gesis Leibniz Institute for the Social Sciences

hhu Heinrich Heine Universität Düsseldorf

How can I enhance my h-index?

Using ML prediction models to discover the influential factors

Introduction:

Assessing future impact is more pivotal for young researchers than seniors because they have smaller amounts of publications and received citations which are the bases for measuring the h-index. Therefore, we require other features to evaluate these researchers or discover the influential factors on their scientific outcomes.

Here are the contributions of this study:

- Defining novel feature sets
- Presenting H-index prediction models for researchers in different career phases
- Examining the temporal extent of the prediction power in the future for different feature categories
- Feature analysis to discover the effect of each feature on the scientific impact

Methodology:

Dataset: Scopus, the bibliographic database containing citations for academic journal articles.

Feature definitions:

The list of features employed to predict the author's h-index

type of feature	name	description
demographic	career_age	years since first publication
	gender	zero for females and one for males
	mobility_score	number of changing the affiliation at the country level
	GPD_current_country	GPD per capita of current affiliation country
paper/venue	primary_author_proportion	proportion of papers being as primary author
	open_access_proportion	proportion of open access papers among all papers
	main_field	the scientific field with the highest amount of publications.
	high_quality_proportion	proportion of publications in high quality journals among all papers
	field_mobility	number of unique disciplines authors has published paper divided to the number of all papers
co-author	max_h-index	maximum h-index of co-authors among all papers
	coauthor_per_paper	number of unique co-authors among all publications divided to the number of all papers
	international_coauthors	number of international coauthors among all papers

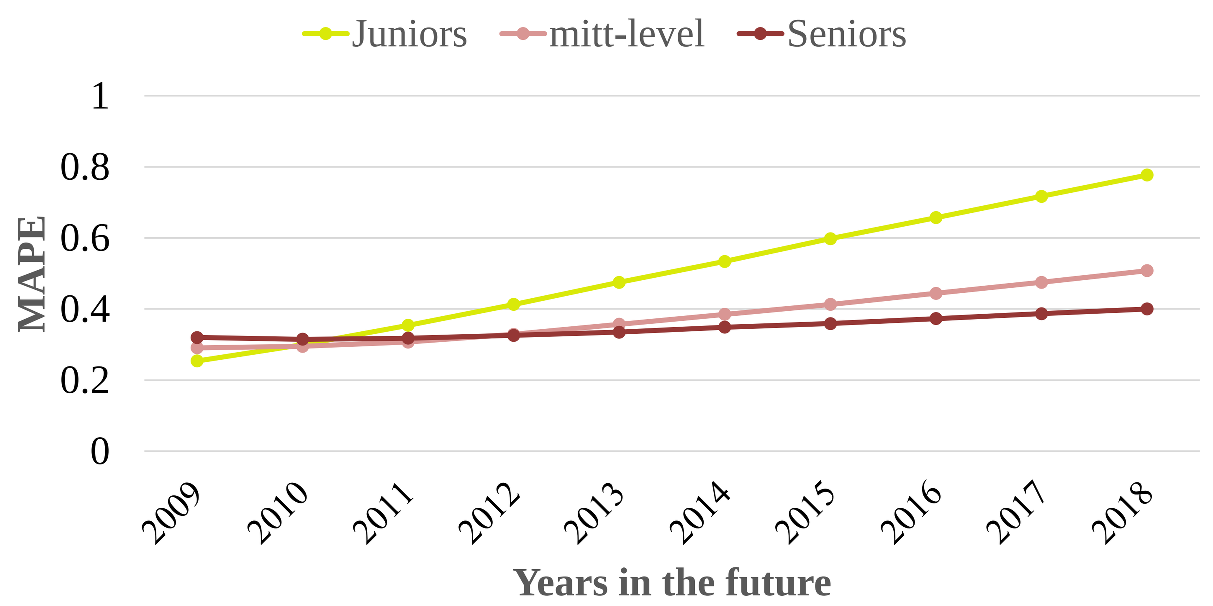
We extracted the values of the features from the first author's publication year till 2008 and predicted the h-index in the next ten years (from 2009 to 2018).

Method: We tackled the prediction task as a regression problem and employed the machine learning approach, XGBoot, to predict the h-index.

Results:

Pearson correlation coefficient between future h-index and features (2009 is the first prediction year)

	h-index in year		
Feature	2009	2014	2018
career_age	0.46	0.39	0.36
gender	0.07	0.06	0.06
mobility_score	0.48	0.48	0.47
GPD_current_country	0.17	0.15	0.13
primary_author_proportion	0.01	0.03	0.04
open_access_proportion	0.09	0.10	0.09
high_quality_papers_proportion	0.85	0.83	0.80
field_mobility	-0.46	-0.48	-0.47
max_coauthor_h-index	0.58	0.57	0.55
coauthor_per_paper	-0.02	-0.003	0.002
international_coauthors	0.23	0.28	0.30



Comparison the prediction performance (MAPE) between three researchers' groups over ten next years

Performance (MAPE) and permutation importance in predicting h-index for three prediction years implemented for three groups of researchers, juniors (maximum 5 years career age), mid-level (career age between 6 and 10 years) and seniors (career age more than 10 years)

career stage	junior			mid-level			senior		
prediction year	2009	2014	2018	2009	2014	2018	2009	2014	2018
feature:									
career_age	0.003	0.006	0.009	0.003	0.0007	0.002	0.006	-0.0003	-0.003
gender	0.0004	0.0004	0.0002	0.001	0.0014	0.002	0.0004	0.0009	0.0001
mobility_score	0.001	0.009	0.017	0.004	0.017	0.027	0.018	0.026	0.033
GPD_current_country	0.012	0.007	0.005	0.014	0.005	0.002	0.014	0.006	0.003
primary_author_proportion	0.077	0.24	0.32	0.11	0.23	0.32	0.08	0.14	0.15
open_access_proportion	0.075	0.163	0.19	0.12	0.267	0.28	0.1	0.177	0.25
main_field	0.021	0.046	0.028	0.027	0.033	0.036	0.027	0.066	0.026
high_quality_papers_proportion	0.078	0.111	0.137	0.124	0.139	0.195	0.161	0.198	0.227
field_mobility	0.219	0.37	0.468	0.389	0.557	0.694	0.454	0.617	0.692
max_coauthor_h-index	0.024	0.037	0.039	0.05	0.044	0.042	0.102	0.109	0.085
coauthor_per_paper	0.06	0.077	0.09	0.07	0.091	0.1	0.055	0.055	0.055
international_coauthors	0.0155	0.0238	0.0355	0.0322	0.0693	0.0863	0.0831	0.1554	0.188
MAPE of the model	0.25	0.53	0.78	0.29	0.38	0.51	0.32	0.35	0.37
size of the sample	584,812			1,063,600			1,354,233		

Conclusion:

- The prediction model with the defined feature set has a better performance for juniors than other researchers in the short term.
- Predicting power for seniors is more stable in the long term.
- Paper-specific features have the most effect, and authors' demographic characteristics minorly influence the scientific impact.
- We still need more features (e.g., textual content of papers, topic authority) to present a prediction model with acceptable performance, especially for young researchers.

Studied by Fakhri Momeni*
Supervisor: Prof. Dr. Stefan Dietze*
* GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667
Email: fakhri.momeni@gesis.org

Member of the Leibniz Leibniz Association