Universiteit Leiden

Opleiding Informatica

Predicting Scientific Impact

Name:	Florijn Burggraaf
Date:	01/08/2018
1st supervisor:	Prof. dr. H. H. Hoos
2nd supervisor:	Dr. C. Luo

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This thesis aims to determine to what extent future scientific impact can be predicted from past scientific impact. H-index, i10-index and cumulative citations for both scholars and publications were used as metrics of scientific impact. The data that was used was retrieved from Google Scholar and the Microsoft Academic Graph. The results show that models can be developed for the prediction of scientific impact that outperform predicting current or average value for prediction distances of at least up to 20 years. Slow-growing metrics are more predictable than fast growing metrics. The impact of the union of the work of several authors is more predictable than the impact of the work of a single author, which is in turn more predictable than the impact of the work co-authored by two authors. Using additional features constructed from citation data improves prediction performance compared to using only past and current values of the metric to predict as features. Predicting the relative increase of the h-index leads to better results than predicting the absolute value for prediction distances up to ten years, while for longer prediction distances predicting the absolute value leads to better results. A comparison of models trained on data from Google Scholar and Microsoft Academic shows that the obtained accuracy scores are to a certain extent dependent upon peculiarities of the data source, such as completeness.

Contents

Chapter 1

Introduction

Various indicators are used to give an impression of the scientific impact of a scholar or publication. The simplest metric is to simply count how often a given author or publication is cited. Other metrics are more sophisticated, such as the h-index, which attempts to measure both productivity and citation impact of an author. These metrics are of interest not only to scholars themselves, but also to universities and funding committees, because they can and do play a role in hiring and funding decisions. However, it is not so much the current scientific impact in itself that is interesting to hiring committees, but more the idea these metrics give of what the scientific impact will be in the future.

This thesis aims to determine to which extent future scientific impact can be predicted from past scientific impact. To achieve this, machine learning models are trained and tested on the citation data of several thousand authors in computer science.

1.1 Limitations

Because the prediction of future scientific impact is too broad a topic to cover in one thesis, the scope of this work is necessarily limited. The aim of this thesis is to determine to which extent scientific impact can be predicted using the method, features and data sources described below. The conclusions of this thesis should therefore not be interpreted as an answer to the question of the predictability of scientific impact in general.

- **Data source:** There are many potential sources available for citation data. This thesis uses Microsoft Academic as a source for citation data of scholars, and Google Scholar as a source for citation data of individual publications.

- **Measuring scientific impact:** Among the many metrics that are used to quantify scientific impact, this thesis uses h-index, i10-index and cumulative citations as scientific impact metrics for scholars, and cumulative citations as metric of scientific impact for individual publications.

- **Features:** One can think of many features that could be relevant to the prediction of future scientific impact. This thesis only considers features that can be constructed from the following data: for each author, for each publication he has authored or co-authored, the cumulative citations by year.

- **Target values:** The experiments in this thesis involve only the prediction of the absolute values of the scientific impact metrics considered.

- **Career age:** The experiments in this thesis count the year of first received citation as year 0, not the year of starting PhD. For tests involving individual publications, the year of publication is counted as year 0, unless noted otherwise.
- **Method:** The main experiments in this thesis are executed using only the regression algorithms available in scikit-learn, without any hyperparameter tuning. Moreover, a model is trained for each combination of career age and prediction distance, so that the end result is a large collection of models. No attempt has been made to develop a single model that handles all predictions, to tune the hyperparameters, to employ regression algorithms from other sources than scikit-learn in the main experiments, to use classification instead of regression to predict future scientific impact or to develop an algorithm for the prediction of scientific impact from scratch.

1.2 Research questions

As this thesis aims to define the extent to which future scientific impact can be predicted from past scientific impact, it answers the following research question:

To what extent is future scientific impact predictable from past scientific impact?

Of course, the answer to this question may differ depending on which metric of scientific impact is used, on the career age of an author or the age of a scientific publication, or on whether we predict the scientific impact of a group of authors, a single authors, or the intersection of the work of multiple authors. Therefore these research questions will also be answered by this thesis:

- *How do the predictability of h-index, i10-index, cumulative citations of authors and cumulative citations of publications compare?*
- *How does the predictability of a single author's research impact compare with that of the union or intersection of two authors?*
- *How is the predictability of future research impact influenced by the career age of an author or age of a publication?*

Past scientific impact can be interpreted as simply meaning "past values of the scientific impact metric to predict" or past scientific impact in a more general sense, including any feature that can be constructed from the citation data of the previous work of the author. Another question to answer in this thesis is therefore:

- *To what extent does the inclusion of additional features based on past citation data improve upon the predictability of a scientific impact metric based on past values of that metric alone?*

As described above, this thesis uses Microsoft Academic as its source for citation data of scholars. However, Microsoft Academic is not complete and not error-free. The same goes for other online citation indices, such as Google Scholar, but the Microsoft Academic Graph API that is used as a source of citation data often contains much less citations for an author than Google Scholar. Because of this, a model trained on Microsoft Academic data will be tested on data from Google Scholar to get an idea of how dependent or independent the model is of the source of data.

- *Is a model for predicting scientific impact based on data from Microsoft Academic applicable to data retrieved from another source?*

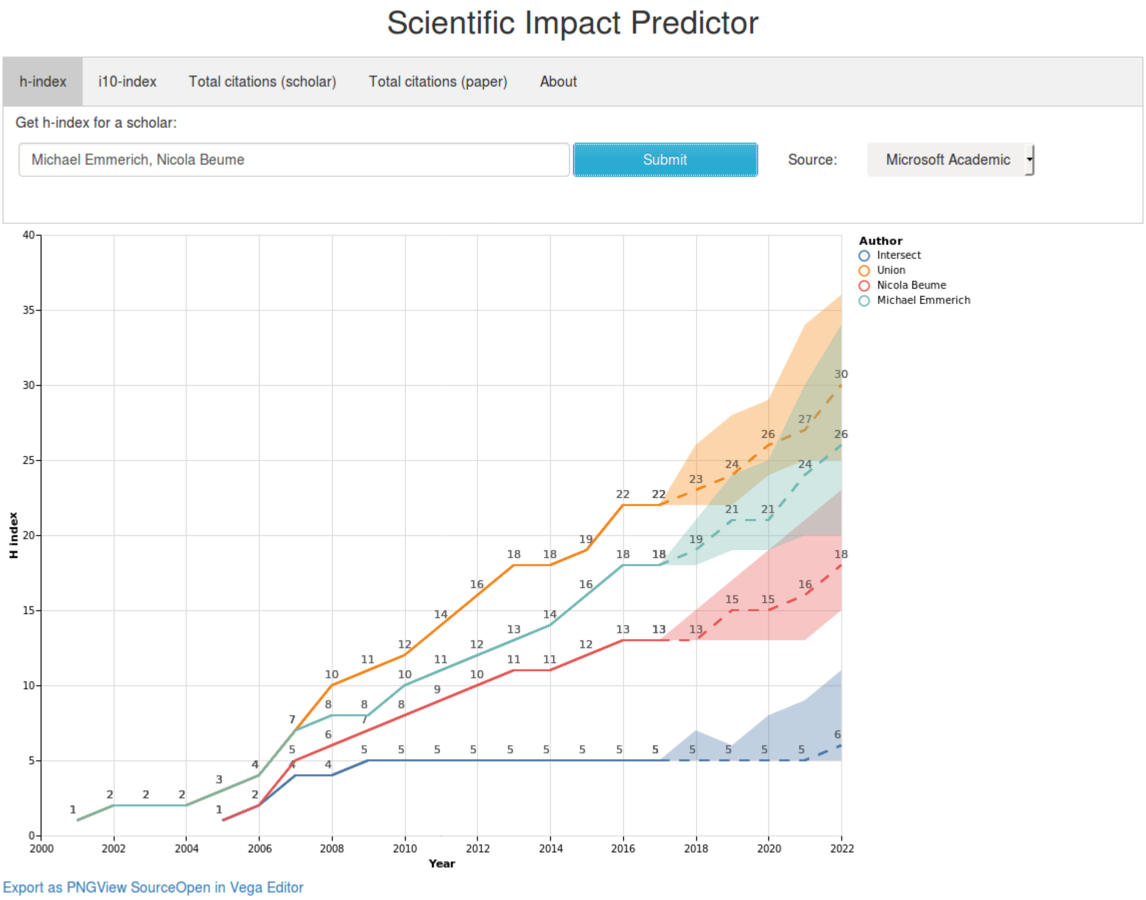
The main experiments in this thesis all involve the prediction of the value of a certain metric of scientific impact. A last test in this thesis is therefore a comparison between the results of predicting the absolute scientific impact in the target year and predicting the relative increase in scientific impact between current year and target year.[1] Among the scientific impact metrics used in this thesis, h-index was chosen for use in this test. The goal of this test can be summarized in the question below:

- *Does prediction of the value of the h-index yield better results than prediction of the relative increase in h-index?*

1.3 Web application

As part of this project, a web application has been developed, a screenshot of which is shown in Figure 1.1. It allows the user to generate graphs of scientific impact over the years (as measured in h-index, i10-index or cumulative citations for authors and cumulative citations for individual publications). It is possible to create a graph of the scientific impact for multiple authors at once. In that case, apart from the individual scientific impact, the scientific impact of the union of their work is also shown, as well as the scientific impact of all publications they have co-authored, if applicable. The user has the option to draw these data from either Microsoft Academic or Google Scholar, although retrieving data from Google Scholar takes some time. Apart from past scientific impact, the model also shows an estimation of future scientific impact over the next five years, with a 90% prediction interval. These predictions are based on gradient boosting regression, employing quantile regression to get a prediction interval. The models are trained on citation data from authors in computer science, drawn from Microsoft Academic. An exception is formed by the models used for the prediction of the cumulative citations of publications, which are trained on data retrieved from Google Scholar.

[1] Relative increase meaning the value of the scientific impact metric in the target year divided by the value of the same metric in the current year. Target year is the year for which we want to predict the scientific impact.



Scientific Impact Predictor

Get h-index for a scholar: Michael Emmerich, Nicola Beume

Source: Microsoft Academic

Figure 1.1: A screenshot of the web application, showing for computer science scholars Michael Emmerich and Nicola Beume their past and estimated future h-index, with a 90% prediction interval.

Chapter 2

Preliminaries

2.1 Machine learning

This thesis employs techniques from machine learning to estimate future scientific impact of researchers or publications. Wikipedia defines machine learning as *"...a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed."*[1] Machine learning can be broadly divided in supervised learning and unsupervised learning. In supervised learning the computer is given combinations of input and corresponding output, and learns a function that maps input to output. This is what is used for most experiments in this project, where the computer is presented with the values of a certain scientific impact metric over a range of years as input, and the corresponding value of that metric for a later year as output. Unsupervised learning differs from supervised learning in that no correct answers are provided to the learning algorithm. Instead, a function is learned based on the structure of the data. In this thesis, unsupervised learning is used to cluster the examples in the data based on their similarities and dissimilarities. A second way to divide machine learning is according to the desired output. Clustering is already mentioned. In classification, the computer is tasked to assign one or more class labels to the input presented. In regression, like classification a supervised task, the output does not consist of discrete labels, but of continuous values. In this thesis various regression methods are employed to estimate future scientific impact.

In machine learning, the learning algorithm is used to find good values for the input parameter. Various learning algorithms also require parameters themselves, for instance to determine the learning speed. These so-called hyperparameters are usually manually chosen by the user. The process of hyperparameter optimization can also be automated, however. This is done in automated machine learning, or autoML in short. There is no universal agreement upon the exact scope of automated machine learning, but basically it is about the automation of certain tasks in machine learning that are usually performed by people. Such tasks include exploratory data analysis, feature transformations, algorithm selection and hyperparameter optimization.

[1] See Wikipedia contributors. (2018, June 20). Machine learning. In Wikipedia, The Free Encyclopedia. Retrieved 13:27, June 20, 2018, from `https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=846671620`

2.2 Machine learning tools

Two different tools for machine learning were used in this thesis. One is Weka, a program that has been developed at the University of Waikato in New Zealand.[6] It contains tools and algorithms for a wide range of machine learning tasks and comes with a graphical user interface. Because of its ease of use, Weka was used in this research for preliminary experiments.
The second tool that was used is scikit-learn, a machine learning library for the programming language Python.[15] Scikit-learn contains a lot of different algorithms for machine learning. In this thesis, many of the algorithms that scikit-learn provides for regression have been applied to the available data and the resulting models have been compared on their effectivity in predicting future scientific impact.
For both Weka and scikit-learn extensions have been developed to enable the application of automated machine learning. In the case of Weka, the tool for automated machine learning is called Auto-WEKA, initially released in 2013.[21] In the case of scikit-learn, auto-sklearn is the extension that provides the possibility of applying automated machine learning.[5] Both Auto-WEKA and auto-sklearn have been used in this research to explore the possibility of obtaining better models for predicting future scientific impact with the aid of automated machine learning.

2.3 Assessing scientific impact

There are various metrics that aim to capture the scientific impact of a scientist or publication. As is already mentioned, a very basic measure is that of cumulative citations, all the citations that an author or publication has received over the years counted together. Because of its basic character and widespread use, cumulative citations is one of the scientific impact metrics of which the predictability is studied in this thesis, both for authors and scientific publications.
While cumulative citations capture the citation impact of an author, the metric does not contain information about his productivity. A metric that aims to capture both citation impact and productivity is the h-index. The h-index was suggested in 2005 by physicist Jorge Hirsch and has since become very popular. It is defined in this way:

If an author has an **h-index** of *h*, it means that he has written at least *h* publications that were each cited at least *h* times.

As an example, imagine an author who has written three publications. The first is cited three times, the second three times and the third only one time. Then the author has at least one publication that is cited one or more times (all three are), and at least two publications that are cited two or more times (namely the first and second publication). He has, however, not written three or more publications that are cited three or more times, and therefore his h-index is two. If the third publication of this author, which was cited only one time, receives two additional citations in the next year, then the author has written at least three publications that are each cited three or more times, and so his h-index increases to three.

Over the years, many modifications of the h-index have been proposed. One of these is the i10-index, introduced in 2011 by Google and used on Google Scholar. The i10-index is a simpler variation on the h-index, and is defined in this way:

The value of the **i10-index** represents the number of publications that an author has written that are cited ten or more times.

In this thesis, the predictability of the i10-index is measured and compared to that of the h-index.

2.4 Assessing prediction accuracy

In this research, various metrics are used to represent the quality of the learned prediction models, each having their own benefits and drawbacks.

R^2-value
The r^2-value or coefficient of determination represents the proportion of the variance for the target variable that is explained by the predictor variables. Its value is calculated according to the formula

R^2 = 1 - (Residual sum of squares / Total sum of squares)

This can be further defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where y$_i$ is the true value for a data point, and f$_i$ the corresponding predicted value. $\bar{y}$ is the average of all true values of the data to be predicted. The R^2-value as used here is therefore an expression of how well the model performs compared to simply predicting the average value of the target variable. The value of R^2, as calculated in this way, can never exceed 1 (a perfect predicting), but can be lower than 0, in cases where the model performs worse than predicting the average value of the target variable.
The r^2-value is a widely popular measure that is used in almost all papers related to estimating future scientific impact,[2] which is why it is also used in this thesis. However, it is sensitive to outliers with high values, because they have a greater than average impact on the residual sum of squares and total sum of squares. If the model predicts a few targets that have extremely high values well, it will mean that the total sum of squares is much higher than the residual sum of squares and the R^2 for the predictions of the model will be relatively high, regardless of how well the other data points are predicted.

Median r^2-value
The median r^2-value is the median value of the r^2-values for each individual prediction. The r^2-value for an individual prediction calculated in this way:

$$R^2 = 1 - \frac{(y_i - f_i)^2}{(y_i - \bar{y})^2}$$

So the calculation of the r^2-value for a single data point is the same as the calculation for

[2] See for instance [1], [2], [4] and [25]

normal r^2 would be in case there was only one data point to predict. Median r^2-value is used in this thesis alongside normal r^2-value because it is more robust to outliers. If a few data points with high true values are mispredicted, it will pull the r^2-value down significantly, even if most data points are predicted quite well. Median r^2-value does not have this problem. However, the median r^2-value is not itself without problems. A few outliers with extremely high values may have a disproportionate impact on the average value of the target variable, and as the median r^2 expresses performance relative to predicting the average such outliers will cause the median r^2-value to be suspiciously high.

Average and median relative error

The relative error for the prediction of a single point is calculated in this way:

$$\text{Relative error} = \frac{|f_i - y_i|}{|y_i|}$$

The average relative error is the average of the relative errors for all predictions. Because in the data there is a huge variation in true values for the data points, absolute error is not really informative about the quality of the model, which is the reason why relative error was chosen instead. In this thesis both average and median relative error are used, because the average relative error is often much higher than the median because for relatively few data points the target value is severely mispredicted. This most often occurs when the true value of the target is 1, which soon leads to a high relative error because the relative error is the error relative to the true value and the true value in such a case is very low. Of all metrics used in this thesis, only median relative error is really immune against distortions by data points that have a very low or high true value, and therefore it is the best metric to evaluate the overall quality of the model.

Chapter 3

Related work

The predictability of future scientific impact has already been addressed often in the scientific literature, especially in more recent years, where a wealth of citation data is freely available online. It is part of the larger field of bibliometrics, concerning the statistical analysis of numbers of written publications and their citations. The term *bibliometrics* was first used in English by Alan Pritchard in a paper published in 1969, titled *Statistical Bibliography or Bibliometrics?* He defined it as "*the application of mathematics and statistical methods to books and other media of communication*".[24] Detailed analysis of citation patterns has been made possible by citation indices, which are bibliographic indices that store citations between publications. The first citation index for papers published in academic journals was the *Science Citation Index (SCI)*, which was officially launched in 1964. Today, there are various citation indices available online, both free (e.g. Google Scholar and Microsoft Academic) and paid (e.g. Scopus and Web of Science).

The ease with which citation data can be obtained has made it possible to develop models that can predict future scientific impact based on bibliometric data. Such models often concern the cumulative citations of individual publications or the h-index of an author.

3.1 Publications on the predictability of cumulative citations of publications

Daniel (2014) compared various articles from the period 2002-2012 about predicting future citations based on what was known before or shortly after publication.[3] The 2002 article proposed a method that could explain only 14% of variance a few years after publication. This climbed during the next years, and the 2012 article explained 90% of the variance in citations counts a few years after publication. Daniel found that this was due to the features used: where early studies focused on content-related features, later studies found author and venue features to be the most predictive.

A paper that makes use of both content-, author- and venue-related features is that of Yan et al. (2011).[25] They have developed a method to predict citation count 1, 5 and 10 years ahead, with an average coefficient of determination (R^2-value) of 0.74.

Another approach was chosen by McNamara et al. (2013), who also tried to predict the future impact of papers, but with their own definition of scientific impact, which is based on the citation count, but with some modifications.[13] For their model they used "features of the paper's neighbourhood in the citation network, including measures of interdisciplinarity". The predictors of high future impact that they found include high early citation counts of the paper, high citation counts by the paper, citations of and by highly cited papers, and interdisciplinary citations of the paper and of papers that cite it.

Wang, Song and Barabsi (2013) developed a model to predict future citations of a paper based on data from the first five years. They focused on estimating a paper's "ultimate impact", the total number of citations a paper will ever acquire.[22]

Stegehuis, Litvak and Waltman (2015) proposed a model based on quantile regression to predict the long-term impact of a recent publications.[18] To do this, they used the impact factor of the journal in which a publication has appeared and the number of citations a publication has received one year after its appearance. Both of these factors were found to be important to the prediction of the future number of citations of a paper.

Recently, there have been attempts to estimate future impact of a publication based on non-traditional bibliometrics, the so-called altmetrics. These altmetrics can be seen as impact measures in and of themselves, but to a certain extent they are also usable for prediction of traditional impact metrics. Eysenbach (2011) found that tweetations (mentions of a publication by tweets) within the first 3 days after publication can predict future citations.[20] He found, for instance, that 75% of highly tweeted article were highly cited, while only 7% of less-tweeted articles were highly cited. Ringelhan, Wollersheim and Welpe (2015) tested whether Facebook likes for unpublished manuscripts that are uploaded to the Internet can predict traditional measures of scientific impact.[17] They concluded that Facebook likes do predict citation count for publications in the field of psychology, but not for publications in the fields of business or life sciences.

3.2 Publications on the predictability of the h-index

In 2005, Jorge Hirsch defined the h-index, a new metric that attempts to measure productivity and citation impact of a scientist.[7] Two years later he established that the h-index over the past twelve years was a good indicator for the h-index over the next twelve years.[8] Assuming a simple linear relation, the correlation coefficient r was 0.91. Mazloumian (2012) confirmed Hirsch's conclusion about the predictive power of the h-index, although he found annual citation count to be an even better indicator for future scientific performance.[12]

Other researchers tried to find a better prediction model for the h-index by incorporating other parameters than the h-index itself. Acuna et al. (2012) developed a model to predict the h-index of neuroscientists, trying out different parameters.[1] Their data set was limited to authors with an h-index greater than 4 and 5-12 years of experience. Their final model was based on five parameters, namely the current h-index, the number of articles written, the years since the author published his/her first article, the number of distinct journals the author has published in and the number of articles in a selection of five prominent journals. The model of Acuna et al. had an R^2 of 0.92 for predicting one year ahead, 0.67 for five years ahead and 0.48 for 10 years ahead. Applying the same method to evolutionary scientists gave somewhat worse results (R^2 of 0.62 for looking five years ahead).

Penner et al. (2013) criticized linear regression models of the h-index, such as Acuna et al. produced.[16] Their first point of criticism was that cumulative achievement models such as the h-index contain intrinsic auto-correlation, which naturally results in a high R^2 value, so that a linear regression model of the h-index such as Acuna et al. use is not so much predicting future impact as it is picking up on a correlation intrinsic to cumulative measures. Apart from that, Penner et al. also criticized that the Acuna model did not differentiate between different age groups. They showed that the predictive value of the model varied widely for different age groups.

Where Acuna et al. based their model on data of neuroscientists, Dong et al. (2016) used data of computer science scholars for their model to estimate the future value of the h-index.[4] Just as Acuna et al., they used a linear regression model based on several factors that were found to correlate well with future h-index. They considered current h-index, average citations per paper, number of coauthors, years since publishing first article and number of publications, and found current h-index to be the most important factor in predicting future h-index, followed by total number of publications and total number of coauthors. They limited the authors considered to those with an h-index above 10. The model of Dong et al. performed significantly better on the data they used than the Acuna et al. model did on their test data. Dong et al. got an R^2 value of 0.91 for predicting the h-index five years ahead, and 0.75 for predicting the h-index ten years ahead.

Ayaz et al. (2018) deviated from previous approaches by imposing no constraints on for example career age or h-index.[2] They looked for the best set of parameters suitable for h-index prediction for computer scientists from all career ages, without enforcing any constraint on their current h-index value. They found that the future h-index of authors with higher current h-index is more predictable than that of authors with a relatively low current h-index, and that the h-index of beginning researchers is very hard to predict, whereas the h-index of researchers having 20-36 years of experience is quite predictable.

Jaeger et al. (2013) wrote a paper on type extension trees (TET), a representation language for count-of-count features such as the h-index.[10] TET-defined features can be incorporated into existing types of predictive models, or predictive models can be built directly on TET-defined features. A TET for the task of h-index prediction was learned and used as a distance measure for k-NN in predicting future h-index. This was done in two ways: using the entire TET, or just the first branch, which contained the relational features necessary to compute the h-index. The predictive performance of these two k-NN algorithms was tested against other approaches, namely predicting current or average value, using k-NN with a linear combination of paper count and citation count as distance, or using a non-linear Support Vector Regressor with paper count and citation count as input. The results were compared on Root Mean Squared Error (RMSE). The TET-based predictors always outperformed the k-NN and SVR that were based on plain counts, and performed better than the prediction of current or average for longer horizons. For longer horizons, the k-NN using the entire TET performed slightly better than the k-NN using the first branch, while for shorter horizons it performed worse.

Chapter 4

Data

4.1 Academic citation databases

There are several online citation indices available for academic publications and literature, such as Scopus, Web of Science, Google Scholar, Baidu Scholar and Microsoft Academic. These indices differ in availability (free or paid), coverage and extra features, such as the availability of an API.

Scopus was launched by Elsevier in 2004. It is a subscription-based service. The same goes for Web of Science, a collection of citation databases currently maintained by Clarivate Analytics. Google Scholar, Baidu Scholar and Microsoft Academic, on the other hand, are freely accessible. Google Scholar was launched in 2004. Baidu Scholar was launched by the Chinese company Baidu in 2014, the year that Google (and also Google Scholar) began being blocked in China. Microsoft Academic dates from 2016 and replaces Microsoft Academic Search, which ended development in 2012.

A comparison of the completeness of Google Scholar, Microsoft Academic, Scopus and Web of Science was conducted in 2016 by Bartosz Paszcza.[14] He compared the coverage of the output of six authors, and found an average coverage of 76.2% by Google Scholar and 76.0% by Microsoft Academic, versus only 66.5% coverage by Scopus and 58.8% by Web of Science. A more comprehensive study by Mike Thelwall (2017) comparing Microsoft Academic and Scopus found 6% more citations for Microsoft Academic compared to Scopus, and 51% more for the current year.[19]

Most data used in this thesis is pulled from Microsoft Academic, because of its combination of free access, the availability of an API that enables the collection of large amounts of data, and favorable evaluations of its coverage relative to other online citation indices. The citation data for individual papers, however, are pulled from Google Scholar, because Google Scholar appears to be more complete than Microsoft Academic and collecting citation data for individual papers from Google Scholar can be done within a sensible timeframe.

Data retrieved from Microsoft Academic via the API has issues when it comes to coverage and accuracy, as will be seen in the next sections. Nevertheless, because it is a free source and data retrieval is very easy, it is an interesting potential source for citation data. Microsoft Academic is very recent and has according to our literature review not yet been used as a source of data in work on the prediction of scientific impact. Part of the goal of this research is therefore to establish how well a model trained on data from Microsoft Academic is usable on data from another source, such as Google Scholar. A second test will determine how well the results of models compare when a model is trained on data from Google Scholar and a model on data from Microsoft Academic, using the same approach. These tests will give insight in how well Microsoft Academic can be used in research related to the prediction of scientific impact.

4.2 The reliability of the Microsoft Academic API

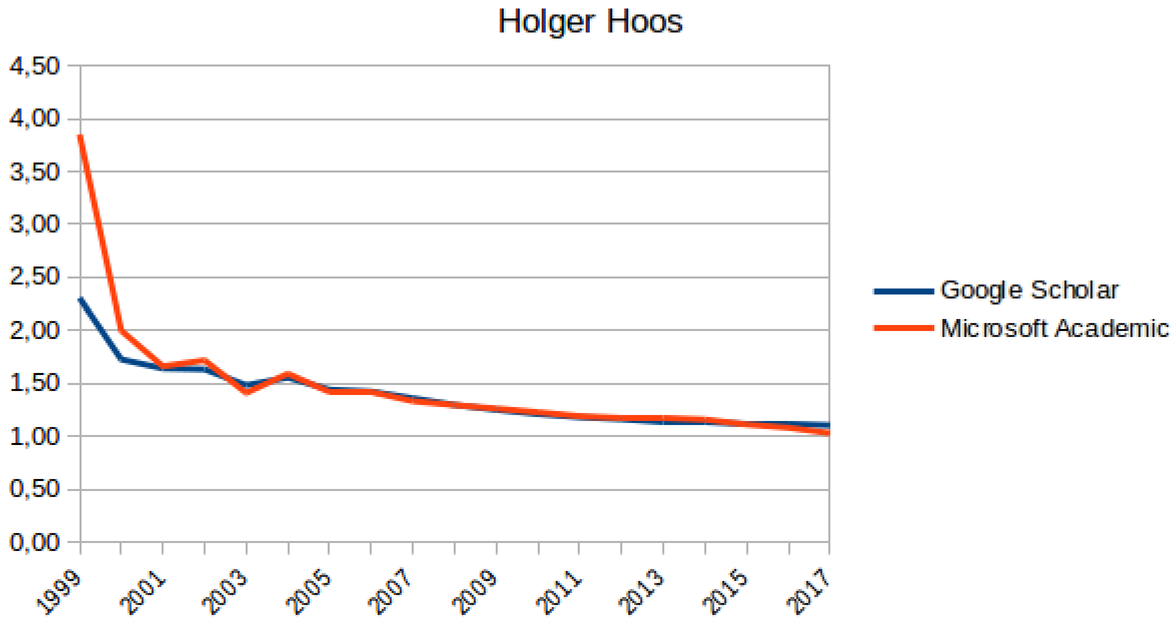
The Microsoft Academic Graph (MAG) was established in 2015, and is currently updated on a weekly basis. Most of its data are derived from web pages indexed by Microsoft's search engine Bing. The data can be accessed either via the Microsoft Academic search engine[1] or the Academic Knowledge API.[2][9]
Microsoft Academic calculates two citation counts for each article. One is the algorithmically verified citation count, and the other the estimated citation count, adding the number of citations that are estimated to exist but not found. Sometimes verified and estimated citation counts are almost the same. In other cases, estimated citation counts may be twice as high as verified citation counts. This difference seems to be field-dependent.[19] Preliminary exploration of Microsoft Academic and Google Scholar has shown that for the field of Computer Science, the estimated citation count (which is shown by the Microsoft Academic search engine) is usually slightly lower than the citation count given by Google Scholar. This estimated citation count is therefore probably a good indicator of actual citation count. However, for this research estimated citation count is not usable, because we need not only present citation count, but also past citation count. To do that, for each publication we need the publication year of each publication that cites it.

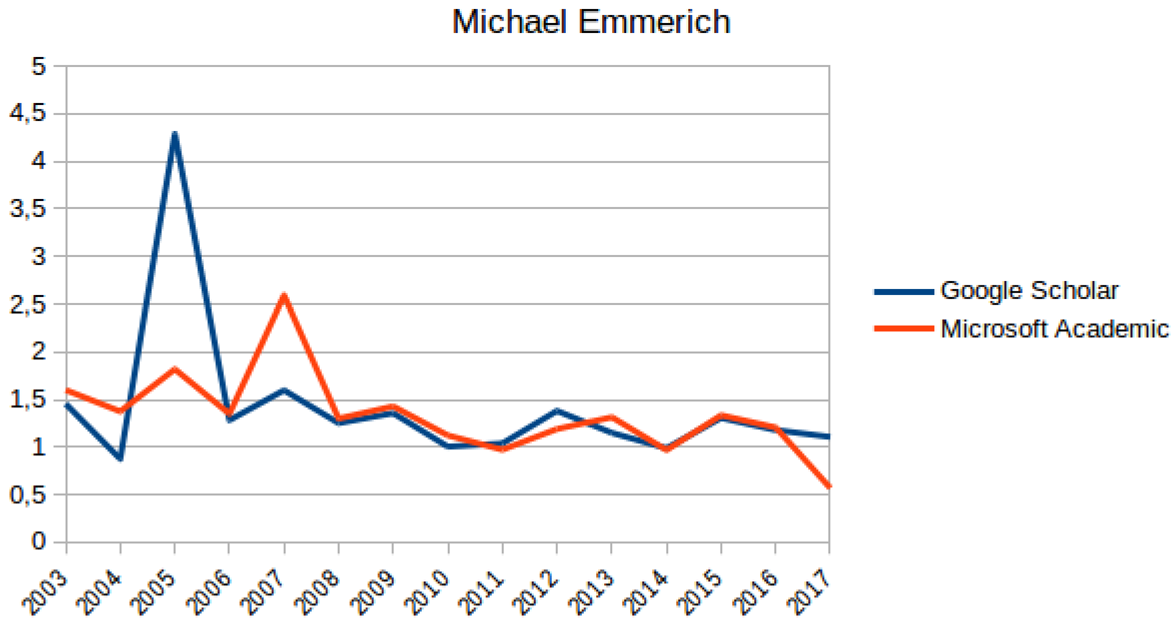
Exploration of the difference between verified and estimated citation count for Microsoft Academic showed that for Computer Science verified citation count is about half the estimated citation count. For instance, the Microsoft Academic search engine gives 971 estimated citations for Michael Emmerich's most cited paper (SMS-EMOA : Multiobjective selection based on dominated hypervolume), while the API only returns 631 articles. Many of the missing articles are very recent. Michael Emmerich's total citation count up until 2017 is given as 3279 by the Microsoft Academic search engine and 3701 by Google Scholar, while an API request for publications citing Michael Emmerich returns only 2087 results. The same goes for other authors. The API returns 12617 publications citing Holger Hoos until the end of 2017, while at that point Google Scholar calculates his cumulative citation count at 20,039.
With such a severe lack of data, it appears that the Microsoft Academic API cannot be used as a source of citation data to be used in this project. However, that is not necessarily the case. While in absolute numbers the Microsoft Academic API is unreliable, it appears that for most years the proportion of missing articles is about the same. Therefore the relative yearly change in metrics like h-index or cumulative citations is about the same, whether calculated based on Google Scholar data or Microsoft Academic API data. The following three graphs are an example of this, showing the relative increase in cumulative citations by year for three authors in Computer Science: Holger Hoos, Lars Kotthoff and Michael Emmerich. As the graphs show, for most years the relative change is about the same. Only in the last two years the values for Microsoft Academic are much lower, giving the impression that new articles are added faster to Google Scholar than to the Microsoft Academic Graph. This is mitigated by limiting our data to citations before 2016. In the first few years, there

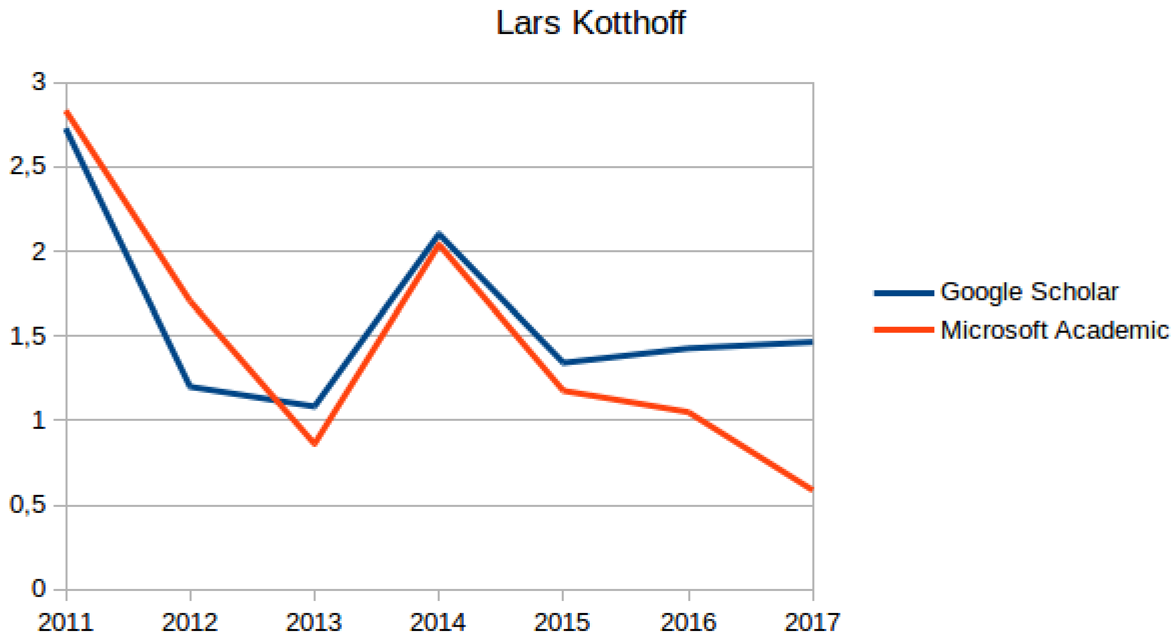
[1] https://academic.microsoft.com/

[2] https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge

can be quite a difference, as seen in the graphs of Michael Emmerich and Holger Hoos. However, in the Emmerich graph the value for the second year is higher for Google Scholar and the value for the third year is higher for Microsoft Academic, while in the Hoos graph the value for the first year is higher for Microsoft Academic, but the second and third year they are about the same. There appears to be no pattern, so this is probably due to the fact that in the first years, one citation more or less makes a big difference in relative change.







While the differences in absolute numbers between Google Scholar and the Microsoft Academic Graph might have a negative effect on the usability of a model trained on data from the Microsoft Academic Graph when applied to data from Google Scholar, the close correspondence in relative changes may mitigate that somewhat. The applicability of both model and approach on Google Scholar data will be tested later in this thesis.

4.3 Collection methods

The citation data of authors was retrieved by querying the Microsoft Academic API. The following query was used.

```
{”path”:”/author/PaperIDs/paper/CitationIDs/citation”,”author”:{”type
    ”:”Author”,”Name”:”’+author+’”},”paper”:{”type”:”Paper”,”select”:[
    ”OriginalTitle”]},”citation”:{”return”:{”type”:”Paper”},”select”:[
    ”PublishDate”,”OriginalTitle”]}}
```

This query returns in JSON-format an array containing the titles of all papers of the author, and for every paper the title and publication date of every paper that cited it.

As Google Scholar does not provide an API, citation data for a publication was collected using the python module Beautifulsoup. Because each publication page contains a graph showing how often it was cited each year, citation data for a publication could be collected by retrieving the url to its Google Scholar page and parsing the page for the information in the graph.

4.4 Data sets

For this thesis, three different data sets were used. One set contains the citation data for 5810 authors in Computer Science and is used for testing the predictability of scientific impact for authors. The second set is a subset of 1023 of these authors and is used for exploratory testing. The third set consists of the citation data of 6700 papers. The table below provides basic information for each data set.

Name	Subject	Source	# samples	Average age of samples	Average value of samples
Data set 1	authors	Microsoft Academic API	5810	24.32	30.20
Data set 2	authors	Microsoft Academic API	1023	27.50	42.65
Data set 3	publications	Google Scholar	6700	15.50	587.50

Table 4.1: Basic information about the data sets used

4.4.1 Data set 1: 5810 authors from Microsoft Academic

The main data set used for predicting h-index, i10-index and cumulative citations of authors, as well as the intersection and union of groups of authors, is derived from Microsoft Academic and consists of the topmost authors in Computer Science and several subfields of Computer Science. The Microsoft Academic search engine offers the possibility to get a list of the 100 topmost authors in a certain field, by rank, citations or h-index, over the past 5 years, 10 years or over all time. Our approach was to select the top 100 authors over all time by rank. We selected first the top 100 authors in Computer Science and all direct subfields of Computer Science as listed by Microsoft Academic. This resulted in a list of 3500 authors. The second step was to select the top 100 authors of the first 67 subfields of Machine Learning as listed by Microsoft Academic. This resulted in another 6700 authors. Because many of these subfields are quite small, this approach made sure that the data set consisted of not only top authors, but also authors that are less often cited overall, but are in the top 100 of a certain field simply because that field is rather small. Combining all authors together and filtering out duplicates resulted in a list of 6782 unique authors. After requesting their citation data using the Microsoft Academic API, for several authors no data were returned, so in the end the data set consists of data for 5810 authors in Computer Science. Because in the last few years the Microsoft Academic API lacks more data than in the years before that, for every author only the citation data up until 2015 are used.

A scatter plot showing the relation between career age (actually year since first received citation) and h-index for the authors in the data set is shown in Figure 4.1. Note that the data set contains many authors with a career age longer than 40 years. This research does not extrapolate beyond the 40^{th} year of someone's career. The average h-index is calculated not only based on the latest h-index of the authors in the data set, but also based on their h-index of earlier years. For instance, the average h-index for year 0 is calculated by the average of the h-indices in year 0 of every author, including those that have a career age longer than one year. No attempt was made to filter out authors that are no longer actively writing new publications, so the model trained on this data set is applicable to active and retired authors alike. The model itself is provided with information that can be used to differentiate to some extent between active and retired authors, because apart from the present scientific impact the scientific impact in past years is also given to the model as features. On top of that, among the additional features used is the number of publications published in the past year, which indicates whether or not an author is still active.

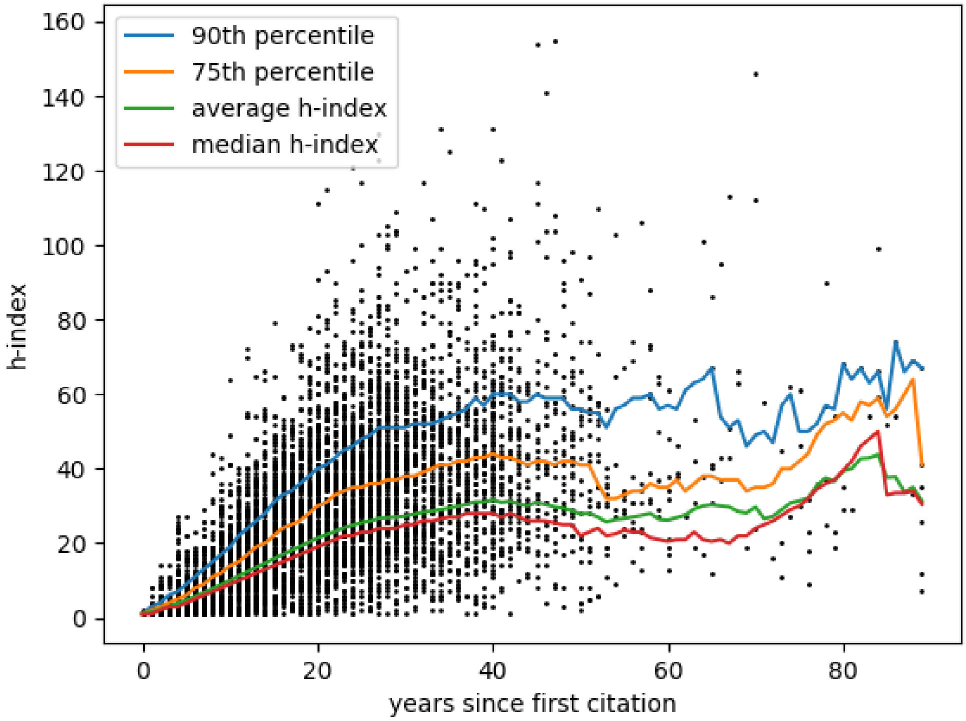


Figure 4.1: Relation between career age and h-index for the authors in data set 1

It is remarkable that there appear to be many authors that have a very low h-index even with a long career. 109 authors had an h-index of 1 in 2015. 470 had an h-index of 5 or lower. There were 55 authors in the data set (about 1%) with a career age longer than 20 years but an h-index lower than 5. One of these authors is Mauro Dell'Amico from the University of Modena and Reggia Emilia. Google Scholar lists 121 publications for this author, starting in 1987, and calculates his current h-index at 29 and his cumulative citations at 4515. Microsoft Academic has five pages for this author.[3]. One page lists 81 publications and 5281 estimated citations. Another page has 14 publications and 197 citations, a third 4 publications and 38 citations, a fourth 1 publication and 29 citations and the fifth page has 1 publication and no citations. When a request was done for the publications citing work of Mauro Dell'Amico, the API returned only those publications citing papers from the second profile page (with 14 publications and 197 citations), instead of from the main profile page. Another example is Kelly Miller from Harvard University, whose first publication dates from 2013, according to Google Scholar. The response from the Microsoft Academic API included among her publications also a publication from Kelly B. Miller from the University of Mexico, Kelly A. Miller from the University of Rochester and Kelly Miller from Howard University. This last publication dates from 1977 and causes the erroneously long career age. Interestingly, in the first case the API showed the same erroneous division of one author in multiple profiles that was seen with the Microsoft Academic search engine, while in the second case, where the information of

[3] see `https://academic.microsoft.com/#/detail/2136425878`, `https://academic.microsoft.com/#/detail/2709446498`, `https://academic.microsoft.com/#/detail/2306102140`, `https://academic.microsoft.com/#/detail/2136425878` and `https://academic.microsoft.com/#/detail/2711127022`

the search engine is correct, the Microsoft Academic API put together papers that according to the information obtained via the Microsoft Academic search engine belonged to different authors. All of this shows us that the data obtained is messy, and may on the one hand wrongly attribute publications to certain authors while on the other hand due to different profiles for one author most publications of an author are not included. That notwithstanding, the average h-index of the data set is still higher than the observation of Malesios and Psarakis (2014) that the average h-index in Computer Science is about 20.[11] To test if examples with erroneously low values and outliers with extremely high values would distort the quality of the model, test results from models trained with the upper and lower ten percent excluded from inclusion in the training data were compared with test results for models where there was no such restriction. The model with no restrictions on the training set resulted in higher r^2-values and lower error rates on the test set. Therefore no restraints have been put on what data from data set 1 to include for training the model.

4.4.2 Data set 2: 1023 authors from Microsoft Academic

A smaller set was used for exploratory testing. This was used to perform many tests in a small time frame, to determine which results were promising enough to execute the test on the larger data set 1. This smaller data set is a subset of data set 1, except for the fact that it was retrieved at an earlier time.

The data set consists of the top 100 authors of all time in Computer Science as listed by Microsoft Academic combined with the top 100 authors of all time in several subfields of Computer Science as listed by Microsoft Academic. These subfields are Artificial intelligence, Computer vision, Computer hardware, Real-time computing, Computer network, Machine learning, Pattern recognition, Data mining, Distributed computing, Multimedia, Knowledge management, Embedded system, Simulation, Library science, Algorithm, Database and Computer security. The resulting set contains 1202 unique authors. For each of these authors, a request was made to the Microsoft Academic Graph via the API to retrieve the needed data. For 1181 authors, the response was received and stored. For 158 of these authors, this response was empty, leaving a set of 1023 items. A scatter plot showing the relation between career age (actually year since first received citation) and h-index for the authors in the data set is shown in Figure 4.2. The same notes that were given for the scatter plot of data set 1 apply here as well.

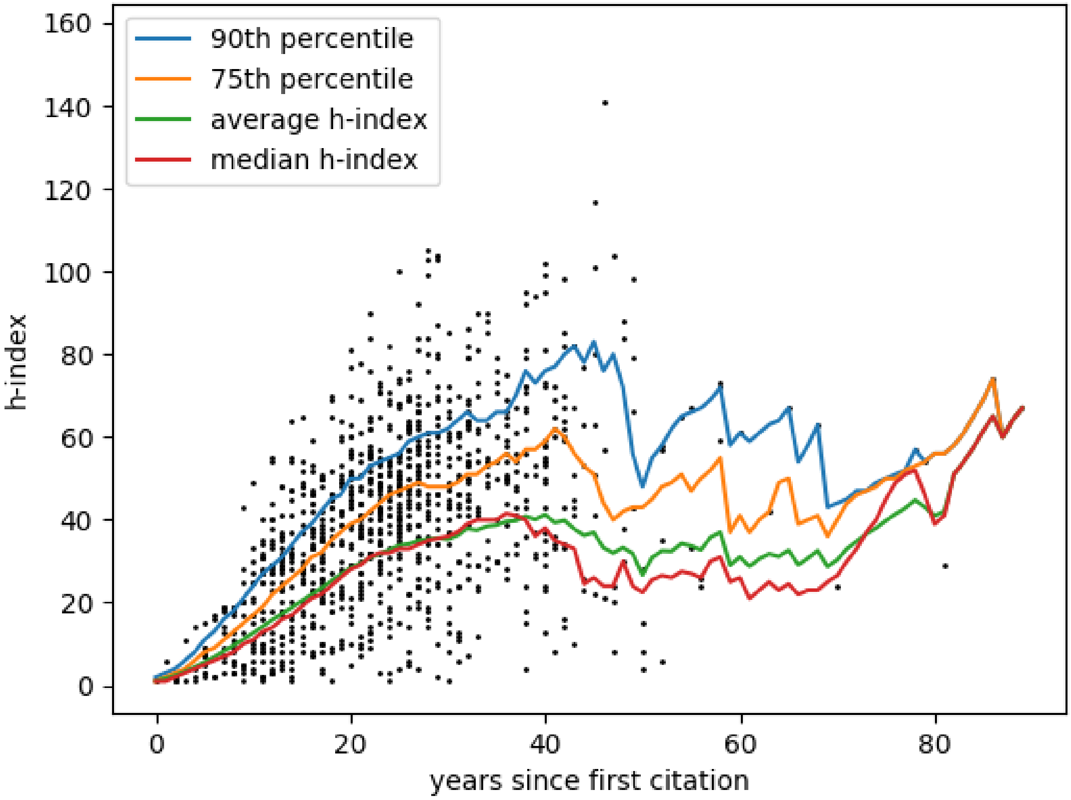


Figure 4.2: Relation between career age and h-index for the authors in data set 2

4.4.3 Data set 3: 6700 papers from Google Scholar

This data set, used for experiments involving the cumulative citations of individual scientific publications, was collected from Google Scholar. To collect this data set, a list of urls was compiled linking to the page of the top 150 authors with label machine learning on Google Scholar.[4] For each of these authors, our scraper visited the page of each of the top 100 articles of this author,[5] provided the author had indeed written that many articles and the article page in question was not yet visited via the page of a co-author. Each article page contains a graph showing citations by year. This was transformed to cumulative citations, extended either backwards to the year of publication (or forwards if the article was already cited before official publication) and cut off at 2017, to exclude the incomplete year 2018. This resulted in a set of 12508 articles. To keep the size of this set comparable to data set 1, the first 6700 articles that were collected in this way were taken as the new data set. A plot showing the relation between age of the article and cumulative citations is shown in Figure 4.3. As can be seen, the vast majority of the articles has a career age of lower than 30 years and is cited between 10 and 1000 times. There are a few extreme outliers with cumulative

[4] For example `https://scholar.google.nl/citations?user=nKC5jkgAAAAJ&hl=en&oe=ASCII`

[5] For example `https://scholar.google.nl/citations?user=nKC5jkgAAAAJ&hl=en&oe=ASCII#d=gs_md_cita-d&p=&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Den%26oe%3DASCII%26user%3DnKC5jkgAAAAJ%26citation_for_view%3DnKC5jkgAAAAJ%3AnatZJ_-F0IUC%26tzom%3D-120`

citations up to 100,000, that cause the average number of cumulative citations to be higher than the 75^{th} percentile.

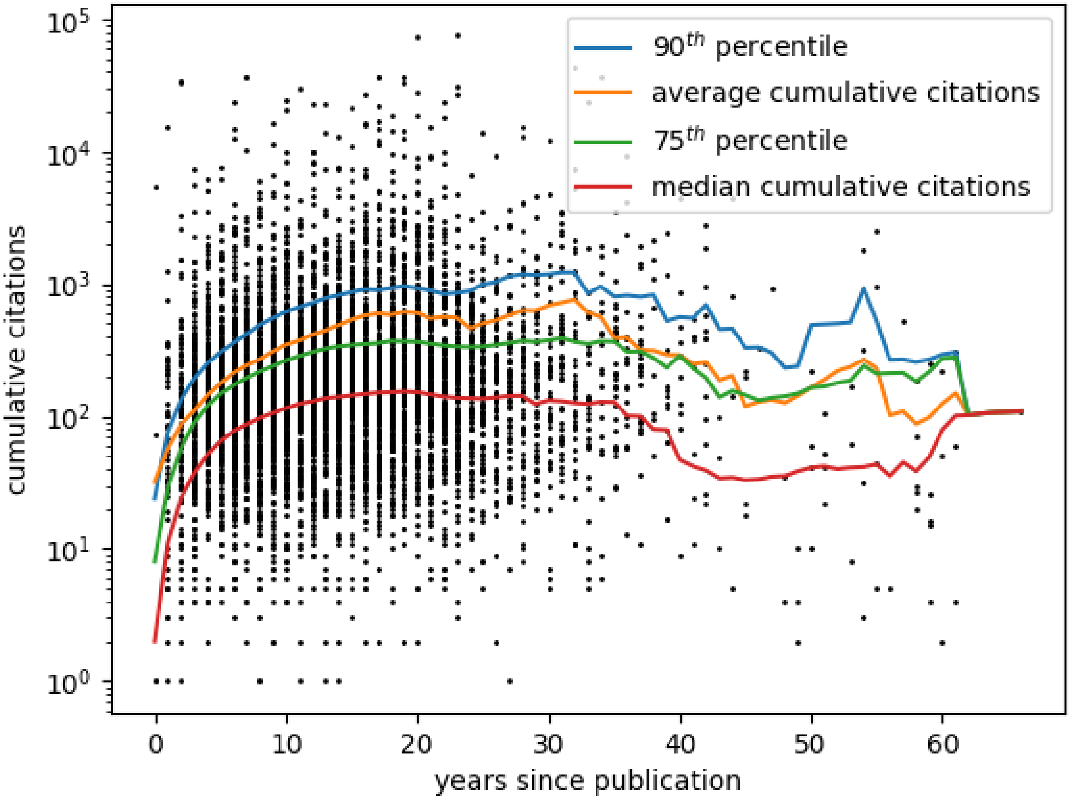


Figure 4.3: Relation between article age and cumulative citations

Chapter 5

Experiments

5.1 Preliminary Experiments

5.1.1 Finding the best regressor

Several experiments were conducted to find out which algorithm was the most effective in estimating future scientific impact. The first experiment was conducted with use of Weka. The data set used here consisted of the top 20 articles for each of the top 20 authors in Machine Learning as given by Google Scholar. The set was used to determine which algorithm was the most effective in predicting the total citations of a paper one year ahead based on the last five years. No attempt was made to differentiate between various ages of a paper. For each paper, 2017 was treated as year 0, the year to be predicted, and 2012-2016 were year 1 to 5, the attributes used to predict. Two methods were used to find the best performing algorithm: AutoWEKA and simply trying all 26 algorithms[1] provided by WEKA that were applicable to the problem without feature engineering. Two-thirds of the data were used for training and one-third as test set.
Both AutoWEKA and manual selection of algorithms returned linear regression as the best performing algorithm for prediction of total citations of a paper one year ahead. The model provided by AutoWEKA had a correlation coefficient of 0.9997 and a relative absolute error of 2.2%. Weka's standard linear regression algorithm performed even better on the test set, with a correlation coefficient of 1 and a relative absolute error of 0.28%. Among the other algorithms tried Simple linear regression and SMOreg (support vector machine for linear regression) performed best. Outside the realm of linear regression the Multilayer Perceptron function had the best results, with a correlation coefficient of 0.9996 and a relative absolute error of 3.6%. As may be expected, for all algorithms year one (the year directly before the year to be predicted) was by far the most important attribute.

The next experiment was done using data set 3, the set with data of 1023 authors drawn form Microsoft Academic. With this data set, the h-indexes of the first ten years of one's

[1] These algorithms were linear regression, simple linear regressian, Gaussian process, SMOreg, IBK, KStar, LWL, additive regression, bagging, CVParameterSelection, MultiScheme, RandomCommittee, RandomizableFilteredClassifier, RandomSubSpace, RegressionByDiscretization, Stacking, Vote, WeightedInstancesHandlerWrapper, InputMappedClassifier, DecisionTable, M5Rules, ZeroR, DecisonStump, M5P, RandomForest, RandomTree and REPTree. Seven of these resulted in a correlation coefficient lower than 0 and were not used in later tests. Those were ZeroR, InputMappedClassifier, WeightedInstancesHandlerWrapper, Vote, Stacking, MultiScheme and CVParameterSelection

career (starting with the first received citation, not with the first year of the PhD) were used to predict the h-index of year 15. As before, all applicable algorithms provided by Weka were tried to see which algorithm performed best in predicting h-index 5 years ahead for a scholar in the tenth year of his/her career. The next step was to test further ahead, predicting year 20 instead of year 15.
When predicting year 15, AutoWEKA selected AdditiveRegression as the best model. Predicting year 20, AutoWEKA selected M5P as the best performing algorithm, with a correlation coefficient of 0.808 on the test set. The results for the five best performing algorithms using normal Weka are shown in respectively Table 5.1 and 5.2.

Algorithm	**r^2 score**
MultilayerPerceptron	0.970
Linear regression	0.968
SMOreg	0.963
M5Rules	0.945
M5P	0.949

Table 5.1: Predicting 5 years ahead from year 10

Algorithm	**r^2 score**
M5Rules	0.825
M5P	0.825
Bagging (REPTree)	0.819
LinearRegression	0.816
RandomTree	0.801

Table 5.2: Predicting 10 years ahead from year 10

Among the algorithms tried, linear regression is the best performing algorithm when predicting one year ahead, occupies the second place when predicting 5 years ahead and the fourth place when predicting 10 years ahead. Apparently linear regression is one of the most effective algorithms in estimating future scientific impact, although its performance relative to other algorithms decreases over time and in all these cases the difference in score between the best performing algorithms is negligible. When predicting far ahead, M5Rules and M5P have a slightly better performance than linear regression. Both algorithms produce multiple linear regression models, and select one based on the value of the h-index in the last known year.

5.1.2 Using automated machine learning

The benefit of using automated machine learning to get better estimations of future scientific impact was tested using autosklearn. The data for this experiment came from data set 2. Year 15 was chosen as the target variable, and the first ten years as the predictor variables.
The "Best regression algorithm" model that will be described in the next section had an R^2-value of 0.880 in that situation, a median R^2-value of 0.917 and a median absolute error of 2.690.
Autosklearn was tested for different learning times, ranging from 5 minutes to 2 hours. Interestingly, 5 minutes was enough to find an algorithm with comparable scores. Improving the time available to the autosklearnregressor did in this experiment not result in better performance.

Learning time	**R^2**	**median R^2**	**median absolute error**
base case	0.880	0.917	2.690
5 m	0.868	0.877	2.566
15 m	0.869	0.860	2.546
1 h	0.870	0.892	2.602
2 h	0.874	0.910	2.183

Table 5.3: Performance scores of autoML-developed models for predicting the h-index 5 years ahead for an author with a career age of 10

5.2 Without additional features

In this section the results are discussed of the experiments that test the predictability of future scientific impact based on past values of said metric, for various scientific impact metrics. All tests are performed on data set 1 and use a training set size of 80%. Three approaches are taken to estimate future predictability and their results are compared on both R^2-value and relative error. The three approaches have in common that no single model is used for all predictions. Instead, for each combination of career age and prediction distance an individual algorithm was trained, with the aim of developing a set of algorithms that has specifically adapted algorithms for each career age/prediction distance combination. Developing a specific model for each career age is done because the relation between scientific impact in for instance the current year and five years ahead is probably very different for a junior scientist compared to a senior scientist. Using specific models for each prediction distance is done to enable the use of linear algorithms without assuming linearity. Note that career age is used in the sense of "years since first received citation", due to the fact that the collected data contain no information about the start year of the PhD. Many of the graphs that are shown further in this work show model performance for predicting up to 20 years ahead. For every prediction distance, the model performance shown in the graph is the average taken over the models for career ages 1 to 20.

The three different approaches mentioned are:

- **Linear regression:** This is a simple approach, using scikit-learn's linear regression algorithm, without hyperparameter tuning. The choice for linear regression is based on the positive results obtained with linear regression when searching for the best regressor (see preliminary experiments).
- **Best regression algorithm:** This approach is comparable to the one above, but instead of only looking at linear regression, various regression algorithms were tried, and in each circumstance the best performing one was selected (the one with the highest R^2-value on the test set. The following algorithms from scikit-learn were tried: LinearRegression(), BayesianRidge(), DecisionTreeRegressor(), RandomForestRegressor(), KNeighborsRegressor(), ElasticNet(), Lars() and OrthogonalMatchingPursuit() .
- **Time series prediction:** This approach is comparable to the "Best regression algorithm" approach above, but instead of taking all past values of the metric into account, the only predictor variables are the values of the past five years. This will enable us

to see if the inclusion of metric values for longer than five years past is still useful for improvement of the predictions, has no effect, or may even be harmful to the quality of the predictions.

5.2.1 The predictability of the h-index

The graphs 5.1 and 5.2 below show the predictability of the h-index, using respectively r^2-value and relative error as measures of the quality of the prediction. They show that there is a decreasing almost linear relationship between r^2 score and how many years ahead one predicts. The relative error also increases almost linearly, although the rate of increase decreases slightly over the years. For the median relative error, we also see a slowing rate of increase, until it reaches a plateau around 0.3. The median r^2 is higher than normal r^2, which suggests that relatively few predictions have a very high absolute error. Median relative error is lower than average relative error. That suggests that there are a few predictions that have a very high relative error, pulling the average up.

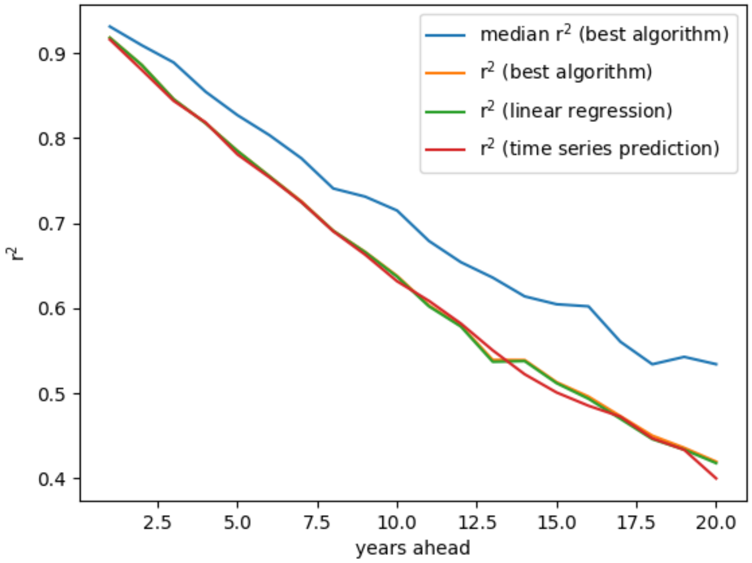


Figure 5.1: r^2-value for predicting the h-index

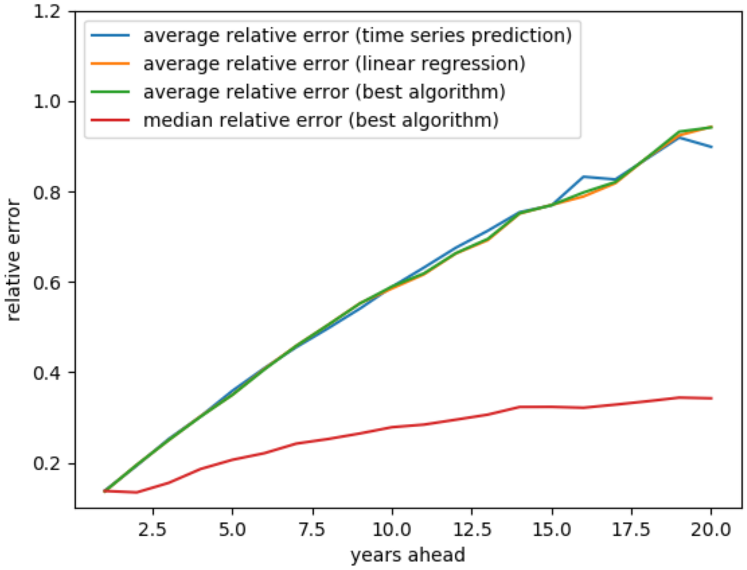


Figure 5.2: relative error for predicting the h-index

Scatterplots

The scatterplots below show the relation between true and predicted values for the test set, for junior scientists (5 years since first citation) and senior scientists (25 years since first citation). For each of these, scatterplots are shown of the predictions 3 and 10 years ahead.

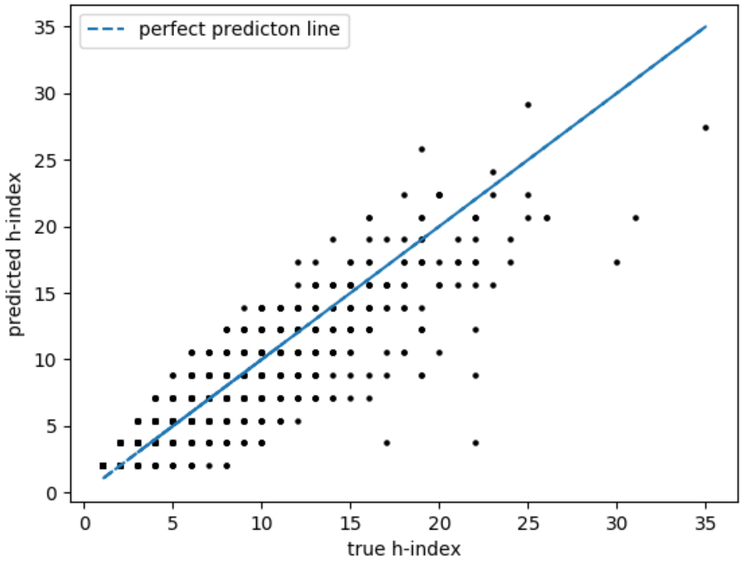


Figure 5.3: Scatterplot of true versus predicted value for the h-index of scientists in the fifth year of their career, predicting 3 years ahead

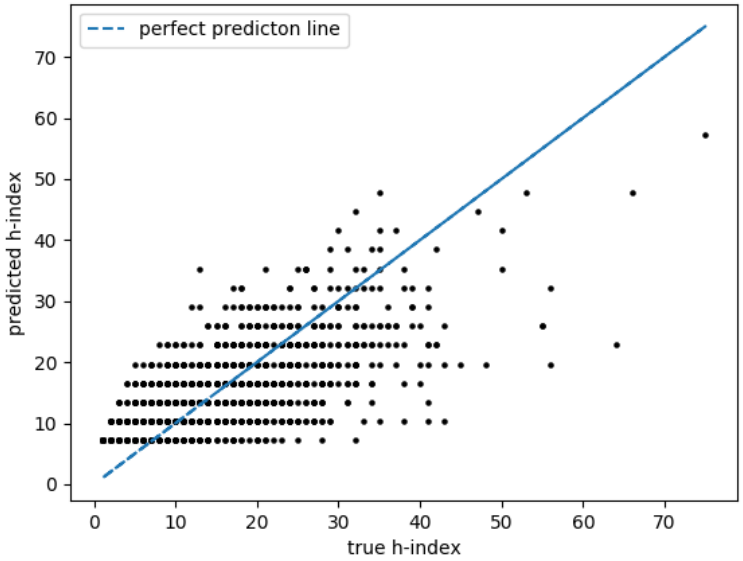


Figure 5.4: Scatterplot of true versus predicted value for the h-index of scientists in the fifth year of their career, predicting 10 years ahead

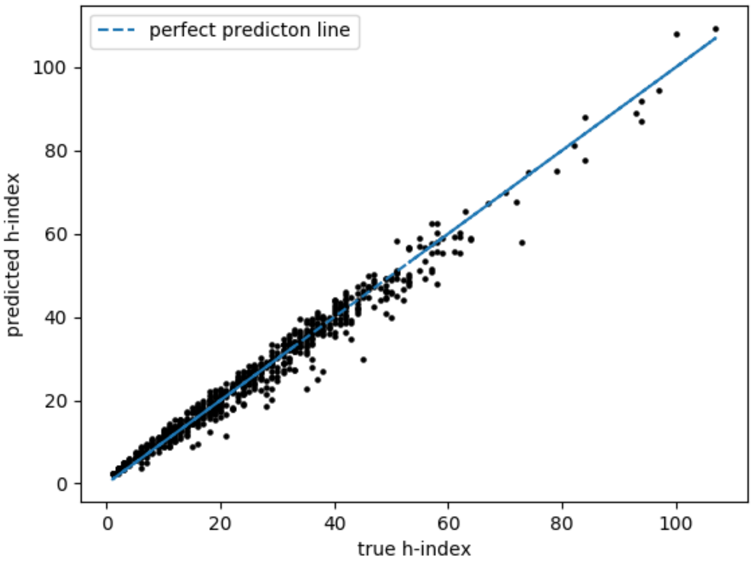


Figure 5.5: Scatterplot of true versus predicted value for the h-index of scientists in the 25^{th} year of their career, predicting 3 years ahead

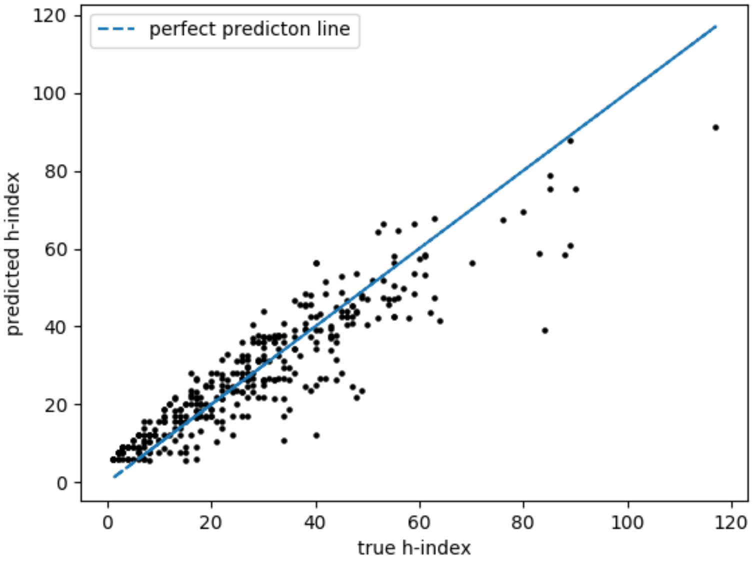


Figure 5.6: Scatterplot of true versus predicted value for the h-index of scientists in the 25^{th} year of their career, predicting 10 years ahead

Cases where the model does not work well

For junior scientists (5 years since first citation), mid-career scientists (15 years since first citation) and senior scientists (25 years since first citation) the results with the greatest relative error were analyzed for predictions 1,3,5 and 10 years ahead. Note that current h-index is used here to refer to the h-index of the last full year that is available for the input features, so if the model is used to predict future scientific performance it will often be the past year, because the current year is then yet incomplete.

Career age	**Years ahead**	**Current h-index**	**True h-index in target year**	**Predicted h-index in target year**
5	1	1	1	1.32
5	3	1	1	2.09
5	5	1	1	2.34
5	10	1	1	6.95
15	1	3	8	3.40
15	3	1	1	1.82
15	5	1	1	2.52
15	10	1	1	5.63
25	1	1	1	1.39
25	3	1	1	2.39
25	5	1	1	3.25
25	10	1	1	5.87

Table 5.4: The predictions with the greatest relative error in predicting h-index for a few sample circumstances

As Table 5.4 shows, in almost all circumstances the greatest relative errors were due to cases were the h-index of an author, at least according to the information contained in our data set, remained 1. In such cases the value predicted by the model is much higher than the true value for the target year. An exception is the prediction 1 year ahead for an author with a career age of 15. In that case, the h-index changed from 3 to 8 in the span of only 1 year, while the model predicted a much slower increase: to 3.40, which would be rounded to 3.

5.2.2 The predictability of the i10-index

The graphs below show the predictability of the i10-index, using respectively r^2-value and relative error as measures of the quality of the prediction. Here we see the same relationship between prediction distance on the one hand and r^2-value and relative error on the other hand that was seen for the h-index. The main difference is that for prediction of the i10-index the average error is greater and the r^2-value lower compared to what was seen for the h-index, showing that the i10-index is harder to predict than the h-index.

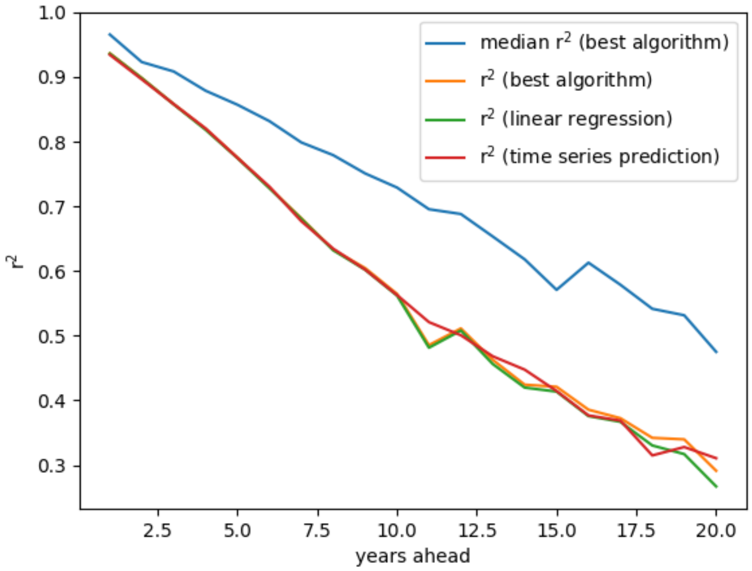


Figure 5.7: r^2-value for predicting the i10-index

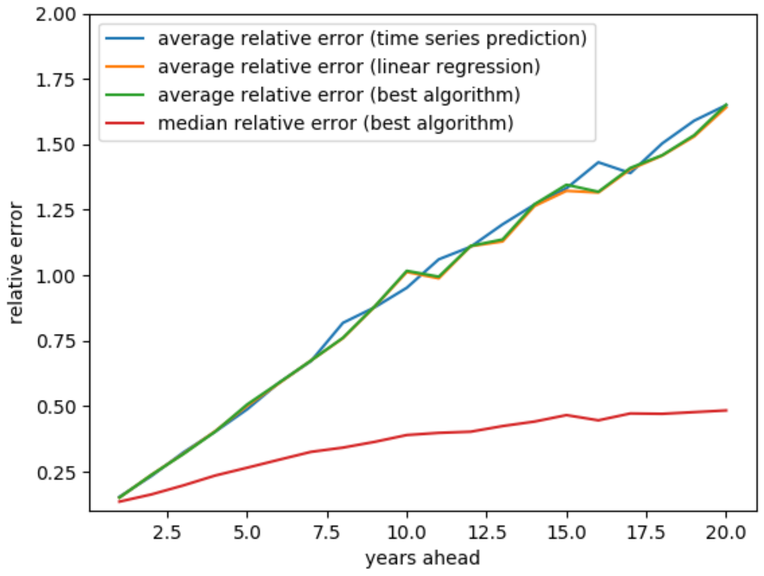


Figure 5.8: relative error for predicting the i10-index

Scatterplots

The scatterplots below show the relation between true and predicted values for the test set, for junior scientists (5 years since first citation) and senior scientists (25 years since first citation). For each of these, scatterplots are shown of the predictions 3 and 10 years ahead.

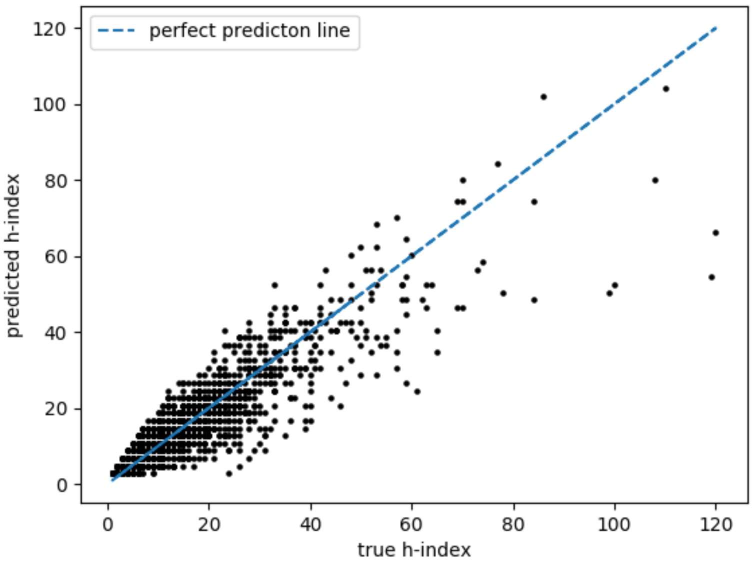


Figure 5.9: Scatterplot of true versus predicted value for the i10-index of scientists in the fifth year of their career, predicting 3 years ahead

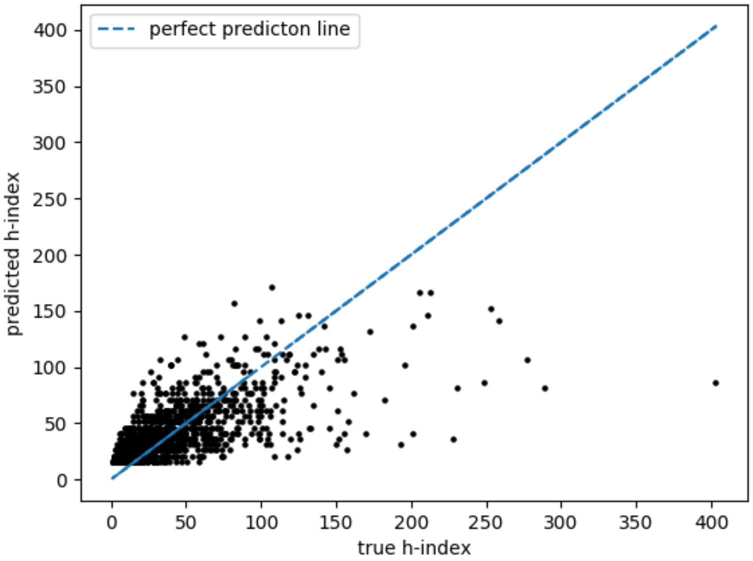


Figure 5.10: Scatterplot of true versus predicted value for the i10-index of scientists in the fifth year of their career, predicting 10 years ahead

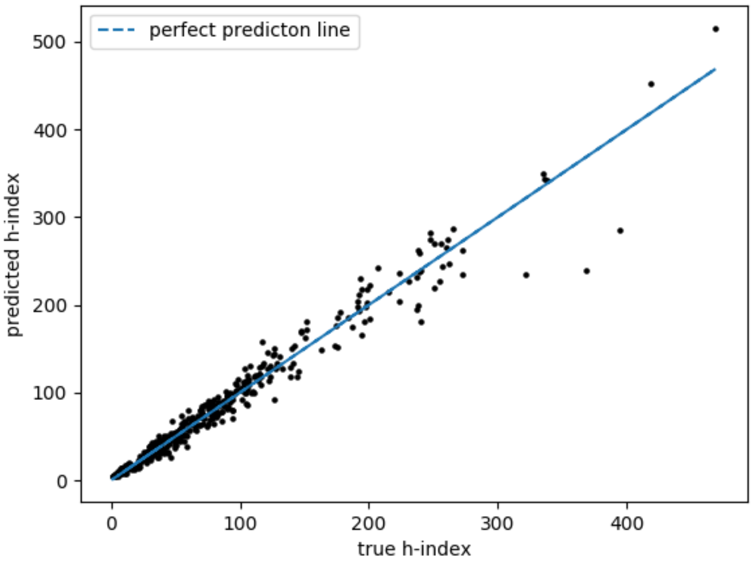


Figure 5.11: Scatterplot of true versus predicted value for the i10-index of scientists in the 25^{th} year of their career, predicting 3 years ahead

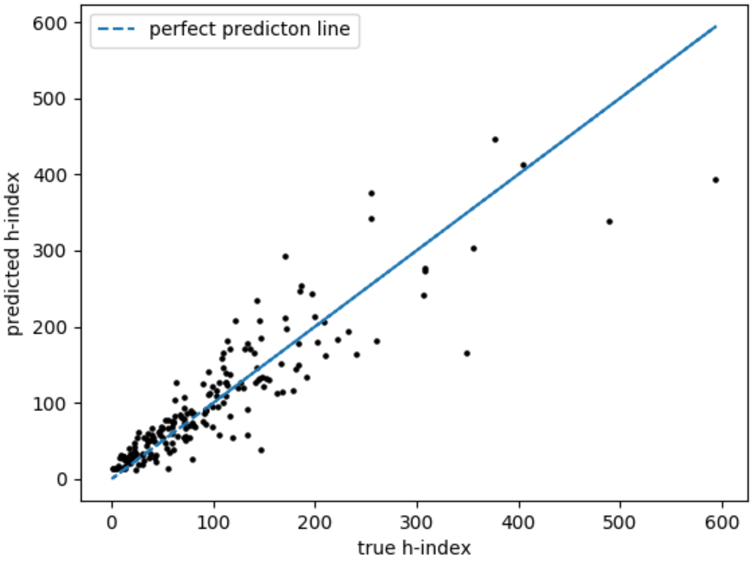


Figure 5.12: Scatterplot of true versus predicted value for the i10-index of scientists in the 25^{th} year of their career, predicting 10 years ahead

Cases where the model does not work well

For junior scientists (5 years since first citation), mid-career scientists (15 years since first citation) and senior scientists (25 years since first citation), the results with the greatest relative error were analyzed for predictions 1,3,5 and 10 years ahead.

Career age	Years ahead	Current i10-index	True i10-index in target year	Predicted i10-index in target year
5	1	1	5	1.42
5	3	1	1	2.80
5	5	1	1	5.19
5	10	1	1	15.67
15	1	1	1	-0.13
15	3	1	1	-0.81
15	5	1	1	-1.96
15	10	1	1	15.28
25	1	1	1	1.81
25	3	1	1	3.99
25	5	1	1	6.23
25	10	1	1	11.64

Table 5.5: The predictions with the greatest relative error in predicting i10-index for a few sample circumstances

As was the case with the predictions of the h-index, the worst relative errors occur when the value of the i10-index remains at 1. An exception is the prediction one year ahead for an author with career age 5. In that case the i10-index suddenly increased from 1 to 5, while the model estimated the i10-index in year 6 for that sample at only 1.42 (which would be rounded to 1).

5.2.3 The predictability of cumulative citations of authors

The graphs below show the predictability of the cumulative citations of authors, using respectively r^2-value and relative error as measures of the quality of the prediction. Comparing Figure 5.13 and Figure 5.14 with what was seen for the h-index and i10-index, we see there is a greater difference between median and normal r^2, and especially between median and average relative error. Apparently, even more than for h-index and i10-index there are relatively few cases with very high absolute error (decreasing the r^2-score) and very high relative error (increasing the average relative error).

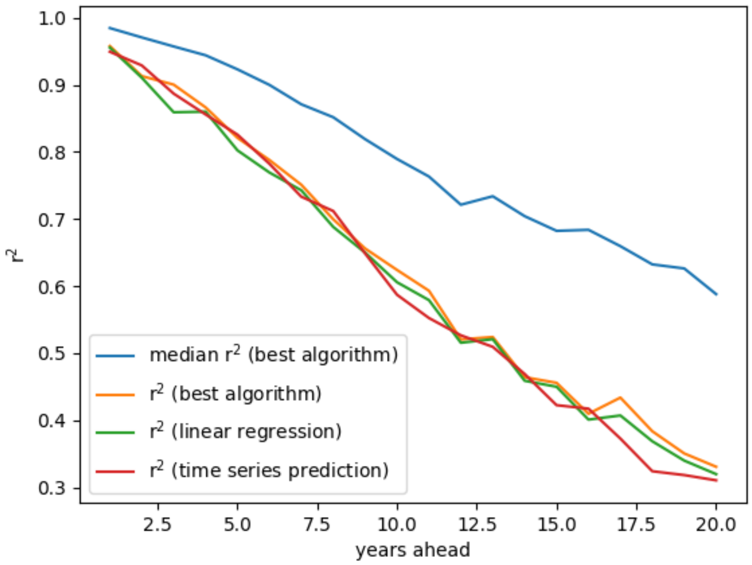


Figure 5.13: r^2-value for predicting the cumulative citations of authors

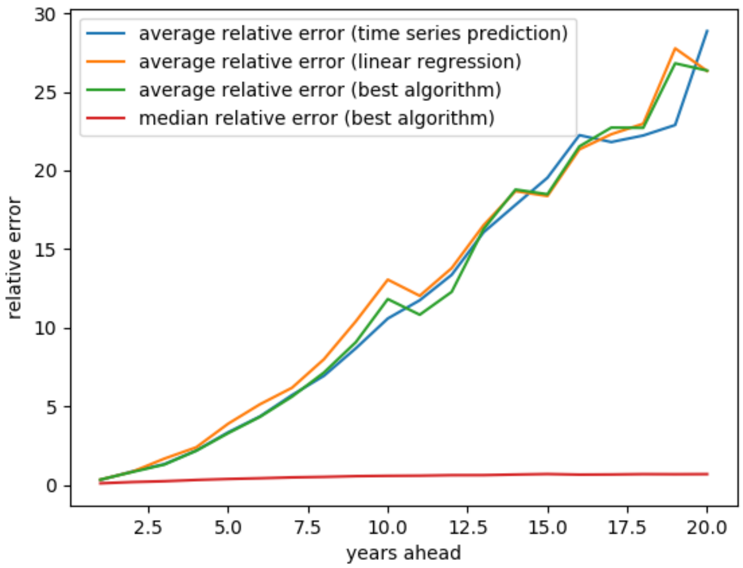


Figure 5.14: relative error for predicting the cumulative citations of authors

Scatterplots

The scatterplots below show the relation between true and predicted values for the test set, for junior scientists (5 years since first citation) and senior scientists (25 years since first citation). For each of these, scatterplots are shown of the predictions 3 and 10 years ahead.

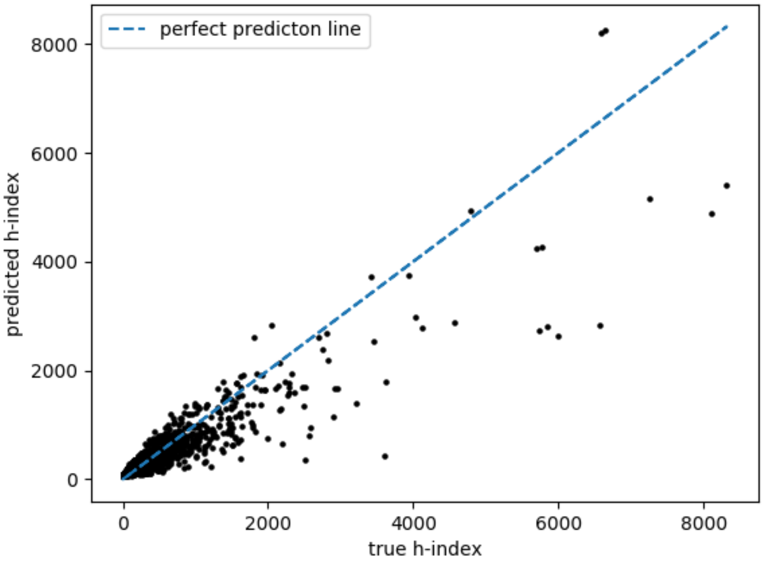


Figure 5.15: Scatterplot of true versus predicted value for the cumulative citations of scientists in the fifth year of their career, predicting 3 years ahead

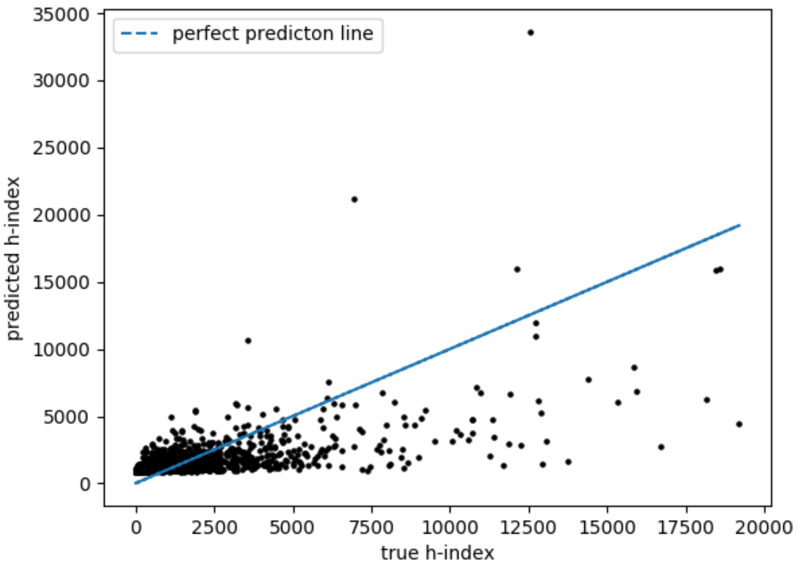


Figure 5.16: Scatterplot of true versus predicted value for the cumulative citations of scientists in the fifth year of their career, predicting 10 years ahead

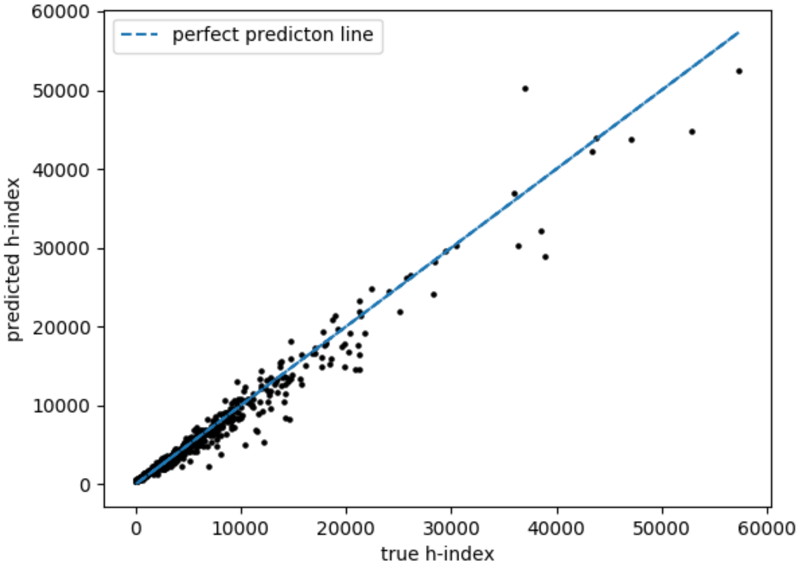


Figure 5.17: Scatterplot of true versus predicted value for the cumulative citations of scientists in the 25^{th} year of their career, predicting 3 years ahead

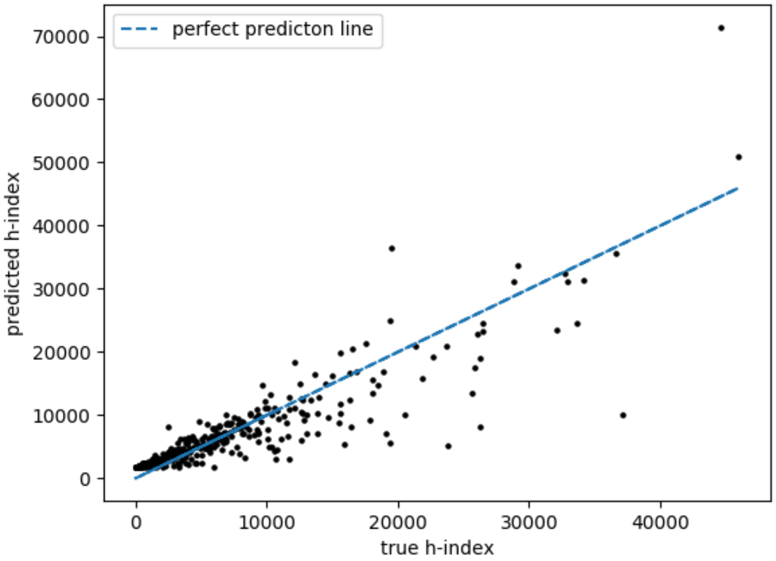


Figure 5.18: Scatterplot of true versus predicted value for the cumulative citations of scientists in the 25th year of their career, predicting 10 years ahead

Cases where the model does not work well

For junior scientists (5 years since first citation), mid-career scientists (15 years since first citation) and senior scientists (25 years since first citation), the results with the greatest relative error were analyzed for predictions 1,3,5 and 10 years ahead.

Career age	Years ahead	Current citations	True citations in target year	Predicted citations in target year
5	1	1	1	6.33
5	3	1	1	58.85
5	5	1	1	119.40
5	10	1	1	767.06
15	1	17	712	28
15	3	1	1	59
15	5	1	1	189
15	10	1	1	922
25	1	1	1	51
25	3	1	1	163
25	5	1	1	361
25	10	1	1	1068

Table 5.6: The predictions with the greatest relative error in predicting cumulative citations of authors for a few sample circumstances

Here we see the same results as with the prediction of h-index and i10-index: The model overestimates samples where the cumulative citations remain constant at only 1. Again, there is one exception when predicting one year ahead for a career age of 15. There the work of an author receives a sudden and drastic increase in citations which the model does not foresee.

5.2.4 The predictability of cumulative citations of publications

The graphs below show the predictability of cumulative citations for publications, using respectively r^2-value and relative error as measures of the quality of the prediction. in Figure 5.19 for the first time there is a clear difference in results between the "Best algorithm" approach and the linear regression and time series prediction approaches. Moreover, the median r^2 is extremely high, due to the fact that it expresses predictive performance relative to predicting the average, and predicting the average is apparently a very bad way to predict the cumulative citations of publications.

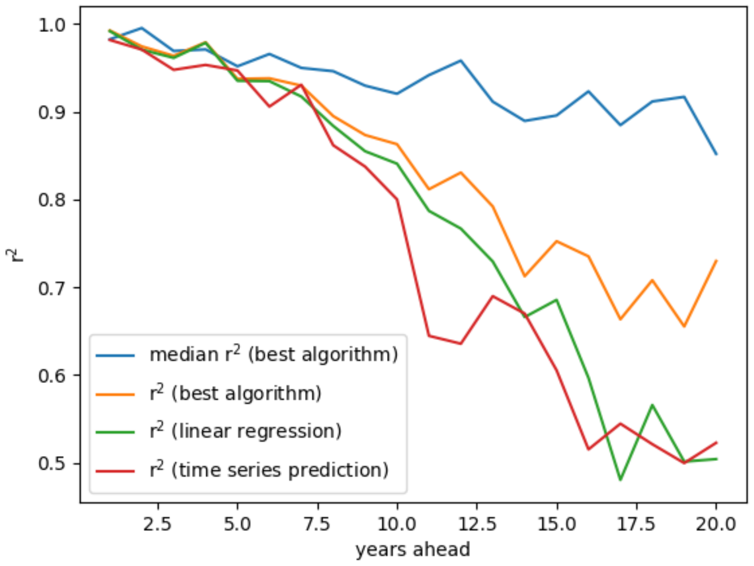


Figure 5.19: r^2-value for predicting the cumulative citations of publications

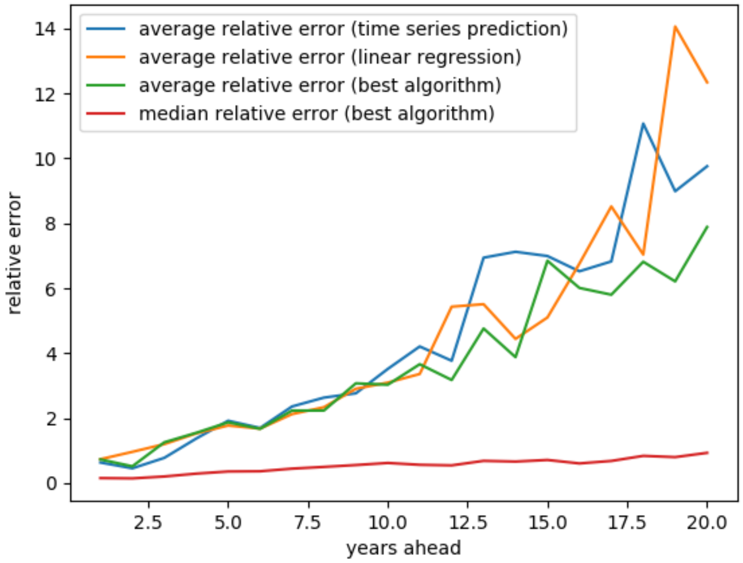


Figure 5.20: relative error for predicting the cumulative citations of publications

Scatterplots

The scatterplots below show the relation between true and predicted values for the test set, for publications with an age of 5 years and publications with an age of 25 years. For each of these, scatterplots are shown of the predictions 3 and 10 years ahead.

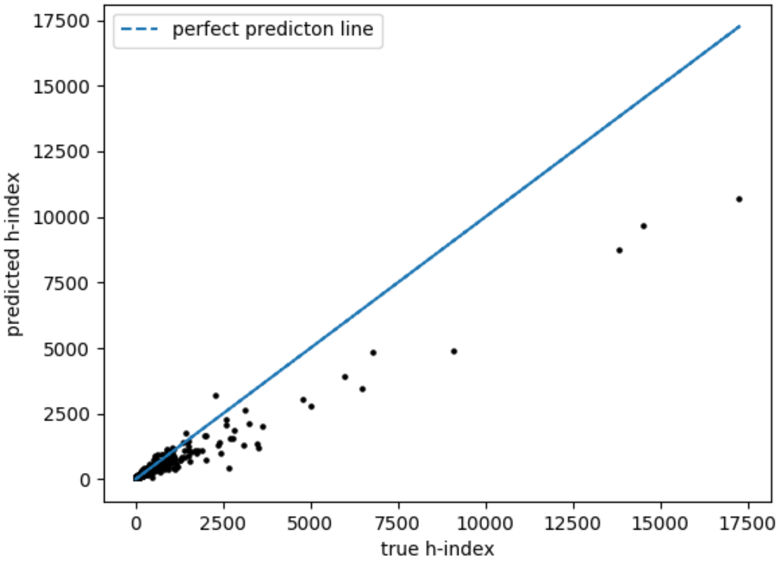


Figure 5.21: Scatterplot of true versus predicted value for the cumulative citations of publications 5 years since publication, predicting 3 years ahead

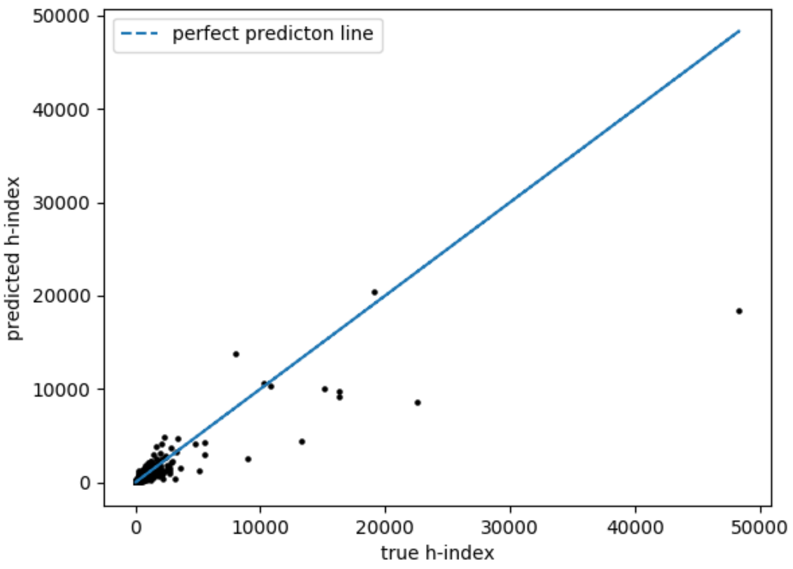


Figure 5.22: Scatterplot of true versus predicted value for the cumulative citations of publications 5 years since publication, predicting 10 years ahead

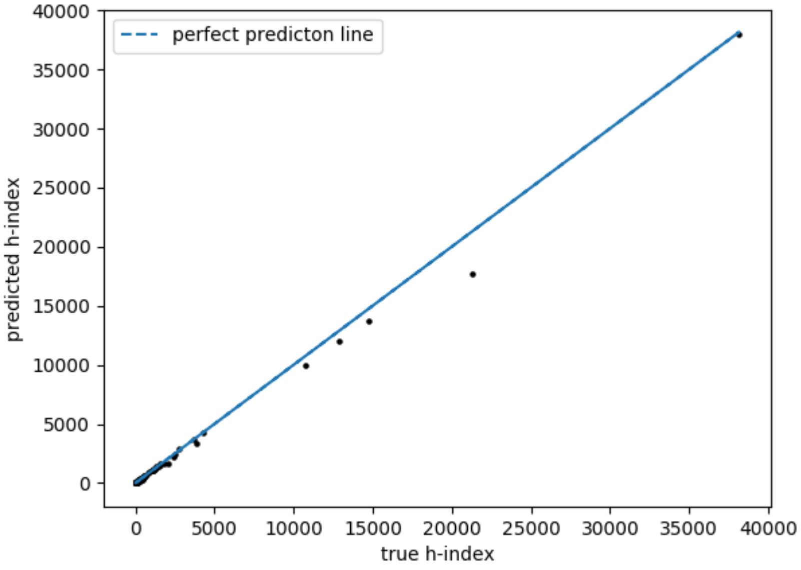


Figure 5.23: Scatterplot of true versus predicted value for the cumulative citations of publications 25 years since publication, predicting 3 years ahead

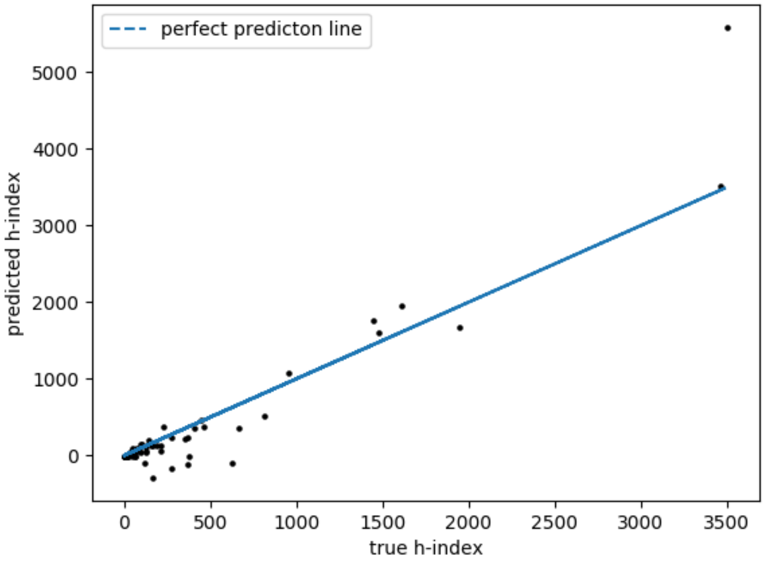


Figure 5.24: Scatterplot of true versus predicted value for the cumulative citations of publications 25 years since publication, predicting 10 years ahead

Cases where the model does not work well
For junior scientists (5 years since first citation), mid-career scientists (15 years since first citation) and senior scientists (25 years since first citation), the results with the greatest relative error were analyzed for predictions 1,3,5 and 10 years ahead. Note that for the cumulative citations of publications, in our model the first year is not the year of first citation, but the year of publication.

Career age	Years ahead	Current citations	True citations in target year	Predicted citations in target year
5	1	0	0	11.27
5	3	0	0	57.81
5	5	0	0	-8.77
5	10	0	0	67.29
15	1	0	0	-15.82
15	3	0	0	-31.99
15	5	0	0	3.60
15	10	0	0	4.45
25	1	0	0	2.93
25	3	0	0	19.37
25	5	0	0	32.24
25	10	0	0	25.14

Table 5.7: The predictions with the greatest relative error in predicting cumulative citations of publications for a few sample circumstances

As can be seen in Table 5.7, all tested cases of greatest relative error occur when the number of citations for a publication stays at 0.

5.2.5 Model performance compared to baselines

The graphs above have shown the obtained r^2-values and relative error for the predictions of our model for various scientific impact metrics. However, part of the reason the model performs so well is that the scientific impact metrics used (h-index, i10-index and cumulative citations) are all cumulative. Therefore they always follow an upward trend and never go downwards, which makes it easy to obtain accurate predictions when predicting only a few years ahead. To get an idea of the real predictive power of the model, the results obtained with the "best regression algorithm" approach are here compared to two baselines: Always predicting the current value, and always predicting the average value in the target year. These tests are performed for h-index and cumulative citations for papers.

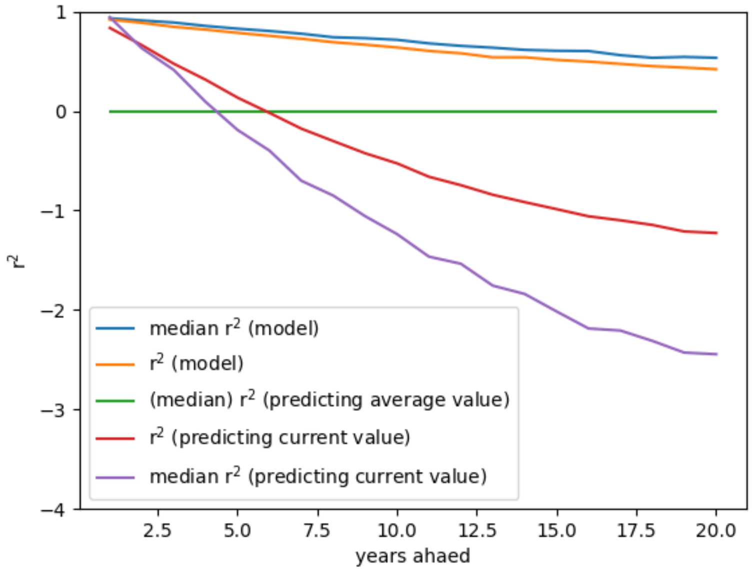


Figure 5.25: Comparison of the r^2-value for the predictions of the model, predicting the average h-index in the target years and predicting the current h-index

Because the r^2-value as used in this thesis expresses the prediction performance relative to predicting the average value, the r^2-value for predicting the average itself is 0. Looking at r^2-value, for all prediction distances up until 20 our model performs better than simply predicting the average value.

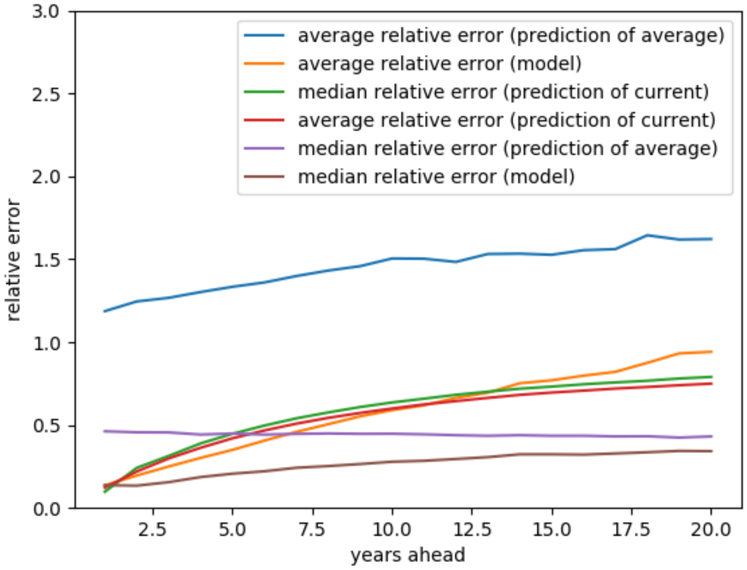


Figure 5.26: Comparison of the relative error for the predictions of the model, predicting the average h-index in the target years and predicting the current h-index

As Figure 5.26 shows, the relative errors for prediction of the average increase only slightly over the years, so for all years the distribution of the target values is about the same. The average and median relative errors for our model start almost at 0, although over the years they gradually approach the scores on these errors for prediction of the average. For a prediction distance of 20 years, however, our model still performs clearly better than prediction of the average.

When predicting more than 10 years ahead, predicting the current value has a lower average relative error than our model, although as we have seen our model still performs better when it comes to r^2-value and median relative error. The reason for this is probably that relative error is calculated as error relative to the true value of the target. Because of the cumulative nature of the h-index, the current value will always be between 0 and the target value, and therefore the relative error for predicting the current value can never exceed 1. Our model, on the other hand, can predict values that are arbitrarily higher than the true target value, and so the relative error can exceed 1.

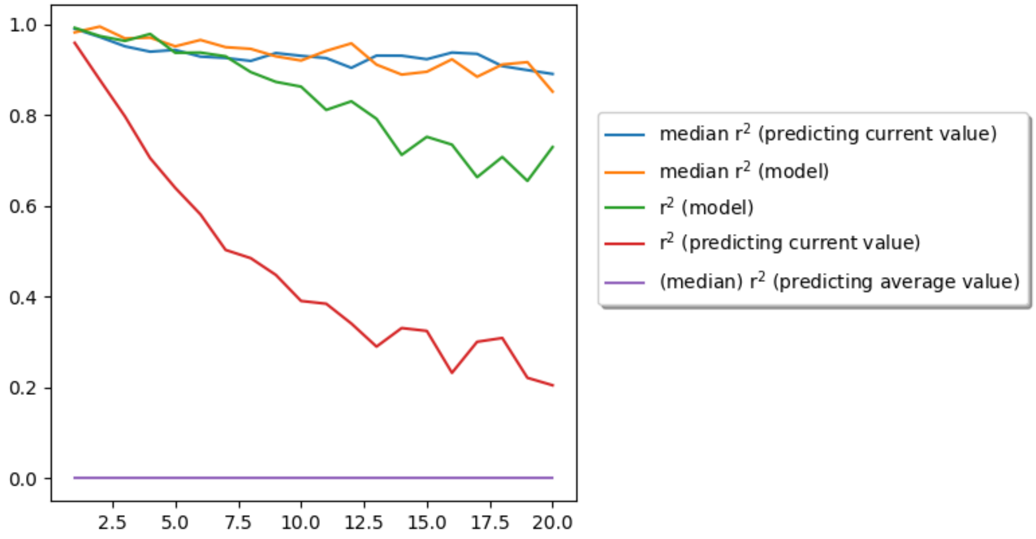


Figure 5.27: Comparison of the r^2-value for the predictions of the model for citations of publications with predicting the average cumulative citations in the target years and predicting the current cumulative citations

In Figure 5.27 it can be seen that the high median r^2-values obtained when predicting cumulative citations for publications are not due to the quality of the model itself, but to the properties of the data. As the r^2-value is an expression of performance relative to predicting the average value of the target variable, it simply means that predicting the average value of the target is a very bad predictor for cumulative citations of publications. In fact, simply predicting the current value results in an equal performance when it comes to median r^2-value, and for predictions further than 20 years ahead even a better performance. Normal r^2-scores are better for our model, meaning that outliers with high values are better predicted by our model than by simply predicting the current value.

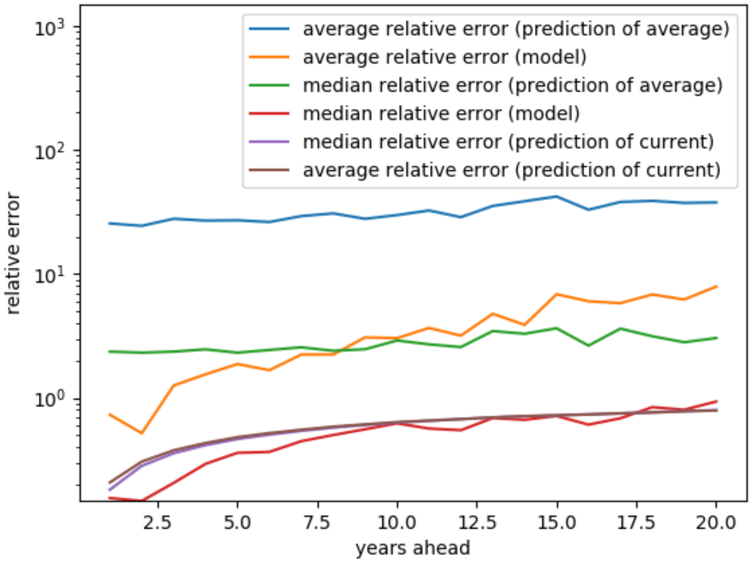


Figure 5.28: Comparison of the relative error for the predictions of the model for citations of publications with predicting the average cumulative citations in the target years and predicting the current cumulative citations. Note that the scale is logarithmic

As with h-index, the relative errors for prediction of the average have a more or less constant value, which is approached by the scores of our model for higher prediction distances.

5.2.6 Comparing the predictability of various impact metrics

Below graphs are shown that compare the predictability of h-index, i10-index, cumulative citations of authors and cumulative citations of papers. They are compared according to four metrics: r^2-value, median r^2-value, average relative error and median relative error. Of these, median relative error is the most informative metric, because out of the four metrics it is the only one that is not unduly influenced by high values of extreme outliers.

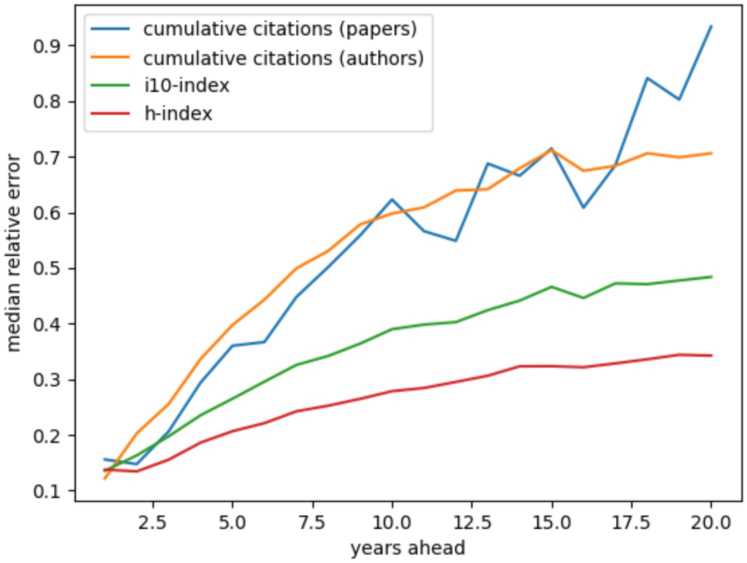


Figure 5.29: comparison of the median relative error for various metrics

As this graph shows, h-index appears to be the most predictable metric. That is no wonder, for it is also the most stable metric of the four. Especially in later years much more is needed to increase the h-index, than to increase i10-index or cumulative citations. It is interesting to note that in the first 20 years of prediction distance, there is not really a difference between median relative error for cumulative citations of authors and of papers. As the cumulative citations for an author can be regarded as the cumulative citations for a group of papers, this gives the impression that the future scientific impact of a group of papers is not more predictable than that of a single paper.

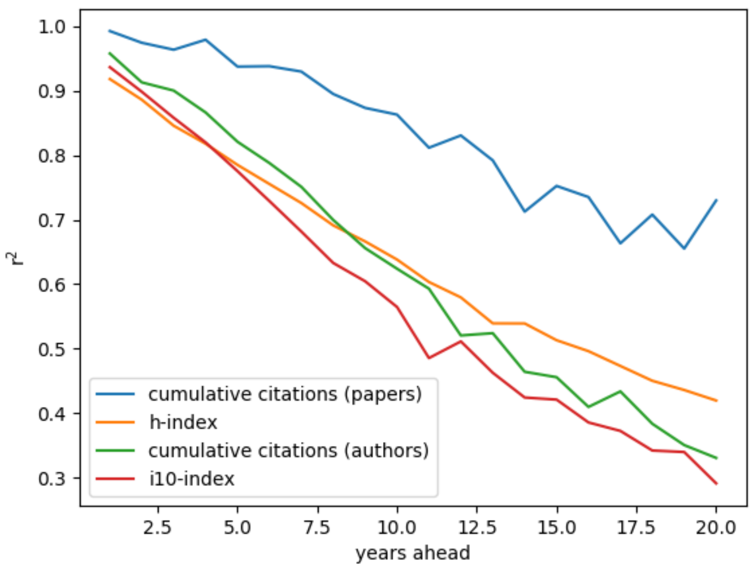


Figure 5.30: comparison of the r^2-value for various metrics

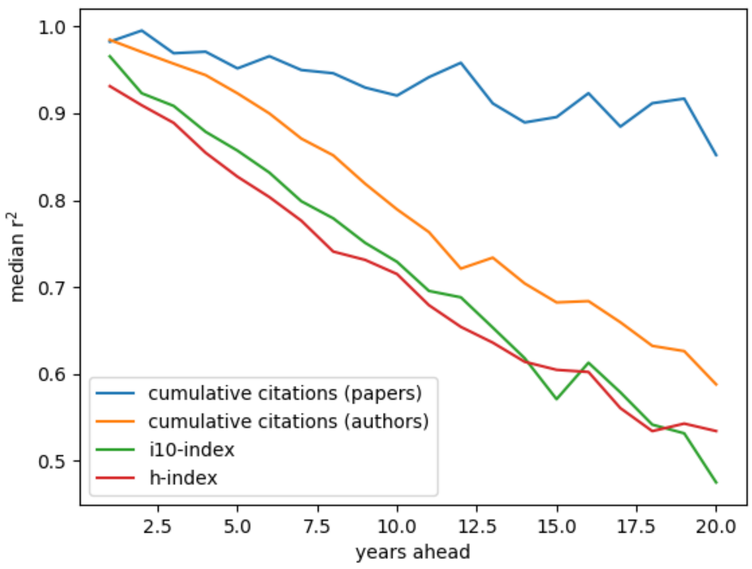


Figure 5.31: comparison of the median r^2-value for various metrics

Figure 5.30 and Figure 5.31 give the impression that cumulative citations for a single paper are much better predictable than cumulative citations for an author or the h-index or i10-index for an author. However, Figure 5.29 has already shown that this is not true. The high value of the r^2 for prediction of the cumulative citations of a paper is due to the way r^2 is calculated. As explained in the "Assessing prediction accuracy" section of the chapter on Preliminaries, the r^2-value is heavily influenced by extreme outliers with high values. If those few outlying values are predicted well, the overall r^2-value will be high, irrespective of the how well the other target variables are predicted.

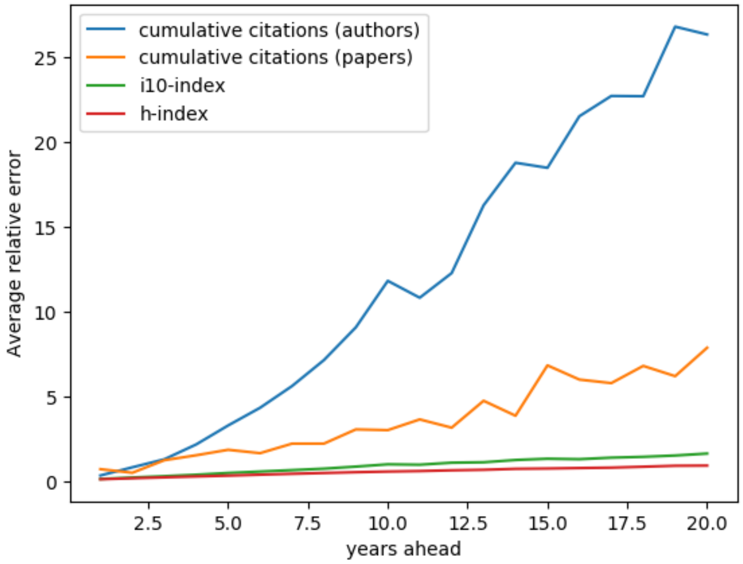


Figure 5.32: comparison of the average relative error for various metrics

In Figure 5.32 we see that the value of the average relative error appears to be related to how high the values of the underlying metrics are. H-index has overall the lowest values, and also the lowest average relative error. i10-index, which has slightly higher values (at least for authors with an i10-index higher than 10) also has a slightly higher average relative error. The same logic goes for the average relative error of predicting cumulative citations of papers and authors, where the highest relative errors are for the cumulative citations of authors, and that metric has indeed higher values than all the others.

5.2.7 The impact of career age on predictability

The impact of career age on predictability was tested by comparing r^2-value and median relative error for predicting the h-index of beginning scientists (a career age of 1), junior scientists (a career age of 5), mid-career scientists (a career age of 15) and senior scientists (a career age of 25). Note that in all these cases career age actually means "years since first received citation".

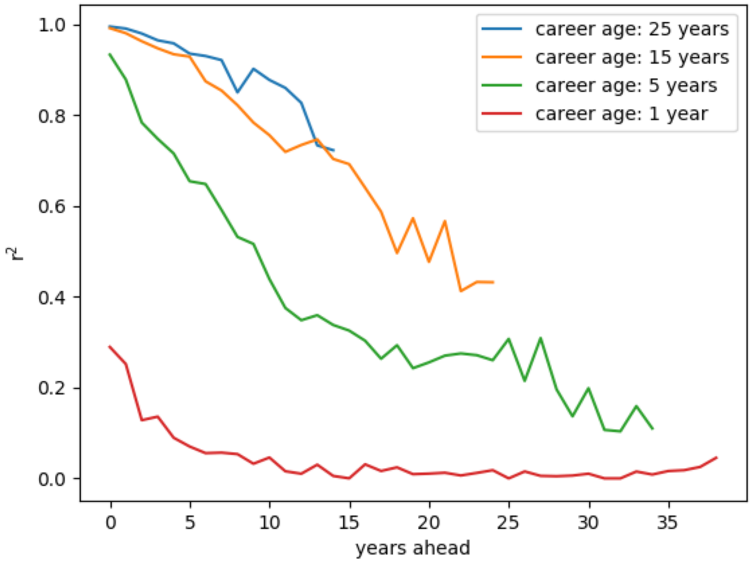


Figure 5.33: comparison of the r^2-value for various career ages

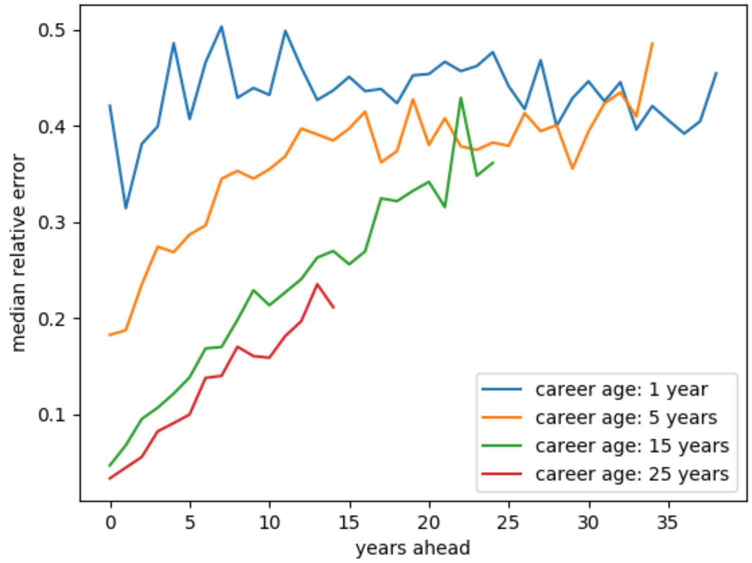


Figure 5.34: comparison of the median relative error for various career ages

Figure 5.33 shows that future h-index is practically not predictable for authors with a career age of only 1, based on past h-index alone. That makes sense, because in such a situation the only parameter that can be used for prediction is the h-index in the first year. That is clearly not enough to say anything meaningful about future h-index. By the time an author is five years into his career, his h-index is more predictable, but only for the next few years. The quality of the prediction rapidly declines. For mid-career and senior authors the situation

is different. Not only can their h-index be predicted with almost perfect accuracy for the next year, but the decline in r^2 over the next years is also less steep. The probable reason for this is that the growth of the h-index is on average much slower for authors with such a long career, because it is probably already quite high.
Additional tests for i10-index and cumulative citations of authors show the same pattern. Increases in career age bring an increase in predictability. The biggest gain in predictive power is when going from a career age of 1 to 5 years. Between a career age of 15 years and a career age of 25 there still is an increase predictive power, but only a slight increase.

5.2.8 The predictability of the union and intersection of authors

The predictability of the h-index for a single author was compared with that of the intersection of two authors and the union of five authors. In the case of intersection, a set of 5000 intersections was created by repeatedly randomly picking two authors from data set 1 and calculating the intersection of their work, provided they had co-authored at least one paper. A set of 5000 unions of authors was constructed by randomly selecting five authors from data set 1, combining their publications together and removing duplicates.

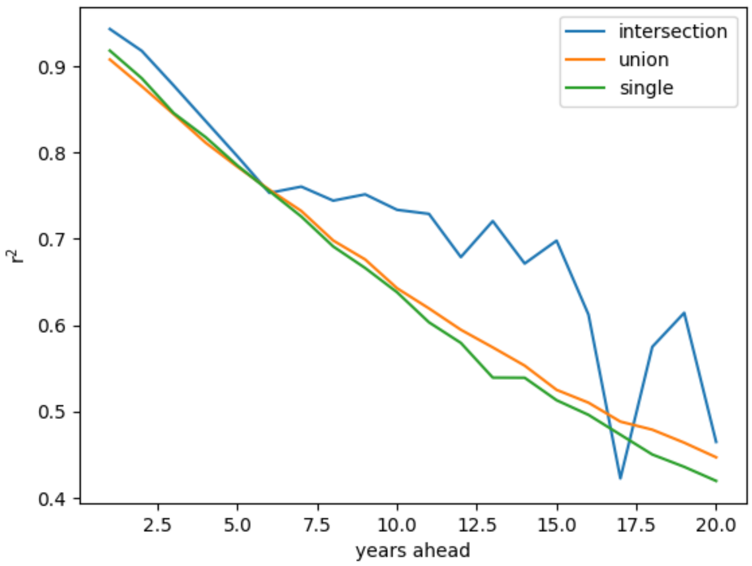


Figure 5.35: comparison of the r^2-value for single authors with that of a union or intersection of authors

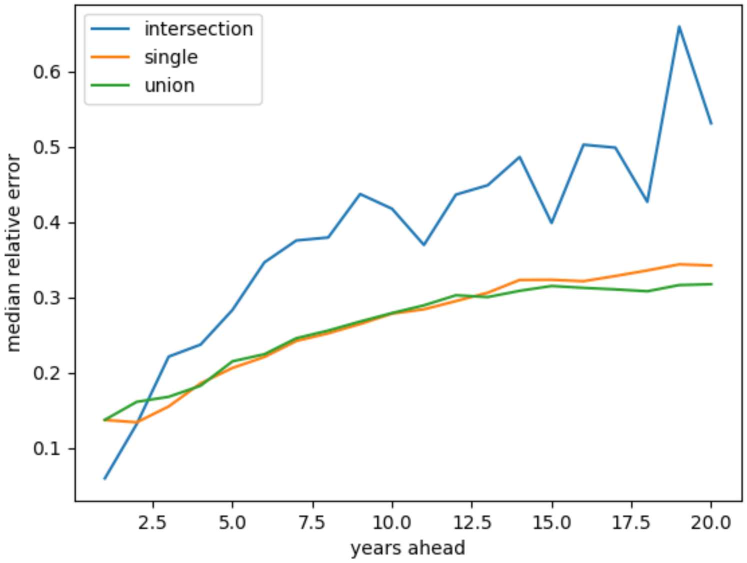


Figure 5.36: comparison of the median relative error for single authors with that of a union or intersection of authors

Figure 5.35 and Figure 5.36 show a slight but clear improvement in results when predicting the h-index of a group of 5 authors instead of a single author. Figure 5.35 is not clear enough to draw conclusions about the predictability of an intersection of authors, but Figure 5.36 shows that the h-index of either a single author or a union of authors is more predictable than that of an intersection of authors.

5.3 With additional features

The benefit of using other citation-related features to improve model performance was tested for the prediction of the h-index. Most of these features were earlier proposed by Weihs and Etzioni (2017). [23]

5.3.1 Clustering

The potential of improving prediction by employing clustering was tested on data set 2, and the conclusions confirmed with a test on data set 1. In these tests, the application of clustering led to worse results than using no clustering. To test the benefit of clustering, the arrays of h-index values over the past years (which are what is used as feature array in the previous experiments) were clustered using k-means clustering, implemented with sklearn.cluster.KMeans. The number of clusters was set to three, to prevent the cluster size from becoming too small. The data set was split in three according to the clusters, and for each cluster a model was trained and tested using 10-fold cross validation. The average r^2 score for these models was compared with the r^2 score obtained when clustering was not used. This test was performed on data set 2 for authors with a career age of 10, predicting 5 and 10 years ahead. The conclusions were confirmed by applying the same procedure to data set 1.

data set	years ahead	r^2 score (base case)	r^2 score (clustering)
2	5	0.84	0.76
2	10	0.66	0.57
1	5	0.84	0.57
1	10	0.68	0.35

Table 5.8: r^2 score for tests with an without clustering

5.3.2 Testing individual additional features

Relative change in h-index over the last two years

Using 10-fold cross-validation, the best performing algorithms when the relative change in h-index over the past two years was added as a feature, were linear regression and Bayesian ridge, both with an r^2-score of 0.869. The same values were obtained when the change in h-index was left out as a feature.

Additional features related to the number of citations

The cumulative citations, the change in citations over the last year, and the mean number of citations per year were all tested.

Additional features	Best algorithm(s)	r^2 score (best algorithm(s)
none	linear regression Bayesian ridge	0.869
cumulative citations	linear regression Bayesian ridge	0.871
change in citations over last year	linear regression Bayesian ridge	0.875
mean number of citations per year	linear regression Bayesian ridge	0.871

Table 5.9: r^2-score and best algorithms for prediction of the h-index with additional features based on number of citations

Additional features related to the number of papers

The number of papers published, the number of papers published in the last one, two and three years, the mean number of citations per paper and the relative change in the mean number of citations per paper over the last two years were all tested.

Additional features	Best algorithm(s)	r^2 score (best algorithm(s)
none	linear regression Bayesian ridge	0.869
number of papers	linear regression Bayesian ridge	0.876
number of papers in last year	Bayesian ridge	0.890
number of papers in last 2 years	linear regression Bayesian ridge	0.886
number of papers in last 3 years	linear regression Bayesian ridge	0.882
mean citations per paper	linear regression Bayesian ridge	0.871
change in mean citations per paper over the past 2 years	linear regression Bayesian ridge	0.870

Table 5.10: r^2-score and best algorithms for prediction of the h-index with additional features based on number of papers

5.3.3 Testing additional features in combination

As a next step, the additional features were used in combination. The following features were added:

- Change in h-index over the last two years
- Cumulative citations
- Change in citations over the last year
- Mean number of citations per year
- Number of papers published
- Number of papers published in the last two years
- Mean number of citations per paper
- Change in the mean number of citations per paper over the last two years

With these additional features, a model was trained to predict the h-index and the results were compared with what was found earlier for models trained without using any additional features. The results can be seen in Figure 5.37 and Figure 5.38. Using extra features results in a clear improvement for any prediction distance, at least if performance is measured by r^2-value or median r^2-value. Measured by median relative error, both models perform approximately the same for the first five years. After that, using additional features actually increases the median relative error. This is probably caused by the fact that the sklearn optimizes the model for r^2 score. The additional features allow for the model to obtain a higher r^2 score, but as we see this is at the cost of median relative error. However, it shows that additional features allow the model to be more optimized, and therefore if the models would be optimized for median relative error, the model with additional features would probably perform better when it comes to median relative error. Measured by average relative error, both models performs approximately the same for all prediction distances, while using additional features leads to better results for all prediction distances when one looks at median absolute error and average absolute error.

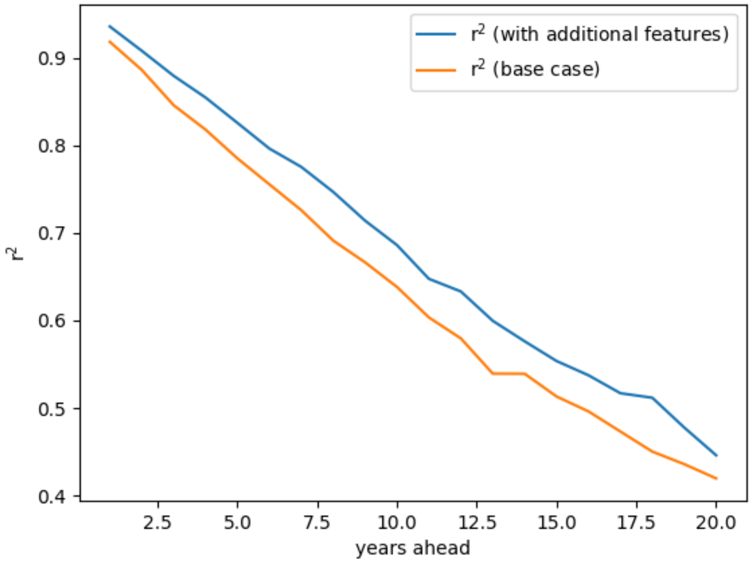


Figure 5.37: r^2-score for using additional features compared with base case

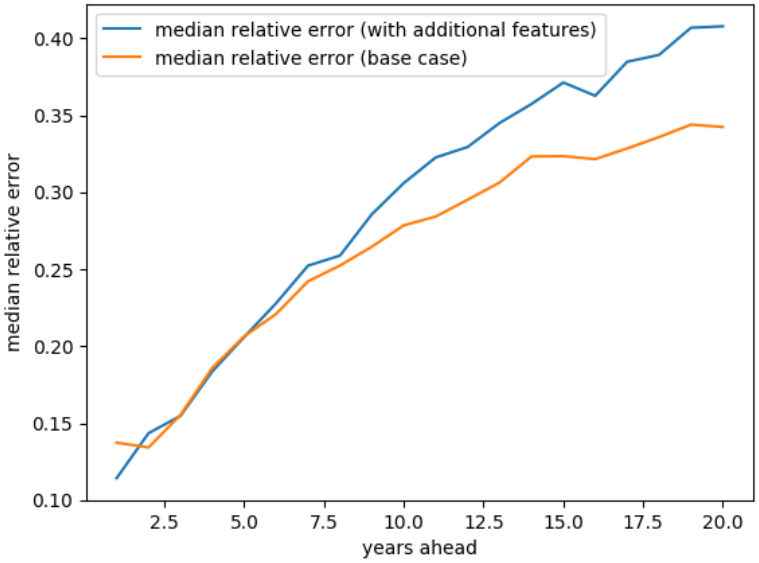


Figure 5.38: median relative error for using additional features compared with base case

5.4 Testing on data from Google Scholar

We tested to what extent a model trained on data from Microsoft Academic was usable for data from other platforms. To this end, we first collected a data set of citations for publications from the Microsoft Academic API. For each author in data set 1, two publications were randomly selected for inclusion in this new data set. Some of these papers were never cited, and others were not cited earlier than 2015 (the last year we take into account). Dropping these, a set of 11,346 publications was left. A scatterplot showing the ages and cumulative citations of these publications in 2015 is shown in Figure 5.39. This data set has significant differences from the set of publications that was scraped from Google Scholar. Overall, the cumulative citations for a publication are much lower. This is evidenced by the median, which stabilizes between 10 and 20 for this data set, whereas for the Google Scholar data set for most years it is around 100.

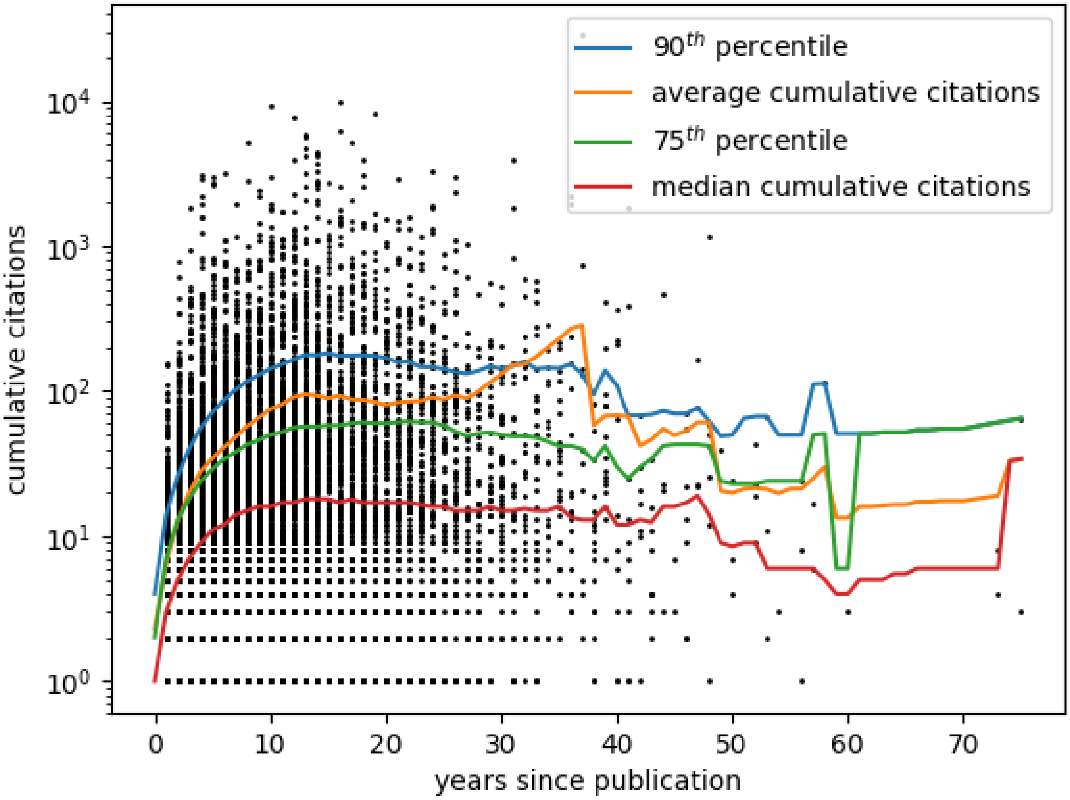


Figure 5.39: Relation between article age and cumulative citations for the data set from Microsoft Academic

This new data set was used to train models for the prediction of cumulative citations for publications, in the same way as was done for the Google Scholar data set. The results were compared with what was obtained for the model trained on the Google Scholar data set. As a next step, the models trained on the Microsoft Academic data set were used to predict citations for the Google Scholar data set, using the entire Google Scholar data set as test set, with one modification. While in the original experiments on the Google Scholar data set, the age of a paper was reckoned from the year of publication, in this case it was reckoned from the year of first citation, as was done for the Microsoft Academic data set. The scores for those predictions were compared with the scores these same models obtained on the Microsoft Academic data set.

The results can be seen in Figure 5.40 and Figure 5.41. Looking at r^2 scores, we see that the results for the Microsoft Academic model on Microsoft Academic data are slightly lower than for the Google Scholar model on Google Scholar data. This makes sense, because as we have seen the Microsoft Academic data sets lack such high outliers as the Google Scholar data set, which can significantly influence the overall r^2 score if predicted right. Because the Microsoft Academic model is not tuned to predict those outliers, the r^2 scores it obtains on the Google Scholar data set are lower. However, even so, the scores for predictions up to 10-15 years ahead are still quite good.

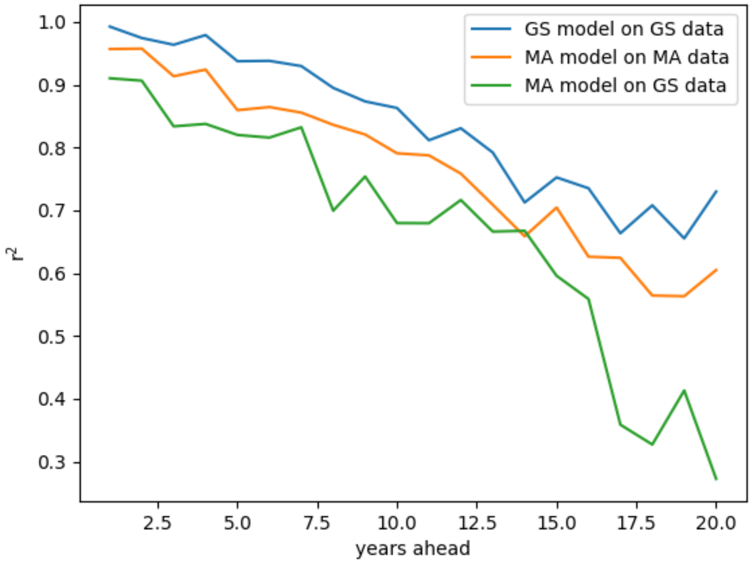


Figure 5.40: R^2-scores for predicting future cumulative citations for publications using data from Google Scholar or Microsoft Academic, and models trained on either of those data sets

When looking at median relative error (see Figure 5.41), the results are quite strange. The median relative error for the Microsoft Academic model on Microsoft Academic data is clearly higher than that for the Google Scholar model on Google Scholar data. Even more strange, the model trained on data from Microsoft Academic performs better on data from Google Scholar than on data from Microsoft Academic, if one takes median relative error as the performance measure. A possible explanation for this is that the Microsoft Academic data set contains quite a lot of publications that have a very low number of citations every for older publications. Because relative error is measured as error relative to the true value, overestimating such examples with low true values will quickly lead to high relative errors.

For average relative error, the best results are again those for the Microsoft Academic model on Google Scholar data, this time followed by the Microsoft Academic model on Microsoft Academic data, while the Google Scholar model on Google Scholar data has the worst results. Looking at median absolute error, by far the best results are for the Microsoft Academic model on Microsoft Academic data, which makes sense because the cumulative citations are on average much lower for publications in this data set.

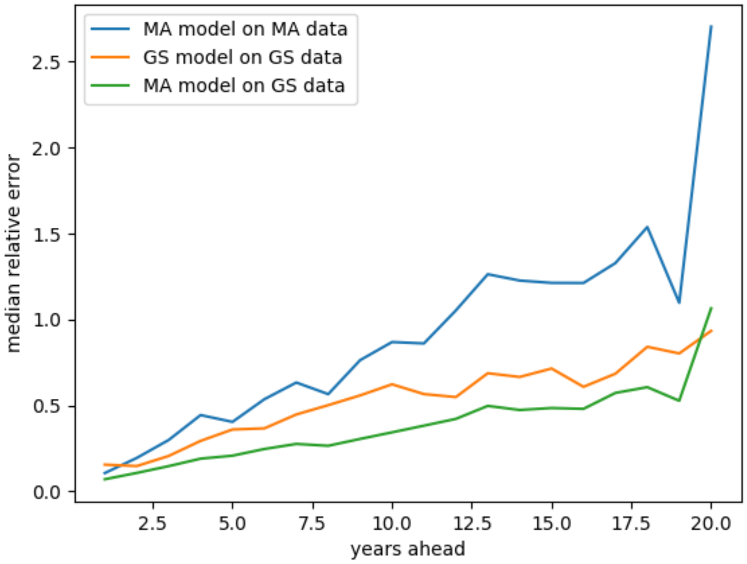


Figure 5.41: Median relative error for predicting future cumulative citations for publications using data from Google Scholar or Microsoft Academic, and models trained on either of those data sets

5.5 Predicting incremental h-index

The same approach that was used to predict the absolute values of the h-index for a certain distance of years ahead has also been used to predict the incremental h-index, or relative increase in h-index. To clarify terms: A model predicts the absolute value of the h-index when it predicts what the value of the h-index will be in the target year. A model predicts the incremental h-index/relative increase in h-index when it predicts the value of the h-index in the target year compared to the value of the h-index in the current year. In the first case, a model might predict the h-index to be 16 in the target year. In the second case, a model might predict the value of the h-index in the target year to be 1.33 times the value of the h-index in the current year.

The scores of the evaluation metrics on the results are very different from those obtained when predicting absolute values of the h-index, in an interesting way. As Figure 5.42 shows, the r^2 scores for predicting the incremental h-index are very low. This means that the predictions of the model for relative increase in h-index are only a limited improvement upon predicting the average relative increase in h-index. However, it does not mean that predicting the incremental h-index really leads to worse results than predicting the absolute values of the h-index. This becomes clear when we look at Figure 5.43 and Figure 5.44. Figure 5.43 shows that simply predicting the average relative increase in h-index yields results comparable to those of the model for small prediction distances. Predicting farther ahead, the median relative error for predicting the average relative increase grows faster than the median relative error for predictions of the model, and eventually predicting the average relative increase yields worse results than predicting the average value of the h-index in the target year.

Another insight that can be learned from Figure 5.43 is that when it comes to median relative error, there is no large difference in performance between predicting the absolute value of the h-index in the target year or the relative increase in h-index between current year and target year. Predicting less than 10 years ahead, predicting the relative increase yields better results, but for longer distances the results for predicting the absolute value are better, and the difference in performance increases with the prediction distance.

Figure 5.44 shows that when it comes to average relative error predicting the relative increase in h-index performs consistently better than predicting the absolute value of the h-index in the target year for all prediction distances up to 20 years. However, it also shows that simply predicting the average relative increase yields results comparable to our model, and in some cases even slightly better. It is therefore no wonder that the r^2 scores for predicting the incremental h-index are so low.

In conclusion, these tests show that predicting the relative increase in h-index yields results with a lower average relative error and for predictions less than 10 years ahead a lower median relative error as well. The tests also show that our approach is not able to improve upon simply predicting the average relative increase in h-index when it comes to average relative error, and when considering median relative error, it yields better results, but only for prediction distances of more than 5 years ahead.

Lastly, simply predicting the average relative increase in h-index yields better results than the predictions of our model for absolute h-index values when considering average relative error, although our model is still the better choice when considering median relative error.

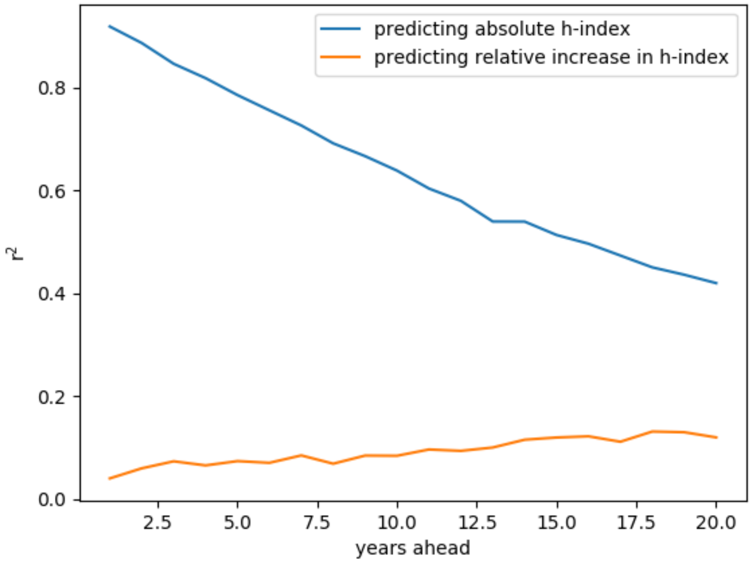


Figure 5.42: Comparison of r^2 scores for predicting the absolute h-index and the incremental h-index

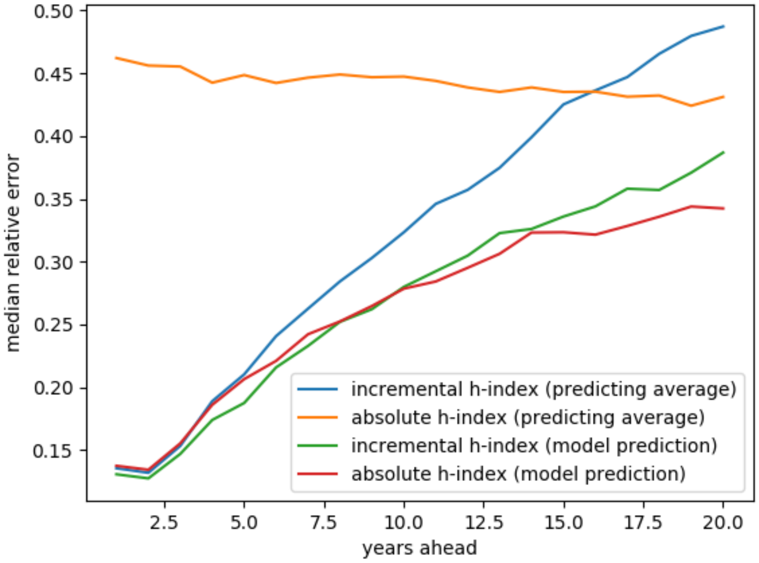


Figure 5.43: Comparison of median relative error for predicting the absolute h-index and the incremental h-index

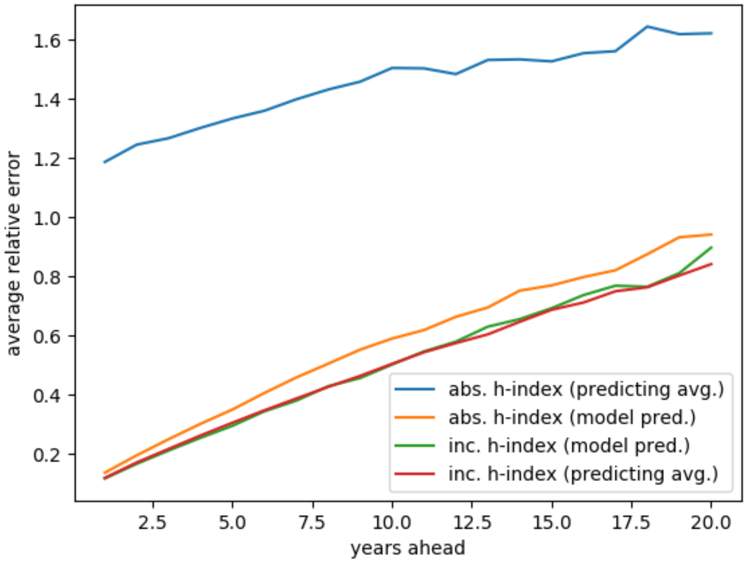


Figure 5.44: Comparison of average relative error for predicting the absolute h-index and the incremental h-index

Chapter 6

Conclusions and future work

The aim of this thesis was to find to which extent future scientific impact was predictable from past scientific impact. The research has shown that past and current h-index, i10-index, cumulative citations of authors and cumulative citations of papers are informative enough about future values of these metrics to provide better predictions than simply predicting the average value for at least 20 years. Because all used metrics are cumulative and contain therefore a certain measure of autocorrelation, the results of the model were also tested against predicting the current value. The model performs better than this baseline as well, except in cases where the used metric was biased towards one of these baselines, as is for instance the case in using median relative error to evaluate predicting the current value.

Comparison of impact metrics
The more stable a metric is and the lower the growth rate, the easier it is to predict. A comparison based on median relative error shows the lowest error rates for h-index, followed by i10-index and cumulative citations. Predictions of cumulative citations for authors (the metric with the highest values on average) score the worst when compared on average relative error. When comparing on median relative error, however, predictions for cumulative citations for authors and papers have about the same error rate for the first 20 years.

Union and intersection
A clear but slight improvement in results is seen when predicting future h-index for a set of five authors instead of a single author. Predicting the h-index of work co-authored by two authors proves to be less easy to predict, at least measured by median relative error.

Career age
A close relation is found between career age and predictability. Predictions for a career age of 1 score very low, but the higher the career age, the more accurate the predictions are. The biggest gain in prediction accuracy occurs during the first few years. After that, the positive impact of a longer career on predictability begins to decrease, although it remains detectable even for the difference between a career age of 15 or 25 years.

Additional features
The benefit of using several additional features that are also based on citation data has been explored for the h-index. The conclusion is that the use of these additional features can indeed improve model performance. The most effective additional feature among the tested features

is the number of new papers an author has published in the last year.

Dependence on the data source

The accuracy scores obtained are to a certain extent dependent on the data set. Using the same approach on citation data for publications from Google Scholar and from Microsoft Academic, the r^2 scores obtained on the Google Scholar data set were higher and the median relative errors lower, while the Microsoft Academic model performed better on average relative error. When testing the model trained on Microsoft Academic data on data from Google Scholar, the r^2 score on the Google Scholar data set is lower, but the Google Scholar data set performs significantly better when looking at relative error. However, all of these differences can be explained by differences in the data sets, making it difficult to draw firm conclusions.

Predicting relative increase

The relative increase in h-index was predicted using the same method applied to the prediction of the value of the h-index. The results show that prediction of the relative increase in h-index yields better results for average relative error, and better results for median relative error for predicting less than 10 years ahead. After that prediction of the value of the h-index yields better results for median relative error.

Future work

In this thesis, the data that was used was limited to the first forty years since the first time an author was cited, and the prediction distance was limited to twenty years ahead. Future work could look into the predictability of scientific impact for longer time frames.

Another topic for future research is the use of automated machine learning. In this paper it was concluded that automated machine learning did not improve upon the results of simple algorithms in a feasible time scale. However, the experiments on autoML were limited. It would be interesting to try a more comprehensive approach, using more data, testing for more different situations and giving automated machine learning more time to find an appropriate model.

As seen in this thesis, most evaluation metrics were influenced by relatively few outliers. Future work could mitigate this by normalizing the scientific impact metrics, or by predicting the relative increase instead of absolute values. That might result in models that are less sensitive to outlying data. The experiment in this thesis involving the prediction of incremental h-index has already shown that under certain circumstances such an approach can indeed improve results.

Bibliography

[1] Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature, 489*(7415), 201202. DOI: 10.1038/489201a

[2] Ayaz, S., Masood, N. & Islam, M.A. (2018). Predicting scientific impact based on h-index. *Scientometrics, 114*(3), 993. DOI: 10.1007/s11192-017-2618-1

[3] Daniel, R. (2014) Predicting Citation Counts. *Research Trends*(37). Retrieved from `https://www.researchtrends.com/issue-37-june-2014/predicting-citation-counts/`

[4] Dong, Y., Johnson, R.A., & Chawla, N.V. (2016). Can Scientific Impact Be Predicted? *IEEE Transactions on Big Data, 2*(1), 18-30. DOI: 10.1109/TBDATA.2016.2521657

[5] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. Blum, M. & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In Cortes. C. et al. (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems, 2* (pp. 2755-2763). Cambridge, MA: MIT Press.

[6] Frank, E., Hall, M. & Witten, I. (2016). The WEKA Workbench. Online Appendix for Witten, I. et al, *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.

[7] Hirsch, J. E. (2005). An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569-16572. DOI: 10.1073/pnas.0507655102

[8] Hirsch, J.E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences, 104*(49), 19193. DOI: 10.1073/pnas.0707962104

[9] Hugh, S.E. & Brändle, M.P. (2017). The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics, 113*(3), 1551-1571. DOI: 10.1007/s11192-017-2535-3

[10] Jaeger, M., Lippi, M., Passerini, A. & Frasconi, P. (2013). Type extension trees for feature construction and learning in relational domains. *Artifical Intelligence, 204*, 30-55. DOI: 10.1016/j.artint.2013.08.002

[11] Malesios, C.C. & Psarakis, S. (2014). Comparison of the h-index for Different Fields of Research Using Bootstrap Methodology. *Quality and Quantity, 48*(1), 521-545. DOI: 10.1007/s11135-012-9785-1

[12] Mazloumian, A. (2012). Predicting scholars scientific impact. *PLoS ONE, 7*(11), e49246. DOI: 10.1371/journal.pone.0049246

[13] McNamara, D., Wong, P., Christen, P. & Ng, K.S. (2013). Predicting High Impact Academic Papers Using Citation Network Features. In Li J. et al (Eds.), *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2013: Trends and Applications in Knowledge Discovery and Data Mining* (pp. 14-25). Berlin: Springer. DOI: 10.1007/978-3-642-40319-4_2

[14] Paszcza, B. (2016). *Comparison of Microsoft Academic (Graph) with Web of Science, Scopus and Google Scholar* (master's thesis). DOI: DOI: 10.13140/RG.2.2.21858.94405

[15] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research*(12), 2825-2830.

[16] Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the predictability of future impact in science. *Scientific Reports* 3, 3052. DOI: 10.1038/s-rep03052

[17] Ringelhan, S., Wollersheim, J. & Welpe, I.M. (2015). I Like, I Cite? Do Facebook Likes Predict the Impact of Scientific Work? *PLoS One, 10*(8), e0134389. DOI: 10.1371/journal.pone.0134389

[18] Stegehuis, C., Litvak, N. & Waltman, L.R. (2015). Predicting the long-term citation impact of recent publications. *Journal of Infometrics 9*(3), 642-657. DOI: 10.1016/j.joi.2015.06.005

[19] Thelwall, M. (2017). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics, 11*(4), 1201-1212. DOI: 10.1016/j.joi.2017.10.006

[20] Thelwall, M., Priem, J., & Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of medical Internet research, 13*(4), e123. DOI: 10.2196/jmir.2041

[21] Thornton, C., Hutter, F., Hoos, H. & Leyton-Brown, K. (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In Ghani, R. et al (Eds.), *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-855). New York, NY: ACM. DOI: 10.1145/2487575.2487629

[22] Wang, D., Song, C. & Barabasi, A. (2013). Quantifying Long-Term Scientific Impact. *Science, 342*(6154), 127-132. DOI: 10.1126/science.1237825

[23] Weihs, L. & Etzioni, O. (2017). *Learning to Predict Citation-Based Impact Measures*. Paper presented at 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada. DOI: 10.1109/JCDL.2017.7991559

[24] Wikipedia contributors. (2018, May 28). Bibliometrics. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:57, June 14, 2018, from `https://en.wikipedia.org/w/index.php?title=Bibliometrics&oldid=843318933`

[25] Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: learning to estimate future citations for literature. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1247-1252. DOI: 10.1145/2063576.2063757