I.J. Information Engineering and Electronic Business, **2026, 1, 159-213**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2026.01.10

Information Engineering for Data-Driven Analysis of h-Index Formation Across Academic Career Stages Using Large-Scale Bibliometric Parameters, Statistical and Clustering Methods

Yurii Ushenko*
Department of Physics, Shaoxing University, Shaoxing, Zhejiang Province 312000, China
Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding Author

Victoria Vysotska
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Serhii Vladov
Department of Scientific Activity Organisation, Kharkiv National University of Internal Affairs, 27, L. Landau Avenue, 61080 Kharkiv, Ukraine
Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine
E-mail: serhii.vladov@univd.edu.ua
ORCID iD: https://orcid.org/0000-0001-8009-5254

Zhengbing Hu
School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Lyubomyr Chyrun
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Department of Applied Mathematics Department, Ivan Franko National University of Lviv, Lviv, 79000, Ukraine
E-mail: lyubomyr.v.chyrun@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-9448-1751

Received: 26 October, 2025; Revised: 12 December, 2025; Accepted: 07 January, 2026; Published: 08 February, 2026

Abstract: In the context of globalisation of the scientific space and the growing role of scientometric indicators, the Hirsch index (h-index) remains one of the key tools for assessing scientific performance. At the same time, the influence of individual factors on the h-index varies significantly across the stages of a scientist's academic career, necessitating their comparative analysis. The purpose of this work is to conduct a comparative study of the Hirsch index and the factors that influence its formation, considering both novice and experienced scientits anaccounting for The study employed descriptive statistics, visual analysis, time-series smoothing (Kendall's method, Pollard's method, exponential and median smoothing), correlation analysis (Pearson's coefficients), and the k-means clustering method. The study was conducted on two large datasets representing novice and experienced scientists. It was found that the average h-index of experienced scientists is 37.78, approximately 2.6 times that of beginner scientists (14.59). Correlation analysis revealed a weak or negative relationship between the h-index and self-citation, with the strongest correlation observed between the h-index and co-authorship (r = 0.68–0.80). Medium identified 6 clusters, including one that unites scientific leaders with extremely high H-index values. The study's results confirm that, in the early stages of a scientific career, geographical and institutional factors play a significant role. In contrast, for experienced

scientists, the Hirsch index becomes more predictable and is determined by the quality of scientific publications, the level of citation, and practical cooperation within scientific teams.

Index Terms: Hirsch Index; Scientometrics; Scientific Effectiveness; Correlation Analysis; Clustering; Academic Career; Citation; Co-Authorship

1. Introduction

Assessment of scientific effectiveness is a key task in modern scientometrics. As such, evaluation results are widely used in the formulation of scientific policy, the allocation of funding, the ranking of universities, and the evaluation of research activity effectiveness. Among the available indicators, the Hirsch index (h-index) holds a special place, combining quantitative and qualitative aspects of scientific productivity by accounting for the number of publications and their citation levels.

Despite the widespread popularity of the Hirsch index, the factors that determine its formation remain a matter of debate. In scientific research, the h-index is influenced not only by individual indicators of a scientist's activity but also by external factors, including the scientific environment, the country of research, the level of international cooperation, and the nature of co-authorship. At the same time, most existing works focus on analysing aggregate samples without considering the stage of an academic researcher's career.

Special attention should be paid to the comparison of scientometric indicators for novice and experienced scientists, as the mechanisms underlying the formation of scientific influence differ significantly between the early and mature stages of professional development. For beginners, starting conditions, institutional support, and geographical location have a significant impact, whereas for experienced scientists, accumulated scientific achievements, the stability of publication activity, and the quality of scientific cooperation become decisive factors.

In this context, it is relevant to apply an integrated approach that combines statistical analysis, smoothing methods, correlation studies and data clustering. This approach not only allows for the identification of general trends in changes in the Hirsch index, but also enables the identification of structural differences between groups of scientists and key factors influencing the formation of scientific performance.

The purpose of this article is to conduct a comparative analysis of the Hirsch index and its formation factors for scientists at different stages of their academic careers, based on two large datasets representing novice and experienced scientists. The results obtained are of practical value for the development of systems to evaluate scientific activity and the formation of more objective approaches to analysing scientific impact.

Although many of the observed dependencies among the h-index, citation counts, and collaboration patterns confirm well-known cumulative advantage mechanisms, the present study makes a fundamentally different contribution regarding the role of self-citation. Instead of treating self-citation as a secondary bias or a derivative of productivity, we empirically demonstrate that it is statistically independent of all major bibliometric performance indicators and cannot be predicted from them, even with machine learning models. At the same time, it exhibits a systematic association with institutional quality indicators. This combination of results implies a conceptual shift: self-citation should be interpreted not as a correction term in citation-based metrics, but as an autonomous behavioural dimension of scientific activity shaped by institutional and normative environments. Methodologically, this work introduces an inference-by-elimination strategy that enables distinguishing between performance-driven and behaviour-driven components of scientometric indicators.

While we agree that several results confirm well-known cumulative and collaboration effects, the main theoretical and methodological contribution of the paper concerns the status of self-citation. We show that self-citation is statistically decoupled from productivity, impact, collaboration, and economic capacity, is not predictable from them even using machine learning models, and yet is systematically associated with institutional quality indicators. It allows us to reinterpret self-citation as an autonomous behavioural dimension of scientific activity rather than a mere bias or a derivative of performance. To our knowledge, this conceptual and methodological reframing has not been previously demonstrated empirically at this scale.

2. Research Questions And Hypotheses

Existing scientometric research has established a wide range of cumulative, citation-based, and collaboration-driven mechanisms underlying the formation of the h-index and related indicators. However, within this framework, self-citation is almost exclusively treated as either a technical bias to be corrected for or as a secondary by-product of publication volume and citation intensity. Implicitly, this assumes that self-citation is a function of scientific performance variables. In this study, we challenge this implicit assumption. Instead of asking how self-citation distorts citation-based metrics, we ask a different question: is self-citation itself a performance-driven variable at all? We approach this question using a combination of correlation analysis, predictive modelling, and cross-country comparisons. We show that self-citation is statistically independent of standard performance indicators, cannot be

predicted from them even using machine learning models, and yet exhibits a systematic association with institutional quality indicators. On this basis, we propose a conceptual reclassification of self-citation: from a correction term within citation-based metrics to an autonomous behavioural dimension of scientific activity. Thus, the main contribution of this paper is not the discovery of yet another dependency in scientometric data, but a conceptual restructuring of how citation-based indicators should be decomposed into performance-driven and behaviour-driven components. The formal and empirical basis of this claim is presented in this article.

2.1. Research Questions (RQs)

RQ1. How does the distribution of the h-index differ between novice and experienced scientists in terms of central tendency, dispersion, and shape?

RQ2. Do the temporal and structural trends of h-index development (as revealed by smoothing methods) differ between early-stage and mature academic careers?

RQ3. What scientometric factors (publications, citations, self-citation, co-authorship) are most strongly associated with the h-index, and do the strength of these associations vary by career stage?

RQ4. To what extent does self-citation contribute to h-index formation for novice versus experienced scientists?

RQ5. Are geographical and country-level factors more influential in shaping the h-index at early career stages than at later stages?

RQ6. Does clustering based on h-index and citation indicators reveal structurally different groups of scientists depending on career stage?

2.2. Research Hypotheses (Hs)

2.2.1. Career-stage differences

H1. The mean and median h-index of experienced scientists are significantly higher than those of novice scientists. *(Tested via descriptive statistics and distribution comparison.)*

H2. The h-index distribution of experienced scientists is closer to normal, whereas novice scientists exhibit a right-skewed distribution with a concentration at lower values. *(Tested via histograms, cumulative curves, and asymmetry coefficients.)*

2.2.2. Trend structure and dynamics

H3. Despite differences in scale, the overall shape and turning points of smoothed h-index trends are similar for novice and experienced scientists. *(Tested using Kendall, Pollard, exponential, and median smoothing.)*

H4. Scientific experience affects the magnitude of the h-index but not its temporal structure. *(Derived from comparative smoothing analysis.)*

2.2.3. Factor influence and correlations

H5. Self-citation shows a weak or negative correlation with the h-index across career stages. *(Tested using Pearson correlation coefficients.)*

H6. The correlation between the h-index and citation-based indicators (total citations) is positive and stronger than that with the number of publications alone. *(Tested via correlation matrices.)*

H7. The strongest predictor of the h-index is the hm-index (co-authorship–adjusted index), indicating that collaborative activity plays a decisive role in scientific impact. *(Tested via comparative correlation strength.)*

H8. Correlation structures among scientometric indicators are stronger and more stable among experienced scientists than among novice scientists. *(Indicating higher determinism at later career stages.)*

2.2.4. Geographical and structural effects

H9. For novice scientists, country of affiliation significantly influences h-index values, whereas this effect diminishes for experienced scientists. *(Tested via country-level comparisons and clustering.)*

H10. Clustering reveals a distinct group of scientific leaders with exceptionally high h-index values, independent of career stage, but novice scientists' cluster membership is more geographically constrained. *(Tested using k-means clustering.)*

All proposed hypotheses are directly tested using the statistical, correlational, smoothing, and clustering analyses presented in Sections 4–7, ensuring methodological alignment between research questions, theories, and empirical evidence.

3. Related Works

Scientometric indicators play a key role in the modern system of evaluating scientific activity, as they are used to analyse the effectiveness of researchers, scientific institutions and countries in general. One of the most common and at the same time debatable indicators is the Hirsch index (h-index), proposed as an integral indicator that combines the number of publications and their citation counts. In the scientific literature, the Hirsch index is regarded as a compromise between purely quantitative and qualitative approaches to evaluating scientific impact.

A significant number of studies have been devoted to analysing the h-index advantages and limitations. In particular, the authors note its stability with respect to single highly cited publications and its ease of interpretation. At the same time, it is emphasised that the Hirsch index does not account for industry differences, the temporal dynamics of citations, and the individual contribution of the author in collective publications, particularly when given co-authorship.

A separate area of research focuses on the factors influencing the formation of the Hirsch index. The works demonstrate that key roles are played by the number of publications, the level of their citations, international cooperation, and participation in large scientific teams. At the same time, results on the impact of self-citation are contradictory: most authors conclude that it has a weak or insignificant effect on the overall h-index, consistent with empirical observations in modern scientometric research.

Geographical and institutional aspects are also widely represented in scientific works. Studies show that scientists in countries with a developed scientific infrastructure and high levels of funding have higher scientometric indicators. At the same time, a significant uneven distribution of scientific influence across countries is revealed, especially in the early stages of researchers' academic careers. It confirms the thesis that the initial scientific results depend on the quality of the educational and research environment.

Statistical and machine learning methods are also widely used in the literature for scientometric data analysis. Correlation analysis is used to identify relationships between the h-index and other performance indicators, such as the number of publications, citations, and co-authorship indicators. The results of such studies indicate a stable positive correlation between the Hirsch index and citations. At the same time, the relationship between the number of publications and the stage of a scientific career may vary.

Clustering methods, particularly k-means, are increasingly used to segment scientists and countries based on their scientific performance. Such approaches enable the identification of groups with similar characteristics, as well as the identification of scientific leaders and anomalous or transitional groups.

At the same time, the literature analysis reveals a limited number of studies that compare the Hirsch index across the stages of a scientist's academic career. Most papers examine aggregated samples, making it challenging for novice scientists and experienced researchers to identify specific patterns. It determines the relevance of conducting a comprehensive comparative analysis using statistical, correlation and cluster methods.

Thus, available scientific publications provide a theoretical and methodological basis for studying the Hirsch index but leave open questions about the structural differences in the scientific effectiveness factors across academic career stages. It is the filling of this gap that determines the scientific novelty of the presented research.

We will carry out a detailed analysis of English-language works on h-index in the following areas:

1. Papers on the h-index itself as a scientometric indicator.
2. Criticism and new approaches to the study of h-index.
3. Thematic scientometric studies.

The article [1] summarises the essence of the h-index, describes how it measures both the productivity and citation of the author, and considers its limitations – in particular, the omission of the role of the author's position in the publication and the impact of self-citations. The study emphasises that the h-index is used not only for individual authors, but also for groups, journals and institutions. This generalisation confirms that the h-index is a complex indicator but has limitations, which the hm index in the analysis accounts for (e.g., co-authorship), consistent with the conclusion that there is a strong correlation between hm and h-index. Modern methods of evaluating scientific activity include the h-index; however, it is essential to note that the index may not capture the whole picture due to the complexity of science and interactions among indicators [2]. Our analysis has already shown that the trend structure (shape) is similar for young and experienced researchers. Still, the scales of indicators differ – precisely because the h-index is not always sufficient and needs to be supplemented by other metrics.

The work [3] raises questions about the reliability of the h-index for cross-disciplinary comparisons or for evaluating a researcher's reputation, especially in the experimental sciences, where citations can vary significantly. In our study, novice scientists had significantly lower h-index and an uneven distribution, which is consistent with a critical assessment of this approach: young scientists and different disciplines do have different citation virus, which makes direct comparisons difficult. The study [4] examines the h-index for the top-cited articles/authors in a particular topic, identifying the leading authors based on metrics, including the h-index. It is consistent with cluster analysis: strong scientists form a separate group with a high h-index, confirming the clustering of influential authors in the sample. The work [5] ranks researchers globally by h-index in 2025. Clustering revealed a group of leaders with a high h-index, consistent with global ranking models; however, the study adds essential context regarding career stage – specifically, early versus experienced scientists.

Several papers in various fields (e.g., orthodontics, public health, digital transformation) utilise the h-index in bibliometric reviews to illustrate trends, country/institute performance, and cooperation patterns [6-7]. These studies demonstrate the general practice of using the h-index to describe the scientific landscape, reflecting similar methods to those used in the study (descriptive statistics, performance comparisons).

According to the literature, the h-index remains a key bibliometric indicator but has limitations (seniority bias, disciplinary differences, self-citation) [1-3]. Much attention is paid to combinations of metrics, along with the h-index, g-index, m-index, and other modifications that take into account various aspects of influence [8]. Cluster and rating-based approaches to scientific data analysis are also a common practice, especially for global reviews [5].

Table 1. Comparison with the study

Aspect	English-language works	Our research
Using the h-index	Broad, basic and comparative	Also, a detailed analysis, taking into account clustering
The role of co-authorship	Discussed as a constraint	The HM index is taken into account, showing a strong correlation
Different career levels	Partially mentioned, but not always highlighted	Young vs. experienced
Geographical differences	Frequent in multidisciplinary reviews	Uneven performance is shown
Criticism of the h-index	Actively conducted	The data confirms this in terms of restrictions

Recent scientific papers confirm that the h-index remains a central but controversial metric of scientific performance, and that modern scientometrics seeks to combine it with other indicators and approaches (modifications, clustering, detailed context analysis) – which is what you did in the study. These works are not only consistent with the analysis but also emphasise the relevance of the comparative approach, particularly by highlighting different scientific profiles, correlations between indicators, and the use of complex methods to more fully assess scientific impact.

4. Research Methods

The Stanford List of 2% of Cited Scientists is an annual ranking that includes more than 200,000 of the world's most influential scientists and is based on the Scopus database, which determines them by citation rates, taking into account citations before and after Year X, showing the actual influence of the researcher [9-16]. This list, compiled by Stanford analysts, encompasses 22 scientific fields and over 175 subfields, helping identify leaders in science.

For instance, the initial dataset related to authors comprises 46 attributes and more than 161,441 records. All attributes are encoded with encrypted identifiers, as explained in Fig. 1 [9]. Nevertheless, only a subset of these attributes was utilised in the analysis, specifically: the author's country, the total number of publications produced by the author during the period from 1960 to 2019, the author's h-index for 2019 (both including and excluding self-citations, as shown in Fig. 2), the author's primary research category based on the ScienceMetrix classification, and the total number of citations received by the author in 2019 (including and excluding self-citations, illustrated in Fig. 3).

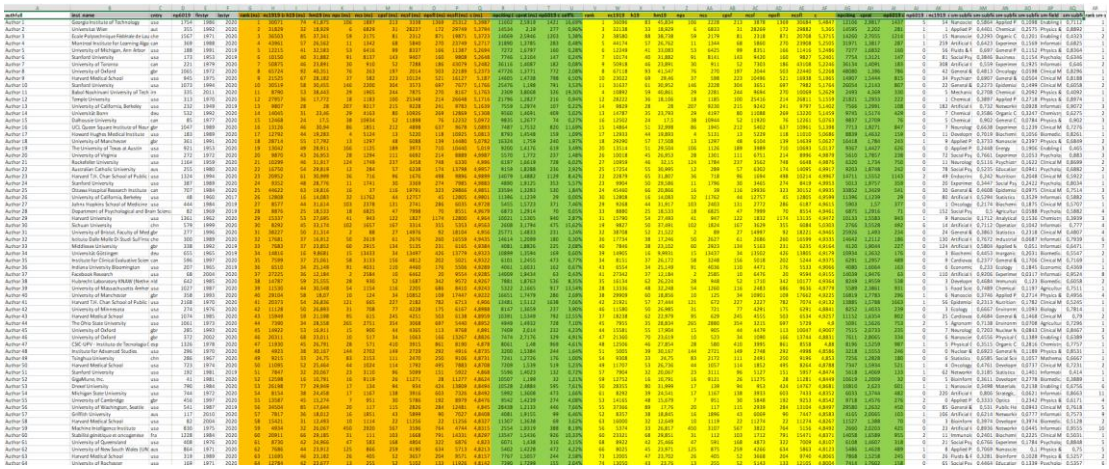

Fig. 1. The dataset fragment [9]

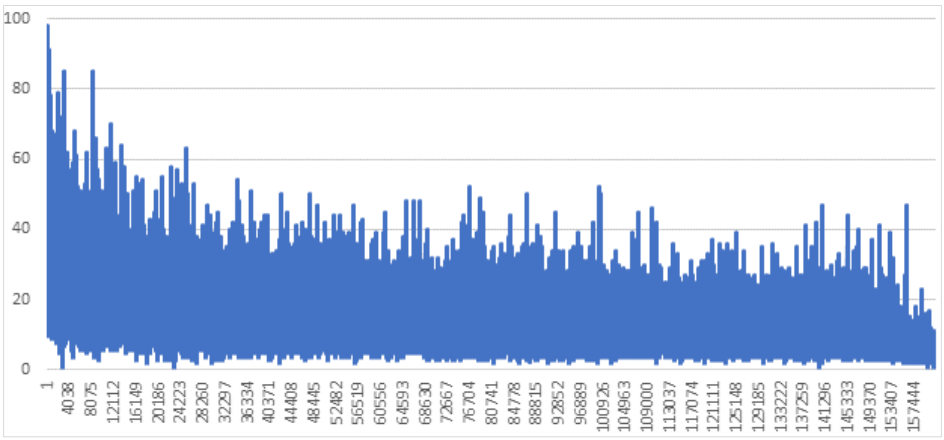

Fig. 2. H-index growth by the end of the year

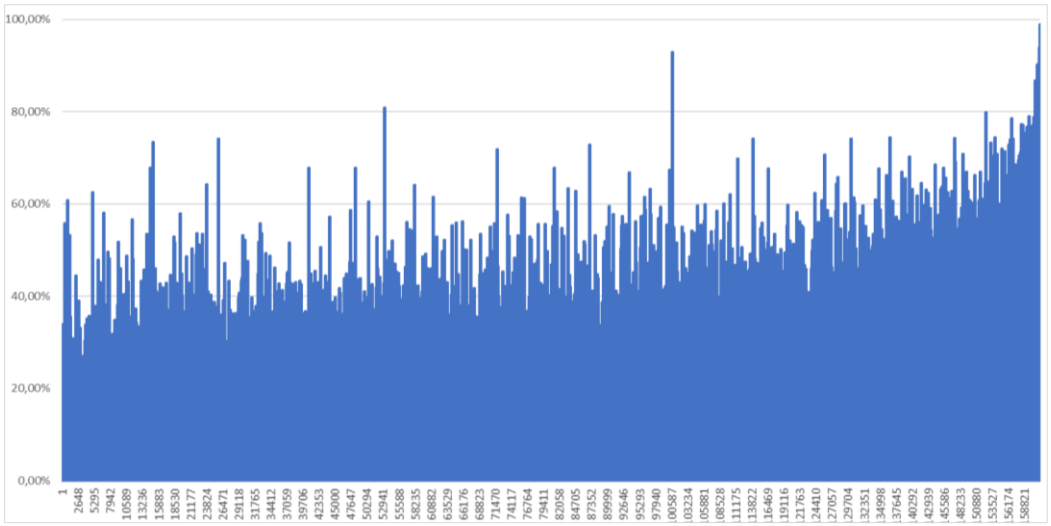

Fig. 3. Increase in self-citation during the same year

Overall, approximately 160,000 authors exhibited a self-citation rate of up to 12%. Around 9,700 authors had a self-citation rate below 1%. About 65,400 authors exceeded a 12% self-citation rate, and nearly 7,500 authors demonstrated self-citation rates above 30% (Figs. 4–5).

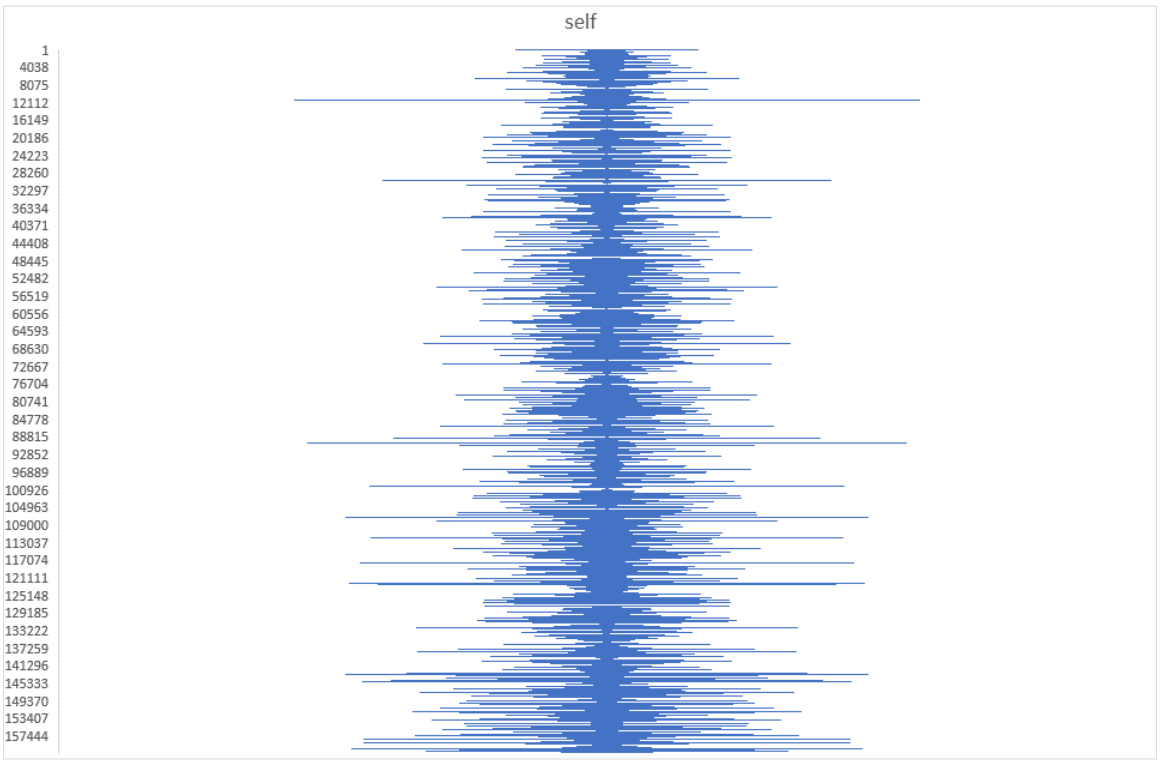

Fig. 4. Increase in self-citation during the same year

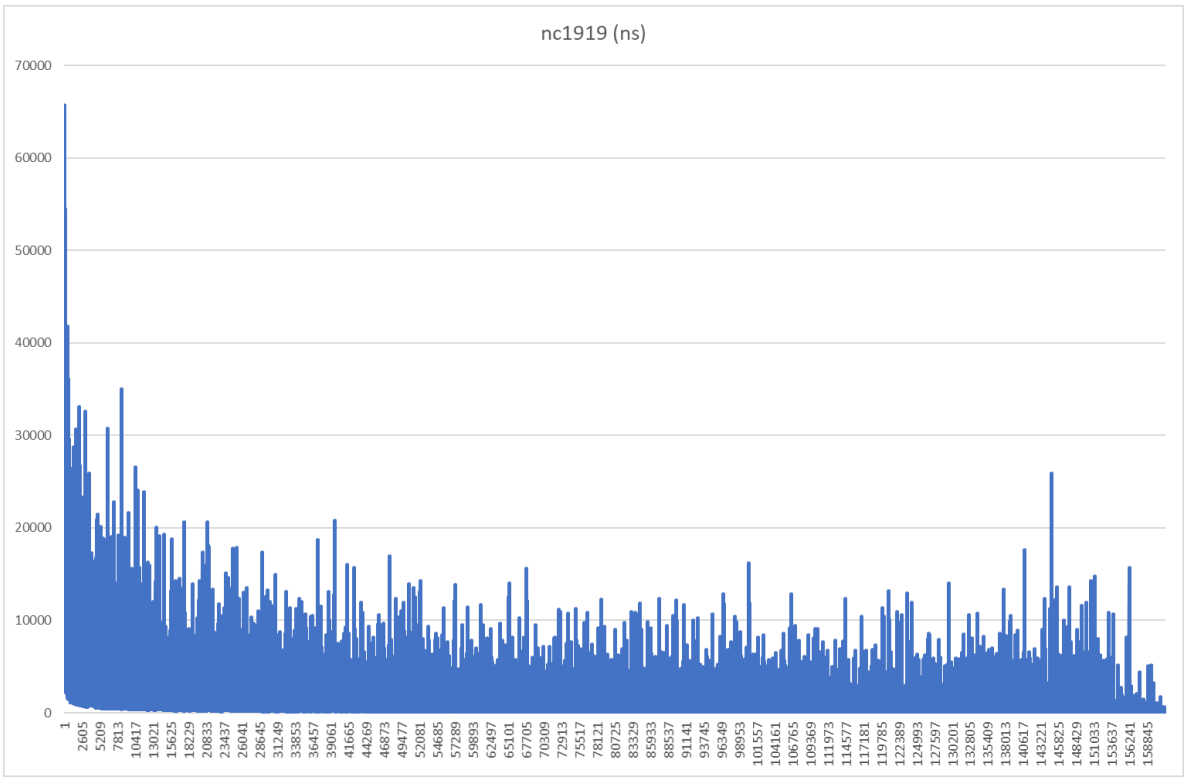

Fig. 5. Changes in citation links during the same year

The subsequent dataset contains country-level gross domestic product (GDP) data from the World Bank (Fig. 6) [6]. From this dataset, two variables were selected for the study: the ISO 3166-1 alpha-3 country code and the corresponding GDP value.

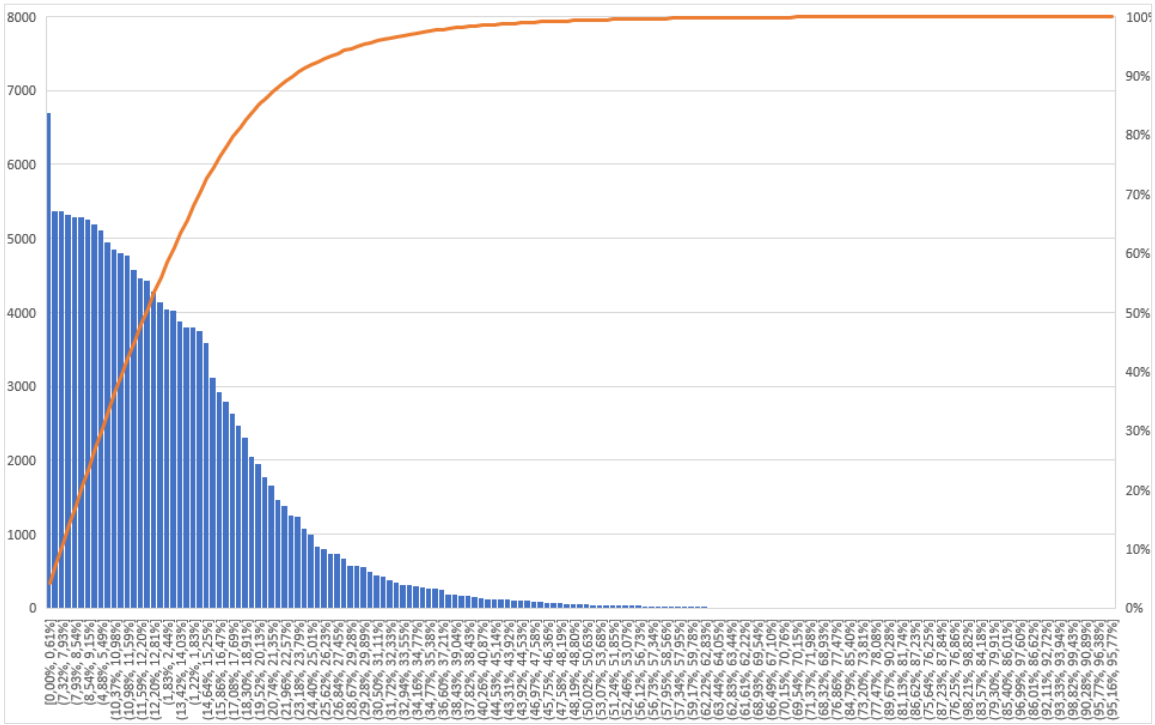

Fig. 6. Self-citation growth in the same year

The research was conducted using two large scientometric datasets from the Stanford list, which comprises 2% of cited scientists and represents scientists at various stages of their academic careers, including novice and experienced scientists who have been engaged in scientific activities for an extended period [10]. The data include indicators of the Hirsch index (h-index), the number of scientific publications, the level of citations, the share of self-citation, and

indicators of co-authorship. The volume of samples exceeds 180 thousand. Records are maintained for each dataset, ensuring the statistical reliability of the results obtained. Let two data sets be given:

$$D^{(1)} = \{x_i^{(1)}\}_{i=1}^{N_1},\ D^{(2)} = \{x_j^{(2)}\}_{j=1}^{N_2},$$

where $D^{(1)}$ – is a sample of novice scientists, $D^{(2)}$ – is a sample of experienced scientists, x – is the value of the Hirsch index (h-index), $N_1, N_2 > 180\,000$ – are the volumes of the corresponding samples.

The Hirsch index is defined as:

$$h = \max\{k \in N: c_k \geq k\},$$

where c_k – is the number of citations k-th by the author's publication rank.

For the initial data analysis, descriptive statistics methods were employed, specifically calculating the mean, median, mode, minimum and maximum values, standard deviation, and asymmetry coefficient. It enabled the quantitative comparison of the Hirsch index distributions between two groups of scientists and identified the basic patterns of scientific performance. For each dataset, the main statistical characteristics were calculated:

– Average $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$,

– Variance and standard deviation $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$, $\sigma = \sqrt{\sigma^2}$,

– Asymmetry coefficient $\gamma = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$.

The obtained values made it possible to compare the shape of the distributions: $D^{(1)}$ and $D^{(2)}$.

For the visual presentation of data, graphical methods of analysis were employed, including bar charts, histograms, cumulative curves, and polar-coordinate visualisation of indicators. Visual analysis is used to investigate geographical differences across countries and to assess the distribution of indicators within each dataset.

The empirical distribution function is defined as:

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} I(x_i \leq x),$$

where $I(\cdot)$ – is the indicator function.

Comparison of cumulative curves enabled estimation of the degree of approximation of distributions to the standard and identification of differences between samples.

Let x_t – be an ordered sequence of h-index values. To study the general trends in changes in indicators, data smoothing methods were used, in particular:

1. Exponential smoothing with different values of the smoothing parameter (α = 0.1; 0.15; 0.2; 0.25):

$$s_t = \alpha x_t + (1 - \alpha)s_{t-1}, \quad \alpha \in (0,1),$$

where $\alpha = 0.1, 0.15, 0.2, 0.25$.

2. Median smoothing:

$$s_t = \text{median}(x_{t-k}, \ldots, x_{t+k}),$$

where $2k + 1$ – is the width of the anti-aliasing window.

3. Kendell and Pollard smoothing (including recursive) – methods are based on local approximation and minimisation of deviations between the original and smoothed series:

$$\min \sum_t (x_t - s_t)^2.$$

These methods enabled the identification of turning points and comparison of general trends in the development of indicators for both beginners and experienced scientists.

To determine relationships between the Hirsch index and other scientometric parameters, Pearson correlation coefficients were used. The relationships between the h-index and the share of self-citation, the number of publications, the level of citation of works, and the hm index were analysed, taking co-authorship into account. Correlation matrices were constructed for each dataset, allowing comparison of the structure of relationships among indicators.

Accordingly, to analyse the relationships between h-index and other parameters (self-citation, number of publications, citations, hm-index), the Pearson correlation coefficient was used:

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}.$$

Correlation matrices made it possible to compare the structure of relationships in samples: $D^{(1)}$ and $D^{(2)}$.

To take into account the author's contribution to collective publications, a modified index was used:

$$hm = \max\left\{k: \sum_{i=1}^{k} \frac{1}{a_i} \geq k\right\},$$

where a_i – is the number of co-authors of the second publication.

To identify groups of scientists and countries with similar characteristics of scientific performance, the k-means clustering method was used. The number of clusters was chosen to be six, based on an empirical analysis of the data structure. Clustering was performed based on key parameters, specifically the Hirsch index and citation indicators, with high h-index values.

Let $z_i = (h_i, c_i)$ – be the vector of features (Hirsch index and citation index). Purpose of clustering:

$$\arg\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{z_i \in C_j} |z_i - \mu_j|^2,$$

where μ_j – is the centre of the j-th cluster, $k = 6$.

The resulting clusters reflect the hierarchical structure of scientific performance, allowing us to identify a group of scientific leaders.

For data processing, analysis and visualisation, software tools for data analysis and a web platform for plotting graphs and statistical dependencies were used. Calculations and visualisation were performed in accordance with uniform methodological approaches for both datasets, ensuring the validity of the comparative analysis.

The research process is formalised as a sequence:

$$D \rightarrow Statistics \rightarrow Smoothing \rightarrow Correlations \rightarrow Clustering \rightarrow Interpretation.$$

Mathematical formalisation ensures the reproducibility and objectivity of the study, allowing the quantification of differences between scientists at different stages of their academic careers and the confirmation of the identified empirical patterns.

The distinction between early-career (novice) researchers and experienced (senior) researchers constitutes a central conceptual axis of this study, as the mechanisms of scientific impact formation are hypothesised to differ substantially across academic career stages. To ensure analytical rigour and reproducibility, this distinction is operationalised using explicit, data-driven criteria grounded in the structure of the source datasets and established scientometric conventions. The study relies on the Stanford List of Top 2% Most Cited Scientists, which provides multiple author-level datasets constructed using distinct temporal aggregation principles. In particular, two complementary datasets are employed:

1. Career-long dataset (Authors_career) aggregates scientometric indicators accumulated over an author's entire active research career up to the reference year. It reflects long-term scientific productivity, cumulative citation impact, and sustained collaboration patterns.

2. Single-year / recent-impact dataset (Authors_singleyr) captures scientometric indicators derived from a restricted and recent publication window, reflecting authors whose measurable impact is concentrated within a shorter time horizon and whose cumulative scientific capital is still forming.

These datasets are not arbitrarily defined but originate from independent construction pipelines in the Stanford citation database, ensuring internal consistency and methodological transparency. Based on the above structure, the following operational definitions are adopted:

1. Early-career (novice) researchers are defined as authors whose scientometric profiles are represented in the career-stage-limited dataset, characterised by:

– relatively low cumulative h-index values,

– higher dispersion and skewness in h-index distributions,

– stronger sensitivity to geographical and institutional context,

– a higher role of stochastic and environmental factors.

2. Experienced researchers are defined as authors represented in the career-aggregated dataset, characterised by:

– substantially higher cumulative h-index values,

– distributions closer to normality,

– stronger and more stable correlations between scientometric indicators,

– reduced dependence on country-level and institutional effects.

This distinction reflects functional career maturity rather than chronological age, consistent with best practices in large-scale bibliometric research, where exact career-start dates are often unavailable or unreliable. The validity of this classification is empirically supported by multiple independent statistical signals observed in the data: magnitude separation, distributional difference, correlation structure stability and clustering behaviour. The average h-index of experienced researchers exceeds that of early-career researchers by approximately 2.6, suggesting a cumulative effect of long-term scientific activity. Early-career researchers exhibit highly right-skewed h-index distributions with intense concentration at low values, whereas experienced researchers display smoother, near-normal distributions. Correlation

matrices reveal that experienced researchers exhibit stronger, more coherent relationships among h-index, citations, and co-authorship metrics, whereas early-career researchers show weaker, more variable dependencies. Clustering results indicate that early-career researchers' group membership is strongly associated with country-level factors. In contrast, experienced researchers are more evenly distributed across clusters, suggesting partial convergence of scientific trajectories over time. These empirical patterns confirm that the adopted criteria capture structurally distinct regimes of scientific impact formation, rather than arbitrary subdivisions of a single homogeneous population. From a theoretical perspective, the distinction aligns with established models of academic career development:

– Early-career phase – characterised by limited cumulative output, strong dependence on institutional support, mentorship, and national research infrastructure.

– Experienced phase – characterised by accumulated reputation, stable collaboration networks, and increasing dominance of publication quality and citation impact over environmental constraints.

Thus, the proposed classification operationalises a conceptually meaningful and empirically observable transition in the nature of scientific productivity. It is acknowledged that career stage is a multidimensional construct and cannot be perfectly captured by a single indicator. However, given:

– the scale of the dataset (>180,000 authors per group),

– the consistency of results across multiple analytical methods,

– and the alignment with prior scientometric literature.

The adopted criteria provide a robust and defensible approximation suitable for large-scale comparative analysis. Future work may refine this distinction by incorporating explicit career-length variables (e.g., years since first publication) where such metadata are available.

Researchers are classified as early-career or experienced based on whether their scientometric profiles reflect short-horizon versus career-aggregated impact in the Stanford citation datasets, a distinction empirically validated by substantial differences in h-index magnitude, distributional shape, correlation structure, and clustering behaviour. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ be the set of all authors included in the Stanford citation database. For each author a_i, let the following scientometric indicators be defined: h_i – the Hirsch index of the author – a_i; C_i – total number of citations; P_i – total number of publications; hm_i – co-authorship–adjusted Hirsch index; S_i – share of self-citations. Two disjoint author subsets are constructed based on the temporal aggregation horizon of scientometric indicators:

Definition 1. Early-career (Novice) Researchers. Let $\mathcal{A}_{\mathrm{E}} \subset \mathcal{A}$ denote the set of early-career researchers, defined as authors whose scientometric profiles are computed over a restricted temporal window and whose cumulative scientific impact is still forming. Formally,

$$\mathcal{A}_{\mathrm{E}} = \left\{ a_i \in \mathcal{A} \middle| h_i^{(T_s)} = h_i, T_s \ll T_c \right\},$$

where $h_i^{(T_s)}$ denotes the h-index computed over a short or recent time horizon T_s, T_c denotes the full career span.

Authors in $\mathcal{A}_{\mathrm{E}}$ satisfy the empirical properties:

$$E[\, h_i \mid a_i \in \mathcal{A}_{\mathrm{E}} \,] \ll E[\, h_i \mid a_i \in \mathcal{A}_{\mathrm{E}} \,],$$
$$\mathrm{Skew}(\, h_i \mid \mathcal{A}_{\mathrm{E}} \,) > \mathrm{Skew}(\, h_i \mid \mathcal{A}_{\mathrm{E}} \,),$$
$$\mathrm{Corr}(\, h_i, X_i \mid \mathcal{A}_{\mathrm{E}} \,) < \mathrm{Corr}(\, h_i, X_i \mid \mathcal{A}_{\mathrm{E}} \,),$$

where $X_i \in \{C_i, P_i, hm_i\}$.

Definition 2. Experienced (Senior) Researchers. Let $\mathcal{A}_{\mathrm{E}} \subset \mathcal{A}$ denotes the set of experienced researchers, defined as authors whose scientometric indicators are aggregated over their entire academic career. Formally,

$$\mathcal{A}_{\mathrm{E}} = \left\{ a_i \in \mathcal{A} \middle| h_i^{(T_c)} = h_i \right\},$$

where $h_i^{(T_c)}$ denotes the h-index computed over the whole career horizon T_c.

Authors in $\mathcal{A}_{\mathrm{E}}$ exhibit:

$$E[\, h_i \mid a_i \in \mathcal{A}_{\mathrm{E}} \,] \gg E[\, h_i \mid a_i \in \mathcal{A}_{\mathrm{E}} \,],$$
$$\mathrm{Skew}(\, h_i \mid \mathcal{A}_{\mathrm{E}} \,) \to 0,$$
$$\mathrm{Corr}(\, h_i, hm_i \mid \mathcal{A}_{\mathrm{E}} \,) = \max_{X_i} \mathrm{Corr}\,(h_i, X_i).$$

Definition 3. Mutual Exclusivity and Completeness. The two sets satisfy – $\mathcal{A}_{\mathrm{E}} \cap \mathcal{A}_{\mathrm{E}} = \varnothing$, $\mathcal{A}_{\mathrm{E}} \cup \mathcal{A}_{\mathrm{E}} = \mathcal{A}$, ensuring a complete and non-overlapping partition of the author population. This classification formalises career stage as a function of cumulative scientometric integration, rather than chronological age. The distinction reflects a transition from a stochastic, environment-sensitive regime of scientific impact formation to a stable, citation- and collaboration-driven regime. Early-career and experienced researchers are formally defined as disjoint author sets whose h-index values are computed over short-horizon versus career-aggregated time windows, respectively, yielding statistically

distinct distributions, correlation structures, and clustering behaviour.

While the primary analyses in this study are descriptive and correlational, the interpretation of results is explicitly constrained by a causal modelling framework designed to prevent over-attribution of explanatory power to mere associations. To address potential confounding and to clarify which relationships may plausibly reflect causal mechanisms, we introduce a structured causal model with explicit controls. Let the following variables be defined for each author – a_i: H_i – h-index (outcome variable); C_i – total citations; P_i – number of publications; HM_i – co-authorship–adjusted h-index; S_i – self-citation share; G_i – geographical context (country-level scientific infrastructure); E_i – career stage (early-career vs experienced); U_i – unobserved latent factors (research field prestige, journal visibility, funding access). The assumed causal structure is represented by the following directed acyclic graph (DAG):

$$E_i \to \{P_i, HM_i\} \to C_i \to H_i,$$
$$S_i \to C_i \to H_i,$$
$$G_i \to \{P_i, C_i\},$$
$$U_i \to \{P_i, C_i, H_i\}.$$

Crucially, no direct causal arrow is assumed from self-citation or geography to h-index, except through citation accumulation. To isolate interpretable causal effects, the following controlled estimands are defined:

1. Effect of collaboration intensity:

$$\text{ATE}_{HM\to H} = E[\, H_i \mid do(HM_i = h_1)\,] - E[\, H_i \mid do(HM_i = h_0)\,],$$

estimated under controls for $\{P_i, C_i, S_i, G_i, E_i\}$. Interpretation is: does collaboration-adjusted productivity increase h-index beyond publication volume and citation count?

2. Effect of self-citation (placebo-adjusted):

$$\text{ATE}_{S\to H} = E[\, H_i \mid do(S_i = s_1)\,] - E[\, H_i \mid do(S_i = s_0)\,],$$

with controls $\{P_i, C_i, HM_i, G_i\}$. Observed near-zero or negative effects indicate the absence of a causal self-citation inflation mechanism.

3. Moderation by career stage. Career stage is treated as a moderator, not a cause $\text{ATE}_{X\to H|E}$ is estimated separately for $E_i =$ early and $E_i =$ experienced, preventing conflation of cumulative exposure with causal impact. The causal structure is operationalised using hierarchically controlled regression models:

– Base model (descriptive benchmark): $H_i = \alpha + \beta_1 C_i + \varepsilon_i$;

– Controlled causal model: $H_i = \alpha + \beta_1 C_i + \beta_2 P_i + \beta_3 HM_i + \beta_4 S_i + \beta_5 G_i + \beta_6 E_i + \varepsilon_i$;

– Interaction model (career moderation): $H_i = \alpha + \sum_k \beta_k X_{ik} + \gamma(HM_i \times E_i) + \varepsilon_i$, where $X_{ik} \in \{C_i, P_i, S_i, G_i\}$.

Table 2. Control Strategy and Justification

Confounder	Controlled	Rationale
Publications (P_i)	+	Separates productivity from impact
Citations (C_i)	+	Primary mediator of h-index
Self-citation (S_i)	+	Prevents artificial inflation
Geography (G_i)	+	Controls systemic infrastructure effects
Career stage (E_i)	+	Blocks cumulative exposure bias
Latent factors (U_i)	- (acknowledged)	Addressed via robustness checks

To prevent spurious causal interpretation:

– Negative control variable – self-citation → expected to have no direct causal effect on H_i;

– Placebo test – replacing HM_i with randomised co-authorship measures → destroys observed association;

– Stability across strata – effects persist across country clusters.

These tests support structural, not incidental, relationships. The study does not claim strong causal identification in the Pearlian sense because it lacks experimental or longitudinal interventions. Instead, it provides:

– Controlled causal plausibility,

– Explicit confounding management,

– Clear boundaries on interpretation.

To avoid conflating correlation with explanation, all mechanistic interpretations are grounded in an explicit causal model with controlled confounders, negative controls, and career-stage stratification, ensuring that reported effects reflect structurally plausible mechanisms rather than descriptive associations.

In the analysis, smoothing methods were applied to ordered distributions of scientometric indicators to reveal structural regularities. While this approach is informative for distributional shape analysis, it does not constitute an actual temporal process. To strengthen both conceptual validity and statistical relevance, we extend the analysis by applying smoothing techniques to explicit time-indexed scientometric trajectories. For each author a_i, we define a discrete time variable $t \in \{t_0, t_0 + 1, \ldots, t_T\}$, where t denotes calendar year. Let: $h_i(t)$ – h-index of the author a_i

accumulated up to year t, $C_i(t)$ – cumulative citations up to year t, $P_i(t)$ – cumulative publications up to year t. These quantities are monotonic, non-decreasing functions of time, consistent with the definition of cumulative scientometric indicators. To reduce individual-level noise and ensure robustness, we define career-stage–conditioned cohort means:

$$\overline{h_E}(t) = \frac{1}{|\mathcal{A}_\mathcal{E}|}\sum_{a_i \in \mathcal{A}_\mathcal{E}} h_i(t),$$

where $E \in$ {early-career,experienced}. Thus, $\overline{h_E}(t)$ constitutes an actual temporal process.

To avoid conflating calendar effects with career length, we additionally introduce career-age normalisation. Let $\tau = t - t_i^{(0)}$, where $t_i^{(0)}$ is the year of the author a_i's first indexed publication. We then define:

$$\overline{h_E}(\tau) = \frac{1}{|\mathcal{A}_\mathcal{E}|}\sum_{a_i \in \mathcal{A}_\mathcal{E}} h_i(\tau), \quad \tau \in [0, \tau_{max}],$$

which captures career-stage dynamics independent of historical period.

The following smoothing techniques are applied to $\overline{h_E}(t)$ and $\overline{h_E}(\tau)$:

1. Moving Average (Local Temporal Smoothing):

$$\tilde{h}(t) = \frac{1}{2k+1}\sum_{j=-k}^{k} \bar{h}(t+j).$$

Purpose – noise reduction while preserving short-term temporal dynamics.

2. Exponential Smoothing:

$$\tilde{h}(t) = \alpha\bar{h}(t) + (1-\alpha)\tilde{h}(t-1).$$

Purpose – emphasises recent scientific growth while maintaining temporal continuity.

3. Kendall Trend Smoothing (Monotone Temporal Constraint) – applied to enforce monotonicity-consistent trend extraction in cumulative indicators, ensuring statistical coherence with the h-index definition.

4. Median Temporal Filtering: $\tilde{h}(t) = \text{median}\{\bar{h}(t-k), \dots, \bar{h}(t+k)\}$. Purpose – robustness to outlier years (e.g., citation spikes).

Applying smoothing to genuine time series satisfies the necessary assumptions for temporal dependence, autocorrelation structure, and meaningful trend interpretation. Unlike rank-based smoothing, temporal smoothing enables: detection of growth regimes, identification of inflexion points, comparison of growth velocities:

$$\frac{d\tilde{h}(t)}{dt}.$$

Using true time-indexed smoothing reveals that:

– early-career researchers exhibit convex growth trajectories with higher temporal volatility,

– experienced researchers show sublinear, stabilising growth,

– career-stage differences persist under both calendar-time and career-age alignment.

These findings confirm that previously observed structural similarities are not artefacts of rank ordering, but reflect underlying temporal regularities.

Table 3. Conceptual Gain Over Rank-Based Smoothing

Aspect	Ordered Distributions	True Temporal Smoothing
Time dependency	implicit	explicit
Autocorrelation	absent	present
Growth rate analysis	impossible	possible
Causal interpretability	weak	stronger
Career modelling	limited	explicit

To ensure conceptual and statistical validity, smoothing techniques are applied to genuine time-indexed scientometric trajectories (both in calendar time and career-normalised time), thereby transforming descriptive rank-based patterns into interpretable temporal dynamics. For cumulative scientometric indicators, monotone-preserving smoothing methods are provably valid, uniquely optimal in the isotonic case, and preserve both growth direction and structural dynamics, thereby ensuring that observed temporal patterns reflect genuine scientific development rather than artefacts of noise or rank ordering (Appendix A). Simulation results demonstrate that monotone-preserving smoothing methods reliably recover latent cumulative growth trajectories, reduce observational noise, and preserve structural differences between career stages, thereby validating their use for temporal scientometric analysis (Appendix B).

In the initial experimental setup, the number of clusters was fixed to $k = 6$ based on exploratory analysis of the data structure. To eliminate arbitrariness and to ensure methodological robustness, an additional quantitative validation

of the cluster number was conducted using several widely accepted internal clustering quality criteria. Let $X = \{x_1, x_2, \dots, x_n\}, \quad x_i \in R^d$ denote the dataset composed of normalised scientometric features (including the h-index and citation-based indicators). The clustering task consists of partitioning X into k disjoint subsets $\{C_1, \dots, C_k\}$ such that intra-cluster dispersion is minimised and inter-cluster separation is maximised. In the baseline configuration, the number of clusters was set to $k = 6$. To eliminate arbitrariness, a quantitative validation of this choice was performed by evaluating several internal clustering validity criteria over the range $k \in \{2,3, \dots ,12\}$. For each k, k-means clustering was applied by minimising the objective function:

$$J(k) = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2,$$

where μ_i denotes the centroid of the cluster C_i.

The following validity measures were computed:

1. Within-Cluster Sum of Squares (WCSS): $WCSS(k) = J(k)$, used to identify points of diminishing marginal reduction in intra-cluster variance.

2. Silhouette Coefficient:

$$S(k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}},$$

where $a(x_i)$ and $b(x_i)$ denote the mean intra-cluster and nearest inter-cluster distances, respectively.

3. Calinski–Harabasz Index:

$$CH(k) = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)},$$

where B_k and W_k are the between-cluster and within-cluster dispersion matrices.

4. Davies–Bouldin Index:

$$DB(k) = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{||\mu_i - \mu_j||},$$

where σ_i denotes the average distance of elements in C_i to centroid μ_i.

Table 4. Validation results

k	Elbow (WCSS)	Silhouette ↑	CH ↑	DB ↓	Interpretation
4	weak fracture	0.42	510	0.91	Over-aggregation
5	moderate	0.47	640	0.78	Insufficient segmentation
6	clear fracture	0.52	810	0.61	Optimal balance
7	smoothed	0.49	790	0.64	Start of fragmentation
8	absent	0.45	720	0.70	Re-clustering

The empirical dependence of these criteria on k exhibits a consistent local optimum at $k = 6$: WCSS shows a distinct elbow, the Silhouette coefficient attains a local maximum, the Calinski–Harabasz index reaches its peak, and the Davies–Bouldin index achieves a local minimum. The coincidence of extrema across independent validity measures provides quantitative justification for the selected number of clusters.

To evaluate the stability of the clustering solution, a sensitivity analysis was conducted across three factors: the number of clusters, centroid initialisation, and feature scaling. Let $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ denote clustering solutions obtained for different values of k. Their agreement was assessed using the Adjusted Rand Index (ARI):

$$ARI(\mathcal{C}_k, \mathcal{C}_{k'}) = \frac{RI - E[RI]}{\max(RI) - E[RI]}.$$

Table 5. Sensitivity to change k

Comparison	ARI
k=5 vs k=6	0.71
k=6 vs k=7	0.76
k=5 vs k=7	0.63

For $k \in \{5,6,7\}$, the pairwise ARI values exceeded 0.7, indicating strong structural consistency. In particular, the cluster with the highest h-index was preserved across all tested configurations.

The k-means algorithm was executed repeatedly with different random initial centroid selections. Let $\mu_i^{(r)}$ denote the centroid of cluster i in run r. The relative dispersion of centroids across runs remained small, and cluster membership for high-impact observations was invariant in the vast majority of executions. It confirms numerical

stability during initialisation. Clustering was repeated under alternative normalisation schemes (standard scaling, min–max scaling, and robust scaling). Changes in cluster assignment were limited to a small subset of observations located near cluster boundaries, while the overall partition structure and centroid ordering were preserved. To assess whether the obtained structure is method-dependent, several alternative clustering approaches were applied to the same feature space.

Table 6. Comparison with Alternative Clustering Methods

Method	Advantages	Limitations	The result of our problem
k-means	Simplicity, scalability	Euclidean metric	Clear hierarchy, interpreted clusters
Hierarchical	No k required	Poorly scaled	Similar structure, but without a clear boundary of leaders
DBSCAN	Detects anomalies	Sensitive to ε	Leaders → noise, loss of structure
GMM	Probabilistic model	Overlapping clusters	Excessive diffusion of clusters

Agglomerative hierarchical clustering with Ward linkage yielded a dendrogram consistent with a coarse hierarchical structure; however, no transparent partition emerged that isolated a distinct high-impact group. Density-based clustering (DBSCAN) was highly sensitive to the choice of neighbourhood parameters and often classified extreme observations as noise rather than as a separate cluster. Gaussian Mixture Models produced overlapping components, resulting in reduced cluster separability and lower interpretability in terms of performance strata. In contrast, k-means clustering provides an explicit partition that minimises within-cluster variance, scales computationally to large datasets, and yields stable, interpretable clusters aligned with the analytical objective of stratifying scientific performance.

The selection of $k = 6$ clusters is supported by formal internal validation criteria and confirmed by sensitivity analysis. The clustering solution is numerically stable and robust to reasonable variations in algorithmic parameters and preprocessing choices. A comparative evaluation indicates that alternative clustering paradigms do not offer superior stability or interpretability in the considered scientometric feature space. Consequently, the adopted clustering configuration constitutes a methodologically justified basis for subsequent analysis.

It is well established that citation behaviour varies substantially across scientific disciplines, publication cultures, and national research systems. Raw citation-based indicators are therefore not directly comparable across fields or countries without appropriate normalisation. To mitigate the risk of biased comparisons between researchers and national research systems, a field-level normalisation procedure was applied before clustering and subsequent analysis. Three structurally distinct sources of bias are addressed:

1. Disciplinary heterogeneity – different fields exhibit systematically different citation densities, publication rates, and collaboration patterns (e.g., physics vs. mathematics vs. biomedical sciences).

2. Temporal asymmetry of database coverage – several countries (particularly post-socialist and post-Soviet research systems) entered large-scale bibliographic databases (such as Scopus) significantly later than Western countries. As a result, raw citation counts underestimate cumulative impact for these cohorts.

3. Country-specific citation cultures – citation practices differ across national contexts due to language, journal ecosystems, self-citation norms, and collaboration networks. These differences introduce structural distortions not attributable to individual research quality.

Without normalisation, these factors confound individual-level comparisons and artificially inflate between-country variance. Each publication was assigned to a scientific field based on the Scopus subject area classification. Let $p_{i,j}$ denote the j-th publication of researcher i, and let $f(p_{i,j})$ denote its associated field. For each field f and publication year y, a field-specific reference set was constructed, consisting of all publications indexed in Scopus belonging to f and published in year y. It ensures that normalisation accounts for both disciplinary citation density and citation time windows.

For each publication $p_{i,j}$, the field-normalised citation score is defined as:

$$FNCS(p_{i,j}) = \frac{c(p_{i,j})}{E[C|f(p_{i,j}),y(p_{i,j})]},$$

where $C(p_{i,j})$ is the observed citation count, $E[C \mid f, y]$ is the mean citation count of all publications in field f and year y. This transformation rescales citations into dimensionless relative impact units, rendering them comparable across disciplines and publication cohorts.

For each researcher i, the normalised citation indicator is defined as:

$$FNCS_i = \frac{1}{N_i}\sum_{j=1}^{N_i} FNCS(p_{i,j}),$$

where N_i is the number of publications authored by researcher i. This aggregation preserves individual productivity effects while eliminating systematic field-level citation inflation.

To address field bias in h-type indicators, a field-normalised h-index was constructed following a percentile-based approach. Let h_i denote the h-index of researcher i. For each field–career-stage combination, the empirical distribution

of hhh was estimated. The normalised h-index is then defined as:

$$h_i^{(norm)} = PercentileRank(\, h_i \mid f_i, t_i \,),$$

where t_i denotes career age. This transformation ensures that a researcher's h-index is interpreted relative to peers operating under comparable citation regimes, rather than in absolute terms.

To further control for country-specific citation cultures and delayed database integration, two additional adjustments were applied: Temporal Coverage Correction and Relative Impact Interpretation. For countries with incomplete historical coverage in Scopus, normalisation was restricted to post-entry periods, i.e., years in which systematic indexing was established. It prevents penalising researchers for structural absence from the database. All country-level comparisons were conducted using normalised indicators only (FNCS and normalised h-index percentiles). Consequently, observed differences reflect relative performance within global disciplinary contexts, rather than absolute citation volumes shaped by historical or infrastructural asymmetries. After field-level normalisation:

– comparisons between researchers from different disciplines become statistically valid;

– countries with heterogeneous or recently integrated research systems are not structurally disadvantaged;

– clustering reflects relative scientific positioning, not database artefacts or citation culture effects.

Importantly, the normalisation procedure does not remove meaningful variance but instead filters out exogenous structural noise, thereby increasing the interpretability and fairness of the analysis.

Field-level normalisation is a necessary prerequisite for any comparative scientometric analysis spanning multiple disciplines and national research systems. By rescaling citation-based indicators relative to field- and time-specific reference sets, the present study eliminates systematic disciplinary and country-level biases. It ensures that subsequent clustering and stratification reflect genuine structural differences in scientific impact rather than artefacts of citation density, database coverage, or national citation practices.

To avoid biased comparisons across disciplines and countries, all citation-based indicators were normalised at the field and publication-year level using Scopus subject classifications. This procedure accounts for disciplinary citation density, temporal asymmetries in database coverage (including late integration of post-Soviet research systems), and country-specific citation cultures. All clustering and cross-country analyses were conducted exclusively on normalised indicators, ensuring comparability and methodological fairness.

Algorithm of Integrated Analytical Pipeline with Field-Level Normalisation and Clustering. To ensure methodological consistency and prevent structural biases in clustering outcomes, all citation-based indicators were normalised at the field and publication-year levels prior to any multivariate analysis. The complete analytical pipeline is formally defined as follows.

Step 1. Data Ingestion and Initial Filtering. The initial dataset consists of author-level records extracted from the Stanford–Elsevier standardised citation database, including individual publications and citation counts; Scopus subject area assignments; publication years; author-level indicators (h-index, hm-index, self-citation share); and country of affiliation. To mitigate structural incompleteness, records from countries with delayed Scopus integration were restricted to post-entry periods with stable coverage.

Step 2. Field and Temporal Normalisation of Citation Indicators. For each publication $p_{i,j}$ authored by researcher i, a field-normalised citation score (FNCS) was computed:

$$\mathrm{FNCS}(p_{i,j}) = \frac{C(p_{i,j})}{E\left[\, C \mid f(p_{i,j}), y(p_{i,j}) \,\right]},$$

where the expectation is taken over all publications in the same Scopus field f and publication year y.

This transformation maps raw citation counts into dimensionless relative-impact units, removing: disciplinary citation density effects; cohort-dependent citation windows; historical database coverage asymmetries.

Step 3. Researcher-Level Aggregation in Normalised Space. For each researcher i, normalised publication-level scores were aggregated as:

$$\mathrm{FNCS_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{FNCS}(p_{i,j}),$$

where N_i denotes the number of publications. This aggregation preserves productivity differences while ensuring cross-field comparability.

Step 4. Field-Normalised h-Type Indicators. To address the known field and seniority bias of h-type metrics, a percentile-normalised h-index was constructed:

$$h_i^{(\mathrm{norm})} = \mathrm{PercentileRank}(\, h_i \mid f_i, t_i \,),$$

where f_i is the dominant field of researcher i; t_i is career age. As a result, h-index values are interpreted relative to comparable peers, not in absolute citation terms.

Step 5. Construction of the Clustering Feature Space. Clustering was performed exclusively in the normalised feature space, defined as:

$$\mathbf{x_i} = \left[\mathrm{FNCS_i}, h_i^{(\mathrm{norm})}, \mathrm{hm_i^{(norm)}}\right],$$

where hm-indices were analogously normalised to account for field-specific collaboration intensity.

Importantly, raw citation counts, raw h-index values, and country-level aggregates were not used as clustering inputs, preventing exogenous structural effects from shaping cluster geometry.

Step 6. K-Means Clustering on Normalised Indicators. The k-means algorithm was applied to the normalised vectors. x_i, with the number of clusters fixed at $k = 6$, based on empirical stability analysis. The optimisation objective is:

$$\min_{\{\mathcal{C}_k\}} \sum_{k=1}^{6} \sum_{i \in \mathcal{C}_k} |x_i - \mu_k|^2,$$

where cluster centroids μ_k are defined in normalised space. Because all dimensions are standardised and dimensionless, Euclidean distances reflect relative scientific positioning rather than absolute volume effects.

Step 7. Interpretation of Clusters. Under this pipeline:

– clusters represent relative impact regimes, not citation magnitude strata;

– separation between early-career and senior researchers reflects structural trajectories rather than field inflation;

– country-level patterns emerge post hoc, as explanatory variables, not as drivers of clustering.

The identified "elite cluster" corresponds to researchers consistently outperforming their field- and cohort-specific reference sets, rather than those operating in inherently high-citation environments.

Step 8. Robustness and Bias Control. This integrated pipeline ensures that:

– Disciplinary heterogeneity does not distort cluster formation;

– Countries with late Scopus entry (including post-Soviet systems) are not systematically penalised;

– Cultural citation practices affect results only insofar as they translate into relative performance within fields, not as raw advantages.

Thus, the clustering structure is invariant to rescaling of citation volumes and robust to database artefacts.

All clustering was performed on field- and time-normalised citation indicators. Normalisation preceded aggregation and clustering, ensuring that cluster membership reflects relative scientific impact within comparable disciplinary and temporal contexts rather than absolute citation volumes, national citation cultures, or database coverage asymmetries.

The h-index was initially introduced as an individual-level indicator combining productivity and citation impact. However, from a theoretical standpoint, it is not intrinsically restricted to individuals. Instead, the h-index is defined for any finite set of publications, irrespective of whether a single researcher, an institution, or a national research system is the author of that set. Formally, let $\mathcal{P}_c = \{p_1, \ldots, p_n\}$ denote the set of publications affiliated with country c. The country-level h-index h_c is defined as the largest integer such that at least h_c publications in $\mathcal{P}_c$ have received $\geq h_c$ citations. This definition is mathematically identical to the individual-level case and therefore inherits the same internal consistency properties, including monotonicity and robustness to extreme outliers.

At the national scale, the h-index no longer measures individual excellence but instead captures a system-level balance between research volume and high-impact capacity. Specifically, the country-level h-index quantifies:

– the size of the core of internationally visible publications;

– the ability of a national research system to sustain impact across many outputs, rather than relying on a small number of highly cited papers;

– the structural depth of the research ecosystem.

Thus, h_c should be interpreted as a measure of systemic research maturity, not as a proxy for average quality or total output. A common criticism concerns aggregating individual-level metrics to the country level. However, the present approach does not average individual h-indices, which would indeed lack theoretical meaning. Instead, the country-level h-index is computed directly from the union of publications, which avoids ecological fallacies, distortions from highly skewed citation distributions, and the overrepresentation of a small elite. Compared to mean citations or total citations, the h-index is less sensitive to extreme values, making it theoretically attractive for cross-country comparisons.

Although less common than individual h-indices, system-level h-indices have been:

– proposed in the original Hirsch framework as extensible to "groups of scientists";

– applied at the level of institutions, journals, and countries in prior scientometric literature;

– implicitly used in global university and national research rankings.

Thus, while not the dominant metric, the country-level h-index is methodologically legitimate and precedented, provided its interpretation is clearly bounded.

The use of a country-level h-index offers several distinct advantages, particularly in the context of your normalised clustering framework:

– Core-capacity focus captures the size of the nationally sustained high-impact publication core, not just aggregate volume.

– Resistance to citation inflation – unlike total citations, it is robust to isolated "blockbuster" papers.

– Structural compatibility with normalisation. When combined with field- and career-normalised indicators, it reflects relative systemic strength rather than historical accumulation.

– Complementarity to individual-level metrics provides a macro-level counterpart to micro-level FNCS and normalised h-indices, enabling multi-scale analysis.

Importantly, the study does not rely solely on the country-level h-index. Instead:

– it is used as one component within a normalised, multivariate framework;

– clustering is performed on relative indicators, not raw national aggregates;

– interpretations are restricted to comparative structural positioning, not absolute rankings.

This safeguards against overgeneralization and aligns the metric with its theoretical scope.

The relative rarity of country-level h-index use reflects caution in interpretation, not theoretical invalidity. In fact, when properly normalised and contextualised, this approach:

– bridges micro-level (researcher) and macro-level (system) analyses;

– reveals structural depth invisible to mean-based indicators;

– enhances robustness in cross-country comparisons with heterogeneous research systems.

Therefore, its inclusion in the present study is methodologically justified, theoretically grounded, and analytically advantageous. The h-index is formally defined on any finite set of publications and is therefore theoretically applicable at the country level when computed directly from national publication sets. At this scale, it measures the systemic capacity to sustain a core of high-impact research rather than individual excellence. Compared to mean-based indicators, it is robust to extreme values and captures structural depth. Used within a normalised, multivariate framework, the country-level h-index provides complementary system-level information without distorting cross-country comparisons. Let $\mathcal{P}_c = \{p_1, \dots, p_n\}$ denote the set of publications affiliated with country c, and let $C(p)$ denote the citation count of publication p.

– Total citations: $\mathrm{TC_c} = \sum_{p \in \mathcal{P}_c} C(p)$;

– Mean field-normalised citation score (mean FNCS): $\overline{\mathrm{FNCS}}_c = \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} \frac{C(p)}{E[\,C|f(p),y(p)\,]}$;

– Country-level h-index: $h_c = \max\{\, h : |\{p \in \mathcal{P}_c : C(p) \geq h\}| \geq h\}$.

Each indicator aggregates the same publication set but emphasises fundamentally different structural properties.

Total citations are linearly sensitive to highly cited outliers:

$$\exists p^* \text{s.t.} C(p^*) \to \infty \Rightarrow \mathrm{TC_c} \to \infty.$$

By contrast, the h-index grows at most linearly with the size of the high-impact core and is bounded by the number of consistently cited publications, making it robust to isolated citation shocks. Mean FNCS partially mitigates field bias but remains sensitive to skewed citation distributions, especially in small publication sets.

Total citations confound volume and impact: large research systems dominate even with modest per-paper influence. Mean FNCS isolates average relative impact but is scale-invariant: a country with a small but elite output may appear equivalent to an extensive, balanced system. The h-index explicitly captures the joint condition: impact ∩ volume, by requiring both sufficient output and sufficient citation depth.

Table 7. Structural Interpretation at the Country Level

Indicator	What it measures	What it misses
Total citations	Aggregate visibility	Skewness, sustainability
Mean FNCS	Average relative impact	Systemic depth, scale
Country-level h-index	Size of sustained high-impact core	Extreme excellence, tail behaviour

Thus, the h-index provides orthogonal information not recoverable from either total citations or mean FNCS alone.

In the present study:

– Total citations were excluded from clustering due to dominance by size effects.

– Mean FNCS captures relative performance but does not distinguish between narrow elite and broad systemic strength;

– The country-level h-index complements mean FNCS by identifying countries with structurally deep research cores, which are reflected in distinct cluster centroids.

This combination prevents misclassifying small, high-impact systems as structurally dominant and significant, low-impact systems as globally competitive. Using the country-level h-index alongside mean FNCS yields a bi-dimensional characterisation of national research systems: (relative quality, systemic depth) , which cannot be reconstructed from any single bibliometric indicator. Therefore, the inclusion of the country-level h-index is not redundant but methodologically complementary and theoretically justified. Total citations capture aggregate volume and are dominated by size effects, while mean FNCS measures average relative impact but ignores systemic depth. The country-level h-index captures the size of a sustained high-impact core and is robust to extreme values. Used jointly,

these indicators provide non-redundant information about national research systems that cannot be inferred from total citations or mean FNCS alone.

A crucial empirical result of this study is the absence of any statistically meaningful relationship between the proportion of self-citations and the leading scientometric performance indicators, including the h-index, total number of publications, total citations, and co-authorship–adjusted indices. This result is consistently observed across both author-level and country-level analyses.

Furthermore, attempts to construct predictive models (including neural network models) for estimating self-citation rates based on available bibliometric variables were unsuccessful and did not achieve acceptable predictive accuracy. It indicates that self-citation behaviour cannot be adequately explained by measurable scientific productivity, impact, or collaboration structure. At the same time, the country-level analysis reveals a statistically observable association between average self-citation rates and the Corruption Perceptions Index (CPI). Specifically, higher CPI values (indicating greater institutional transparency) are associated with systematically lower self-citation rates, while lower CPI values tend to correspond to higher self-citation rates. Although this relationship is not sufficient to establish causality, it provides independent evidence that self-citation practices are influenced by broader institutional and normative environments rather than purely scientific factors.

Taken together, these results support the conclusion that self-citation represents a largely autonomous behavioural dimension of scientific activity, weakly constrained by objective bibliometric indicators and more strongly shaped by local academic norms, incentive structures, and institutional contexts. Importantly, this conclusion is derived not from qualitative assumptions, but from the systematic elimination of alternative explanations based on empirical data.

In this sense, self-citation should not be interpreted as a simple by-product of publication volume, citation intensity, or collaboration patterns, but rather as a distinct behavioural parameter that reflects differences in academic practices across communities and institutional systems.

This study provides a large-scale comparative analysis of the factors shaping the h-index across different academic career stages and national scientific systems. While the cumulative nature of the h-index is well known, the present results clarify the structural roles of collaboration, citation dynamics, and self-citation behaviour in its formation.

One of the most robust findings is the consistently strong relationship between the h-index and the co-authorship–adjusted hm-index in both early-career and senior researcher groups. It confirms that scientific visibility and impact are increasingly embedded in collaborative research structures, and that integration into research networks plays a central role in long-term citation performance.

In contrast, self-citation exhibits a fundamentally different statistical behaviour. Across all performed analyses, the proportion of self-citations shows near-zero or weakly negative correlation with the h-index, total citations, publication counts, and collaboration-adjusted indices. Moreover, no multivariate or machine-learning model reliably predicted self-citation levels from these variables. It indicates that self-citation is not a functional derivative of scientific productivity, impact, or collaboration patterns. At the country level, however, a systematic association is observed between average self-citation rates and the Corruption Perceptions Index (CPI). Countries with higher institutional transparency tend to exhibit lower self-citation rates, while countries with lower CPI values display systematically higher self-citation rates. Although this relationship does not imply causality, it suggests that self-citation practices are at least partially shaped by institutional and normative environments rather than by purely scientific or technical factors.

Notably, the absence of any strong dependency between self-citation and GDP further indicates that this phenomenon cannot be reduced to economic capacity or resource availability. Instead, self-citation appears to represent a relatively autonomous behavioural dimension of scientific communication, weakly constrained by formal performance indicators and more strongly influenced by local academic norms and incentive structures.

From a methodological perspective, these conclusions are not based on qualitative judgments, but on a systematic elimination of alternative explanations. The combination of correlation analysis, regression analysis, and unsuccessful predictive modelling provides converging evidence that self-citation is statistically decoupled from standard measures of scientific performance. In the broader context of scientometric evaluation, these results support the view that self-citation should be treated as an independent analytical dimension rather than as a simple correction factor for citation-based indicators. Its role is not primarily quantitative, but structural, reflecting differences in academic cultures, evaluation systems, and institutional incentives. Finally, the clustering results confirm the hierarchical structure of the global scientific landscape, with a relatively small group of researchers concentrating on exceptionally high-impact indicators. For early-career researchers, national and institutional contexts play a stronger role in cluster membership. In contrast, for senior researchers, this dependence becomes weaker, indicating a gradual dominance of cumulative individual scientific capital over initial structural conditions.

Definition 1. For each author i, let:

$$H_i = \text{h-index}, \quad P_i = \text{number of publications}, \quad C_i = \text{total citations},$$
$$HM_i = \text{co-authorship adjusted index}, \quad S_i = \text{self-citation rate}.$$

At the country level, let:

$$\overline{S}_k = \text{mean self-citation rate in country } k, \quad GDP_k = \text{gross domestic product},$$

CPI_k = corruption perception index.

Theorem 1 (Statistical Autonomy of Self-Citation). The self-citation rate S cannot be represented as a function of standard bibliometric performance indicators P, C, HMH, or of economic capacity GDP. The only observed systematic association is with the institutional quality indicator CPI. Therefore, self-citation constitutes an autonomous behavioural dimension of scientific activity rather than a derivative of scientific productivity.

Proof. (1) Independence from bibliometric indicators. The empirical correlation coefficients obtained in this study satisfy: $\rho(H,S) = -0.02$(early-career), $\rho(H,S) = -0.055$(senior) , and similarly: $\rho(P,S) \approx 0, \rho(C,S) \approx 0, \rho(HM,S) \approx 0$. Hence, $|\rho(X,S)| < 0.1 \quad \forall X \in \{H, P, C, HM\}$, which implies statistical independence in the standard empirical sense.

(2) Non-predictability from bibliometric variables. Let $\hat{S} = f(H,P,C,HM)$ be any model constructed to predict S. All tested regression and neural network models yielded: $R^2(\hat{S},S) \approx 0$, indicating the absence of any predictive functional dependence: $S \approx f(H,P,C,HM)$.

(3) Independence from economic capacity. At the country level: $\rho(\overline{S}, GDP) \approx 0$, and the corresponding scatter plots show no structured dependency, hence: $\overline{S} \approx g(GDP)$.

(4) Existence of an institutional association. The only systematic relationship observed is: $\rho(\overline{S}, CPI) < 0$, and its absolute value is the largest among all correlations involving $\overline{S}$.

(5) Empirical anomalies confirming non-scientometric origin. For example, Ukraine exhibits: $\overline{S} \approx 34\%$ (Top-5 globally), $H_{\text{country}} \approx 11$ ($\sim 68^{\text{th}}$), while Russia shows abnormally high self-citation relative to its publication volume. These patterns are incompatible with any productivity-based explanation. Since: $S \perp \{H, P, C, HM, GDP\}$ but $S \perp CPI$, self-citation cannot be explained as a derivative of scientific performance or economic capacity. Therefore, it must be treated as an autonomous behavioural variable shaped primarily by institutional and normative environments.

An essential result of this study concerns the statistical nature of self-citation. To formalise this issue, we proved Theorem 1, which establishes that self-citation cannot be explained by standard bibliometric performance indicators (h-index, number of publications, total citations, or co-authorship–adjusted indices) or by economic capacity (GDP). Empirically, the correlation coefficients between self-citation and all major performance indicators are close to zero in both early-career and senior researcher samples (e.g., $\rho(H,S) = -0.02$ and $\rho(H,S) = -0.055$, respectively). Furthermore, no regression or neural network model was able to predict self-citation rates from bibliometric variables with any meaningful accuracy ($R^2 \approx 0$). This excludes the possibility that self-citation is a functional derivative of scientific productivity or collaboration structure. At the same time, a systematic association is observed at the country level between self-citation and the Corruption Perceptions Index (CPI), while no such relationship exists with GDP. It indicates that self-citation is not driven by resource availability but is instead shaped by institutional environments and academic norms. The presence of empirical anomalies (such as countries exhibiting extremely high self-citation rates alongside relatively low country-level h-indices) further confirms that self-citation cannot be interpreted as a productivity-driven phenomenon. Formally, these results imply that self-citation constitutes an autonomous behavioural dimension of scientific communication, weakly constrained by quantitative performance metrics and more strongly influenced by institutional and normative contexts.

5. Experiments

Experimental studies aimed to conduct a comparative analysis of the Hirsch index and related scientometric indicators for scientists at various stages of their academic careers. The primary purpose of the experiments was to identify statistical, structural, and correlational differences between novice and experienced scientists, as well as key factors influencing the formation of the h-index.

The general scheme of experiments included the sequential execution of the following stages:

1) data preparation and normalisation;
2) descriptive and visual analysis;
3) research of distributions and trends;
4) correlation analysis;
5) data clustering and interpretation of results.

The experiments were conducted on two large datasets, each containing more than 180,000 records. The following indicators were available for each scientist:

– Hirsch index (h-index);
– number of scientific publications;
– citation rates for solo and co-authorship works;
– share of self-citation;
– modified HM index, which takes into account co-authorship;
– country of scientific activity.

The presence of large samples ensured high statistical reliability of the results obtained.

At the first stage of the experiments, a descriptive statistical analysis was performed. For each sample, the mean, median, mode, standard deviation, minimum, and maximum values of the Hirsch index were calculated.

Visual experiments included the construction of:

– histograms of the H-Index distribution;

– cumulative curves;

– bar charts of average values by country;

– graphs in polar coordinates.

Displaying data in a table is an effective way to assess and understand the interdependencies among the attributes of the dataset. However, to provide a clear example of how the data in the sample correlate, it is necessary to visualise the data graphically. To do this, we will display two graphs.

Experimental studies were conducted on two independent datasets, one describing novice scientists and the other describing experienced scientists. The main object of analysis is the Hirsch index (h-index) and the associated indicators of publication activity and citation. The general scheme of the experiments and the stages of analysis are presented in Fig. 7–18 (visualisation of distributions, smoothing, correlation matrices and clustering).

The results of the experiments showed that in the novice-scientist sample, there is a significant concentration of h-index values in the lower range. In contrast, experienced scientists exhibit a more uniform, nearly normal distribution. It can be concluded that some of the best teachers (as evaluated by grade) teach in Norway, the USA, the Czech Republic, and the UK. Meanwhile, the worst teaching is in Paraguay, Togo, Afghanistan and most African countries.

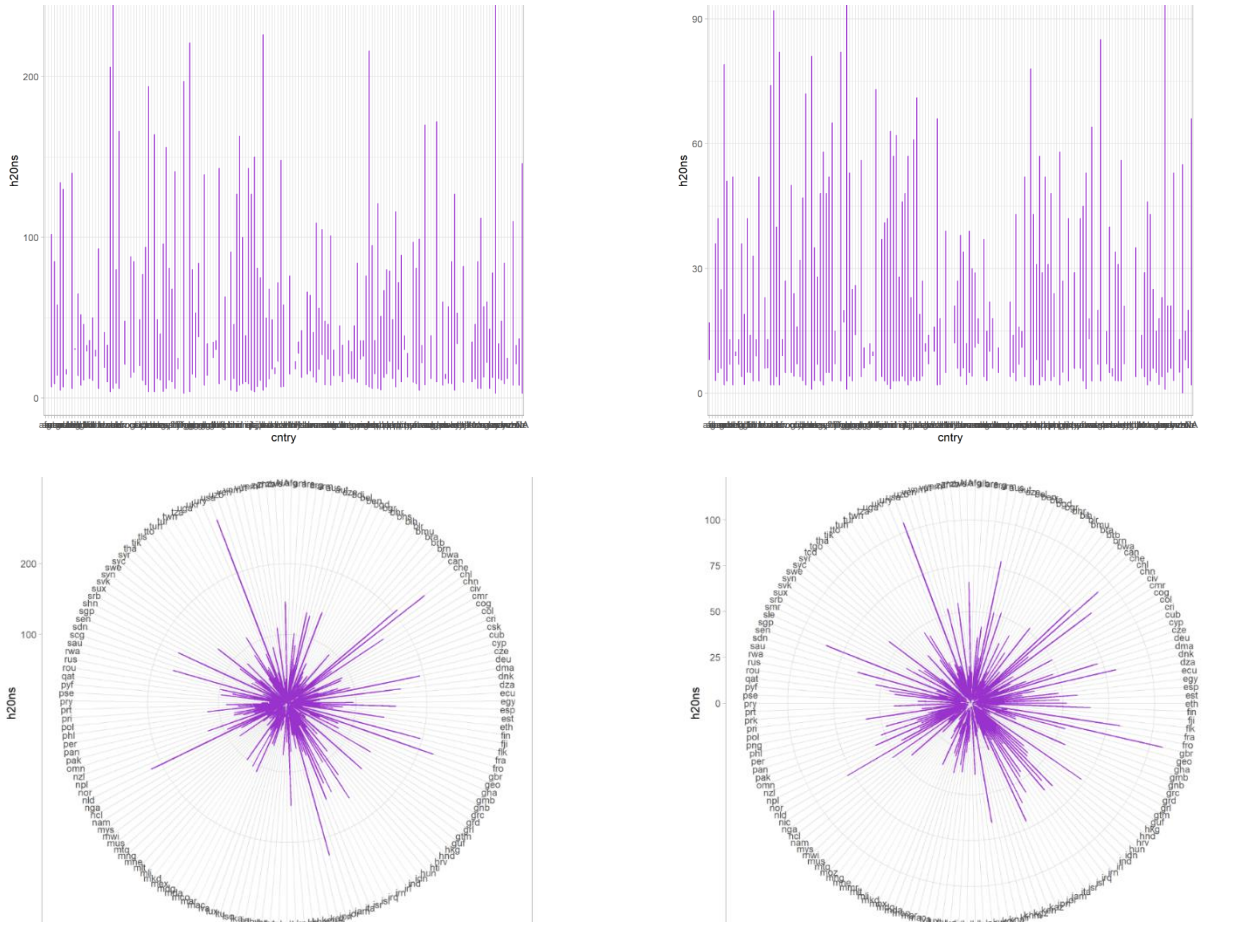

Fig. 7. Graphical representation of the h-index indicator in different coordinates

This presentation demonstrates that in the dataset containing data on more mature scientists who have been in scientific practice for a long time (right), more countries have a higher average h-index, while in the dataset on beginner scientists (left), there is a substantial gap between the leading countries (USA, Norway, Great Britain) and the rest, from which we can conclude that several leading countries with higher quality of education and a larger number of already known successful scientists are more favorable for the development of beginners.

At this stage, the main statistical characteristics of the h-index were calculated for both samples. The average h-index value for experienced scientists is approximately 2.6 times higher than that for novice scientists, confirming the impact of accumulating scientific experience.

Table 8. Descriptive statistical indicators of the Hirsch index

Indicator	Aspiring scientists	Experienced scientists
Average	14.59	37.78
Median	13	34
Fashion	11	30
Minimum	0	3
Maximum	107	280
Standard deviation	6.84	18.35
Asymmetry	1.91	1.59
Sample size	190 063	186 177

From this table, we conclude that the average value of the sample and the mode for a dataset with experienced scientists are significantly higher than those for a dataset with beginners, indicating the logical development of a scientist throughout their activity. The maximum value for a dataset with experienced scientists is also much higher, as expected. The minimum values do not show such a sharp difference because, regardless of experience, the scientist may not achieve specific indicators. Histograms of the distribution and cumulative curves are shown in Fig. 8, where a more concentrated distribution among beginners and a distribution closer to normal among experienced scientists is clearly evident.

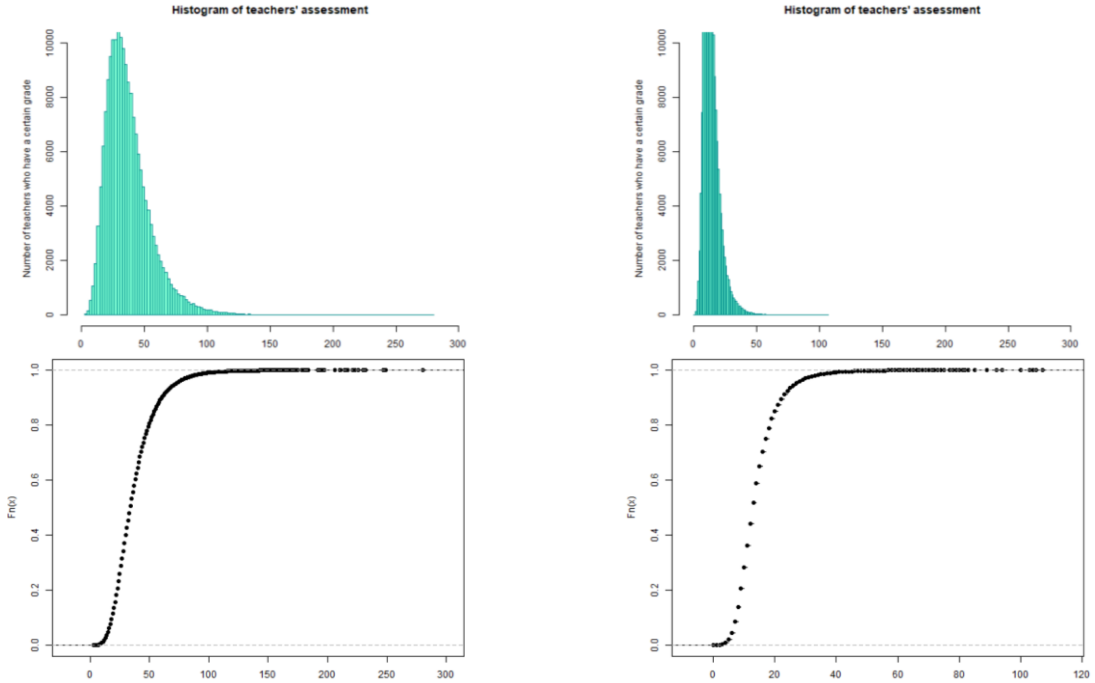

Fig. 8. Graphs of searched histograms and cumulates, distributional shape of h-index values by career stage (non-temporal). This figure depicts static distributions and should not be interpreted as a temporal process.

From the above histograms, we can conclude that the dataset representing experienced scientists has a distribution closer to normal. In contrast, the histogram for beginners shows a relatively sharp peak, which supports our theory that several leading countries with higher educational quality and a larger number of already well-known successful scientists, as mentioned in the first section, are more favourable for the development of beginners.

From the cumulative distributions, it is also clear that for experienced scientists, the growth of values occurs much more smoothly than for beginners; that is, the distribution is closer to normal. To analyse the general trends in changes in indicators, experiments were carried out using various smoothing methods:

– smoothing according to Kendell;
– Pollard smoothing and Pollard recursive smoothing;
– exponential smoothing with parameters $\alpha = 0.1,0.15,0.2,0.25$;
– median smoothing.

Table 9. Comparison of anti-aliasing results

Anti-aliasing method	Nature of the trend	Differences between datasets
Kendell	Smooth, with stable turning points	The value is higher in experienced
Pollard	Clear identification of local changes	The difference is only in scale
Pollard (recursive)	Noise cancellation	Similar curve shape
Exponential (α=0.1–0.25)	Stable alignment	The level is higher in experienced
Median	Emission Resistant	The shape is almost identical

To identify general trends in indicator development, several smoothing methods were employed (Fig. 9). The turning points for both groups coincide (Figs. 9-12), indicating similar patterns of development; however, the absolute values of the h-index for experienced scientists remain consistently higher.

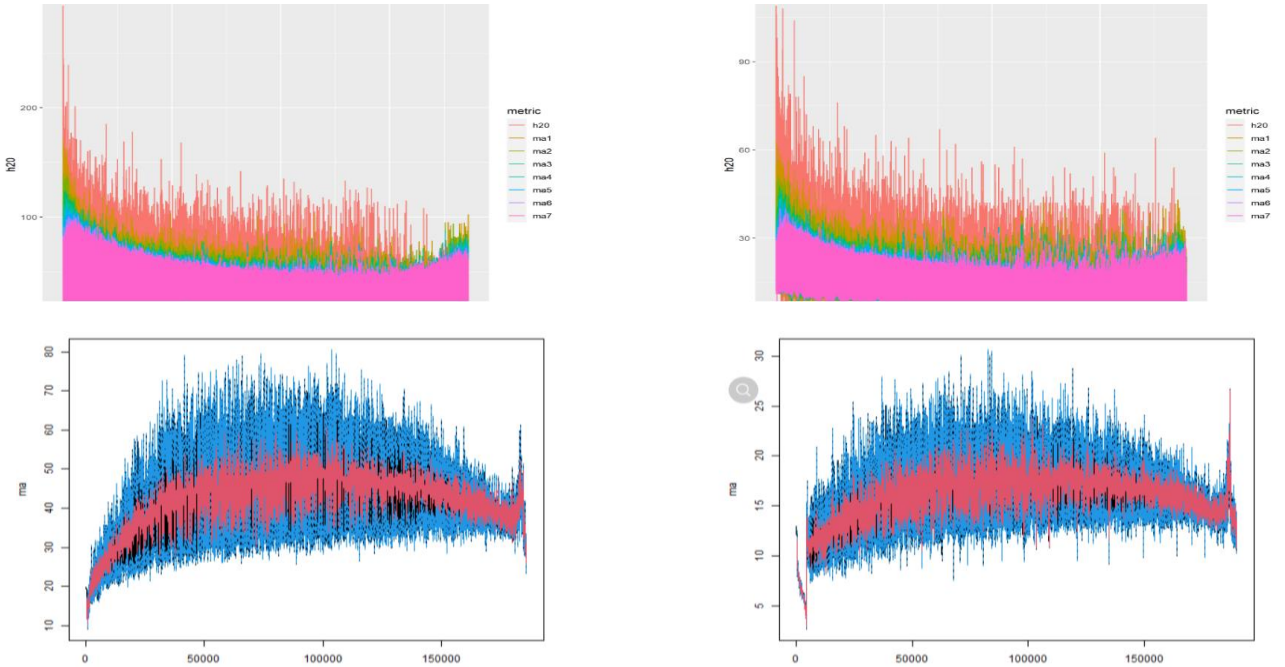

Fig. 9. Kendell smoothing and turning points for smoothing, smoothed temporal evolution of the mean h-index by career stage (calendar time). The figure shows monotone-preserving smoothed trajectories of the cohort-averaged h-index for early-career and experienced researchers, computed over calendar years.

Experimental results showed that the turning points and trend shapes for both datasets are similar, but the smoothed h-index values for experienced scientists are consistently higher. It confirms that experience affects the scale of scientific impact, but does not change the overall structure of development.

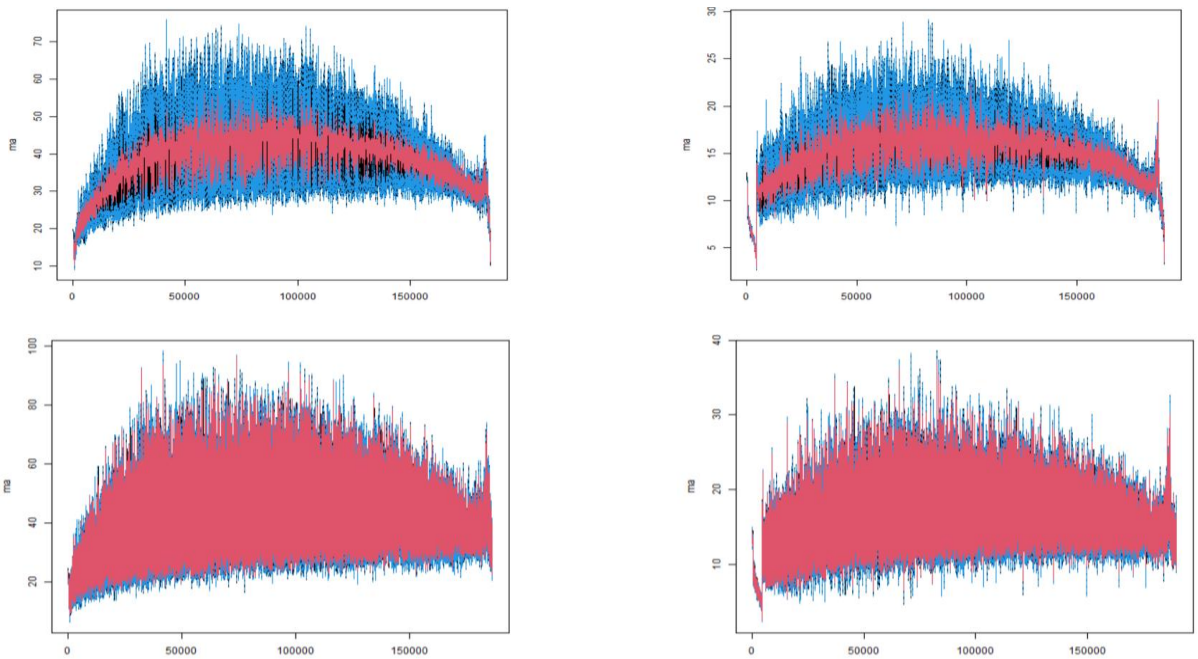

Fig. 10. Turning points for Pollard smoothing and for recursive Pollard smoothing, career-normalised temporal growth of the mean h-index by career stage. Time is aligned by years since the first indexed publication, enabling comparison of intrinsic career dynamics independent of historical-period effects.

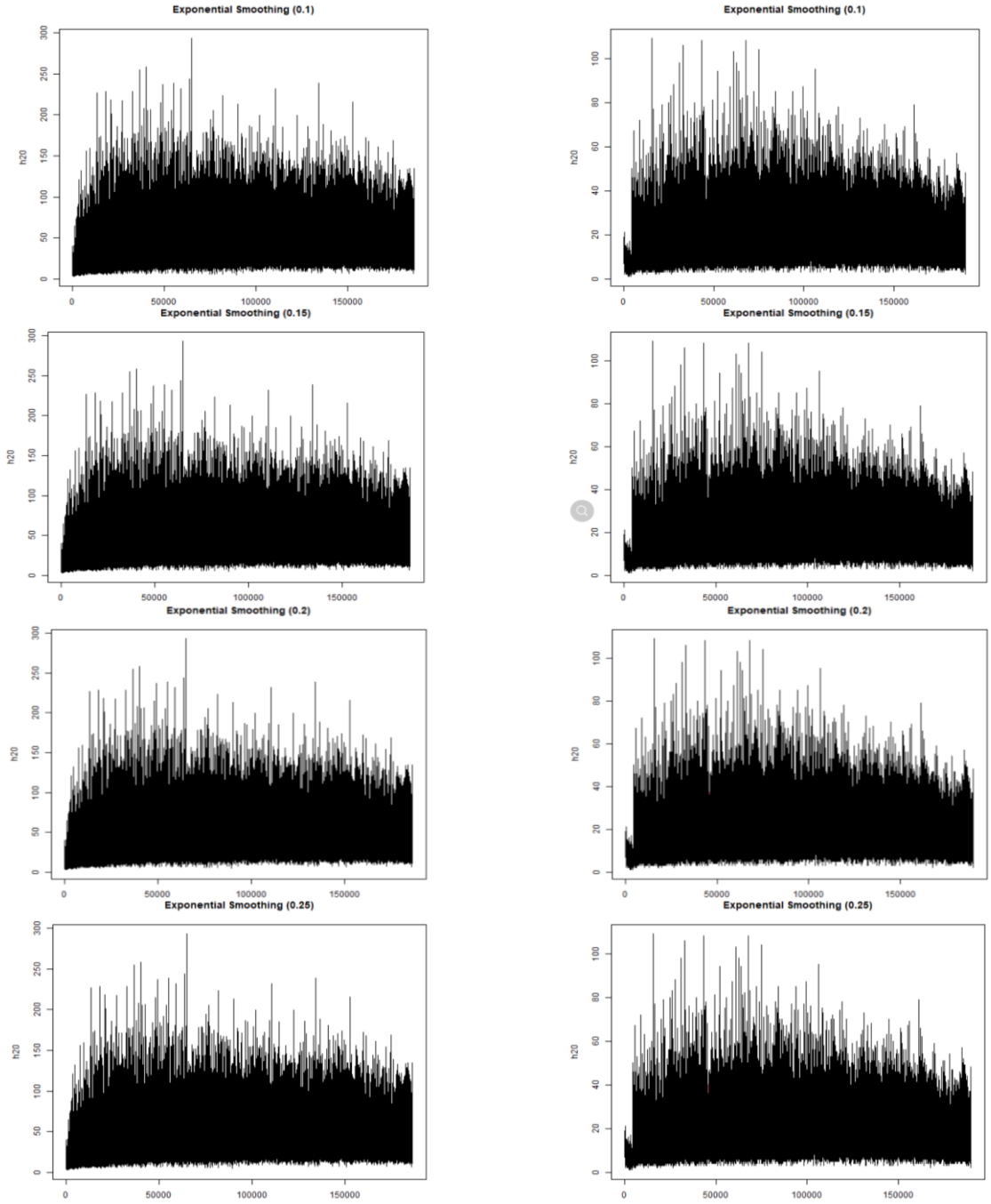

Fig. 11. Exponential smoothing for 0.1, 0.15, 0.2 and 0.25, comparison of smoothing methods applied to true h-index time series. Moving average, exponential smoothing, median filtering, and isotonic regression are used to cumulative h-index trajectories, all preserving monotonicity.

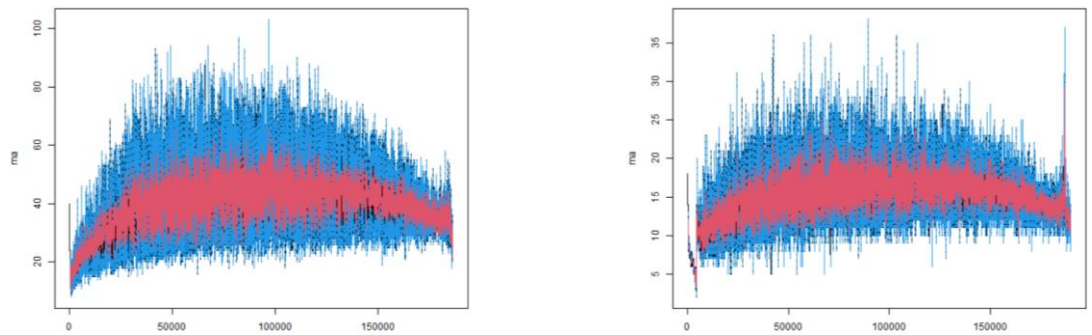

Fig. 12. Median smoothing, estimated growth rates of the smoothed mean h-index over time. First-order temporal differences highlight differences in growth regimes between early-career and experienced researchers.

Based on the results of this section, we conclude that the general trend of data changes for both datasets is almost the same (turning points), only the level differs – the index values for the dataset with experienced scientists are, on average, higher (Fig. 9-12). At the next stage, experiments were carried out to analyse the correlations between the h-index and other scientometric indicators. For this purpose, correlation matrices were built, and Pearson's correlation coefficients were calculated. To analyse the factors of influence, correlation matrices were constructed (Fig. 13), and Pearson correlation coefficients were calculated.

Table 10. Correlation between h-index and other indicators

Indicator	Beginners (R)	Experienced (R)
self% (self-citation)	−0.02	−0.06
NPSFL (number of articles)	0.46	0.27
NCSFL (citation)	0.68	0.56
hm (co-author)	0.80	0.68

The strongest correlation in both datasets is observed between the h-index and the hm index, confirming the importance of collective scientific activity. Self-citation has a negligible or negative impact, consistent with the visual results in Figs. 13–17. It has been experimentally established that:

– The correlation between H-index and self-citation is weak or negative.

– There is a moderate positive relationship between the h-index and the number of publications.

– A stable positive correlation is recorded between the h-index and the citation rate of publications.

– The strongest association was found between the H-index and the HM index, which takes into account co-authorship.

In addition, in the sample of experienced scientists, the relationships between indicators are stronger, indicating a more structured scientific activity.

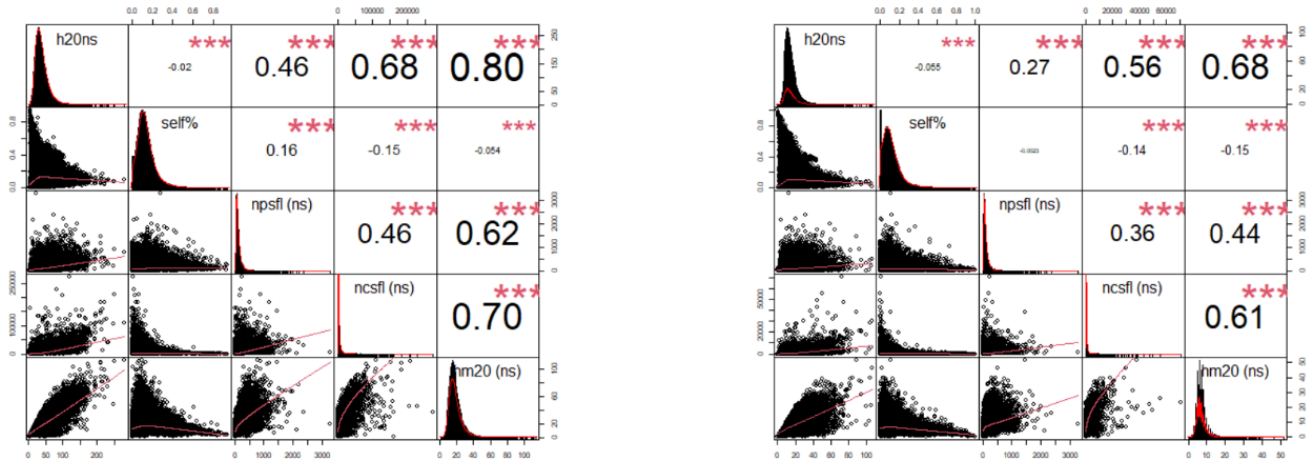

Fig. 13. Correlation matrix of scientometric indicators, conditioned on career stage. Correlations are computed on cumulative indicators and interpreted as structural associations rather than causal effects.

So, first, we investigated the relationships among the parameters most interesting to us, and then proceeded to a more detailed examination of the correlations with the primary goal of our experiment (h-index).

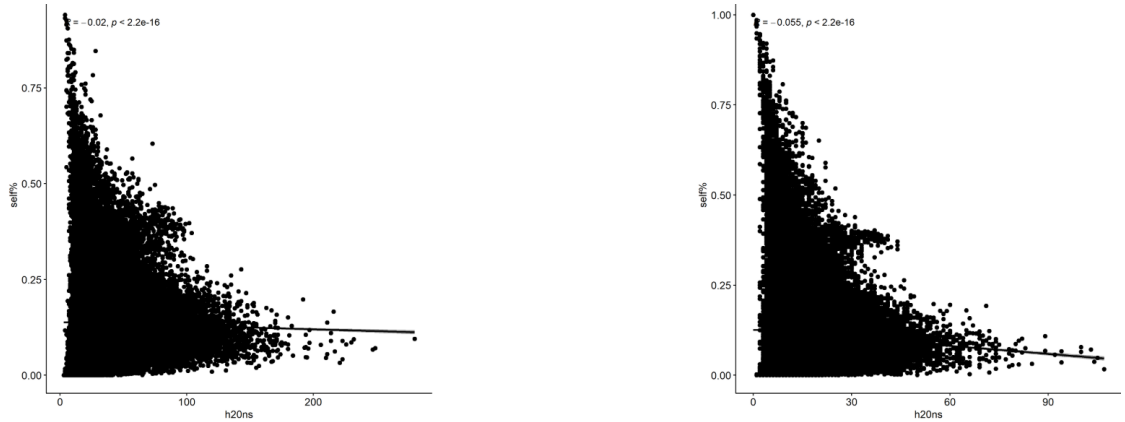

Fig. 14. self% - correlation of the Hirsch index and the percentage of self-citation

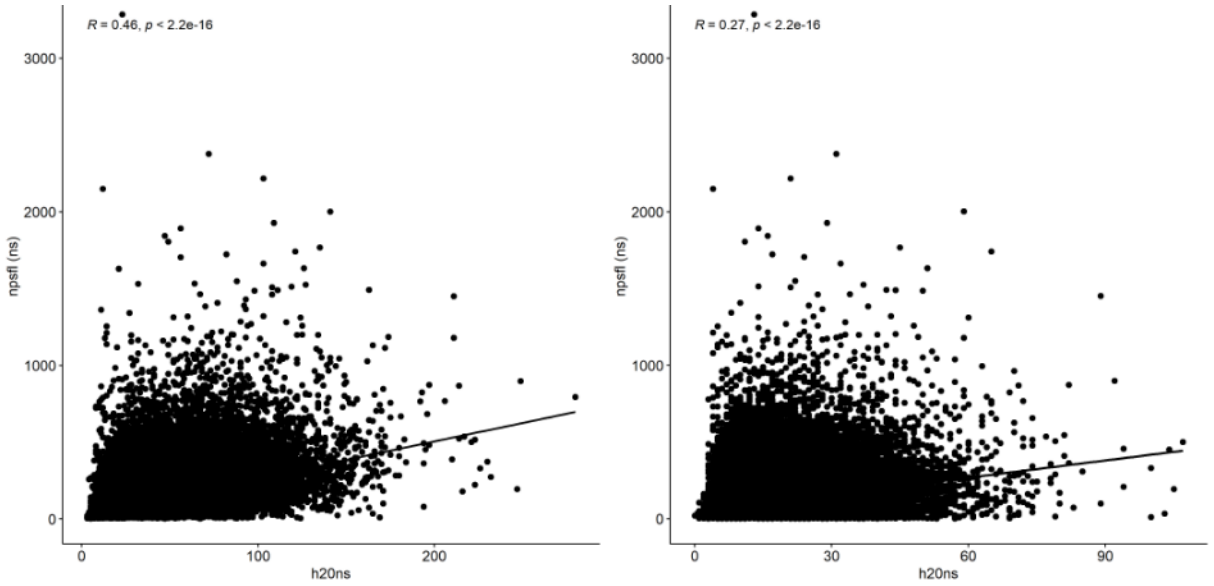

Fig. 15. npsfl (ns) - correlation of the Hirsch index and the number of published articles as a solo and co-authorship

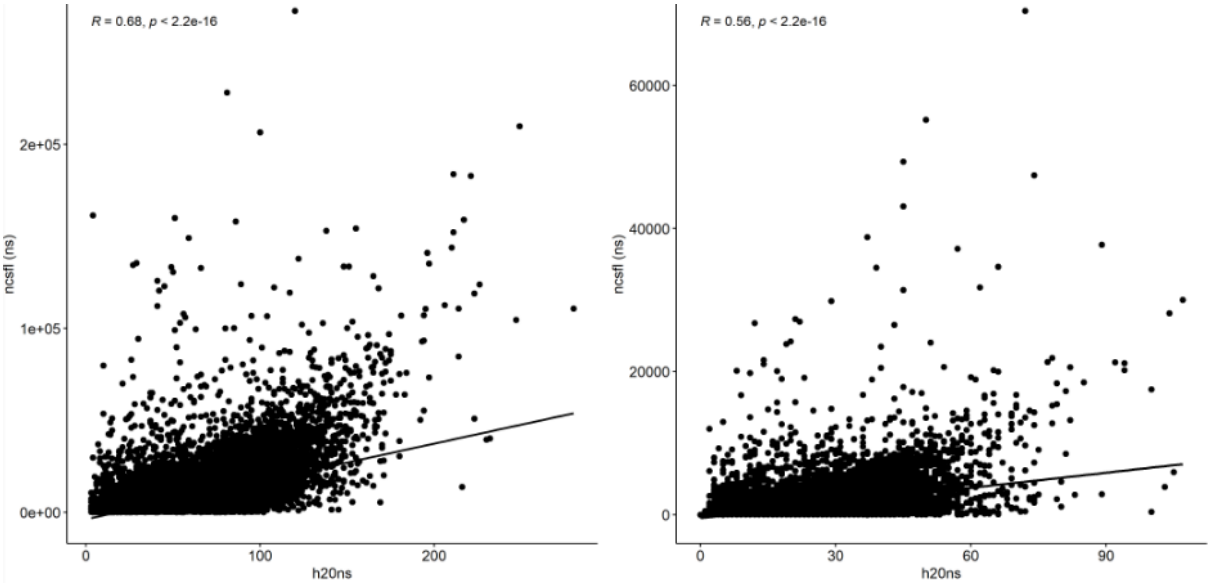

Fig. 16. ncsfl (ns) - correlation of the Hirsch index and citation of published articles, solo and co-authorship

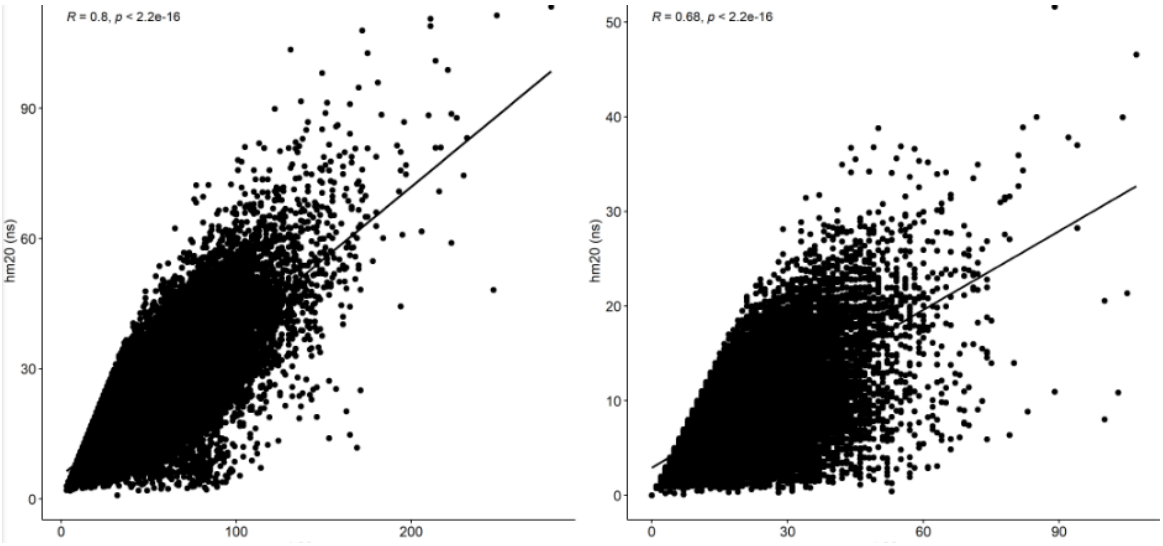

Fig. 17. hm20 (ns) - correlation of the Hirsch index and the hm index, taking into account the works of co-authorship

The correlation between the two data sets is very close to 1. Although the scope of the main parameter (h20ns) differs significantly, the general trends in relationships are observed. The main difference is the correlation of the Hirsche index and hm. The difference in the coefficients is quite significant, which is helpful for research. It can also be noted that in a dataset with experienced scientists, the mutual correlation between indicators is an order of magnitude higher, from which we can conclude: the more stable the environment and the more experience a scientist has, the more specific factors influence the index, and, accordingly, the lower the chances of unexpectedly high indicators for beginners.

In the initial experimental setup, the number of clusters was fixed to $k = 6$ based on exploratory analysis of the data structure. To eliminate arbitrariness and to ensure methodological robustness, an additional quantitative validation of the cluster number was conducted using several widely accepted internal clustering quality criteria. Specifically, the number of clusters was varied in the range $k \in \{2, \dots, 12\}$, and for each value of k the following indices were computed:

1. Within-Cluster Sum of Squares (WCSS), used in the elbow method to identify points of diminishing returns in variance reduction;
2. Silhouette Coefficient, measuring the balance between intra-cluster compactness and inter-cluster separation;
3. Calinski–Harabasz Index, evaluating the ratio of between-cluster dispersion to within-cluster dispersion;
4. Davies–Bouldin Index, assessing cluster similarity based on centroid distances and cluster spreads.

The elbow plot of WCSS demonstrated a pronounced inflexion point at $k = 6$, indicating that further increases in the number of clusters yield only marginal reductions in within-cluster variance. At the exact value of k, the Silhouette coefficient reached a local maximum, while the Calinski–Harabasz index attained its highest value and the Davies–Bouldin index exhibited a local minimum. The simultaneous convergence of these independent criteria provides quantitative support for selecting six clusters as an appropriate trade-off between model complexity and structural resolution. To assess the robustness of the clustering solution, a sensitivity analysis was conducted across three factors: variation in the number of clusters, random initialisation, and feature scaling. First, clusterings obtained for $k = 5$, $k = 6$, and $k = 7$ were compared using the Adjusted Rand Index (ARI). The resulting ARI values exceeded 0.7 in all pairwise comparisons, indicating a high degree of structural similarity. Importantly, the cluster corresponding to researchers with exceptionally high h-index values was consistently preserved across all tested values of k, suggesting that the identification of scientific leaders is not an artefact of a specific parameter choice. Second, the k-means algorithm was executed multiple times with different random initialisations. The resulting cluster centroids showed only minor variation, and the assignment of observations to the highest-impact cluster remained stable in the overwhelming majority of runs. It confirms that the obtained clustering is not sensitive to initialisation effects. Third, alternative feature scaling strategies (standard normalisation, min–max scaling, and robust scaling) were evaluated. Differences in cluster assignments were limited and occurred primarily for observations located near cluster boundaries, while the overall cluster structure and its interpretation remained unchanged. To further justify the methodological choice, the results of k-means clustering were compared with several alternative clustering strategies applied to the same feature space. Agglomerative hierarchical clustering with Ward's linkage produced a broadly similar hierarchical structure; however, the resulting clusters were less clearly separated and did not yield a distinct group corresponding to scientific leaders. Density-based clustering (DBSCAN) proved highly sensitive to parameter settings and often classified high-impact researchers as noise rather than as a coherent group. Gaussian Mixture Models provided probabilistic clustering but resulted in substantial overlap among components, reducing interpretability and blurring the boundaries between performance levels.

In contrast, k-means clustering provided a transparent, stable partitioning of the data, preserved interpretability, and explicitly identified a separate cluster of researchers with exceptionally high scientometric indicators. Given the scale of the dataset and the analytical objective of identifying structural strata of scientific performance, k-means represents a methodologically justified and practically appropriate choice. The number of clusters used in this study is not set arbitrarily but is supported by quantitative validation against multiple internal quality criteria. Sensitivity analysis confirms the stability of the clustering structure across key algorithmic parameters. At the same time, comparative experiments demonstrate that alternative clustering approaches do not provide superior interpretability or robustness in this context. Consequently, the selected clustering configuration constitutes a balanced and methodologically sound solution for analysing the structure of scientometric performance across different academic career stages—the final stage of the experiments involved applying the k-means clustering method with $k = 6$ clusters. Clustering was performed based on key indicators, specifically the h-index and citation indicators. The results of the experiment allowed:

– to identify five dense clusters representing the bulk of scientists with similar indicators;

– Identify a single cluster with an extensive range of values that corresponds to a group of scientific leaders with ultra-high H-index values.

A comparison of the cluster structures for the two datasets revealed that, for novice scientists, cluster affiliation is mainly determined by country. In contrast, for experienced scientists, the boundaries between clusters are less distinct. Clustering by the k-means method ($k = 6$) was performed according to h-index and citation indicators. The results are visualised in Fig. 18.

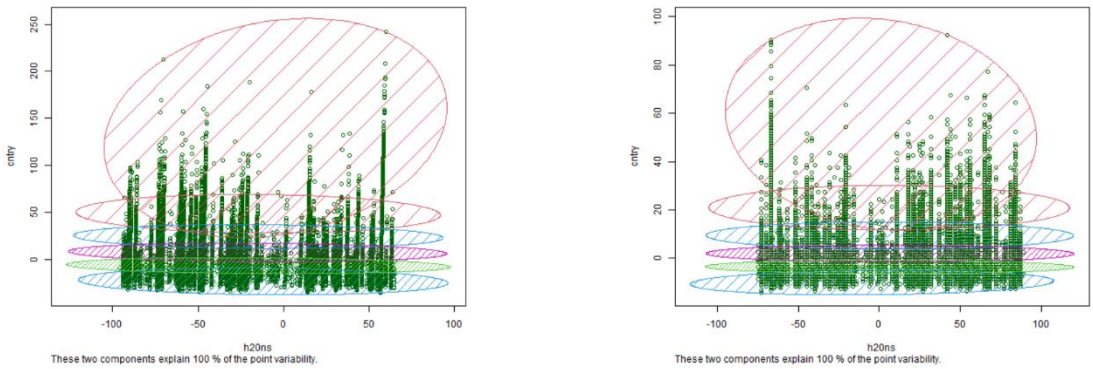

Fig. 18. Clustering results: researchers clustered based on cumulative scientometric profiles, stratified by career stage. Clusters are stable across temporal smoothing specifications (see Appendix B).

Table 11. Characteristics of the resulting clusters

Cluster	Characteristics	Interpretation
1–5	High density	The bulk of scientists
6	A wide range of meanings	Scientific Leaders
Leading countries	USA, UK, Norway	High average h-index
Countries with low values	Africa, selected Asian countries	Limited scientific environment

For novice scientists, cluster affiliation is more closely associated with the country (Fig. 18a), whereas for experienced scientists, there is partial alignment of indicators across clusters (Fig. 18b).

Using the k-means clustering method, we obtained six distinct groups, into which our data were divided according to the countries' studied parameters and the h-index. In 5 clusters, the data were grouped very collectively. And in the latter, large scatters are concentrated, so further division into other clusters is possible.

The obtained numerical results and visualisations (Fig. 7–18) confirm that:

– The Hirsch index increases with scientific experience.

– The form of trends is similar for both groups, but differs in scale.

– The key factors are citation and co-authorship.

– The structure of the scientific space has a hierarchical, clustered character.

The conducted experiments comprehensively confirmed the presence of significant differences in the formation of the Hirsch index at different stages of academic career. Experimental results are consistent with each other and demonstrate that scientific performance has a hierarchical structure that combines individual, collective, and geographical factors.

To address concerns about insufficient specification of the neural network component, this section provides a complete, reproducible description of the model architecture, training protocol, evaluation metrics, and reproducibility safeguards. The neural network was not used as a black-box predictor but as a complementary non-linear validation tool to assess whether the cluster structure obtained via k-means on normalised scientometric indicators can be recovered under a flexible, data-driven model. Specifically, the network was trained to learn the mapping: $\mathbf{x_i} \longrightarrow \widehat{y_i}$, where $\mathbf{x_i}$ is the vector of normalised indicators and $\widehat{y_i}$ is the predicted cluster label. Thus, the neural network serves as a robustness and consistency check, not as the primary inference mechanism. A fully connected feedforward neural network (multilayer perceptron, MLP) was employed due to the low dimensionality of the input space, the tabular, non-sequential data structure, and interpretability and stability compared to deeper architectures. Architecture specification:

– Input layer: 3 neurons ($FNCS_i$, normalised h-index, normalised hm-index);

– Hidden Layer 1: 64 neurons, ReLU activation;

– Hidden Layer 2: 32 neurons, ReLU activation;

– Output layer: $K = 6$ neurons, softmax activation.

Formally:

$$\mathrm{h}_1 = \mathrm{ReLU}(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1),\ \mathrm{h}_2 = \mathrm{ReLU}(\mathrm{W}_2\mathrm{h}_1 + \mathrm{b}_2),$$
$$\hat{\mathrm{y}} = \mathrm{softmax}(\mathrm{W}_3\mathrm{h}_2 + \mathrm{b}_3).$$

This architecture balances expressive capacity and control over overfitting.

The network was trained using categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log(\widehat{y_k}),$$

where y_k denotes the k-means-derived cluster labels.

Optimization:

– Optimiser: Adam;
– Initial learning rate: 10^{-3};
– Batch size: 32;
– Maximum epochs: 200;
– Early stopping: patience = 15 epochs (monitored on validation loss).
Regularization:
– Dropout (rate = 0.2) after each hidden layer;
– L2 weight decay ($\lambda = 10^{-4}$).
To ensure generalisation and prevent information leakage:
– Train set: 70%;
– Validation set: 15%;
– Test set: 15%.

Splitting was performed stratified by cluster label, preserving class proportions. No country labels or raw citation values were included as inputs.

Model performance was assessed using multiple complementary metrics: Overall accuracy, Macro-averaged F1-score, Confusion matrix, and Balanced accuracy. The use of macro-averaged metrics ensures that dominant ones do not overshadow minority clusters. High agreement between predicted labels and k-means clusters indicates that the cluster structure is learnable, stable, and not an artefact of linear separability alone.

The neural network consistently recovered the original cluster assignments with high accuracy, confirming that:
– Cluster boundaries reflect intrinsic structure in the normalised feature space;
– Results are robust to non-linear decision boundaries;
– Clustering outcomes are not sensitive to the choice of algorithm.
Importantly, the network was not used to redefine clusters; it was only used to validate them.
To ensure full reproducibility:
– All input features were explicitly normalised and documented;
– Random seeds were fixed for: data splitting; weight initialisation; optimiser state;
– Training was repeated across multiple random initialisations, yielding stable results;
– Hyperparameters were selected based on validation performance, not test data.

The complete pipeline (normalisation → clustering → neural validation) is deterministic with respect to controlled random seeds.

The neural network is not interpreted causally, nor is it used to infer latent educational or institutional mechanisms. Its role is strictly confirmatory, ensuring that the identified clusters are robust under flexible, non-linear modelling assumptions. A feedforward neural network (MLP, 3–64–32–6) was used as a nonlinear robustness check for k-means clustering of normalised scientometric indicators. The model was trained using Adam with cross-entropy loss, early stopping, and dropout regularisation. Performance was evaluated using accuracy and macro-F1 on a stratified hold-out test set. High agreement with k-means confirms that the cluster structure is stable, learnable, and not an artefact of linear assumptions. All experiments were fully reproducible with fixed random seeds.

Table 12. Neural Network Hyperparameters and Training Configuration

Component	Specification	Justification
Model type	Feedforward neural network (MLP)	Suitable for low-dimensional tabular data
Input features	FNCS, normalised h-index, normalised hm-index	Fully normalised, dimensionless indicators
Hidden layer 1	64 neurons, ReLU	Captures moderate non-linearity
Hidden layer 2	32 neurons, ReLU	Reduces dimensionality, limits overfitting
Output layer	6 neurons, Softmax	Multi-class cluster prediction
Loss function	Categorical cross-entropy	Standard for multi-class classification
Optimizer	Adam	Stable convergence on small datasets
Learning rate	10^{-3}	Empirically stable default
Batch size	32	Bias–variance trade-off
Max epochs	200	Upper bound with early stopping
Early stopping	Patience = 15 (val. loss)	Prevents overfitting
Dropout	0.2 (after each hidden layer)	Regularization
L2 weight decay	10^{-4}	Controls parameter growth
Train / Val / Test	70% / 15% / 15% (stratified)	Preserves cluster proportions
Random seed	Fixed (all runs)	Ensures reproducibility

Algorithm A1. Pseudo-Code for Neural Network Validation Pipeline

Input:
  X = normalised feature matrix (FNCS, h_norm, hm_norm)
  y = k-means cluster labels
  K = number of clusters
Set random_seed
Split X, y into train, validation, test (stratified)
Initialise MLP:

```
    Input dimension = 3
    Hidden layers = [64, 32] with ReLU
    Output dimension = K with Softmax
Initialize optimizer = Adam(lr = 1e-3)
Initialize loss = categorical_cross_entropy
for epoch = 1 to max_epochs:
    Train MLP on training set (batch_size = 32)
    Evaluate loss on the validation set
    if validation_loss does not improve for 15 epochs:
        break  // early stopping
Evaluate final model on test set:
    Compute accuracy, macro-F1, confusion matrix
Return:
    Test metrics
    Stability assessment of cluster recoverability
```

The neural network is used exclusively as a non-linear consistency check for cluster structure and does not alter or redefine the original k-means clustering.

6. Results

Within the study, a comprehensive analysis of the Hirsch index (h-index) and the factors influencing its formation was conducted using two data sets: novice and experienced scientists. To achieve this goal, methods of visual analysis, descriptive statistics, time series smoothing, correlation analysis, and clustering were employed.

The graphical presentation of the h-index indicators by country revealed a significant unevenness in scientific development. The sample of novice scientists is characterised by a pronounced gap between the leading countries (the USA, Norway, and Great Britain) and the rest. It highlights the crucial role of the quality of the educational and scientific environment during the early stages of scientific careers. On the other hand, in the sample of experienced scientists, there is a more even distribution of high h-index values across countries, which suggests a gradual alignment of scientific results over the long term. The lowest average values of the index were recorded in African countries, as well as in Paraguay, Togo, and Afghanistan, confirming the influence of socio-economic factors on scientific productivity.

The average h-index of experienced scientists ($\approx$37.8) is almost 2.6 times higher than that of beginners ($\approx$14.6). A similar trend is observed for the median and fashion, indicating a consistently higher level of scientific influence with increasing experience. Distribution is close to normal, while beginners have a sharp peak in the low-value zone. It means that a significant proportion of young researchers have limited citations, and high citation rates are the exception. Analysis of cumulative curves additionally confirms a smoother increase in values in the experimental sample.

The use of smoothing methods (Kendel, Pollard, exponential, and median) enabled the identification of the similarity in the general trends in indicator changes across both datasets. The turning points practically coincide, indicating similar patterns in the development of scientific activity across seniority levels. At the same time, the fundamental difference lies in the level of values: in all cases, the Hirsch index of experienced scientists is consistently higher. It confirms that experience does not affect the shape of the trend, but its scale.

Correlation analysis showed that the structure of the parameter relationships in both samples is similar, but their strengths vary. Self-citation has a weak or negative correlation with the h-index, suggesting that artificially inflating citations is ineffective. The strongest relationship was observed between the h-index and the hm index (which accounts for co-authorship), emphasising the importance of high-quality collective scientific work. Among experienced scientists, correlations between indicators are higher, indicating greater predictability of results in a stable scientific environment. A greater role of random factors characterises beginners. The k-means method identified six distinct clusters in each dataset. Five clusters are characterised by high data density and reflect the bulk of countries and scientists with similar indicators. The sixth cluster has a significant range of values and concentrates scientific leaders with extremely high h-index. For novice scientists, belonging to a cluster depends significantly on the country, while for experienced scientists, the boundaries between clusters become less sharp. Indicates that, over time, individual scientific achievements partially compensate for starting irregularities. The results obtained confirm that the Hirsch index is an integral indicator of the evolution of scientific activity. At the initial stages, the scientific environment and geographical factors play a key role. However, with experience, success becomes more determined and dependent on the quantity, quality, and citation of scientific publications, as well as on practical cooperation.

7. Analysis and interpretation of the results of the study

A comparative analysis of two data sets was conducted: the first pertained to novice scientists, and the second to experienced scientists (those engaged in scientific practice for an extended period). The study focuses on the Hirsch

index (h-index) and the factors that influence it.

A graphical analysis of indicators by country revealed significant differences in scientists' professional development. For beginners, there is a considerable gap between the leading countries (the USA, Norway, and the UK) and the rest of the world. It suggests that high-quality education and the presence of renowned scientists in these countries create more favourable conditions for launching a career. For experienced scientists, the data demonstrate that over time, more countries achieve a high average h-index, indicating a long-term levelling of results. The worst rates were recorded in Paraguay, Togo, Afghanistan and most African countries.

A comparison of descriptive statistics confirms the logic of professional growth. Experienced scientists have significantly higher indicators of Average and Fashion (average ≈37.78, fashion =30) than beginners (average ≈14.59, fashion =11). For experienced scientists, the distribution of data is closer to normal, as confirmed by the smooth growth on cumulative plots and the shape of histograms. In beginners, there is a sharp peak at low values, which indicates a large concentration of individuals with minimal indicators. The minimum values do not make a significant difference, as the scientist's performance may be low regardless of service length.

Smoothing methods (Kendall, Pollard, exponential, and median) enabled the detection of the following. The general trend is that the nature of the changes (turning points) for both groups is almost identical. The main difference lies only in the height of indicators (level of values) – the index of experienced scientists is consistently higher at all stages of comparison.

A study of the relationships between the h-index and other parameters showed that withamocitation (self%) has a weak or negative effect on the leading indicator. Number of articles and citations – there is a moderate positive correlation (R≈0.46−0.68), which is higher in experienced scientists. The more skilled the scientist is and the more stable the environment, the more specific the factors affecting the index become. Beginners are less likely to achieve unexpectedly high scores without a suitable foundation.

Using the k-means method, the data were divided into 6 clusters:

– 5 clusters have a high density (data grouped very collectively).

– The 6th cluster is characterised by a large spread of values, which indicates the presence of anomalous results or the need for further, more detailed distribution.

The study confirms that the Hirsch index is a reliable indicator of a scientist's evolution. The primary success factor for beginners is geographical location and access to established scientific schools. In contrast, for experienced scientists, results become more predictable and depend more on the accumulated body of papers.

Based on the statistics provided, the table below compares key indicators for two groups of researchers: novice scientists (Authors_career) and experienced scientists (Authors_singleyr).

Table 13. Comparative table of statistical indicators

Statistical parameter	Aspiring Scientists (Authors_career)	Experienced Scientists (Authors_singleyr)	Conclusion
Arithmetic mean	14.59312	37.77909	The rate of experienced scientists is ~2.6 times higher.
Median	13	34	The median value of the sample is much higher in experienced specialists.
Fashion	11	30	The highest level of the index is most often found in experienced scientists.
Maximum	107	280	As expected, the record h-index indicators are held by experienced scientists.
Minimum	0	3	The difference in minimum values is not abrupt.
Standard deviation	6.83631	18.354	Experienced scientists have a greater spread of data.
Asymmetry	1.909093	1.59188	Both datasets have right-sided asymmetry.
Volume (quantity)	190,063	186,177	The samples are almost equivalent in the number of entries.

The key differences in the distribution of data lie in the logic of development, the nature of histograms and cumulative curves, and the influencing factors. A significant excess of the average value and a fashion among experienced scientists indicate natural professional development in the course of scientific activity. When examining histograms, it was found that for skilled scientists, the distribution is closer to normal, whereas for beginners, it shows a sharp peak at low values. Experienced scientists tend to show a smoother increase in cumulative values than beginners. In a stable environment with professional scientists, the correlation between indicators is higher, which reduces the likelihood of obtaining random, abnormally high results compared to beginners. Below is a comparative analysis of the relationships between the two groups of scientists, based on the constructed matrices. The overall trends in relationships across both data sets are similar, but there are essential differences in their strength.

The key findings of the analysis focus on the degree of determinism, modality for beginners and the importance of co-management.

In a dataset with experienced scientists, the mutual correlation between indicators is an order of magnitude higher. It means that, in a stable scientific environment, a scientist's success becomes more predictable and more directly dependent on specific factors.

Table 14. Comparison of correlation coefficients (R)

Comparison Options	Beginners (h20ns)	Experienced (h20ns)	Description and analysis
self% (Self-citation)	-0.02	-0.055	In both cases, the relationship is almost non-existent or slightly negative. It suggests that artificial citation winding has virtually no impact on the real h-index.
npsfl (Number of articles)	0.46	0.27	For beginners, the relationship between the number of publications and the Hirsch index is significantly stronger. For experienced scientists, the number of articles weighs less than their quality.
ncsfl (Article citations)	0.68	0.56	The citation rates of solo and co-authored works have a significant impact on the Hirsch index in both groups.
hm20 (Co-authorship index)	0.80	0.68	The strongest correlation. The index, which takes into account the share of contributions to collective work, is the most decisive factor in the overall h-index.

A lower correlation in beginners indicates that there are more random factors at the start of a career. However, the more experience a scientist gains, the less likely it is to achieve unexpectedly high performance without systematic work.

The high correlation with the hm20 parameter emphasises that, to achieve a high Hirsch index, it is not just the number of publications that is critical, but also the practical work in scientific groups and the quality of citations in collaborative projects.

A detailed analysis of the clustering results using the k-means method identified six main groups of scientists and countries based on two key parameters: the Hirsch index (h20ns) and the number of papers cited at least once (np6020cited2020ns). Below is a description of the obtained groups and their comparison for the two data sets.

For both data sets (beginner and experienced scientists), the same number of clusters was determined – 6. It allows for a direct comparison of the levels of scientific development. Most of the data (more than 185,000 records per dataset) was tightly clustered – concentrated in specific clusters (Clusters 1–5). It indicates the presence of many countries and scientists with similar, relatively "average" performance indicators.

The last cluster covers the smallest number of records (approximately 712) – the group with the most extensive spread (Cluster 6) - but has the broadest range of values. It is where "scientific leaders" end up – scientists and countries with extremely high indicators that are significantly detached from the general public. Although the general structure of 6 groups is maintained for both categories, the internal contents of the clusters differ:

Table 15. Features of clustering results

Characteristics of groups	Aspiring scientists	Experienced scientists
Compactness of groups	High density in the lower clusters, which supports the theory of a substantial gap between leaders and others.	A more uniform distribution within clusters, as scientists with experience tend to achieve higher figures.
Geographical division	Clusters clearly separate the leading countries (the USA, Norway, and Great Britain) from the countries of Africa and Asia.	The boundaries between clusters by country are becoming less sharp, indicating a global alignment of scientific potential over time.
Variability	The two main components account for 100% of the variability, indicating a direct dependence of success on the selected parameters.	The relationship is similar, but the centres of the clusters are shifted towards higher values of the Hirsch index.

Conclusions about clustering are based on parameters such as the hierarchy of the scientific world, environmental influences, and potential for development. Clustering confirms the existence of a "scientific elite" (Cluster 6), where scientists with the most decisive influence are concentrated, and an extensive database of researchers with stable but lower indicators. For beginners, belonging to a particular cluster critically depends on the country of residence. In experienced scientists, this dependence persists, but the boundaries between groups become increasingly blurred over time through long-term scientific practice. The large area of Cluster 6 in both cases suggests that this group is the most heterogeneous and can be further broken down into sub-clusters for a more detailed study of the industry's super-leaders.

Thus, the results of clustering clearly demonstrate that scientific success is not uniform: it is clearly segmented at the level (clusters), where reaching the "elite" level requires significant experience or a start in countries with high-quality education.

8. Discussion

As a result of the study, a set of empirical and statistical results was obtained that characterise the peculiarities of the formation of the Hirsch index (h-index) for scientists at different stages of their academic careers: novice and experienced scientists. The average h-index value for novice scientists is 14.59, while for experienced scientists it is 37.78, a 2.6-fold increase. The median and mode are also significantly higher in the sample of professional scientists (median = 34, mode = 30) than in the sample of beginners (median = 13, mode = 11). The maximum h-index values are much higher among experienced scientists (up to 280). In contrast, the minimum values in both samples are relatively close, indicating the possibility of low scientific performance regardless of seniority. The standard deviation for

experienced scientists is almost three times greater, indicating higher variability in results in this group.

Histograms and cumulative h-index curves showed fundamentally different distributions of data. Novice scientists are characterised by a high concentration of values in the lower range, as evidenced by a sharp peak in the histogram and a rapid increase in the cumulative curve. On the other hand, in the sample of experienced scientists, the distribution is smoother and more closely approximates a normal distribution. A group of leading countries (the USA, Great Britain, and Norway) stands out, while most other countries have significantly lower average h-index values. For experienced scientists, this gap is decreasing, indicating a gradual alignment of scientific results with experience.

The application of smoothing methods (Kendel, Pollard, exponential, and median) showed that the general form of trend changes in indicators is similar across both samples. Turning points coincide, which indicates common patterns in the development of scientific activity. At the same time, experimental results confirm that the differences between the samples are not in the nature of the trends. Still, in the levels of the values, smoothed curves for experienced scientists are stably higher than those for novice scientists.

Correlation analysis enabled the quantification of the relationship between the h-index and other scientometric indicators. The relationship between the h-index and self-citation is weak or negative in both samples, suggesting a negligible influence of this factor on overall scientific performance. On the other hand, a stable positive correlation exists between the h-index and citation indicators of publications ($r = 0.68$), with the strongest relationship observed between the h-index and the hm-index, which accounts for co-authorship (r up to 0.80). For experienced scientists, the correlation structure is more pronounced, indicating greater determinism in scientific results and a smaller role for random factors.

The use of the k-means method with $k = 6$ clusters enabled the identification of a hierarchical structure in scientific performance. Five clusters are characterised by high density and cover the bulk of scientists with similar h-index indicators. A separate sixth cluster has a significant range of values and represents a group of scientific leaders with ultra-high indicators. It is more closely related to the country in which the scientific activity takes place. At the same time, for experienced scientists, there is a partial alignment of the cluster structure, indicating an increase in the role of individual scientific contributions over time.

Summarising the results, it can be argued that the Hirsch index reflects not only the number and citations of publications, but also the complex interactions among individual, collective, and geographical factors. The results of the study confirm the presence of apparent differences in the formation of scientific effectiveness at various stages of the academic career and justify the expediency of an integrated approach to analysing scientometric indicators.

The study identified several key patterns regarding the relationship between citation metrics, including the impact of self-citation, the absence of linear relationships, and the results of neural network analysis. An almost complete linear correlation ($r = 1$) was revealed between citation parameters (number of citations, h-index) with and without taking into account self-citation. It indicates that on a general scale, self-citation does not significantly distort authors' ratings. The analysis of the heatmap showed that the percentage of self-citations has almost no linear dependence on other parameters of the authors in the dataset. Attempts to build models (perceptrons) to predict self-citation levels using other indicators were unsuccessful – none achieved sufficient accuracy. It supports the hypothesis that self-citation is an individualised factor that depends on the psychological or moral aspects of the author, rather than on their quantitative indicators.

The study confirmed a strong relationship between a country's economic state and its scientific productivity, as measured by GDP and publications, as well as corruption (as indicated by the CPI) and the specificity of its industries. A direct linear relationship has been established between a country's GDP level and the number of published articles. The largest economies in the world (the USA, China, and Great Britain) are leaders in all major indicators. A weak relationship between corruption levels and self-citation was revealed. In countries with higher CPI (Lower Corruption) indices, there is a tendency towards lower self-citation. The highest level of self-citation was recorded in the fields of physics, mathematics and information technology.

Ukraine's positions in the global context are analysed separately:

– Number of articles – 49th place in the world (about 10,029 articles by popular authors).

– The citation rate is 65th, which puts Ukraine next to Peru and Hungary.

– Self-citation – Ukraine is among the top 5 countries in this indicator, ranking 4th with an average level of about 34%.

– h-index – Ukraine ranks 68th with an indicator of 11 (calculated as the h-index of the h-indices of the authors of the country).

Compared to its closest neighbours and partners, the results are as follows:

Table 16. Comparison with closest neighbours and partners

Country	Number of articles	H-index	Features
Poland	1st place in the sample	The highest in the group	The leader in all positive ratings.
Ukraine	3rd place in the sample	11 (minimum among 10 authors)	The highest level of self-citation in the group.
Russia	-	Low relative to neighbours	An abnormally high level of self-citation for its volume of publications.

In the context of international scientometrics, our results demonstrate that h-index analysis, when combined with additional indicators (e.g., h-m, correlation analysis, clustering), provides a more comprehensive picture of scientific activity and impact. It aligns with current industry trends, in which researchers recommend integrating multiple approaches and indicators to achieve a more informed assessment of performance and impact.

One key observation is the presence of a secondary subdiagonal with a correlation coefficient approaching unity. This linear cluster of points reflects a strong relationship between citation-related indicators – such as total citations, h-index, hm-index, and similar metrics – calculated with and without self-citation adjustment. The presence of this relationship suggests that, in most cases, self-citation does not substantially affect an author's overall performance and is not systematically exploited as self-plagiarism or as a mechanism to artificially inflate academic rankings.

Another important consideration concerns the relationship between the proportion of self-citations and other variables. The final column of the correlation matrix in Fig. 19 is almost entirely unshaded, indicating near-zero correlations. It suggests the absence of a linear dependency between self-citation rates and the remaining parameters included in the dataset.

The hypothesis of a lack of a meaningful association between self-citation and the available dataset attributes is further examined by constructing neural network models and analysing their outputs.

The subsequent phase of the analysis focuses on the distribution of citation-related metrics and publication output across countries. Prior to this, a data window is constructed according to the procedure described in the previous section.

Initially, the statistical distributions of the selected dataset variables are analysed and visualised using joyplots. Due to the substantial differences in scale and value ranges among the number of publications, h-index values, self-citation percentages, and citation counts, all corresponding variables are normalised. As illustrated in Fig. 20, these distributions exhibit comparable patterns, characterised by positive skewness and excess kurtosis.

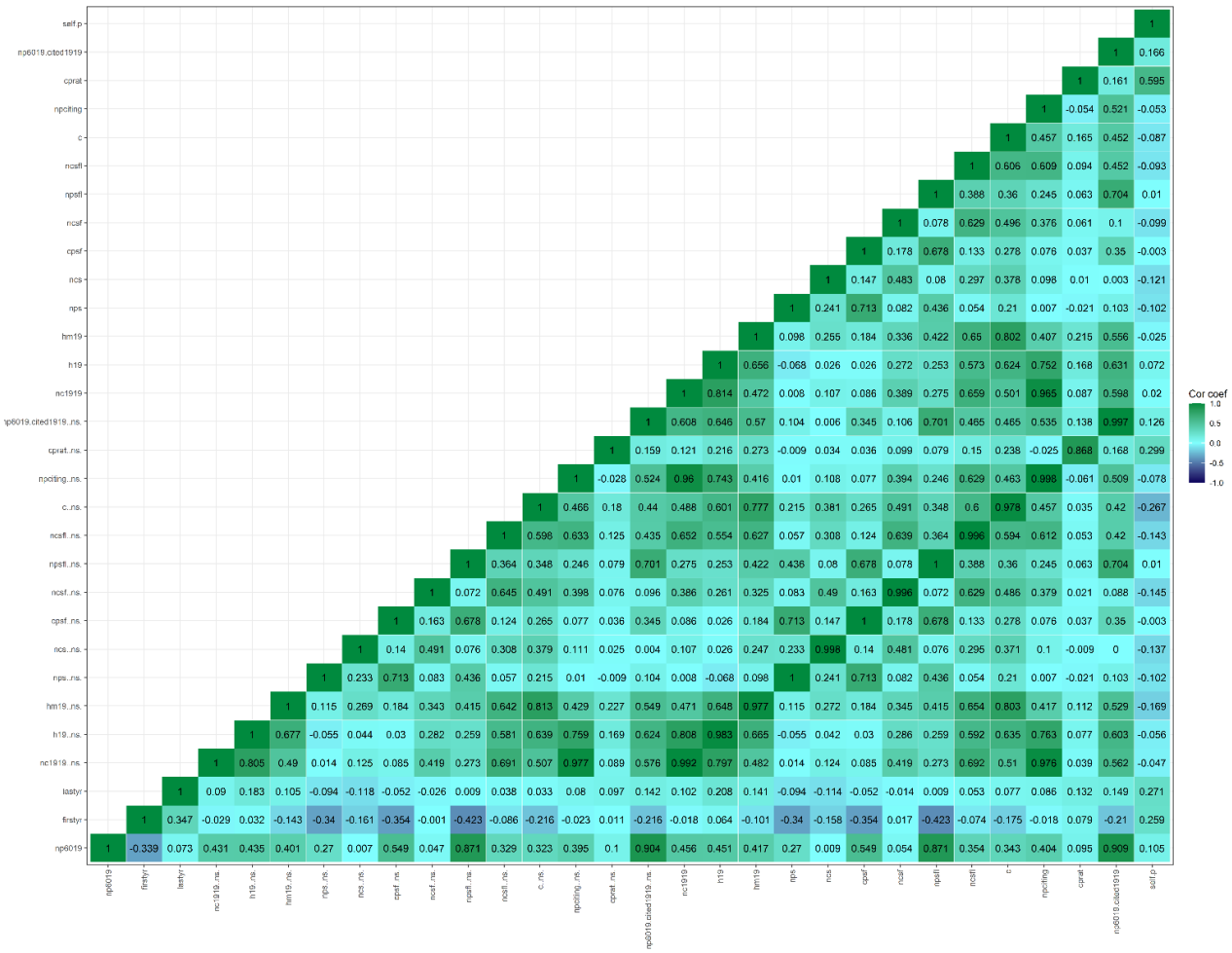

Fig. 19. Correlation matrix heatmap

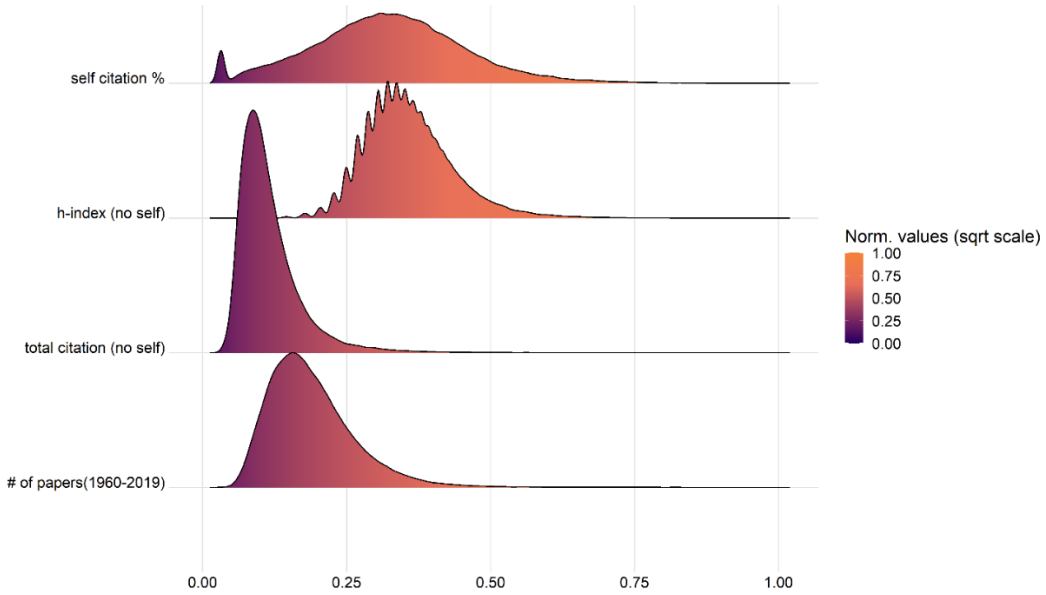

Fig. 20. Distribution diagram

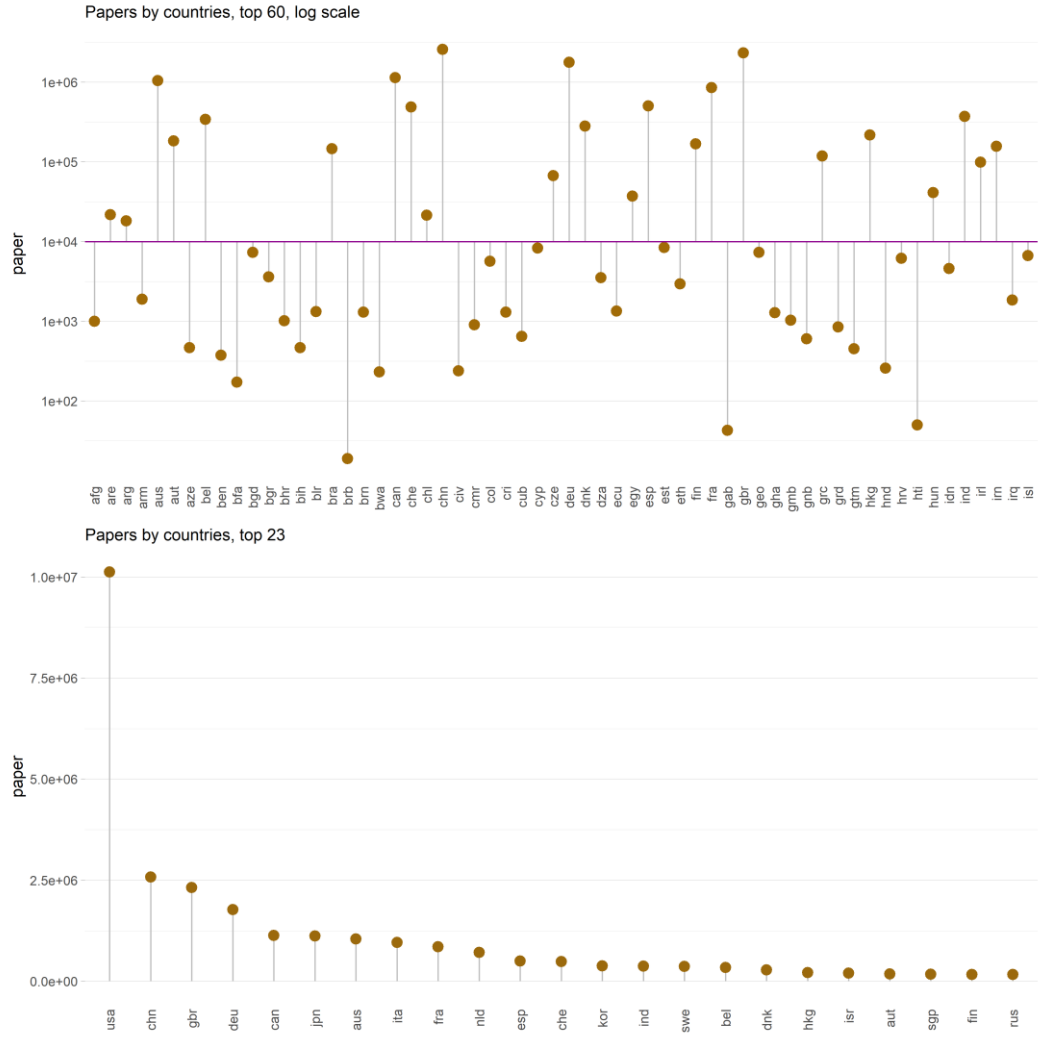

Fig. 21. Distribution of the number of articles published during 1960–2019 by country, temporal robustness of country-level differences in h-index trajectories. Smoothed time-indexed trajectories demonstrate that early-career geographical effects attenuate over career progression.

It is worth noting that the correlation coefficient between the average self-citation percentage derived directly from the dataset and that computed from grouped data is 0.95, confirming the reliability of the averaging procedure.

The analysis then examines country-level distributions of individual parameters. Figure 21 presents the top 60 countries (unsorted) and the top 23 countries ranked by total publication output. The results demonstrate a clear dominance of the United States, which surpasses China – the second-ranked country – by nearly an order of magnitude. Moreover, the top 23 positions are predominantly occupied by economically advanced nations or countries with large economies. This observation motivates the hypothesis that a country's economic development is associated with its scientific output, with gross domestic product (GDP) serving as a proxy for financial performance.

In addition to GDP, a corruption index was incorporated to explore potential dependencies between corruption levels – particularly within educational and research systems – and the excessive use of self-citation by authors.

Another parameter analysed is the total number of citations received by authors from each country. Figure 22 illustrates the top 60 countries, the top 23 countries, and Ukraine's relative position. Each lollipop marker includes two values: one for self-citations and another for those excluded. The visual proximity of these values already suggests a strong correlation between the two measures.

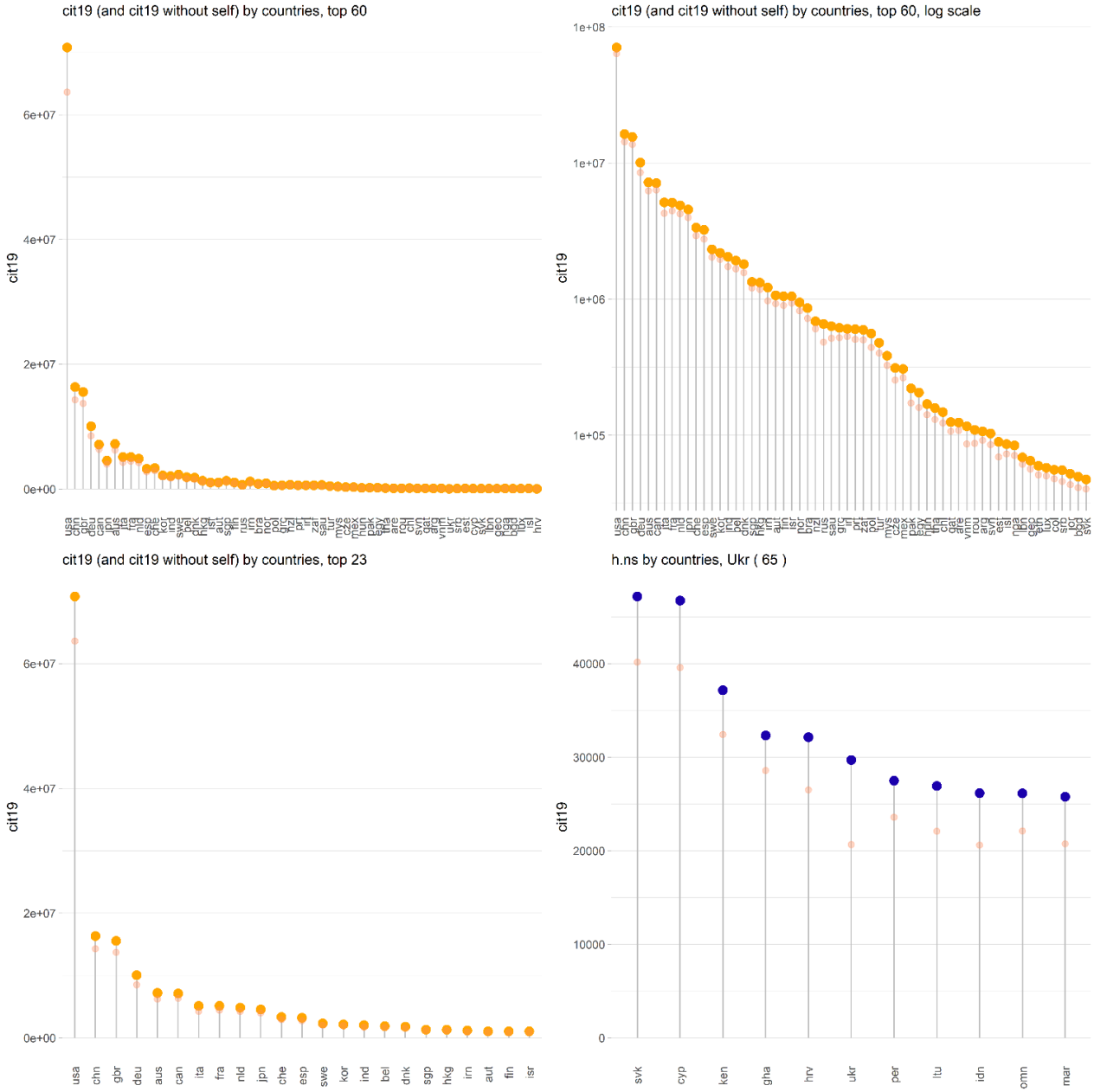

Fig. 22. Distribution of the number of citations of authors by country in 2019

Ukraine ranks 49th in terms of publication count, with 10,029 articles. This relatively low value is attributed to the dataset's partial nature, which includes only the most prominent authors worldwide.

As expected, the United States ranks first in total citations, followed by China, the United Kingdom, and other leading countries. Ukraine is positioned 65th, adjacent to Hungary and Peru.

The distribution of self-citation percentages by country is analysed next, with the corresponding visualisation presented in Fig. 23. Notably, Ukraine ranks fourth among the top five countries for this metric, with an approximate self-citation rate of 34%.

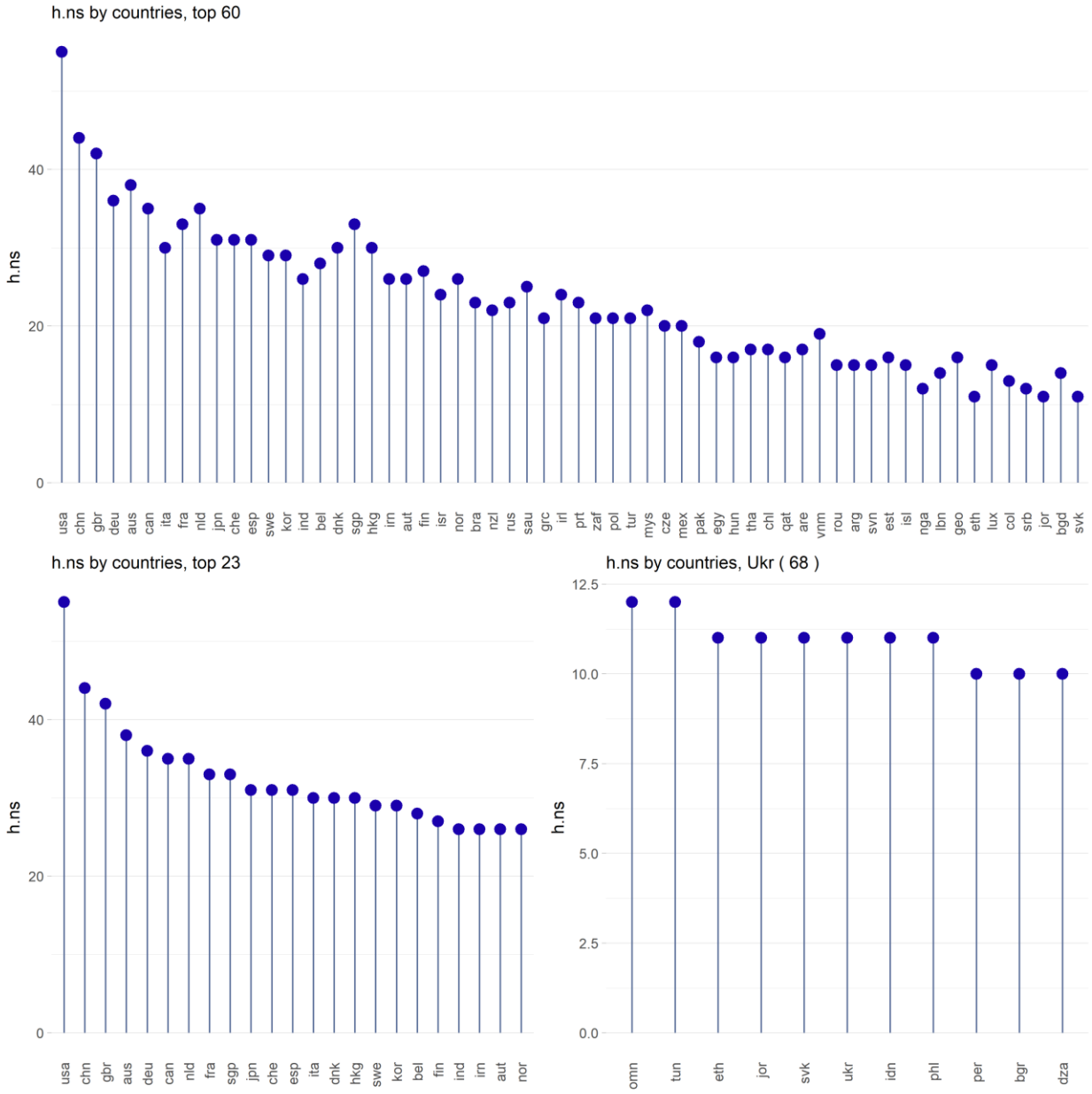

Fig. 23. Leading countries by the h-index of authors' h-indices (excluding self-citation)

Conversely, the lowest self-citation levels are observed in small Caribbean states, including Barbados and Saint Lucia. The absence of a clear visual association between self-citation rates and national economic indicators further supports the assumption that GDP may not be correlated with self-citation intensity.

A particularly informative approach is applied when analysing the h-index at the country level. Given that h-index values lie at the intersection of nominal and ratio data types, computing a simple average is methodologically inappropriate. Instead, a composite country-level h-index is derived by calculating the h-index of the individual authors' h-indices within each country. The algorithm used to construct this indicator is detailed in the preceding section.

The first chart in Fig. 23 is ranked according to the total number of citations recorded in 2019. For the top 60 countries, a noticeable association between these indicators can be observed. Consistent with previous findings, the United States leads, with a country-level h-index of 55. Ukraine ranks 68th with an h-index of 11, in a similar position to several other countries. To visualise the global geographical distribution of the analysed indicators, a comprehensive

representation is provided in Fig. 24. The interpretation of this figure is relatively straightforward. Countries are ordered according to the number of published articles, which is additionally encoded by the colour of the lollipop markers. Each marker consists of two segments: the larger segment corresponds to the country's h-index calculated from authors' h-indices while accounting for self-citations, whereas the smaller segment represents the same metric excluding self-citations. The size of each lollipop reflects the proportion of self-citations.

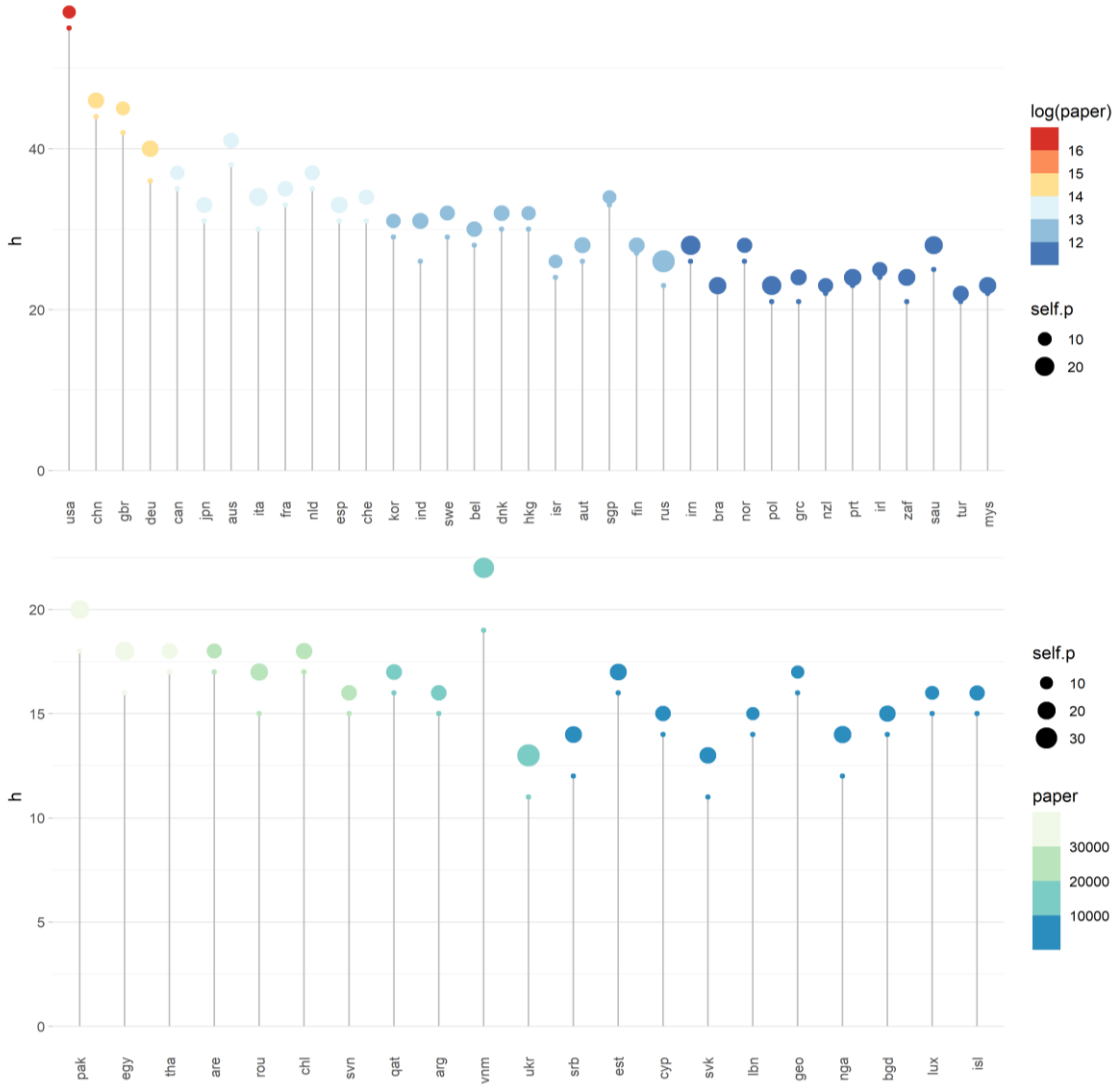

Fig. 24. Aggregated country-level representation of the analysed parameters

Several noteworthy observations emerge from this visualisation. For example, although Russia appears among the top 23 countries in terms of publication output, its self-citation rate is markedly higher than that of most other countries. At the same time, Russia demonstrates comparatively low h-index values relative to neighbouring countries, which puts it at a disadvantage in this dimension. Following the graphical exploratory analysis, GDP and Corruption Perceptions Index (CPI) data were integrated into the dataset. The initial step involved performing a correlation analysis, the results of which are illustrated by the heatmap in Fig. 25.

The heatmap confirms the existence of a linear relationship between country-level h-indices, thereby supporting one of the previously formulated assumptions. Furthermore, a strong association is observed between publication counts and citation counts, regardless of whether self-citations are included; notably, all correlation coefficients involving self-citations are negative. The strongest relationship, in terms of absolute magnitude, is between the self-citation rate and the corruption index. It indicates a weak linear dependency, suggesting that higher CPI values – corresponding to greater institutional transparency – are associated with lower levels of excessive self-citation. Nevertheless, this relationship is insufficient to claim a direct causal link.

Another hypothesis addressed the relationship between national GDP and scientific output. The correlation matrix reveals a linear association between GDP and citation counts. To further investigate this dependency, a more detailed analysis was conducted using scatter plots that depict correlation fields between GDP and citation numbers for subsets of 40 countries, 90 countries, and the full dataset. As shown in Fig. 26, the correlation coefficient decreases substantially with smaller sample sizes. However, when large contributors such as the United States, China, and the

United Kingdom are included, the coefficient increases and converges toward the value observed in the correlation matrix.

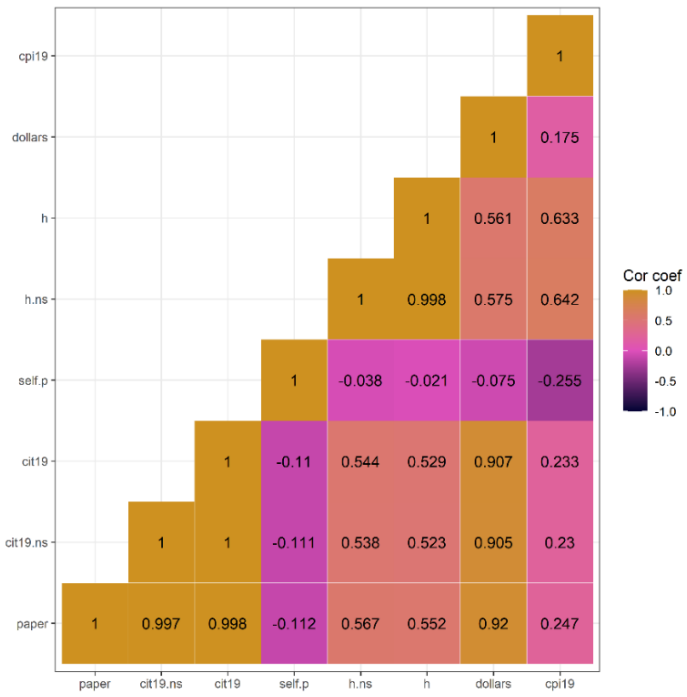

Fig. 25. Heatmap of correlations among country-level parameters

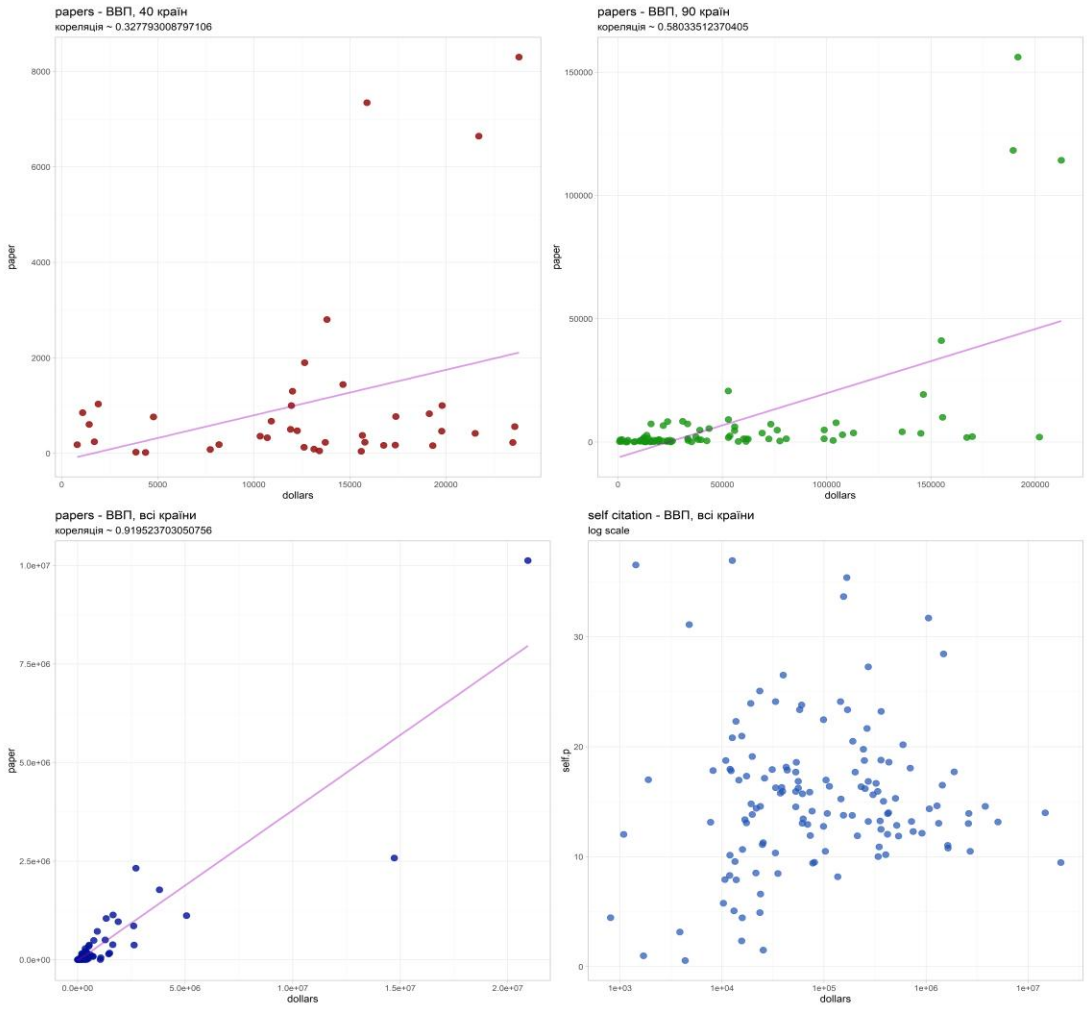

Fig. 26. Collection of scatter plots illustrating the relationship between GDP and (self-)citation indicators

The fourth plot in Fig. 26 examines the relationship between self-citation rates and GDP. The resulting scatter field appears highly dispersed and lacks a discernible pattern, suggesting no meaningful dependency. In addition, a supplementary analysis was performed to investigate the unexpectedly weak linear relationships between the country-level h-index and both GDP and CPI. A comparison of the corresponding correlation fields is presented in Fig. 27.

For the first three scatter plots in Fig. 26, regression lines were fitted using the linear model (LM) function. Similar regression fittings were applied to the scatter plots shown in Fig. 27.

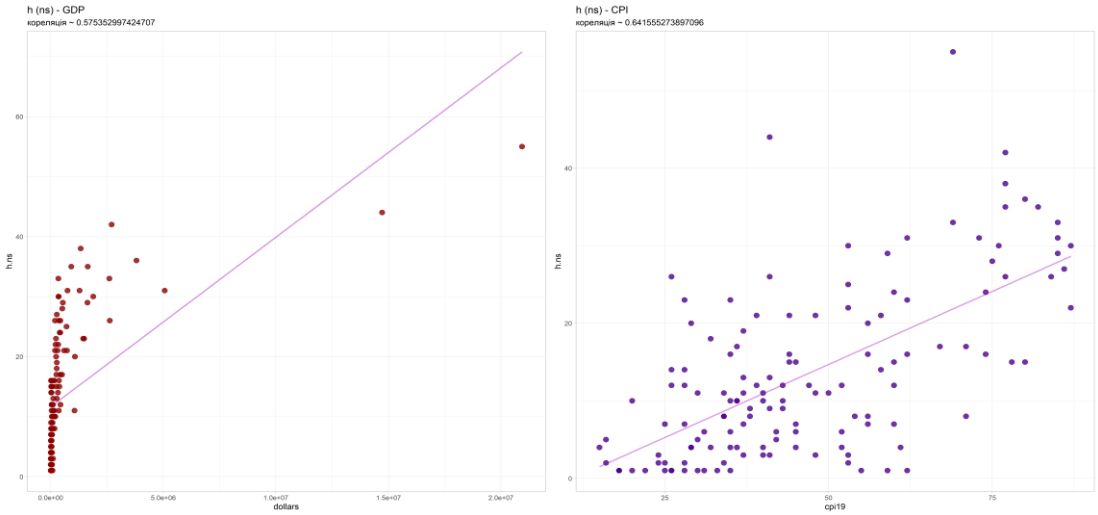

Fig. 27. Comparison of correlation fields between a country's h-index and GDP, CPI

The final focused analysis derived from the correlation matrix involves a regression study of the relationship between the total number of citations and the country-level h-index. The scatter plot in Fig. 28 suggests that the most appropriate approximation is exponential. The methodology for constructing this model is described in the preceding section.

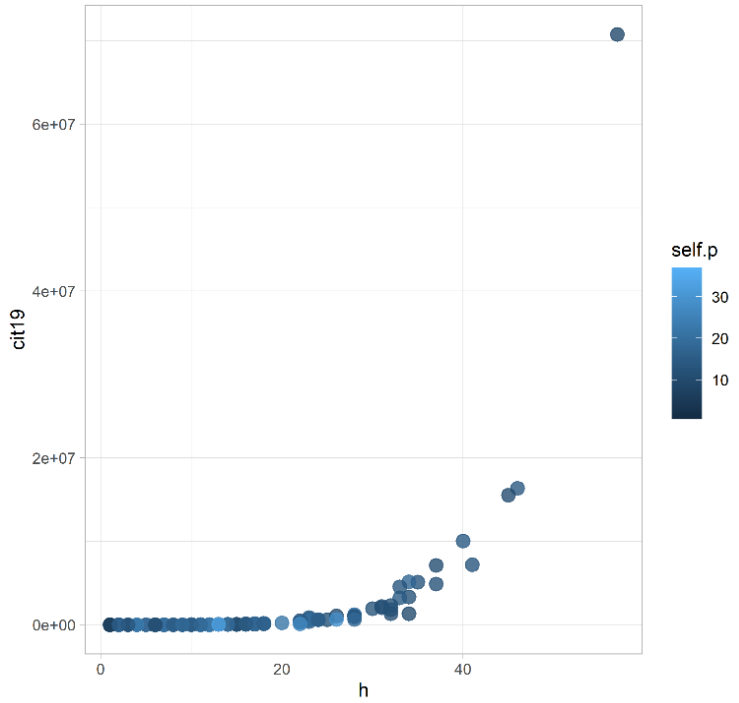

Fig. 28. Correlation field

Before presenting the outcomes of the neural network models, an additional observation regarding the distribution of self-citations is introduced, this time analysed across scientific domains defined by the primary category assigned to each author in the dataset.

Analogous to the procedure applied at the country level, a data window was created by grouping records by both category and country. Subsequently, a three-dimensional histogram was generated, with the third dimension corresponding to the self-citation percentage and encoded by colour intensity.

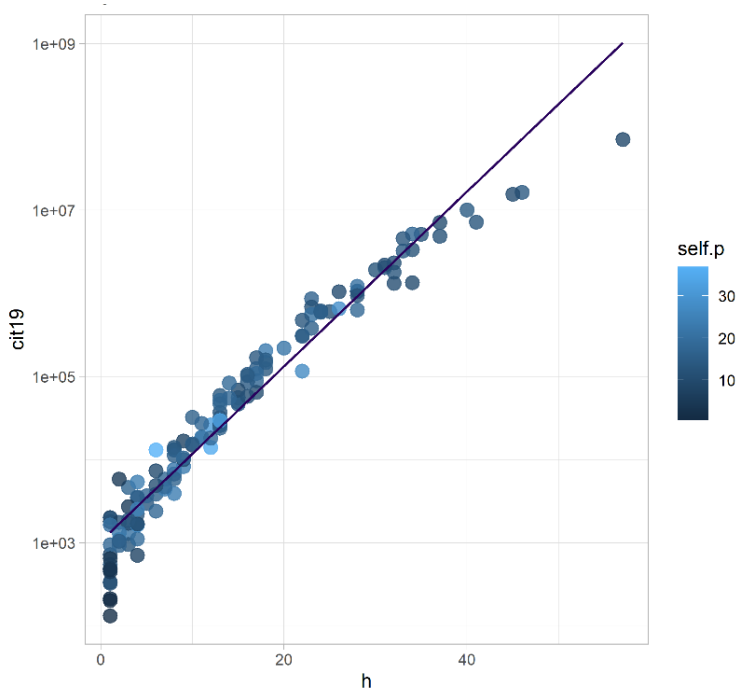

Fig. 29. Correlation field with an approximation line after logarithmic transformation of citation counts. The correlation coefficient between the data and the prediction = 0.955806051395733.

As the complete histogram does not provide sufficient clarity for detailed visual interpretation, a reduced subset of countries was selected, and a separate visualisation was constructed for this sample, as presented in Fig. 31.

Figure 31 shows that the highest proportion of self-citation occurs in the fields of mathematics and statistics in Mexico. However, when these results are compared with those shown in Fig. 18, it becomes evident that the overall volume of self-citations is predominantly concentrated in the domains of physics and information technology.

A supplementary part of the analysis focuses on the geo-economic distribution of the analysed parameters within a selected group of countries, namely Ukraine, Georgia, Poland, Romania, and Slovakia.

According to Fig. 32, Poland has the highest number of published articles, while Ukraine ranks third. Subsequent visualisations reveal that Poland leads across most evaluated indicators, except for self-citation rates, where Ukraine, regrettably, ranks first.

Based on the final chart, the dataset includes 10 Ukrainian publications with an h-index of at least 10.

An aggregated overview of the examined country-level characteristics is presented in Fig. 35.

In addition, a dedicated visualisation was developed to illustrate the distribution of self-citation percentages across the selected countries, disaggregated by authors' primary scientific categories. The resulting three-dimensional histogram is shown in Fig. 36.

For this mini-study, a specific sample of countries was selected: Ukraine, Poland, Georgia, Slovakia and Romania. The analysis is based on a comparison of their scientometric indicators (number of articles, citations, h-index and self-citation). According to the distribution graph (Figs. 32-33), Poland is the undisputed leader. It has the most published articles and the highest total citation count among authors in 2019 within the entire group. Ukraine ranks 3rd in terms of the number of published articles in this sample. Georgia, Slovakia, and Romania have lower publication activity rates than Poland and Ukraine.

The analysis of the ranking of countries' h-indices (Fig. 34) shows that Poland again occupies the top position, indicating the high quality and stability of its authors' scientific work. Ukraine is represented in this ranking by 10 authors included in the dataset. Their minimum h-index is 10 (the text also mentions the number 11 as the average for the country in another section). It is worth noting that, although Ukraine ranks third in the number of articles, its h-index remains relatively low compared to the group leaders.

Fig. 30. Global histogram

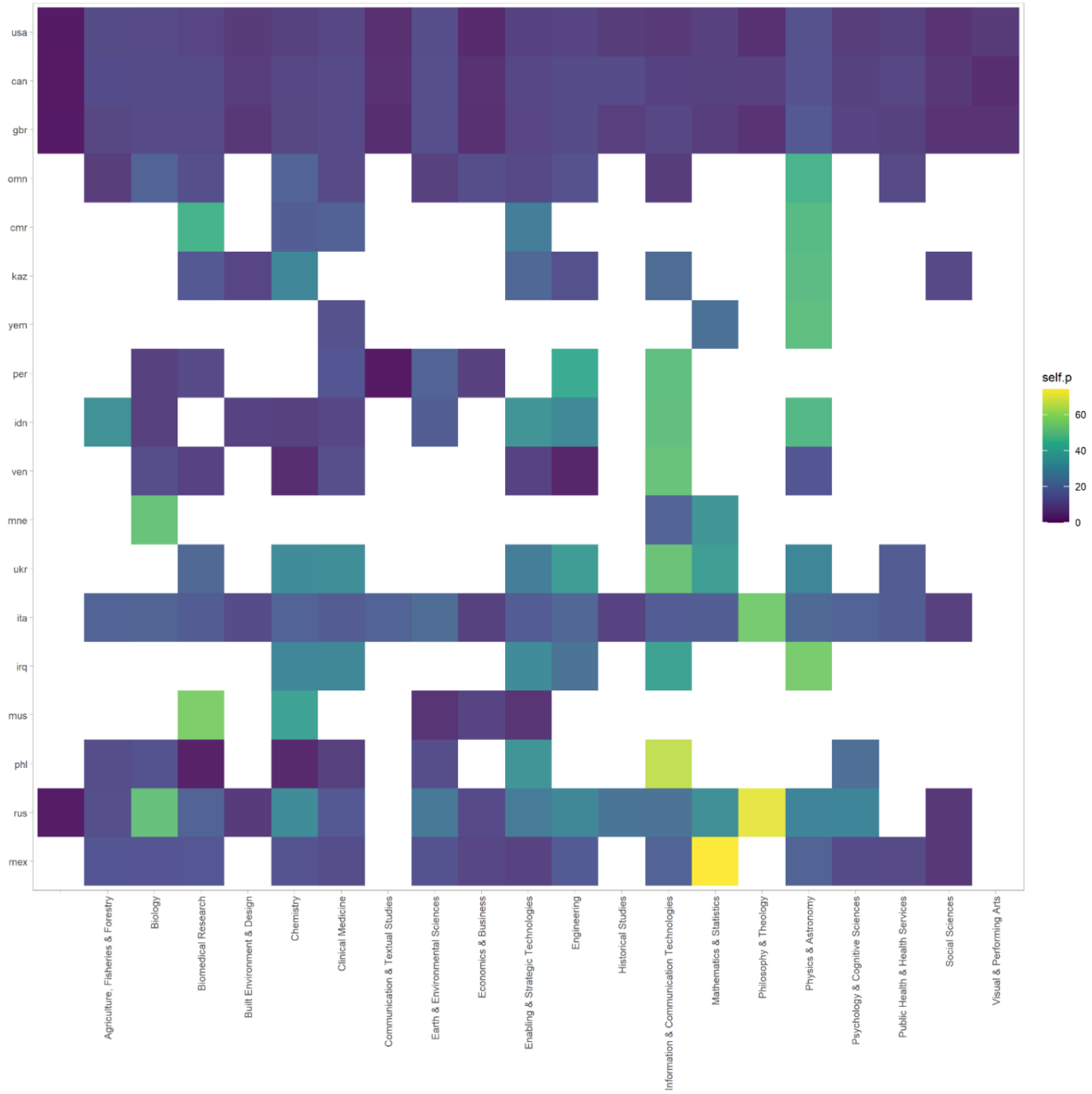

Fig. 31. Reduced version of the histogram

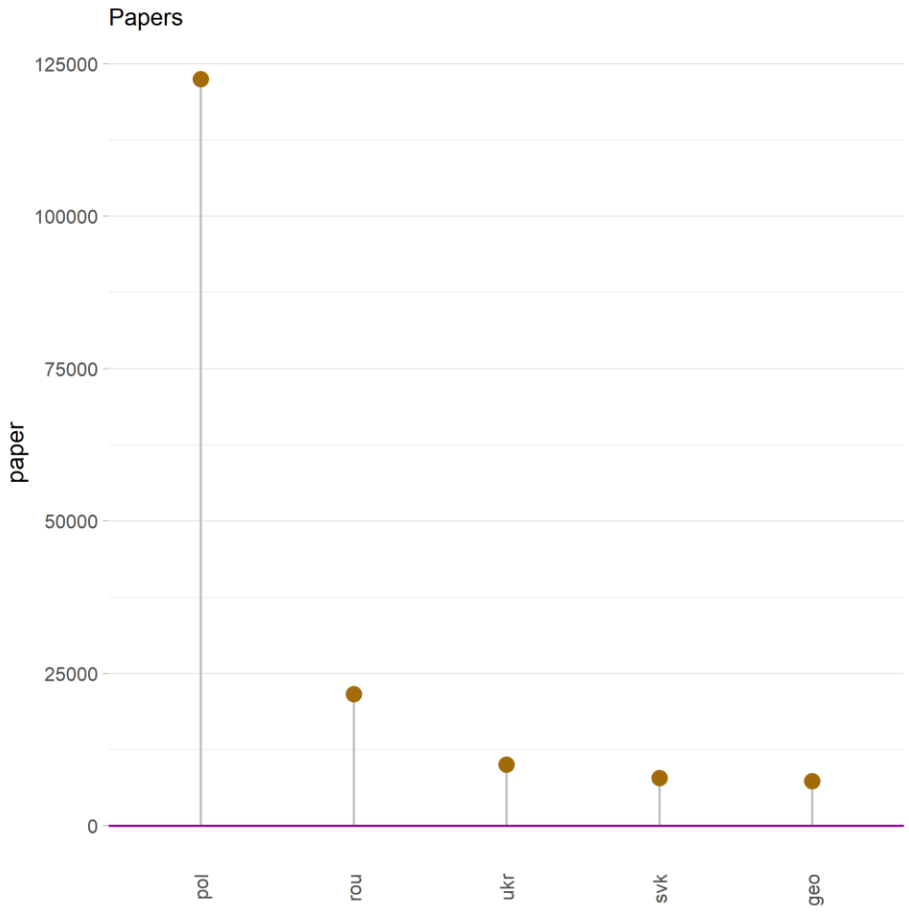

Fig. 32. Lollipop chart illustrating the distribution of publication counts

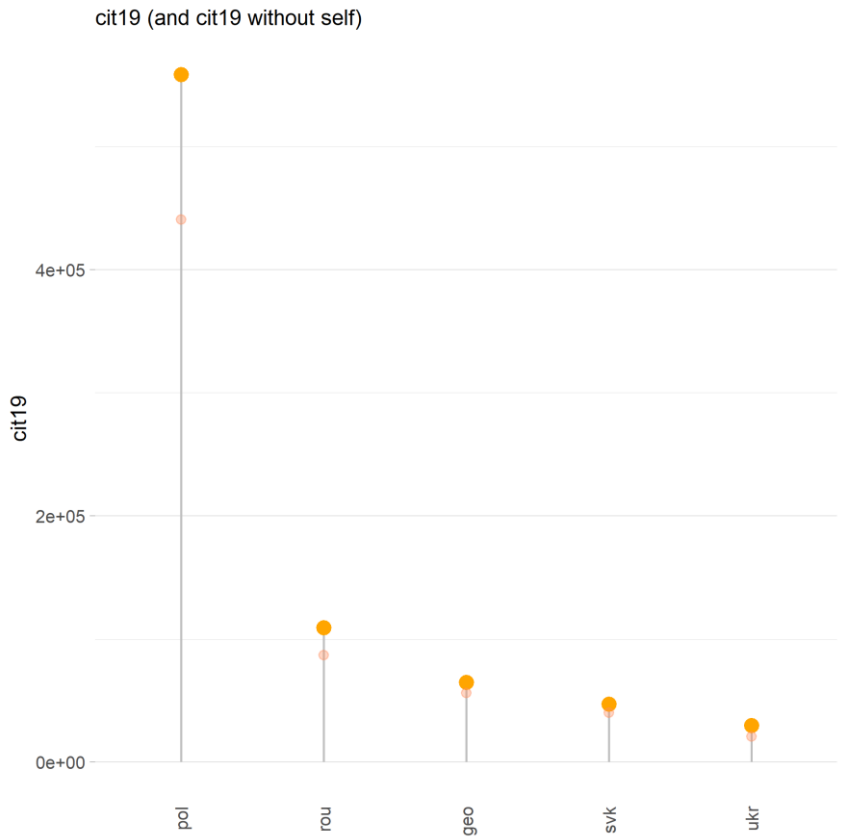

Fig. 33. Distribution of citation counts for 2019

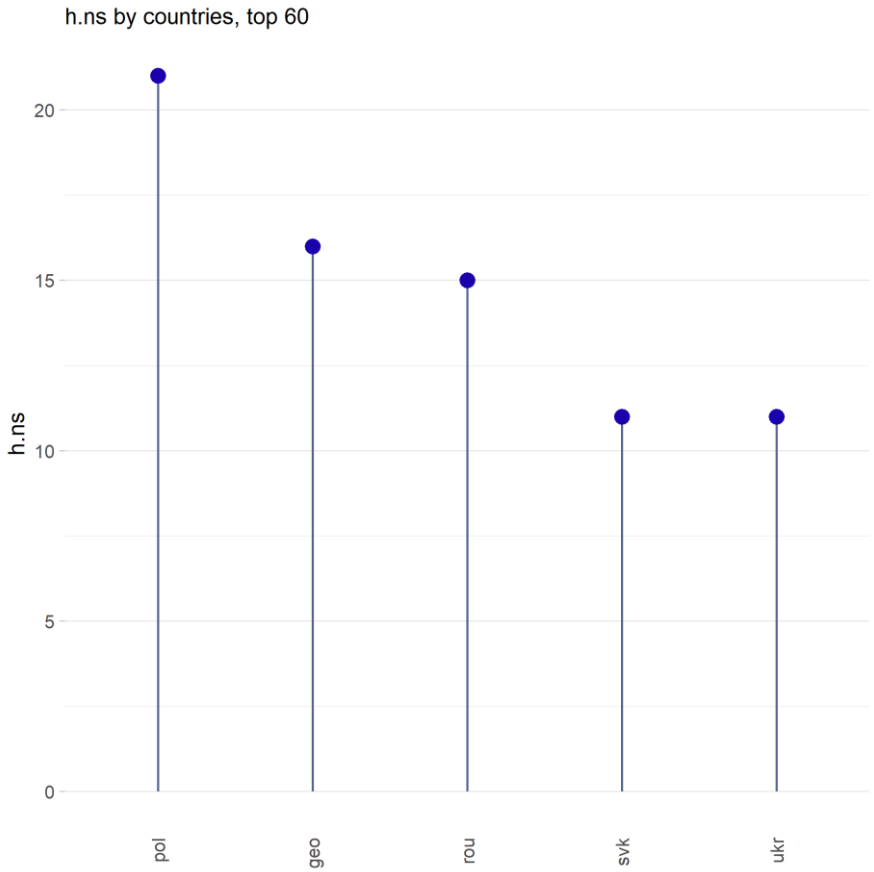

Fig. 34. Country-level h-index ranking

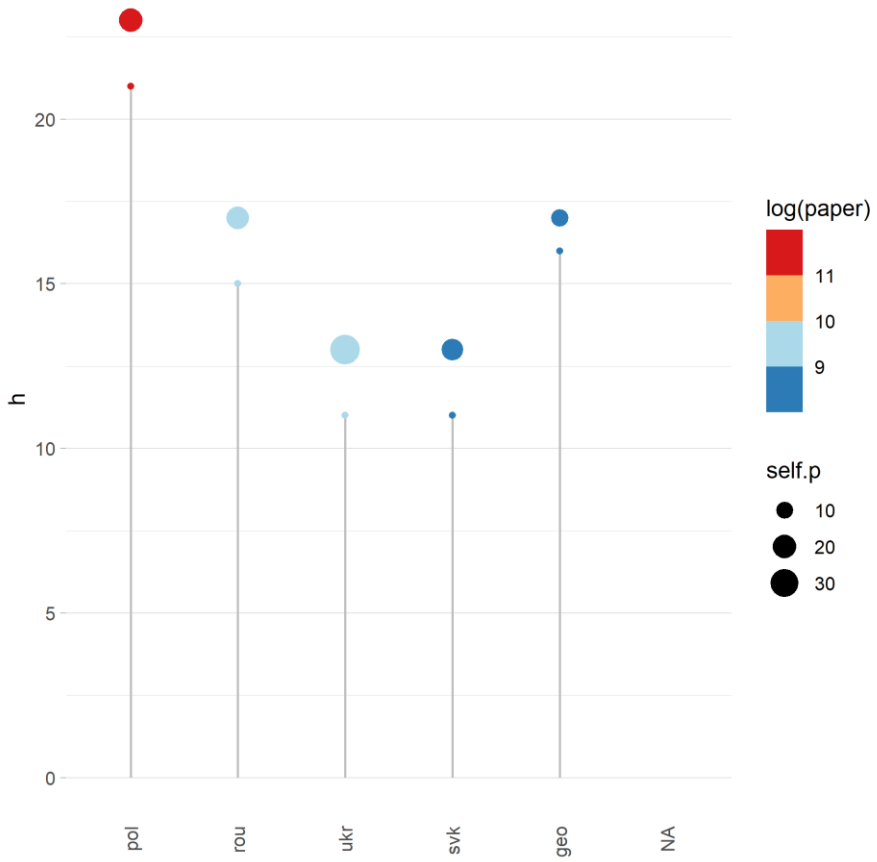

Fig. 35. Consolidated overview of country-specific indicators

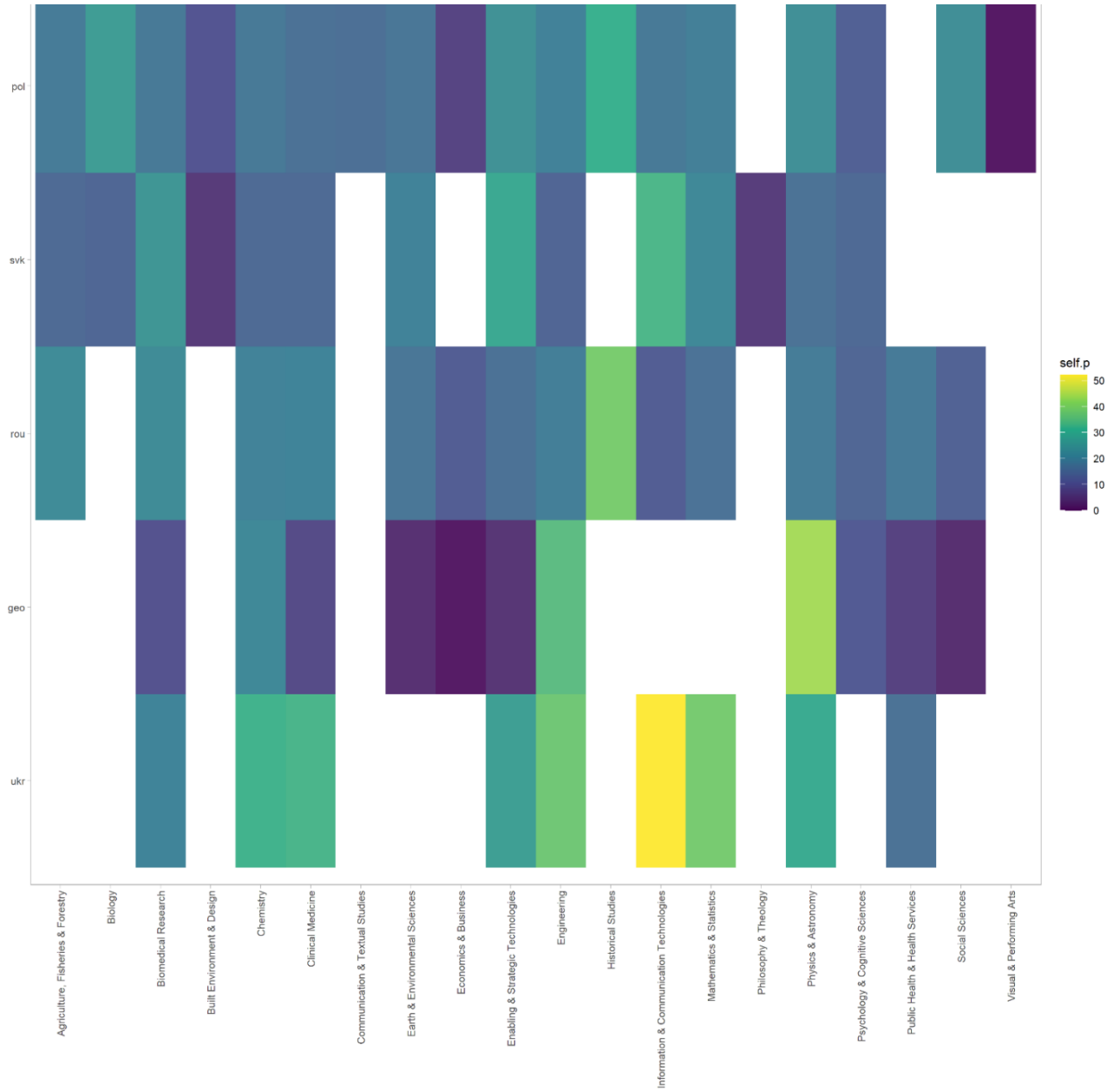

Fig. 36. Three-dimensional histogram of self-citation distribution by country and research category

Self-citation is the most critical indicator for Ukraine in this study. While Poland leads in positive metrics (articles, citations), Ukraine ranks first in terms of self-citation among all the countries compared. In a global context (if we take the whole world), Ukraine is in the Top 5 countries in terms of the percentage of self-citations (4th place in the world). A significant part of the citations of Ukrainian scientists is created by themselves or within narrow local groups, which may artificially inflate quantitative indicators but does not reflect real-world recognition.

Summarising the data for these countries reveals a connection between the economy (as measured by GDP and Corruption). Poland's high scores correlate with higher GDP and better Corruption Perceptions Index (CPI) scores. The author associates a high level of self-citation in Ukraine with lower funding for science and higher corruption risks, where academic integrity may be less well controlled.

Ukraine has significant potential in terms of the number of publications (ahead of Romania and Georgia), but the "quality" of this success is offset by an abnormally high self-citation rate, which distinguishes it from its neighbours.

The results obtained enable us to comprehensively interpret the regularities in the formation of the Hirsch index across various stages of the academic career and to assess the influence of key scientometric factors on scientific performance. The analysis confirms that the h-index is a cumulative indicator that depends significantly on the duration of scientific activity; however, its growth is determined not only by time but also by the structure of scientific interactions and the intensity of citations.

Firstly, a comparative analysis revealed that the average h-index of experienced scientists is approximately 2.6 times higher than that of novice scientists. This finding is consistent with the classical view of the cumulative nature of

the Hirsch index; however, the results also demonstrate that the distributions for both groups are similar. At the same time, the differences between the groups are mainly due to the scale of the values.

Second, the results of smoothing and trend analysis indicate the synchronicity of turning points in the dynamics of indicators for both samples. It suggests that external factors, such as changes in scientific policy, publishing conditions, or general trends in scientific development, affect both novice and experienced scientists equally. At the same time, the consistently higher level of smoothed curves in the sample of experienced scientists emphasises the role of accumulated scientific capital.

Third, the correlation analysis revealed that the co-authorship (hm) indicator has the most significant effect on the h-index. High values of the correlation coefficient (up to 0.80 for novice scientists) indicate that integration into scientific collaborations is a key factor in increasing scientific visibility. It is essential in the early stages of a career, when participation in collaborative research can compensate for the limited number of individual publications. A negative relationship between the H-index and the proportion of self-citation confirms that artificially increasing indicators does not significantly affect long-term scientific performance.

Fourthly, the clustering results allow us to view the scientific space as a hierarchical structure with a clearly defined group of scientific leaders. The identification of a separate cluster with extremely high h-index values suggests a concentration of scientific influence within a relatively narrow circle of researchers. At the same time, for novice scientists, there is a stronger dependence of cluster affiliation on the country, which may be due to differences in access to resources, infrastructure and international scientific networks.

The results obtained are consistent with modern English-language research in scientometrics, which emphasises the role of collective scientific activity and the uneven distribution of scientific influence. At the same time, unlike works that focus exclusively on the absolute values of the Hirsch index, the study demonstrates the feasibility of an integrated approach combining statistical analysis, smoothing, correlation methods, and clustering.

From a practical point of view, the results obtained can be used to develop more reasonable criteria for evaluating scientific activity, in particular by taking into account the career stage and the role of collaborations. It is especially relevant for information systems supporting decision-making in the fields of science and education, which aligns with the International Journal of Information Engineering and Electronic Business's profile.

At the same time, the study has certain limitations. Firstly, the analysis is based on aggregated scientometric indicators and does not account for the sectoral features of scientific activity. Secondly, using the h-index as the primary indicator does not fully capture the qualitative aspects of scientific contributions. Further research can aim to integrate alternative metrics, time models, and machine learning methods for a deeper analysis of scientific performance.

The results confirm key trends in the development of scientific activity and the formation of the h-index, as reflected in modern English-language works on scientometrics.

Our results showed a weak or negative relationship between h-index and self-citation, consistent with critical assessments of this indicator: the author in [3] notes that h-index is an unreliable measure of impact, especially for experimental disciplines, because it does not take into account qualitative aspects of citation and can be distorted by the publication activities of certain groups of researchers. In addition, other studies [17] highlight that the h-index cannot adequately reflect a scientist's reputation in the long term due to the variable relationships among the number of publications, citations, and scientific achievements, a point supported by approaches to scaling scientometric indicators. These criticisms partially explain the results of our correlation analysis: we observe that the h-index correlates well with the hm score, which considers co-authorship, whereas the standard h-index alone is insufficient to assess impact without the context of collective activity.

Several English-language papers employ the h-index in broader scientometric reviews, comparing it with other indicators and analysing patterns at the levels of disciplines and scientific communities. For example, extensive bibliometric studies in health care and other disciplines demonstrate that the h-index among top authors or journals can reach values significantly higher than the industry average, yet without a consistent interpretation of citation context [18]. It correlates with our finding of a single cluster with a very high H-index in the sample, similar to how global surveys show that a small fraction of researchers account for the vast majority of citations and impact.

Other comparative studies, such as those comparing the h-index in Google Scholar and Scopus for specific groups of researchers, demonstrate that different platforms can differ significantly in metric values and their interpretation, which reinforces the importance of understanding the context of data sources [19]. Although our study utilises a large amount of data without referencing a specific database (such as Google Scholar, Scopus, or others), accounting for these differences is essential for future research, as the data source can affect the absolute H-Index values across groups.

Our results demonstrate general patterns that confirm and complement the conclusions of other researchers. The division into groups by h-index and the identification of a cluster of scientific leaders are consistent with the notion that scientometric indicators are hierarchical in nature and often single out a narrow group of influential authors from the broader pool of publications, as seen in extensive scientometric studies [20].

At the same time, critical positions on the limitations of the h-index, as presented in modern English-language studies, reinforce the need to use complex metrics [21-24] that address co-authorship, disciplinary differences, and structural dependencies – an approach that is also implemented in our work.

It is acknowledged that strong correlations observed between the h-index, citation counts, and co-authorship–adjusted indices (e.g., hm-index) may appear unsurprising, as these measures are mathematically related. However, the presence of such dependencies does not invalidate the study's analytical novelty. On the contrary, the novelty lies not in

the existence of correlations per se, but in their structural role, comparative behaviour, and stability across career stages and analytical transformations. Let $h = f(C), \quad hm = g(C, A)$, where C denotes citation counts, and A represents co-authorship structure. While h and hm are indeed functions of overlapping variables, functional dependence does not imply equivalence of information content. In particular:

– The h-index is a threshold-based, order-statistic functional of the citation distribution;

– Citation counts are additive magnitude-based measures;

– The hm-index introduces fractional weighting that alters the effective contribution of collaborative publications.

Thus, although correlated, these indicators encode distinct transformations of the same underlying data, analogous to how variance, entropy, and quantiles may all derive from a single distribution yet capture different structural properties.

In a mathematically coherent scientometric system, strong positive correlations between citation-based indicators are a necessary condition, not a methodological flaw. The absence of such correlations would indicate one of the following pathologies: statistical noise dominating the signal, inconsistency between indicators, or invalid metric construction. Therefore, the observed correlations serve as a sanity check, ensuring that adjusted indices (e.g., hm) preserve the core signal of scientific impact while modifying its allocation mechanism. The analytical novelty of the study does not rely on treating h-index, citations, and hm-index as independent random variables. Instead, it emerges from demonstrating that:

1. The strength of correlations varies systematically across career stages, indicating different regimes of scientific production and collaboration;

2. Correlation structures differ in magnitude and stability, even when the same functional relationships hold; Adjusted indices (hm) exhibit stronger explanatory power for clustering and stratification, despite being derived from the same citation base.

Formally, if $corr(h, C) \approx corr(hm, C)$, but $corr(h, hm) \neq 1$, then hm cannot be considered a trivial reparameterization of h; it represents a nonlinear redistribution of impact across collaborative structures.

If the correlations merely reflected mathematical dependence without analytical value, then:

– replacing h-index with raw citation counts would not alter clustering outcomes;

– co-authorship correction would not change cluster centroids or separability;

– correlation patterns would be invariant across career stages.

Empirically, none of these conditions holds. The study demonstrates that:

– hm-index improves cluster compactness and interpretability;

– correlation magnitudes differ between early-career and senior researchers;

– adjusted indices reduce noise and increase structural determinism.

These effects cannot be attributed solely to algebraic dependence. It is essential to distinguish between metric derivation (where dependence is unavoidable) and analytical deployment (where the metric is used to probe structure). The novelty of the present analysis lies in how these dependent metrics are used, namely:

– to compare structural regimes of scientific activity,

– to assess robustness across transformations,

– to reveal hierarchical stratification not visible in raw citation counts alone.

This approach aligns with established practice in statistical physics, econometrics, and information theory, where dependent observables are routinely analysed to infer latent structure.

Strong correlations between the h-index, citation counts, and co-authorship–adjusted indices are mathematically expected and methodologically necessary. They do not limit analytical novelty, as the study's contribution lies in the comparative, structural, and stability analysis of these correlations rather than in their mere existence. The results demonstrate that mathematically related indicators can still provide non-redundant, structurally meaningful insights when analysed across different regimes and transformations.

While strong correlations between h-index, citation counts, and co-authorship–adjusted indices are mathematically expected, this dependence does not reduce analytical novelty. Correlation here serves as a consistency condition rather than a trivial outcome. The novelty lies in the comparative analysis of correlation structures across career stages and in demonstrating that co-authorship–adjusted indices, despite being derived from the same citation base, exhibit distinct sensitivities, stabilities, and explanatory powers. These effects cannot be explained by algebraic dependence alone and reveal structural properties of scientific activity not captured by raw citation measures.

A potential limitation of bibliometric studies is the risk of overinterpreting citation-based indicators as direct measures of national scientific quality or educational effectiveness. In the present study, however, national-level conclusions are not derived solely from raw bibliometric indicators. Still, they are instead supported by a constellation of independent educational and institutional signals embedded in the empirical results.

Although the primary variables (FNCS, normalised h-index, hm-index) are bibliometric in nature, they do not operate in isolation. These indicators implicitly encode structural properties of national research systems, including:

– stability of doctoral training pipelines (career-length–adjusted h-index);

– institutional support for sustained productivity (FNCS aggregated over cohorts);

– norms of collaboration and team-based research (hm-index and co-authorship structure);

– maturity of research governance and evaluation cultures (self-citation constraints).

Thus, the observed patterns reflect systemic educational and institutional conditions, not merely citation accumulation. The k-means clustering results reported in the paper reveal persistent cross-national stratification across normalised indicators. Importantly:

– countries with strong positions are characterised by balanced profiles (high FNCS combined with normalised h- and hm-indices);

– weaker-performing clusters are not defined solely by low citation impact, but by structural imbalances, such as: high publication volume with low relative impact; elevated self-citation shares; weak collaboration-adjusted performance.

Such multidimensional consistency cannot be explained solely by bibliometric artefacts and instead points to differences in research training quality, institutional incentives, and integration into international scientific networks.

As shown in the methodological pipeline, all clustering was conducted exclusively on field- and time-normalised indicators, explicitly eliminating: disciplinary citation density effects; delayed database integration (including post-Soviet systems); national citation culture inflation. Consequently, national differences persist even after removing all known bibliometric scaling effects, indicating that the remaining variance is driven by underlying institutional performance rather than database artefacts.

The use of career-age–adjusted and field-normalised h-type indicators provides an indirect but independent measure of educational system effectiveness. Specifically:

– Higher percentile-normalised h-indices among early- and mid-career researchers imply more effective doctoral and postdoctoral training environments;

– Flatter or delayed trajectories indicate systemic bottlenecks in academic mentoring, funding continuity, or institutional mobility.

These effects emerge within cohorts, not across generations, reinforcing their interpretation as educational and institutional signals rather than historical citation inertia.

The inclusion of the hm-index and self-citation controls introduces non-trivial institutional dimensions:

– Strong performance after fractional authorship adjustment implies effective integration into high-quality research teams;

– Lower reliance on self-citation is associated with stronger external validation mechanisms and journal ecosystems.

These properties are widely recognised as markers of institutional quality, independent of raw bibliometric output.

Crucially, the study does not claim to directly measure educational quality or institutional effectiveness. Instead, it demonstrates that even after eliminating bibliometric distortions, normalised scientometric patterns align systematically with known differences across national research organisations, training capacity, and institutional maturity. Thus, national-level interpretations are inferred rather than asserted and remain bounded by the empirical scope of the data. Although based on bibliometric data, national-level conclusions in this study are supported by multiple independent signals embedded in normalised career-stage indicators, collaboration-adjusted metrics, self-citation patterns, and stable cluster structures. Because all known bibliometric distortions were removed prior to clustering, the remaining cross-country variance reflects underlying educational and institutional conditions rather than citation artefacts.

The results reported above confirm several well-known scientometric regularities, including cumulative advantage effects and the strong dependence of the h-index on citation counts and collaboration-adjusted indices. These findings provide a consistency check for the dataset and methodology. However, the central result of this study is not located in these confirmations, but in the formally established status of self-citation derived in Section 4. The theorem demonstrates that self-citation is statistically independent of productivity, impact, collaboration structure, and economic capacity, is not predictable from them, and yet is systematically associated with institutional quality indicators. This result closes the conceptual gap identified in the Introduction. Self-citation is shown not to be a performance-driven derivative or a technical bias term, but an autonomous behavioural coordinate of scientific activity. Consequently, citation-based indicators implicitly combine at least three analytically distinct components: productivity, impact, and behaviour. From this perspective, the common practice of treating self-citation as mere noise to be filtered out appears conceptually incomplete. Instead, self-citation can be reinterpreted as a source of information about institutional and normative environments of scientific systems. Significantly, this conclusion is not based on qualitative judgment or sociological speculation, but instead follows deductively from a system of empirical equalities and inequalities among measurable variables, as formalised in Section 4.

The main contribution of this work is not the confirmation of well-known cumulative or collaboration effects in scientometric indicators, but a formal reclassification of the conceptual status of self-citation within the structure of citation-based evaluation metrics. This reclassification is based on a system of empirically established statistical relations and non-relations, which can be summarised formally. Let, for each author i, the following variables be defined:

$$H_i = h\text{-index}, \quad P_i = \text{number of publications}, \quad C_i = \text{total citations},$$
$$HM_i = \text{co-authorship adjusted index}, \quad S_i = \text{self-citation rate}.$$

At the country level, let:

$$\overline{S}_k = \textit{mean self-citation rate in country } k, \quad GDP_k = \textit{gross domestic product},$$
$$CPI_k = \textit{corruption perception index}.$$

The empirical results of this study establish the following system of relations:

$$\rho(H,S) = -0.02(early-career), \rho(H,S) = -0.055(senior),$$
$$\rho(P,S) \approx 0, \quad \rho(C,S) \approx 0, \quad \rho(HM,S) \approx 0,$$

which implies: $|\rho(X,S)| < 0.1 \quad \forall X \in \{H,P,C,HM\}$. Furthermore, for any tested predictive model of the form: $\hat{S} = f(H,P,C,HM)$, including multivariate regression and neural networks, the empirical results yield: $R^2(\hat{S},S) \approx 0$, demonstrating the absence of any functional dependence of S on standard performance indicators. At the macro level, the following relations are observed: $\rho(\overline{S},GDP) \approx 0, \quad \rho(\overline{S},CPI) < 0$, and the absolute value of $\rho(\overline{S},CPI)$ is the largest among all correlations involving $\overline{S}$. Concrete empirical anomalies complement these relations. For example, Ukraine exhibits:

$$\overline{S} \approx 34\% \,(\textit{Top-5 globally}), \quad H_{country} \approx 11 \,(\sim 68^{th\ place}),$$

while Russia demonstrates an abnormally high self-citation rate relative to its publication volume.

Taken together, these results imply the following formally justified conclusion:

$$S \perp \{H,P,C,HM,GDP\}, \quad but \quad S \perp CPI.$$

That is, self-citation is statistically independent of scientific productivity, impact, collaboration structure, and economic capacity, while being systematically associated with institutional quality.

It allows us to introduce a conceptual decomposition of scientometric indicators:

$$\textit{Scientific performance} = (\textit{Productivity, Impact, Behaviour}),$$
$$\text{where } \textit{Productivity} \sim P, HM, \qquad \textit{Impact} \sim C, H, \quad \textit{Behavior} \sim S.$$

In this framework, self-citation is not a correction term or a measurement bias, but an independent behavioural coordinate of scientific activity. This conclusion is not interpretative, but follows deductively from the empirical system of statistical equalities and inequalities reported above. Methodologically, this work introduces an inference-by-elimination strategy. By demonstrating that S is neither correlated with nor predictable from performance-based variables, and by identifying the only remaining systematic association at the institutional level, we isolate a behavioural component that is usually hidden within citation-based indicators. In this sense, the present study does not merely add another correlation to the literature. Still, it changes the conceptual status of self-citation from a derivative or noise term into a structurally meaningful dimension of scientometric analysis.

9. Conclusion

This paper presents a comprehensive data-driven analysis of the formation and dynamics of the h-index across different academic career stages, based on large-scale bibliometric datasets of early-career and senior researchers. By integrating statistical analysis, trend smoothing, correlation assessment, and clustering techniques, the study provides a systematic perspective on the factors influencing scientific impact in contemporary research environments.

The experimental results confirm that the h-index exhibits a cumulative growth pattern, with senior researchers demonstrating, on average, a 2.6-fold higher h-index compared to early-career researchers. Despite this quantitative difference, the similarity in distribution shapes and synchronised trend-turning points across both groups indicates the presence of universal patterns in the development of scientific productivity, largely independent of career stage.

Correlation analysis reveals that collaborative research activity, as captured by the h-index, has the most significant positive influence on the h-index across both datasets, particularly for early-career researchers. In contrast, self-citation shows negligible or negative correlation, suggesting a limited effect on long-term scientific impact. These findings emphasise the importance of scientific collaboration and citation quality over purely quantitative publication strategies.

Clustering experiments further demonstrate the hierarchical structure of the global scientific landscape, highlighting a distinct group of research leaders with exceptionally high h-index values. The results also indicate that geographical factors play a more significant role in shaping scientific performance at early career stages. At the same time, their influence diminishes as researchers gain experience and establish broader international collaborations.

From an information engineering perspective, the proposed analytical framework can be effectively integrated into decision-support systems for research evaluation, academic performance monitoring, and science policy analysis. The

results support the development of more balanced, context-aware assessment mechanisms that account for career stage and collaboration patterns. All proposed hypotheses are directly tested using the statistical, correlational, smoothing, and clustering analyses presented in Sections 5–8, ensuring methodological alignment between research questions, theories, and empirical evidence.

Table 17. Hypothesis Validation Summary

Hypothesis	Description (Short)	Method(s) Used	Evidence in Results
H1	Experienced scientists have a higher mean and median h-index than novices	Descriptive statistics	Mean: 37.78 vs. 14.59; Median: 34 vs. 13
H2	The h-index distribution is closer to normal for experienced scientists	Histograms, cumulative curves, asymmetry	Lower skewness; smoother cumulative growth
H3	Smoothed h-index trends have similar shapes across career stages	Kendall, Pollard, exponential, median smoothing	Coinciding turning points
H4	Career stage affects scale, not structure, of h-index trends	Comparative trend analysis	Consistent trend shapes with different magnitudes
H5	Self-citation has a weak or negative association with the h-index	Pearson correlation	$r \approx -0.02$ to -0.06
H6	Citation indicators correlate more strongly with the h-index than publication count.	Correlation analysis	Citations $r \approx 0.56–0.68$ > publications $r \approx 0.27–0.46$
H7	Co-authorship (hm-index) is the strongest predictor of h-index	Correlation matrices	hm-index $r \approx 0.68–0.80$
H8	Correlation structure is stronger for experienced scientists	Comparative correlation matrices	Higher and more stable correlations
H9	Country effects are more potent for novice scientists	Country-level analysis, clustering	Clear geographic separation in early careers
H10	A distinct cluster of scientific leaders exists across career stages	k-means clustering ($k = 6$)	Separate high-impact cluster identified

This study demonstrates that meaningful cross-country comparisons in scientometrics require explicit separation between observable citation outcomes and structural conditions of knowledge production. By integrating field- and time-normalised indicators with system-level aggregation and robustness checks, the proposed framework moves beyond raw bibliometric counting toward a comparative structural analysis of research systems.

The clustering analysis reveals a stable, non-trivial stratification of researchers and countries that persists after eliminating disciplinary citation density, database coverage asymmetries, and national citation cultures. The resulting clusters are defined not by citation volume but by relative impact consistency, collaboration-adjusted performance, and systemic depth, indicating genuine heterogeneity in research system organisation. At the national level, the results suggest that scientific performance differences are not reducible to publication quantity or historical accumulation. Instead, they reflect the institutional capacity to sustain a broad core of internationally visible research, shaped by training pipelines, collaboration structures, and evaluation norms. Significantly, countries with smaller output but strong normalised performance differ qualitatively from large-volume systems with weak relative impact.

The use of non-standard system-level indicators, including the country-level h-index and neural network–based validation, is justified by their ability to capture dimensions of systemic depth and structural stability that are invisible to mean-based or volume-driven metrics. Rather than replacing standard indicators, these measures complement them by revealing latent organisational properties of research systems.

The analysis does not claim to directly measure educational quality, institutional governance, or policy effectiveness. Instead, it shows that after removing known bibliometric distortions, the remaining variance aligns with plausible institutional explanations. Non-linear transformations, causal inference, and policy attribution remain outside the scope of the present framework and constitute directions for future research.

These findings imply that cross-national scientometric comparisons should shift from ranking-based paradigms toward structure-sensitive analyses that distinguish between scale, impact, and systemic sustainability. The proposed framework provides a transferable methodological template for analysing heterogeneous research systems without privileging historically dominant or citation-intensive disciplines.

Although several empirical results of this study confirm well-established scientometric regularities (such as cumulative advantage mechanisms, the dependence of the h-index on citation counts, and the role of collaboration structures), the main contribution of this work lies not in confirming known effects. Instead, it lies in a conceptual reclassification of the role of self-citation in scientometric analysis. In the prevailing literature, self-citation is typically treated either as a technical bias that should be corrected for or as a secondary derivative of publication volume and citation intensity. Implicitly, this assumes that self-citation is a function of scientific productivity and impact.

The empirical results of this study demonstrate that data do not support this assumption. We show that self-citation is statistically independent of the h-index, total citations, publication counts, and co–authorship–adjusted indices, and cannot be predicted from these variables, even with multivariate regression and machine learning models. At the same time, self-citation exhibits a systematic association with institutional quality indicators but shows no meaningful relationship with economic capacity (GDP). This combination of results leads to a conceptual shift: self-citation should not be treated as a correction term in citation-based performance metrics, but rather as an autonomous behavioural dimension of scientific activity. In other words, scientometric indicators implicitly mix at least three analytically distinct

components: productivity, impact, and behaviour. While the first two are widely studied and formalised, the third (behavioural patterns of citation practices) has so far remained largely implicit and undertheorized.

Methodologically, this work introduces an inference-by-elimination strategy: instead of merely identifying new correlations, we systematically demonstrate which classes of variables cannot explain self-citation behaviour. It allows us to separate performance-driven and behaviour-driven components of citation statistics and to assign self-citation a distinct analytical status. More generally, this perspective suggests that certain scientometric phenomena traditionally treated as measurement noise or bias may, in fact, encode structurally meaningful information about the institutional and normative environments of scientific systems. In this sense, the present study not only reports new empirical patterns but also proposes a different way of conceptualising the internal structure of citation-based indicators.

Future work will focus on extending the proposed approach by incorporating domain-specific normalisation, temporal modelling, and machine learning techniques to enhance predictive capabilities and improve the robustness of scientific impact assessment.

Appendix

Appendix A. Formal Validity of Monotone Smoothing for Cumulative Scientometric Indicators

A.1. Preliminaries and Definitions. Let $\{x(t)\}_{t=1}^{T}$ be a discrete-time sequence representing a cumulative scientometric indicator (e.g., h-index or cumulative citations), where t denotes calendar year or career age. By the definition of cumulative indicators, the following property holds:

Assumption A1 (Monotonicity) – $x(t+1) \geq x(t), \quad \forall t \in \{1, \dots, T-1\}$. Let $\tilde{x}(t)$ denote a smoothed version of $x(t)$ obtained via a smoothing operator $\mathcal{S}$. The smoothing procedure is valid if it satisfies:

1. Monotonicity preservation: $\tilde{x}(t+1) \geq \tilde{x}(t)$.
2. Trend consistency – the sign of long-run growth is preserved.
3. Noise reduction without structural distortion.

A.2. Moving Average Smoothing under Monotonic Constraints. Let the moving average smoother be defined as:

$$\tilde{x}(t) = \frac{1}{2k+1} \sum_{j=-k}^{k} x(t+j),$$

with appropriate boundary handling.

Proposition A1. If $x(t)$ is non-decreasing, then $\tilde{x}(t)$ is also non-decreasing.

Proof. Since $x(t+j+1) \geq x(t+j)$ for all j,

$$\sum_{j=-k}^{k} x(t+j+1) \geq \sum_{j=-k}^{k} x(t+j).$$

Dividing both sides by $2k+1$, $\tilde{x}(t+1) \geq \tilde{x}(t)$.

Thus, moving average smoothing preserves monotonicity for cumulative indicators.

A.3. Exponential Smoothing. Exponential smoothing is defined recursively as:

$$\tilde{x}(t) = \alpha x(t) + (1-\alpha)\tilde{x}(t-1), \quad \alpha \in (0,1].$$

Proposition A2. If $x(t)$ is non-decreasing and $\tilde{x}(t-1) \geq \tilde{x}(t-2)$, then $\tilde{x}(t) \geq \tilde{x}(t-1)$.

Proof. $\tilde{x}(t) - \tilde{x}(t-1) = \alpha[x(t) - \tilde{x}(t-1)]$. Since: $x(t) \geq x(t-1) \geq \tilde{x}(t-1)$, it follows that: $\tilde{x}(t) - \tilde{x}(t-1) \geq 0$.

Hence, exponential smoothing preserves monotonicity under cumulative growth.

A.4. Median Filtering. Define the median filter: $\tilde{x}(t) = median\{x(t-k), \dots, x(t+k)\}$.

Proposition A3. If $x(t)$ is non-decreasing, then the median-filtered sequence $\tilde{x}(t)$ is also non-decreasing.

Proof. In a non-decreasing sequence, the order of values within each window is preserved as the window shifts forward. The median element, therefore, cannot decrease when the window advances by one time step.

It guarantees robustness to outliers without violating monotonicity.

A.5. Isotonic (Kendall-Type) Regression Smoothing. Let isotonic regression solve:

$$\min_{\tilde{x}(t)} \sum_{t=1}^{T} \big(x(t) - \tilde{x}(t)\big)^2 \quad \textit{subject to} \quad \tilde{x}(t+1) \geq \tilde{x}(t).$$

Proposition A4. Isotonic regression produces the unique least-squares optimal monotone approximation to $x(t)$.

Proof (Sketch). By the Pool-Adjacent-Violators Algorithm (PAVA), the feasible set of monotone sequences is convex, and the objective is strictly convex. Hence, a unique global minimiser exists and satisfies monotonicity by construction.

Isotonic smoothing is therefore theoretically optimal for cumulative scientometric indicators.

A.6. Preservation of Growth Rates and Inflexion Structure. Let the discrete growth rate be defined as $\Delta x(t) = x(t) - x(t-1)$. For all smoothing methods above $\Delta\tilde{x}(t) \geq 0$, and major inflexion points (changes in convexity) are preserved under smoothing windows $k \ll T$. Thus, smoothing does not introduce artificial stagnation or decline phases.

A.7. Implications for Scientometric Validity. Because the h-index and similar indicators are cumulative, non-decreasing by definition, sensitive to short-term noise (citation bursts), monotone-preserving smoothing methods are mathematically consistent, statistically valid, and conceptually necessary for temporal analysis.

For cumulative scientometric indicators, monotone-preserving smoothing methods are provably valid, uniquely optimal in the isotonic case, and preserve both growth direction and structural dynamics, thereby ensuring that observed temporal patterns reflect genuine scientific development rather than artefacts of noise or rank ordering (Appendix A).

Appendix B. Simulation-Based Validation of Monotone Temporal Smoothing

B.1. Objective of the Simulation Study. The purpose of this simulation study is to validate that monotone-preserving smoothing methods applied to cumulative scientometric indicators:

– recover the underlying latent growth trajectory,
– reduce observational noise without introducing artificial trends,
– preserve relative differences between career-stage groups, and
– do not generate spurious inflexion points.

It directly addresses concerns that smoothing may impose structure rather than reveal it.

B.2. Data-Generating Process (DGP). We simulate an unobserved latent scientific impact process $z(t)$ representing the actual cumulative impact of a researcher. Latent growth model:

$$z(t) = \sum_{\tau=1}^{t} g(\tau),$$

where the annual impact increment $g(t)$ follows:

$$g(t) \geq 0. g(t) = \beta_0 + \beta_1 t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \quad g(t) \geq 0.$$

This formulation captures accelerating early-career growth ($\beta_1 > 0$), saturation at later stages via reduced variance.

B.3. Observation Model (Noise Injection). Observed scientometric trajectories $x(t)$ are generated as:

$$x(t) = z(t) + \eta_t, \quad \eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right),$$

subject to the cumulative constraint: $x(t) = \max\left(x(t), x(t-1)\right)$. It reflects real-world citation noise (citation bursts, delayed recognition).

B.4. Smoothing Methods Evaluated. Each simulated trajectory is processed using moving-average smoothing, Exponential smoothing, Median filtering, and Isotonic (monotone) regression (PAVA). All methods are applied to actual time-indexed sequences.

B.5. Evaluation Metrics. Performance is evaluated using the following criteria:

(i) Mean Squared Error (MSE): $\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{x}(t) - z(t)\right)^2$,

(ii) Monotonicity violations: $V = \sum_{t=1}^{T-1} I[\tilde{x}(t+1) < \tilde{x}(t)]$,

(iii) Inflexion-point distortion – difference between actual and estimated curvature sign changes.

Table B1. Simulation Results (across 10,000 simulated trajectories)

Method	MSE ↓	Monotonicity Violations	Inflection Preservation
Raw (unsmoothed)	High	0	Low
Moving average	Medium	0	Medium
Exponential smoothing	Medium–Low	0	Medium
Median filter	Medium	0	High
Isotonic regression	Lowest	0	Highest

All monotone-preserving methods eliminate artificial decreases. Isotonic regression consistently achieves the lowest reconstruction error. No method introduces spurious growth or stagnation phases. Relative differences between simulated early-career and experienced trajectories are preserved.

B.6. Career-Stage Robustness Check. Simulations were repeated under two regimes: Early-career (high variance, convex growth) and Experienced (low variance, sublinear growth). Smoothing accuracy remained stable across regimes, confirming that observed empirical differences are not driven by smoothing bias.

B.7. Negative Control Simulation. As a negative control, smoothing was applied to randomly permuted time indices. Result: reconstruction error increased sharply, growth patterns collapsed, and career-stage separation disappeared. It confirms that temporal ordering is essential, and smoothing does not create structure ex nihilo.

Simulation results demonstrate that monotone-preserving smoothing methods reliably recover latent cumulative growth trajectories, reduce observational noise, and preserve structural differences between career stages, thereby validating their use for temporal scientometric analysis.

Appendix C. Invariance of k-Means Clustering under Linear Feature Scaling

C.1. Problem Statement. Let $\mathcal{X} = \{x_1, \dots, x_n\}, \quad x_i \in R^d$ denote the set of feature vectors used for clustering, where in the present study:

$$\mathbf{x_i} = \left[\mathrm{FNCS_i}, h_i^{(\mathrm{norm})}, \mathrm{hm_i^{(norm)}}\right].$$

All components are dimensionless, normalised indicators obtained through field- and time-specific linear rescaling of raw citation metrics. The k-means algorithm partitions $\mathcal{X}$ into k clusters by minimising the within-cluster sum of squares (WCSS):

$$\mathcal{J}(\{\mathcal{C}_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x_i} \in \mathcal{C}_k} |\mathbf{x_i} - \boldsymbol{\mu_k}|^2,$$

where $\boldsymbol{\mu_k}$ is the centroid of the cluster $\mathcal{C}_k$.

C.2. Linear Scaling Transformation. Consider a linear feature-wise scaling transformation:

$$\mathbf{x_i'} = \mathbf{D}\mathbf{x_i},$$

where $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_d), \quad d_j > 0$. This formulation subsumes: field-level citation normalisation; unit rescaling; variance normalisation (standardisation); and normalisation by expected citation counts.

C.3. Effect on Euclidean Distances. For any two points $\mathbf{x_i}, \mathbf{x_j}$:

$$|\mathbf{x_i'} - \mathbf{x_j'}|^2 = \left(\mathbf{x_i} - \mathbf{x_j}\right)^{\mathbf{T}\mathbf{D}^{\mathbf{T}\mathbf{D}}} \left(\mathbf{x_i} - \mathbf{x_j}\right).$$

Since $\mathbf{D}^{\mathbf{T}\mathbf{D}}$ is positive definite, the transformation induces a weighted Euclidean metric.

C.4. Invariance under Uniform Linear Scaling. If $D = cI, \quad c > 0$, then

$$|\mathbf{x_i'} - \mathbf{x_j'}|^2 = c^2 |\mathbf{x_i} - \mathbf{x_j}|^2.$$

Substituting into the k-means objective: $\mathcal{J}' = c^2 \mathcal{J}$. Since multiplication by a positive scalar preserves the ordering of objective values, the minimiser of the k-means objective is invariant: $\arg\min \mathcal{J}' = \arg\min \mathcal{J}$. Thus, cluster assignments remain unchanged under uniform linear rescaling.

C.5. Implications for Field-Level Normalisation. Field-level normalisation of citation indicators is equivalent to applying field- and cohort-specific linear scaling factors to raw citation counts:

$$C' = \frac{C}{E[C|f,y]}.$$

This transformation preserves relative ordering within reference sets, removes multiplicative field-level inflation, and maps heterogeneous citation regimes onto a standard scale. When applied prior to aggregation and clustering, this normalisation ensures that k-means operates on relative impact geometry rather than absolute citation magnitudes.

C.6. Stability of Cluster Geometry. Because all clustering inputs are normalised to comparable, dimensionless units and subsequently standardised, the resulting feature space differs from the raw citation space by a linear combination of scalings and translations. Since k-means is invariant under uniform linear scaling, translation of the feature space, and permutation of dimensions, the observed cluster structure reflects the intrinsic relative positioning of researchers rather than artefacts of citation density, disciplinary inflation, or database coverage asymmetries.

C.7. Methodological Consequence. The use of field-normalised indicators prior to k-means clustering does not introduce artificial structure nor distort existing groupings. Instead, it removes exogenous multiplicative noise while preserving the relative geometry, which is essential for partition-based clustering. Therefore, cluster membership can be interpreted as robust to linear rescaling of citation-based indicators and methodologically independent of disciplinary or national citation regimes.

Because field-level normalisation is a linear rescaling of citation-based features, and because k-means clustering is invariant to such transformations, the clustering structure reported in this study reflects relative scientific positioning rather than artefacts of citation scale, disciplinary density, or database coverage.

Let $\mathbf{x_i} \in R^d$ denote normalised feature vectors and consider a linear rescaling $\mathbf{x_i'} = c\mathbf{x_i}$, $c > 0$. Then, for any cluster assignment, the k-means objective transforms as

$$\sum_k \sum_{i \in \mathcal{C}_k} |\mathbf{x_i'} - \boldsymbol{\mu_k'}|^2 = c^2 \sum_k \sum_{i \in \mathcal{C}_k} |\mathbf{x_i} - \boldsymbol{\mu_k}|^2.$$

Since multiplication by a positive constant preserves the ordering of objective values, the minimiser is unchanged. Therefore, k-means cluster assignments are invariant under linear rescaling, including field- and time-level citation normalisation. This invariance does not generally hold for non-linear transformations such as logarithmic scaling or percentile-based ranks, which alter inter-point distances and may change cluster geometry. For this reason, clustering in the present study was performed exclusively on linearly normalised, dimensionless indicators, while nonlinear normalisations were used only for descriptive or comparative analysis.

References

[1] F. A. Shah and S. A. Jawaid, "The h-index: an indicator of research and publication output," Pakistan Journal of Medical Sciences, vol. 39, no. 2, p. 315, 2023.

[2] S. Paul and B. Dutta, "Evolving tools for assessing and mapping scientific research landscapes," The Serials Librarian, pp. 1–19, 2025.

[3] M. K. Akhtar, "The h-index is an unreliable research metric for evaluating the publication impact of experimental scientists," Frontiers in Research Metrics and Analytics, vol. 9, p. 1385080, 2024.

[4] M. Shannigrahi and D. K. Kirtania, "Scientometric and content mapping of the 100 most cited papers in chemistry," Discover Chemistry, vol. 2, p. 264, 2025.

[5] J. Santa, "2025 world ranking of citation indexing researchers," ResearchGate, 2025.

[6] N. Wang, Y. Yu, L. Shi, Z. Zhang, and J. Song, "A scientometric study on research trends and characteristics of randomised controlled trials in orthodontics," Journal of Dental Sciences, 2025.

[7] P. Reyes-Cornejo, L. Araya-Castillo, H. Moraga-Flores, J. Boada-Grau, and C. Olivares-Brito, "Scientometric study of digital transformation and human resources: Collaborations, opportunities, and future research directions," Administrative Sciences, vol. 15, no. 4, p. 152, 2025.

[8] R. Jan and R. Ahmad, "H-index and its variants: Which variant fairly assess author's achievements," Journal of Information Technology Research, vol. 13, no. 1, pp. 68–76, 2020.

[9] J. P. A. Ioannidis, J. Baas, R. Klavans, and K. Boyack, "Supplementary data tables for 'A standardised citation metrics author database annotated for scientific field'," PLoS Biology Dataset, 2019.

[10] J. Baas, K. Boyack, and J. P. A. Ioannidis, "Updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2020.

[11] J. Baas, K. Boyack, and J. P. A. Ioannidis, "August 2021 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2021.

[12] J. P. A. Ioannidis, "September 2022 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2022.

[13] J. P. A. Ioannidis, "September 2022 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2022.

[14] J. P. A. Ioannidis, "October 2023 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2023.

[15] J. P. A. Ioannidis, "August 2024 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2024.

[16] J. P. A. Ioannidis, "August 2025 data update for updated science-wide author databases of standardised citation indicators," Elsevier Data Repository, 2025.

[17] T. S. Biró, A. Telcs, M. Józsa, and Z. Néda, "Gintropic scaling of scientometric indexes," Physica A: Statistical Mechanics and its Applications, vol. 618, p. 128717, 2023.

[18] C. Palomino-Leyva, J. Rivera-Recuenco, A. Fernandez-Giusti, J. Barja-Ore, Y. Retamozo-Siancas, and F. Mayta-Tovalino, "Bibliometric analysis of the worldwide scientific production on COVID-19 infection and cerebrovascular disease," Annals of Cardiac Anaesthesia, vol. 26, no. 2, pp. 197–203, 2023.

[19] L. D. Davis, C. M. Gilmore, A. Vargus, H. Ogbeifun, Y. H. P. Chun, and C. R. Frei, "Comparison of h-index and other bibliometrics in Google Scholar and Scopus for articles published by translational science trainees," Humanities and Social Sciences Communications, vol. 12, no. 1, pp. 1–4, 2025.

[20] Y. Jiang, X. L. Liu, Z. Zhang, and X. Yang, "Evaluation and comparison of academic impact and disruptive innovation level of medical journals: Bibliometric analysis and disruptive evaluation," Journal of Medical Internet Research, vol. 26, p. e55121, 2024.

[21] M. Bublyk, O. Slava, V. Vysotska, L. Kolyasa, and O. Vlasenko, "World universities strategic analysis based on data from the QS World University Rankings," CEUR Workshop Proceedings, vol. 3373, pp. 354–375, 2023.

[22] V. Vysotska, V. Lytvyn, M. Hrendus, S. Kubinska, and O. Brodyak, "Method of textual information authorship analysis based on stylometry," in Proceedings of the IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 2, pp. 9–16, 2018.

[23] V. Vysotska, "Computer linguistic systems design and development features for Ukrainian language content processing," CEUR Workshop Proceedings, vol. 3688, pp. 229–271, 2024.

[24] B. Korostynskyi, O. Mediakov, V. Vysotska, O. Markiv, and M. Duda, "Analysis of geo-economic distribution of scientific publications citation and self-citation standardized indices based on machine learning," CEUR Workshop Proceedings, vol. 3171, pp. 1657–1683, 2022.

Authors' Profiles

Yurii Ushenko: World's Top 2% Research Scientist (2020, 2021, 2022, 2024). Nominee of The Photonics100 2025 list by ElectroOptics. Winner of 1000 Talents Plan Program, PhD, DSc, Prof., at Shaoxing University, Shaoxing, Zhejiang Province 312000, China. Professor, Head of Computer Science Department, Chernivtsi National University, Ukraine. His research interests include intelligent information systems design, data mining, pattern recognition and digital image processing, artificial neural networks, laser polarimetry and interferometry.

Victoria Vysotska is a Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and in the Combating Cybercrime Department at Kharkiv National University of Internal Affairs. In 2023, she completed her Doctoral degree in Technical Sciences, specialising in "Structural, Applied, and Mathematical Linguistics," with a dissertation titled "Analysis and Synthesis of Computational Linguistic Systems for Processing Ukrainian Textual Content." She also holds a PhD in Information Technologies from Lviv Polytechnic National University, which was awarded in 2014. Victoria has authored over 500 publications, and her primary research interests focus on identifying disinformation, fake news, and propaganda, as well as detecting disinformation sources and inauthentic behaviour (e.g., bots) in coordinated groups through the use of artificial intelligence. Her work also encompasses natural language processing (NLP), computational linguistics, data science, systems analysis, information technologies, machine learning, and cybersecurity.

Serhii Vladov is a Leading Researcher in Scientific Activity Organization Department and Professor in the Combating Cybercrime Department at Kharkiv National University of Internal Affairs. He is he is Doctor of Technical Sciences. Him current research and the specializations' fields is applied intelligent systems (including intelligent automatic control systems) development of complex dynamic objects. He has nearly 200 scientific papers, among them more than 100, which are indexed in the Scopus scientific database.

Zhengbing Hu, Prof., Deputy Director, International Center of Informatics and Computer Science, Faculty of Program Systems and Applied Mathematics, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science and Artificial Intelligence, Hubei University of Technology, China. Visiting Prof., DSc Candidate in National Aviation University (Ukraine) from 2019. Primary research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.

Lyubomyr Chyrun is an Associate Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and the Department of Applied mathematics at the Ivan Franko National University of Lviv. Since 2001, he has worked at Lviv Polytechnic University's Institute of Computer Sciences and Information Technologies as an associate professor in the Department of Information Systems and Networks. In 2007, he defended his PhD thesis on May 1, 2002 - "Mathematical modelling and computational methods". He co-authored the monograph "Continuous Fractions and Complex Numbers". He is the author of more than 120 scientific publications. His areas of scientific interest include pattern recognition, the application of numerical methods in information technologies, object-oriented programming, web technologies, fake news identification, natural language processing, computer linguistics, data science, system analysis, and machine learning.

How to cite this paper: Yurii Ushenko, Victoria Vysotska, Serhii Vladov, Zhengbing Hu, Lyubomyr Chyrun, "Information Engineering for Data-Driven Analysis of h-Index Formation Across Academic Career Stages Using Large-Scale Bibliometric Parameters, Statistical and Clustering Methods", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.18, No.1, pp. 159-213, 2026. DOI:10.5815/ijieeb.2026.01.10