, for example, in assigning research fundings or allocating research resources, where the future impact of existing works shall be irrelevant. Figure 3b shows the ability of using rp.rp5$_{i\tau_1}$ to predict rp.rp5$_{i\tau_2\setminus\tau_1}$, where the latter is based on only the future papers, i.e. those written in the $(\tau_1, \tau_2]$ time interval. As expected, we see lower a predictive power than predicting rp.rp5$_{i\tau_2}$, but rp.rp5$_{i\tau_1}$ can still explain most of the variations in rp.rp5$_{i\tau_2\setminus\tau_1}$ in relatively short horizons.

Predictive models

For both the publication and scholar impact, we've studied how well the performance by age τ_1 predicts the cumulative achievement by age τ_2. We demonstrate that rank percentile indicators (rp.c$_{j\tau}$ and rp.rp5$_{i\tau}$) have high predictive powers. Meanwhile for the scholar impact, we further investigate the prediction of performance in the subsequent $\tau_2 - \tau_1$ ages, i.e. using rp.rp5$_{i\tau_1}$ to predict rp.rp5$_{i\tau_2\setminus\tau_1}$.

We now formulate these prediction tasks as supervised learning problems, and examine how much improvement we can get by having an extensive list of features and complex fitting algorithms. We consider the following fitting procedures; these models are ordered by increasing complexity, starting from simple baselines and ending with complex machine learning models:

- Baseline: a simple linear regression of rp$_{\star\tau_2}$ upon rp$_{\star\tau_1}$.

- Simple Markov model (sm): a linear regression model of rp$_{\star\tau_2}$ upon features only including rp$_{\star\tau_1}$ and the change of rp over the last 2 ages.

- Penalized linear regression models, including ridge[35], LASSO[36], elastic net (enet)[37] and the Gamma LASSO (gamlr)[38]: linear models with different penalties on the regression coefficients. All of these methods shrink the coefficients towards zero, and the later three methods can provide sparse solutions (shrink the coefficients to exactly zeros).

- Ensemble methods of regression trees, including random forest (rf)[39] and extreme gradient boosting trees (xgbtree)[40]: rf is a bagging of regression trees with a subset of randomly selected features chosen at each split to avoid overfitting. xgbtree is a fast implementation of gradient boosting on regression trees with a model formation that emphasizes the role of regularizations to avoid overfitting.

- Deep neural networks (dnn): feedforward networks with multiple hidden layers and using dropout and l_2 regularization to avoid overfitting.

Features and model fitting

A crucial step for supervised machine learning models is feature engineering. The features that we create are based on the citation histories and are characterized into either author related or paper related features. For example, to predict the author impact rprp$5_{i\tau_2}$, an author related feature can be the number of papers that the author publish by age τ_1, and a paper related feature can be the average citations of these papers. The 30 features for predicting the publication impact and the 42 features for predicting the author impact, are listed in tables S2 and S3 respectively.

The features are created only using the citation information available by τ_1 and the response is specified at τ_2. With a pair of features and response, all the models are trained and evaluated on the testing set, where the training and testing are randomly split in a 9-1 ratio. We consider 5 different values of τ_1 representing different stages of a publication or a scholar, that are $\tau_1 \in \{5, 10, 15, 20, 25\}$, and for each τ_1 we have $\tau_2 = \tau_1 + 1, \cdots, 30$, which results in 75 pairs of (τ_1, τ_2) in total. Hence all the models are trained 75 times independently. Meanwhile, the machine learning models all rely on some hyper-parameters that require careful tuning. We use randomized searching with multiple iterations over a pre-specified parameter space, and the optimal set of hyper-parameters is chosen to minimum the 10-fold cross-validation error.

If we ignore the fact that we are modeling rp.$\mathrm{c}_{j\tau}$ and rp.rp$5_{i\tau}$ at fixed time points, and view them as time series. Both series turn out to be non-stationary, based on statistical tests such as the Dicky-Fuller test[41] and the KPSS test[42], where the details are discussed in the supplemental material section B. It's often encouraged to work on stationary series, for example in the autoregressive and moving average models, and a typical practice is to model differenced series. We follow the same logic here, and use the differenced series as the responses, e.g. $\Delta\mathrm{rp.c}_{j\tau_2} = \mathrm{rp.c}_{j\tau_2} - \mathrm{rp.c}_{j\tau_1}$.

Results

We consider four evaluation metrics for the predictive models, that are the R squares (R^2), root mean squared error (RMSE), root median squared error (RMEDSE) and mean absolute error (MAE). The R^2 for all the models are shown in figure 5, while the rest of the metrics are displayed in figure S6, S7 and S8. We see that the baseline model predicts the cumulative impacts well, and the improvement of using more features and machine learning models have little benefit. Meanwhile, the baseline model can still provide reasonable predictions for the future scholar impact, although the machine learning models can provide better results when τ_1 is relatively large. However, in such scenarios, the simple Markov model performs closely to the machine learning models, and hence the extensive list of features and complex non-linearity have little impact.

Discussion

Comparing the scientific impacts of papers or scholars can assist the process of making academic decisions. The question is how to carry out the comparison accurately. A benchmark specifies a certain field, a specific publication year or a single document type. Rank percentiles based on the number of citations are reasonable metrics to compare two entities in different benchmarks[31]. In this paper, we discuss a general framework to construct rank percentile indicators that is flexible in terms of the choice of evaluation metric. For paper impact, we introduce rp.$\mathrm{c}_{j\tau}$ that measure the performance of paper j at age τ based on the number of citations. Meanwhile, for scholar impact, we consider rp.$\mathrm{c}_{i\tau}$, rp.$\mathrm{h}_{i\tau}$ and rp.rp$5_{i\tau}$ that correspond to the evaluation metric being the number of citations, h-index and an aggregation of paper performance measured using rp.$\mathrm{c}_{j\tau}$. We do not claim a single best metric that shall be applied, although we show some advantages of using rp.rp5$_{i\tau}$ over the others. Once constructed, the rank percentile indicator itself is highly interpretable. It tells the rank of a paper or a scholar relative to others in the benchmark with the same age. Hence we can compare say two scholars in the same area but starting their careers at different years.

We further study the predictive power of the rank percentile indicators. We show that both the paper impact rp.c$_{j\tau}$ and the scholar impact rp.rp5$_{i\tau}$ have high predictive powers. Meanwhile, rp.c$_{j\tau}$ is stable over age, meaning that the ranking of a paper is likely to stay the same along different stages. Predicting cumulative impact rp.c and rp.rp5 is of interest in making academic decision such as hiring new faculties. Assigning research funding often requires foreseeing the future impact of a scholar's future works. We see that rp.rp5 still has reasonably high predictive power in predicting the future impact, although it's not as much high as predicting the cumulative impact. Furthermore, we formulate the prediction tasks into supervised learning problems. We show that an extensive list of features and complex machine learning models bring negligible improvement in predictions. Both the cumulative impact and future impact can be predicted well using simple linear regression models.

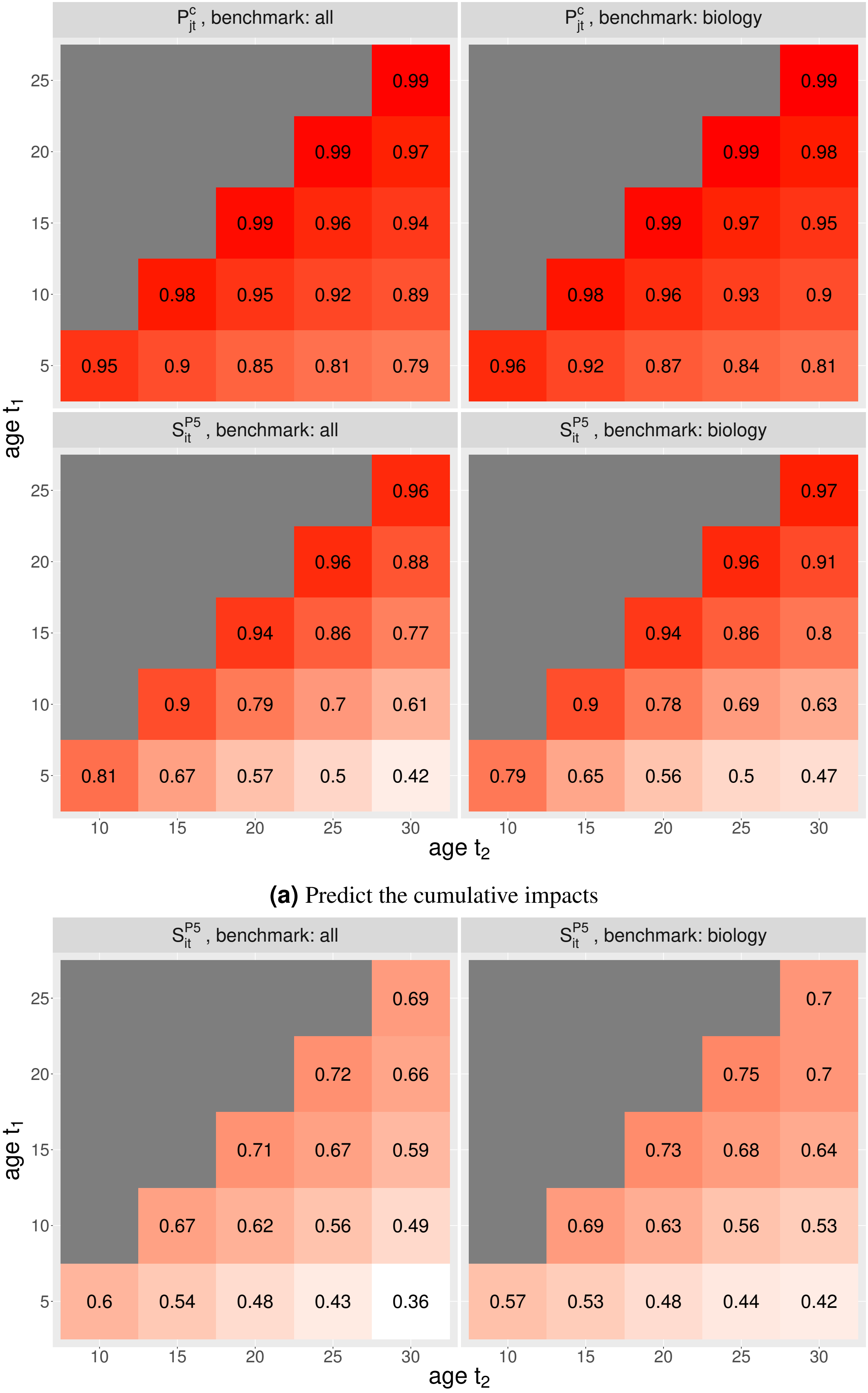
Methods

Data description We focus on authors from all fields in the top universities across the States, who have a Google Scholar account by the end of year 2016. The dataset includes the entire citation history for each publication of these authors. This gives us 14668 authors in total, and together they contribute to more than 1.3 million publications that receive around 100 million citations in total. We've used the benchmark 'biology' throughout our discussion. A scholar and all her corresponding publications are flagged as being in the field of biology, if the areas of interests that she lists on Google Scholar contain any of the keywords: 'biology', 'genetic', 'neuroscience' and 'cell'. An exploratory description of the dataset is shown in figure S1.

Training the machine learning methods All the methods are trained in *R*[43] with the help of the package *mlr*[44], which provides a pipeline of training, validating and testing the model. LASSO, ridge, elastic net, random forest and xgbtree are inbuilt learners of the package. Gamma LASSO and deep neural network are trained using packages *gamlr*[38] and *keras*[45] respectively.

All of the machine learning methods require the tuning of hyper-parameters. The process involves deciding the searching space of parameters and evaluating the sets of parameters using the validation data. The optimal model is the one that minimizes the validation error. Table S1 shows the hyper-parameter(s) for each machine learning model considered in this paper. It's worth noticing that the parameter space can be huge for method like xgbtree where we have an extensive list of tunable parameters. Randomized search with multiple iterations shall be preferred over the grid search in such scenario. Another choice can be using the Bayesian optimization that searches over the parameter space based on the performance gain. Meanwhile, parallel computing can further reduce the computing time.

The data and code are publicly available at `https://github.com/sentian/impact-ranking`.



(a) Predict the cumulative impacts

(b) Predict the future scientific output

Figure 3. The Pearson's correlation between RP indicators at two different ages.

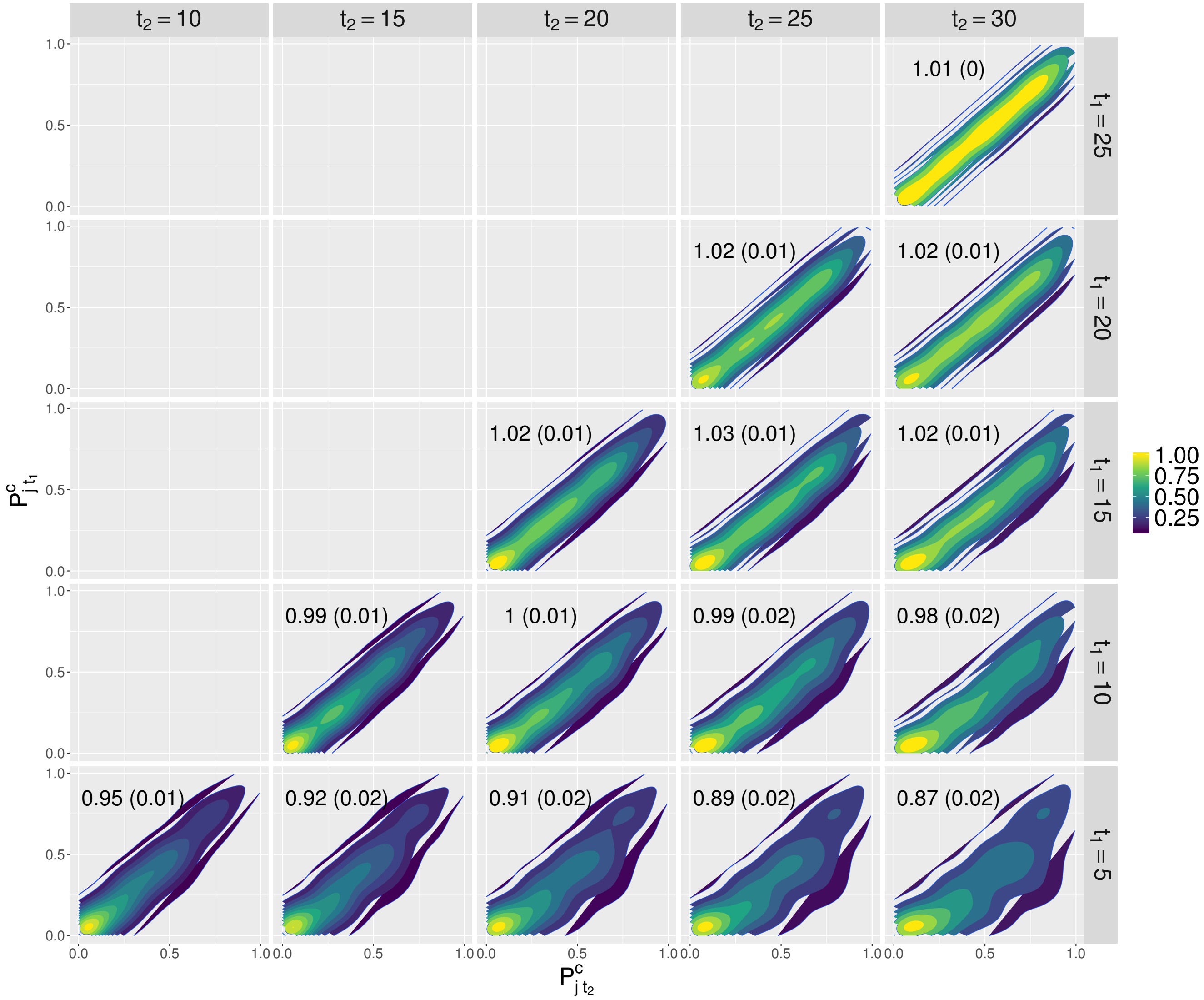


Figure 4. Kernel density estimation of the scatters of $\mathrm{P}^c_{jt_1}$ and $\mathrm{P}^c_{jt_2}$. Meanwhile, we fit a simple linear regression of $\mathrm{P}^c_{jt_2}$ upon $\mathrm{P}^c_{jt_1}$. The estimated coefficient and the corresponding standard error (in the bracket) are displayed in each facet. The benchmark here includes all publications in biology and written in 1980.

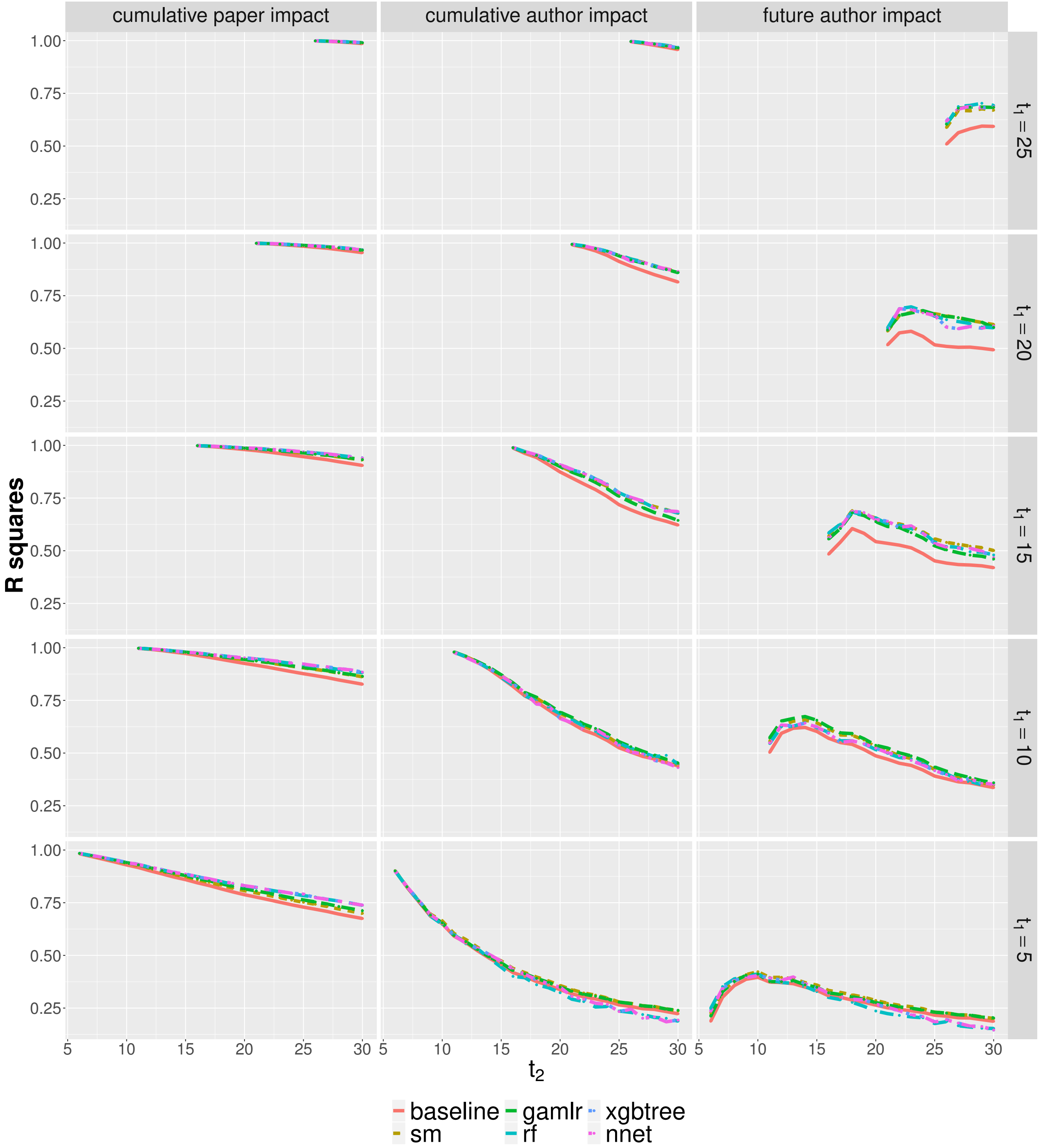


Figure 5. Testing R squares of the predictive models. LASSO, ridge and elastic net are outperformed by Gamma LASSO, and hence are ignored for a better visualization.

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. 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).

3. Garfield, E. & Merton, R. K. *Citation indexing: Its theory and application in science, technology, and humanities* (Wiley New York, 1979).

4. Garfield, E. The history and meaning of the journal impact factor. *Jama* **295,** 90–93 (2006).

5. Egghe, L. Theory and practise of the g-index. *Scientometrics* **69,** 131–152 (2006).

6. Connor, J. Google scholar citations open to all. *Google Scholar Blog* (2011).

7. Chen, P., Xie, H., Maslov, S. & Redner, S. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* **1,** 8–15 (2007).

8. 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).

9. Ma, N., Guan, J. & Zhao, Y. Bringing PageRank to the citation analysis. *Information Processing & Management* **44,** 800–810 (2008).

10. Senanayake, U., Piraveenan, M. & Zomaya, A. The pagerank-index: Going beyond citation counts in quantifying scientific impact of researchers. *PLoS One* **10,** e0134794 (2015).

11. 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).

12. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *science* **286,** 509–512 (1999).

13. 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).

14. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **74,** 47 (2002).

15. Hajra, K. B. & Sen, P. Modelling aging characteristics in citation networks. *Physica A: statistical mechanics and its applications* **368,** 575–582 (2006).

16. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Physical Review E* **62,** 1842 (2000).

17. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342,** 127–132 (2013).

18. Wang, J., Mei, Y. & Hicks, D. Science communication. Comment on "Quantifying long-term scientific impact". *Science* **345,** 149–149 (2014).

19. Wang, D., Song, C., Shen, H.-W. & Barabási, A.-L. Response to Comment on "Quantifying long-term scientific impact". *Science* **345,** 149–149 (2014).

20. Fu, L. D. & Aliferis, C. *Models for predicting and explaining citation count of biomedical articles* in *AMIA Annual symposium proceedings* **2008** (2008), 222.

21. 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).

22. Ibáñez, A., Larrañaga, P. & Bielza, C. Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics* **25,** 3303–3309 (2009).

23. Mazloumian, A. Predicting scholars' scientific impact. *PloS one* **7,** e49246 (2012).

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

25. 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.

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

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

28. 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).

29. Vinkler, P. *The evaluation of research by scientometric indicators* (Elsevier, 2010).

30. Schubert, A. & Braun, T. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics* **9,** 281–291 (1986).

31. 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).

32. Mingers, J. & Leydesdorff, L. A review of theory and practice in scientometrics. *European journal of operational research* **246,** 1–19 (2015).

33. 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).

34. 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).

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

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

37. 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).

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

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

40. 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.

41. 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).

42. 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).

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

44. Bischl, B. *et al.* mlr: Machine Learning in R. *Journal of Machine Learning Research* **17,** 1–5 (2016).

45. 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 Materials: On the use of rank percentile in evaluating scientific impacts, and its predictability

Panos Ipeirotis, Sen Tian

A The robustness of rp.rp5

The evaluation metric to formulate rp.rp5$_{i\tau}$ is an aggregation of the performances of all the publications that the author publish by age τ. And to be more specific, we use rp.c$_{j5}$ to evaluate publication j, and the evaluation metric is given as the sum of all papers published by τ, i.e. m.rp5$_{i\tau} = \sum_{j=1}^{N_\tau^{(i)}}$ rp.c$_{j5}$ (the notation m.rp5 is to distinguish from other metrics we are going to study below). rp.c$_{j5}$ only utilizes the citation history of a publication in the first 5 years. This may not seems to be adequate to represent the quality of a paper. However, as we show in figure 4, rp,c is highly stable over ages, i.e. rp.c$_{j5}$ is likely to be close to rp.c$_{j10}$ for majority of the papers.

We can consider other ways of summarizing the performance of a paper, and we show that not only such metrics have high correlations with m.rp5, but the rank percentile indicators based on these metrics are not statistically significantly different from rp.rp5. Consider the entire history that we have for publication j: rp.c$_{j1}$, rp.c$_{j2}$, $\cdots$, rp.c$_{jT_j}$. Besides evaluating the publication at certain age, we can use the summary statistic to take advantage of the entire history. For example, we can take the best performance throughout its career, i.e. $\max\limits_{t=1,\cdots,T_j}$ rp.c$_{jt}$, and aggregate them to formulate the evaluation metric, i.e. m.rpmax$_{i\tau} = \sum\limits_{j=1}^{N_\tau^{(i)}} \max\limits_{t=1,\cdots,T_j}$ rp.c$_{jt}$. The correlation between m.rp5 and m.rpmax at each fixed age $\tau = 1,\cdots,30$ is displayed in figure S4, and we can see that these two metrics are highly correlated. Meanwhile, other choices of the evaluation metrics also have high correlations with m.rp5 according to the figure. Furthermore, we construct rp.rpmax based on m.rpmax and study its difference with rp.rp5. We perform a paired t-test on the two samples rp.rpmax and rp.rp5 at each fixed age $\tau = 1,\cdots,30$. The p-values that we get are all extremely close to 1, which indicates that the difference between rp.rpmax and rp.rp5 is not significant. Similar results are obtained for other types: rp.rpmean, rp.rpmedian and rp.rp10.

B Stationarity test

Two commonly used statistical tests for stationarity are the Dicky-Fuller test[1] and KPSS test[2]. Two tests formulate the hypothesis testing problem differently. Dicky-Fuller test assumes a unit root is present in the series. A unit root means 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 doesn't 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.

We apply these tests on each individual rp series. The test statistics are shown in figure S5. The dashed lines indicate the critical values at 5% level. KPSS test indicates that rp$^{(c)}$ and rp.rp5$^{(c)}$ are non-stationary, and we don't have enough evidence to reject them being $I(1)$ according to the Dicky-Fuller test. Meanwhile, the differenced series are stationary based on both tests.

References

1. 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).

2. 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).

C Tables and figures

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. Hyper-parameter(s) of the machine learning models.

Feature	Description
pub_cit_cumulative	total citations of publication j
pub_cit_yearly	citations of publication j at age τ_1
pub_cit_peryear	average citations of publication j per age
pub_rp_cumulative	rank percentile indicator calculated based on total citations, i.e. $\text{rp.c}_{j\tau_1}$
pub_rp_yearly	rank percentile indicator calculated based on yearly citations
aut_cit_cumulative	total citations of author i
aut_cit_yearly	citations of author i at age τ_1
aut_npub_cumulative	total number of publications of author i
aut_npub_yearly	number of publications of author i at age τ_1
aut_cit_perpaper	average citations per paper of author i
aut_h_index	h-index of author i
aut_g_index	g-index of author i
aut_maxcit_pub	largest citations that a single paper of author i has received
aut_rprp5_cumulative	author rank percentile calculated based on all papers, i.e. $\text{rp.rp5}_{i\tau_1}$
aut_rprp5_yearly	author rank percentile calculated based on papers written at τ_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. The features are created at τ_1.

Feature	Description
aut_cit_cumulative	total citations of author i
aut_cit_yearly	citations of author i at age τ_1
aut_npub_cumulative	number of publications of author i
aut_npub_yearly	number of publications of author i at age τ_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	author rank percentile calculated using all publications, i.e. $\text{rprp5}_{i\tau_1}$
aut_rprp5_yearly	author rank percentile calculated using publications written in age τ_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 in age τ_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 in age τ_1
*_delta	the difference over the last two ages for each of the above features

Table S3. Features for predicting the impact of author i. The features are created at τ_1.

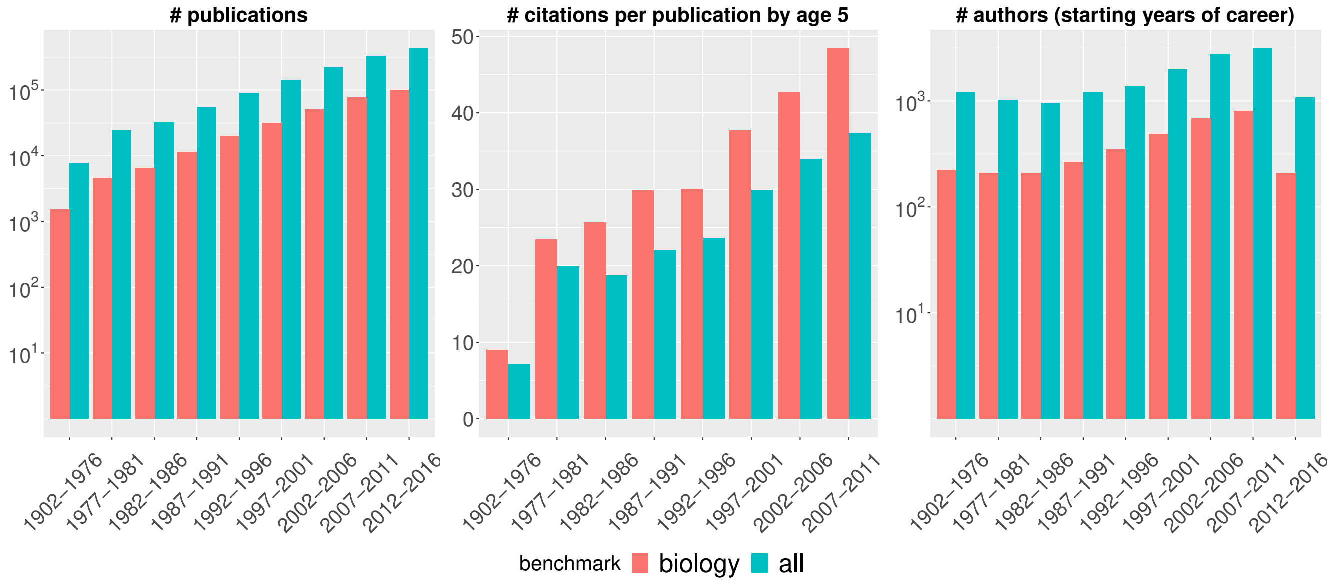


Figure S1. **Left panel**: number of publications; **middle panel**: average number of citations per publication by age 5, for papers published in a certain period; **right panel**: number of authors who start their careers in a certain period. Newer publications have more citations (on average) than the older ones, due to the seniority effect of an author. The authors in our dataset remain active by year 2016 and their publications in the early stage of career is usually less recognizable and attracts less citations, than those published more recently. Meanwhile, publications in biology generally have more citations per paper, which corresponds to the fact that biology is a highly productive field and it attracts a large number of citations.

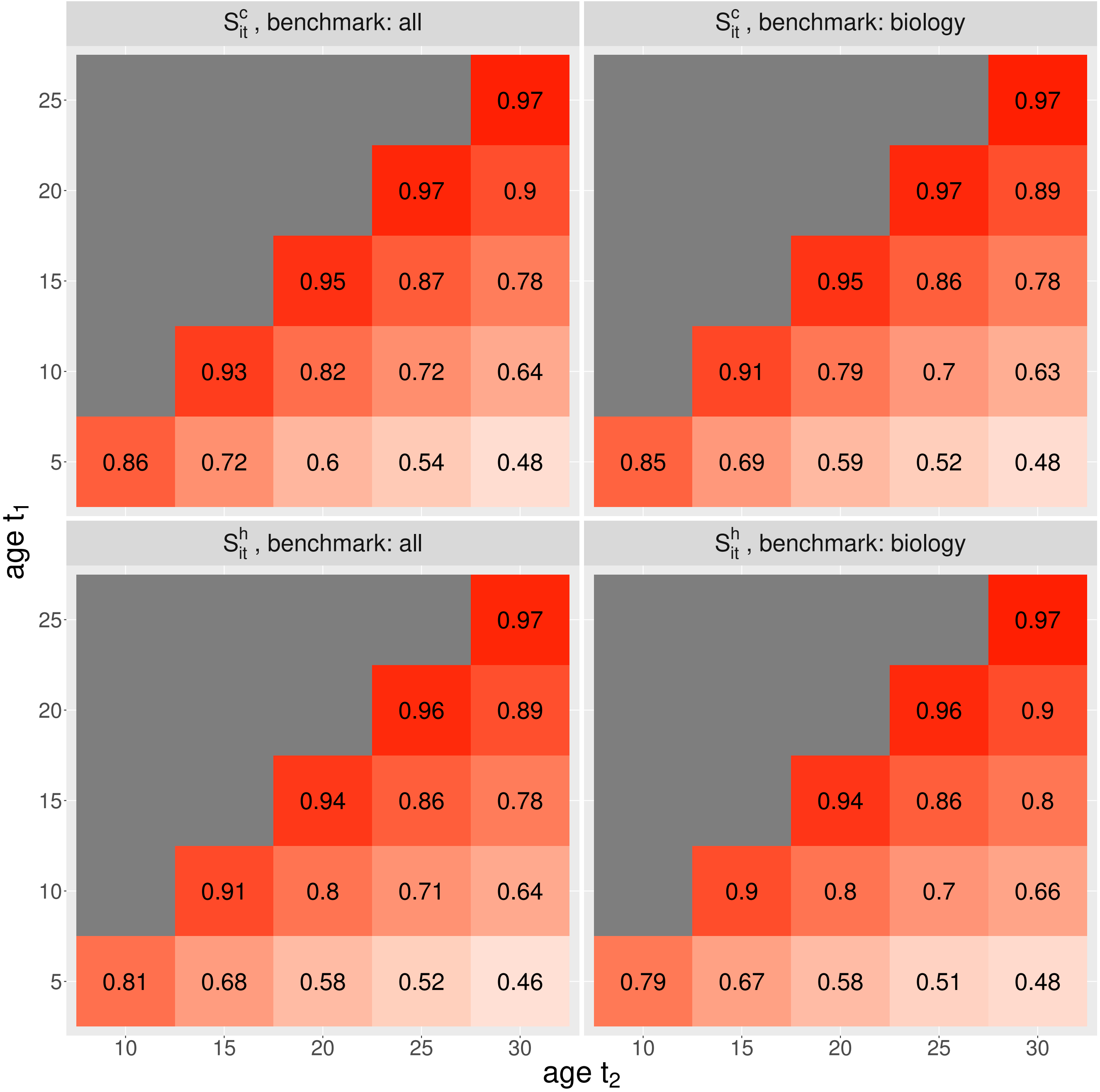


Figure S2. The Pearson's correlation between rp indicators at two different ages. The benchmark either includes all publications or only publications in biology.

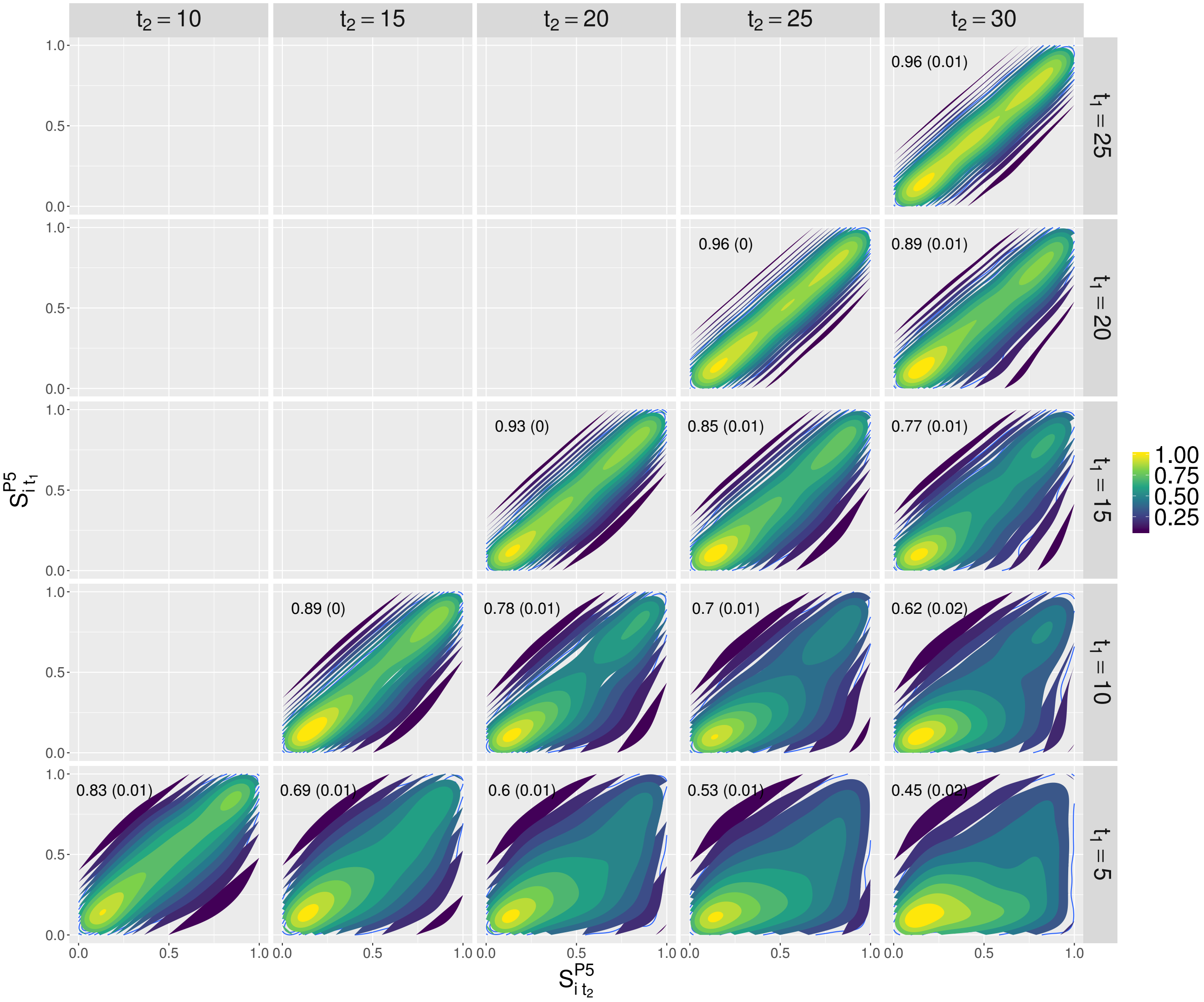


Figure S3. Kernel density estimation of the scatters of S_{it_1} and S_{it_2}. Meanwhile, we fit a simple linear regression of S_{it_2} upon S_{it_1}. The estimated coefficient and the corresponding standard error (in the bracket) are displayed in each facet. The benchmark here is all.

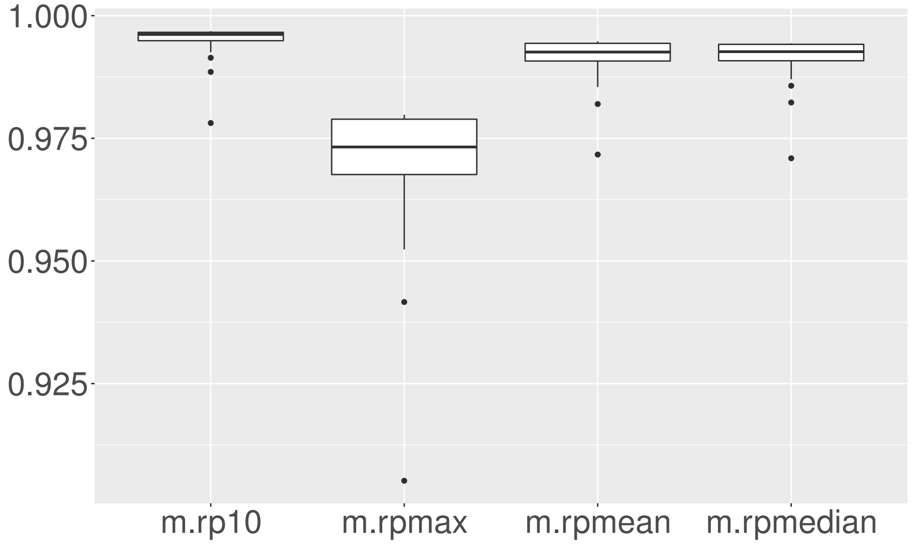


Figure S4. Pearson's correlation between m.rp5 and other evaluation metrics.

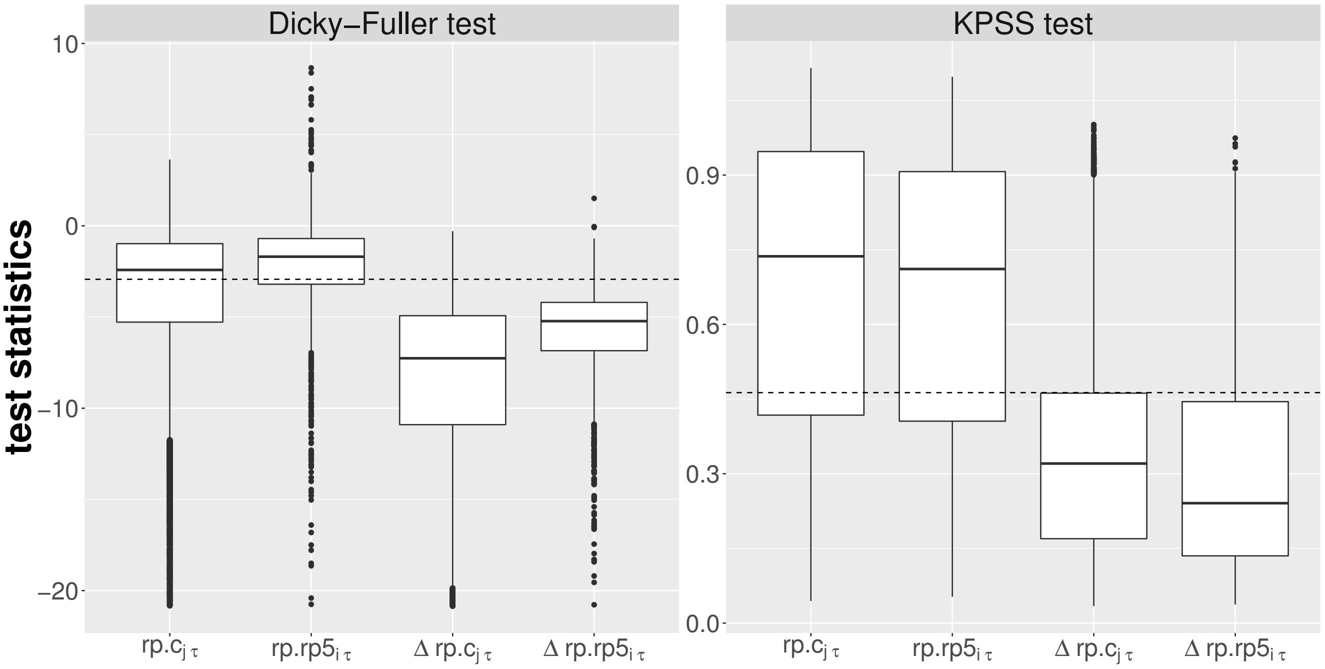


Figure S5. Statistical tests for the stationarity of rp series. Each test is applied on every individual rp series, and the test statistics are presented. Meanwhile, the 5% critical value for each test is shown as the dashed horizontal line. Both tests suggest that publication rp.c and scholar rp.rp5 are likely to be non-stationary. Meanwhile, both tests show evidence that the differenced series can be stationary.

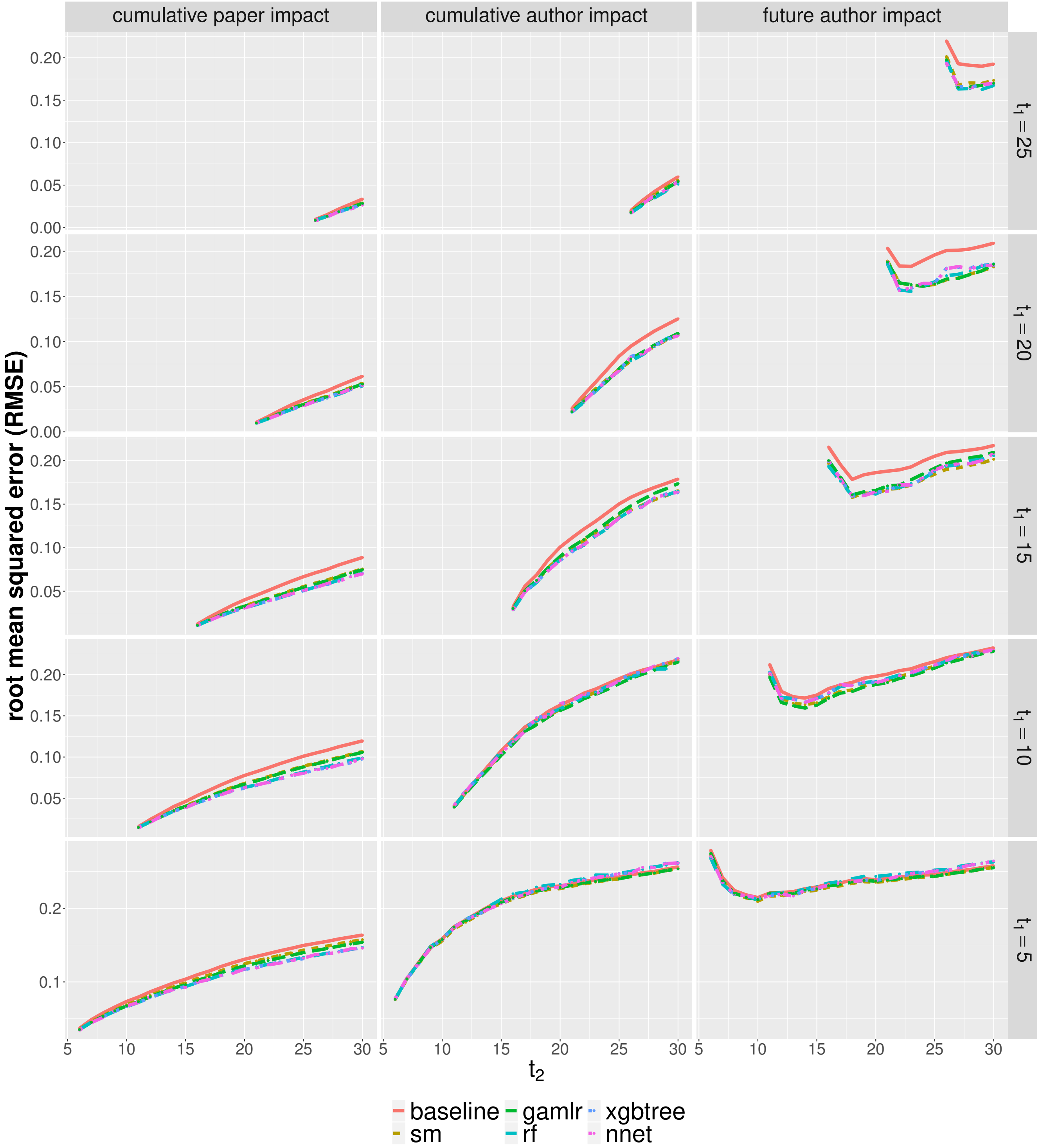


Figure S6. Root mean squared error (RMSE) of the predictive models. LASSO, ridge and elastic net are outperformed by Gamma LASSO, and hence are ignored for a better visualization.

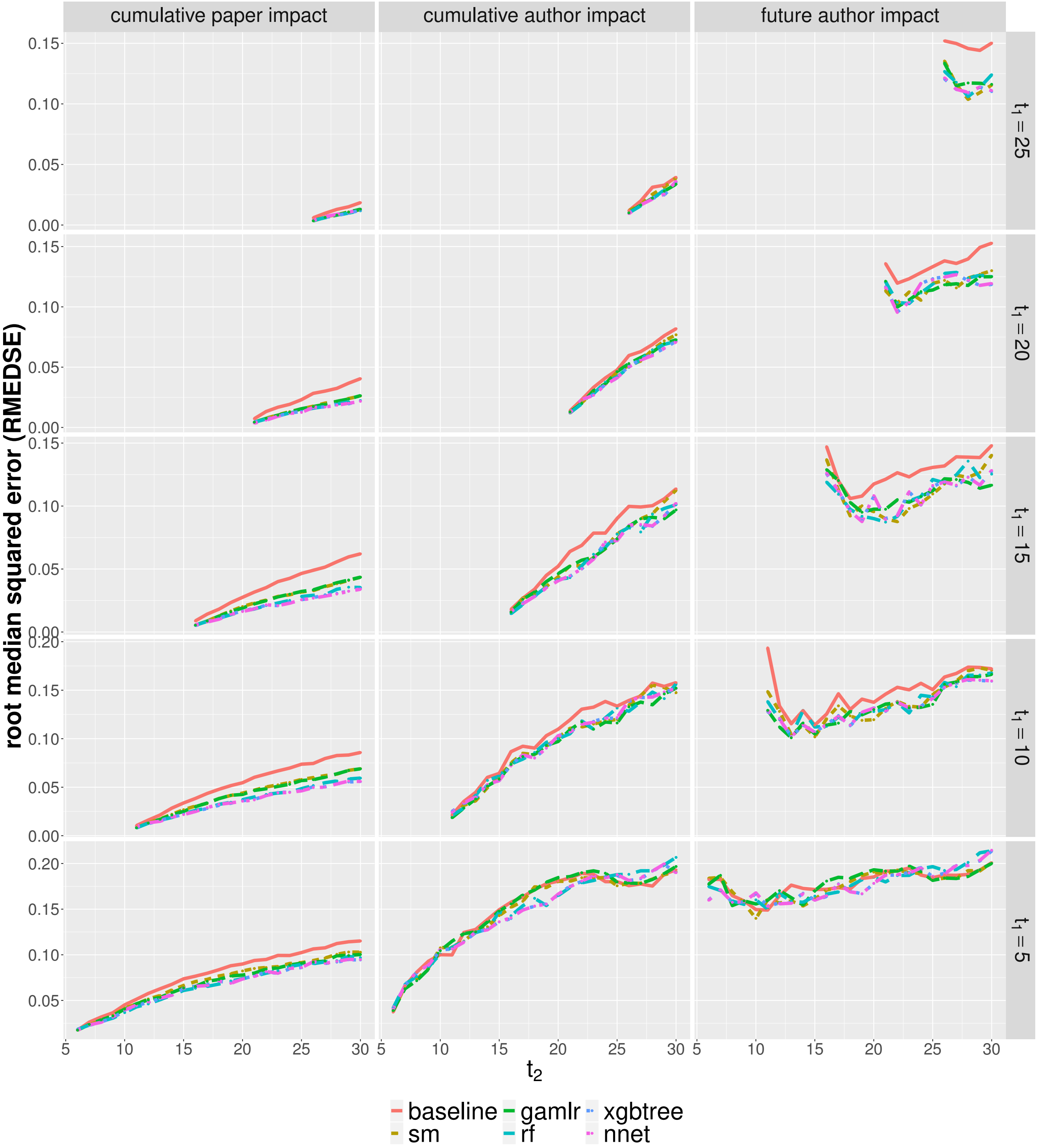


Figure S7. Root median squared error (RMEDSE) of the predictive models. LASSO, ridge and elastic net are outperformed by Gamma LASSO, and hence are ignored for a better visualization.

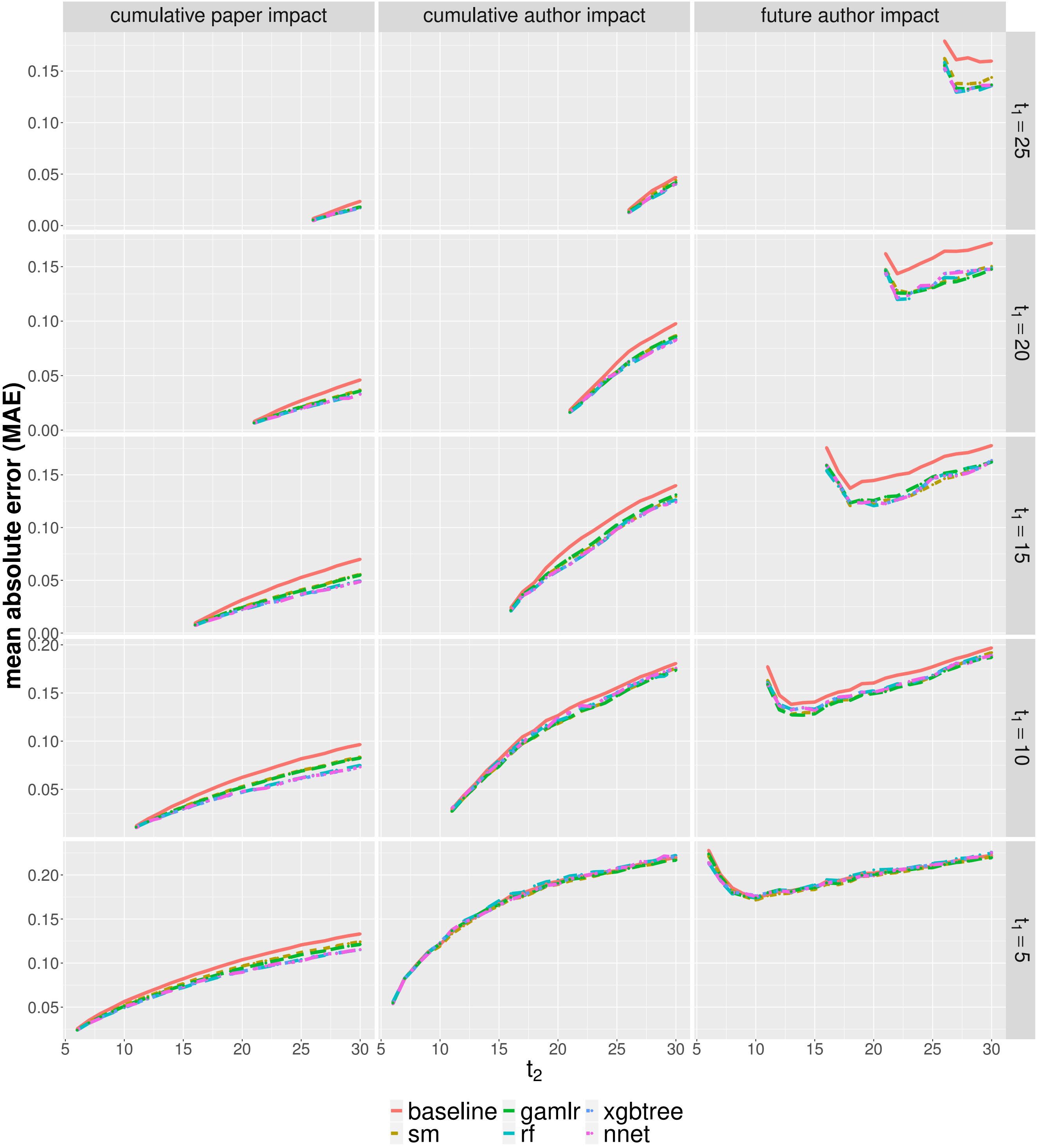


Figure S8. Mean absolute error (MAE) of the predictive models. LASSO, ridge and elastic net are outperformed by Gamma LASSO, and hence are ignored for a better visualization.