LUISS

Degree Program in Data Science and Management

Course of Big Data and Smart Data Analytics

Data-Driven Analytics of Scientific Publications for Journal Rating and Field Characterization

Prof. Irene Finocchi

SUPERVISOR

Prof. Alessio Martino

CO-SUPERVISOR

Dmitrii Treshchev 790461

CANDIDATE

Academic Year 2024/2025

Table of contents

List of tables

List of figures

Acronyms

IF Impact Factor

JCR Journal Citation Reports

SJR SCImago Journal Rank

KM knowledge management

IC Intellectual Capital

PSJR Prestige SCImago Journal Rank

API Application Programming Interface

CSV Comma-Separated Values

CNN Convolutional Neural Network

AI Artificial Intelligence

PCA Principal Component Analysis

CS Computer Science

Chapter 1

Introduction

Honest and transparent evaluation of scientific journals plays an important role in modern science, as journal indicators serve as fundamental benchmarks for assessing both the quality and significance of published studies and the overall productivity of researchers and institutions. These rankings form the basis of scientific certification systems, which directly influence key academic decisions: career growth and promotion, the distribution of competitive research grants, certification assessments, and comparisons of scientific productivity between institutions.

Quantitative metrics such as Impact Factor (Garfield, 1972) or H-index (Hirsch, 2005) are used as one of the quality assessments, but the interpretation of such indicators depends on the subject area and the structure of scientific fields. Qualitative metrics, such as SCImago Journal Rank, use different methods to calculate the ranking of scientific journals, for example, using the 'prestige' of the journal. SCImago Journal Rank is provided by the SCImago portal, which in turn uses data from the Scopus database and enables the comparison and analysis of more than 34,000 journals across different countries and scientific fields. In this study, an analysis of journal data within fields and subject categories was carried out, allowing differences between fields to be identified and the dynamics of thematic coverage to be characterised.

First, a methodology was developed for downloading and processing data on scientific journals from the SCImago portal, and the choice of time periods for analysis was justified, reflecting the key stages in the evolution of scientific fields. Then, various tools and metrics were applied: comparison of datasets with different sampling percentages (top-%), calculation of the percentage coverage by subject categories within a field, and the Jaccard coefficient for statistical evaluation of the similarity between datasets. In addition, graphs of

connections between categories within each field were constructed, where the weights of the edges were determined by the Jaccard coefficient between subjects, and signature tables were formed based on the structural characteristics of each graph. These tools made it possible to calculate analytical indicators and perform a comparative analysis.

The analysis revealed that the structural differences between the studied scientific fields remain consistent across all examined years, confirming the stability of their internal organisation and thematic focus. Thus, this study shows that combining different methods – top-percent rankings, coverage and overlap metrics, and graph-based features – creates a reliable and repeatable set of tools for describing and comparing scientific fields.

Chapter 2

Literature Review

This chapter discusses the main indicators for assessing the quality of scientific journals and publication platforms. In particular, the most common bibliometric indicators are analysed: Impact Factor (JCR), h-index, as well as indicators based on citation prestige, such as SJR (SCImago).

Glänzel and Moed (2002) define and analyse the Impact Factor (IF) within the Journal Citation Reports (JCR), which was introduced by Eugene Garfield (1972) as ‘a fundamental citation-based measure of the significance and performance of scientific journals’, which attempts to describe the average citation rate of a journal's articles within a short time window. Impact Factor is calculated for year T as the ratio of the number of citations received in year T by journal publications that appeared in T-1 and T-2 to the number of journal publications for those two years that are considered ‘citable items’. Due to the fixed 2-year time window, IF measures the short-term visibility of a journal rather than its long-term impact. A short time window works well where citation is frequent, and less well in disciplines with a longer citation accumulation cycle.

One of the problems discussed in the article is the asymmetry of the numerator and denominator. Citations are sometimes counted towards a broader set of materials than those publications that are included in the denominator as ‘citable items’ and, "in a sense, references to these documents are ‘free’" and can significantly inflate the IF for some journals. In addition, the IF is systematically skewed by the citation rules of different disciplines and document types, and also depends on the quality of the reference data.

Furthermore, the authors also highlighted some attempts to improve and correct Impact IF limitations. They considered replacing the average value with other characteristics, such as

the approach proposed by Pinski and Narin (1976), which suggests an 'influence methodology' where citations are assigned a weight depending on the impact of the citing journal. Attempts were also made to increase the citation window to more than two years and the age of articles at which articles and works cease to be cited. Methods were considered that divided documents into types, where IF was calculated separately for each group of documents (letters, reports, reviews, etc.). However, all these attempts were unsuccessful due to difficulties in interpretation or complexity of calculation and did not become particularly widespread among participants. In either case, IF has established itself as the standard for evaluating journals due to its ease of interpretation, relative stability, and regular availability through JCR.

Hirsch (2005) introduced a new tool for evaluating the contribution of scientific works called the h-index. The index is equal to h if a scientist has h publications, each of which has been cited at least h times, and all other publications have no more than h citations. The author emphasises that the h-index is less sensitive to the biases characteristic of other metrics, where one 'big-hit' article can dramatically increase the total number of citations, but will not necessarily increase h, and vice versa, a large number of low-cited articles increases publication activity but has little effect on h unless the number of works exceeding the threshold increases.

Hirsch links the h-index to other citation metrics and notes that the total number of citations usually increases along with the h-index in a predictable pattern. The higher the h, the greater the total number of citations. He also notes that with relatively stable publication activity, the h-index tends to grow at a fairly steady rate over time, so the rate of growth can be used to indirectly compare researchers with different levels of experience. It also highlights the need for cautious interpretation, as the value of h and its growth rate depend on the discipline and its citation norms, as well as on co-authorship practices - especially in large collaborations, where contributions and citations are distributed differently.

Bontis and Serenko (2009) in their study on the evaluation of academic journals on knowledge management and intellectual capital (KM/IC) attempted to create a rating system for evaluating these KM/IC academic journals, which had not been evaluated previously. While their first study used expert surveys as the sole evaluation method, their current study also used

the explicit preference method, also known as the citation method. The methodology describes that citation data was collected using Google Scholar (https://scholar.google.com) and the Publish or Perish tool (https://harzing.com/resources/publish-or-perish). The h-index and g-index are used as evaluation metrics. The results of the study showed that the two approaches to journal evaluation produce very similar results and a high correlation with minor deviations. Limitations are also discussed in the paper, such as citation frequency depending on the age of the journal, the volume of publications, self-citations, and the specifics of Google Scholar indexing. The rating is not a true measure, but only a comparison tool.

The authors also note that it is impossible to apply the IF in their study, even taking into account that the corpus of cited reports consisted of 7,500 journals across 200 disciplines and did not include new KM/IC journals. The use of the IF is also limited by its short citation window, as it is calculated using citations to items published in the previous two years. In addition, high IF indicators for a publication may be the result of a few excessively cited articles, while the rest of the works are cited infrequently or not at all.

Borja González-Pereira, Vicente P. Guerrero-Bote, Félix Moya-Anegón (2010) describe a new methodology for determining the prestige and value of journals. SCImago Journal Rank (SJR) is based not only on the number of citations, but also on the quality of the citing journals and their 'weight' through a model of eigenvectors. The idea was inspired by Larry Page and Sergey Brin's PageRank (1998) and applied to the Scopus database of works and citations.

Their method creates an oriented graph from journal to journal, where connections are normalised by the number of links in the citing journal, and uses a time window of three years. The time window is chosen as the minimum, as it covers citation peaks, but at the same time shows the dynamics of communication across different areas of Scopus. A maximum limit of 33% on self-citation in a journal has also been introduced to prevent artificial inflation, while maintaining a natural level of self-citation. Thus, the Prestige SCImago Journal Rank (PSJR)

is calculated as the overall prestige of the journal, then it is normalised to pure publications to obtain the average prestige per article and highlight the SJR.

For statistical comparison of SJR, authors select IF with a 3-year citation window for a more accurate comparison, as mentioned above, SJR uses a 3-year time window. The comparison results showed that the distribution of ratings between SJR and IF are generally similar and correlate with each other, but SJR shows a higher concentration of prestige in fewer journals. The authors also note that journal rankings can vary significantly depending on the field. The main feature of SJR is that it measures the influence of citations, as opposed to popularity and the number of citations, so SJR can be used as an alternative or additional tool for assessing the significance of scientific journals.

Chapter 3

Data and Preprocessing

This chapter outlines the scope and timeframe of the study and describes the data preparation procedures. The analysis focuses on three subject areas – Computer Science, Mathematics, and Decision Sciences – and four selected years (2012, 2019, 2022, and 2024). These years represent key time points capturing major developments: from the AlexNet breakthrough – a deep learning model introduced by Krizhevsky et al. in 2012 – to the rise of generative AI research, and up to the most recent year available.

The data was collected manually from the SCImago portal (https://www.scimagojr.com) due to the lack of an API: for each year, a sequential download was performed by subject area and subject category, sorted by SJR. The CSV files were then merged into a Pandas DataFrame. The final datasets were structured by year and area for ease of visualisation and further comparative analysis.

3.1 Definition of subject area and year interval

Three scientific fields were selected for this work: Computer Science, Mathematics, and Decision Sciences. This choice was made because these fields are most closely related to the development and application of artificial intelligence methods, including approaches based on data analysis and modeling. Four different years were also identified for the study, allowing us to examine the dynamics of changes over time and compare the results between selected time periods.

The starting year for the analysis was chosen as 2012 in connection with the release of the AlexNet neural network (Krizhevsky et al., 2012), which won ImageNet (large scale visual

recognition challenge) and revived interest in CNN and AI in general after a period of stagnation in research in these areas. This event had a noticeable impact on the Computer Science community.

The next key date chosen is 2019, a pre-COVID-19 year with complete and stable data. A comparison of publication dynamics before and after 2019 shows fluctuations in the number of articles due to pandemic restrictions on publications and conferences (Figure 3.1).

The year 2022 was chosen as another point of analysis because it was during this period that the release of ChatGPT (OpenAI, 2022) became a notable event in the development of artificial intelligence research. This event contributed to increased attention from the scientific community to generative AI approaches and technologies, as well as heightened interest in their capabilities and practical applications. The last year of analysis is 2024 - it is the most relevant at the time of the study and, at the same time, the last year for which data is available for correct filtering within the set used.

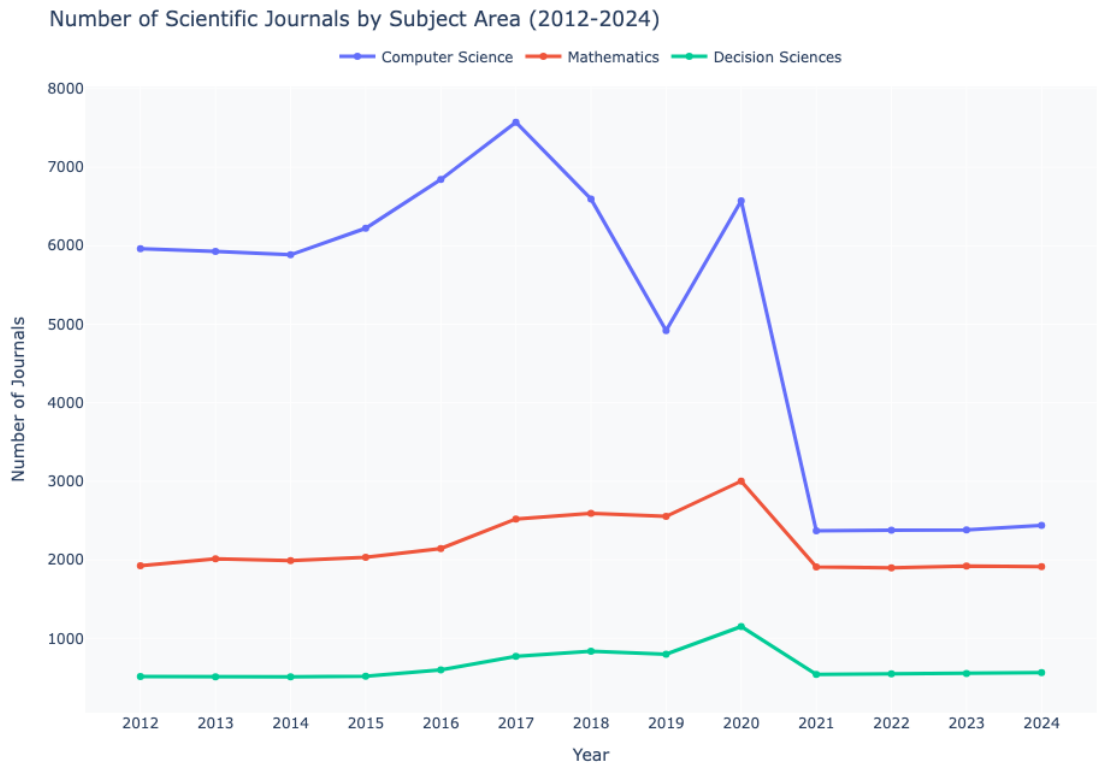

Figure 3.1: The graph shows the total number of journals on the SCImago from 2012 to 2024.

3.2 Data collection and processing

The data for the current work was taken from the SCImago Journal & Country Rank (SJR) portal. It does not provide a special API or other tools for accessing portal data from other applications; the data was downloaded manually via the website. Figure 3.2 shows the SCImago portal with various filters. For the current work, the selected filters were: *Subject Area*, *Subject Category*, and *Journal Evaluation Year*. As mentioned above, only three scientific areas and four different years were used.

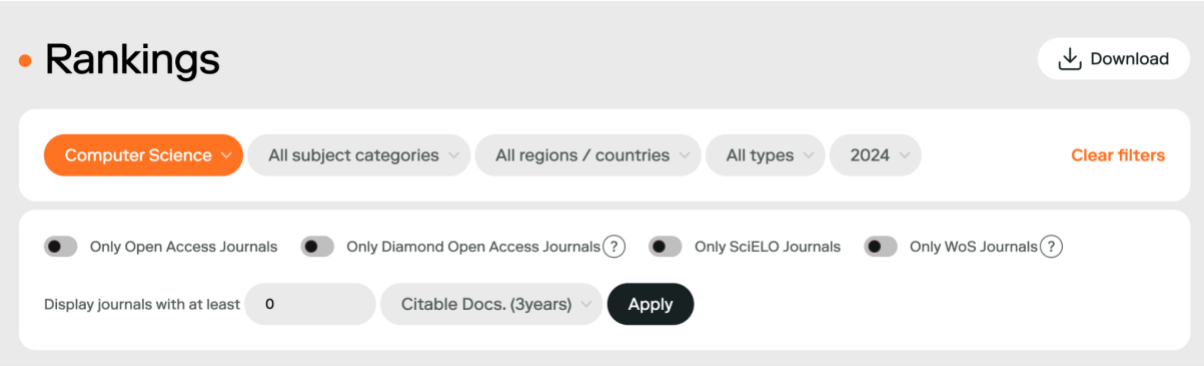

Figure 3.2: SCImago filter view

The data collection method itself involves iterative downloading of data from the SCImago website. First, the entire subject area is downloaded without selecting a subject category for the required year. Journal lists are filtered by descending SJR rating (the most prestigious journals at the top). Next, within the subject area, the subject category for the required year is selected and downloaded one by one, filtering the rating in the same way by descending SJR score. After completing all the steps, the procedure is repeated for each selected subject area. Once the data download process was complete, all previously saved CSV files were sequentially read and merged into a Pandas DataFrame structure for further analysis (Figure 3.3). The DataFrame for a subject area contains all journals of all categories within that area for the selected year, filtered by SJR rating. The DataFrame for a subject category contains only journals of the selected subject category, also filtered by SJR rating.

All datasets used in this study contained missing values. Since the corresponding fields were not used in the subsequent analysis, these missing entries did not affect the final results. Therefore, records with missing values in fields required for the computations were removed when necessary, while missing values in auxiliary columns were kept as NaN/“unknown”. No fully duplicated records (identical rows) were found in the datasets, so no additional deduplication step was required.

	Rank	Sourceid	Title	Type	Issn	Publisher	Open Access	Open Access Diamond	SJR	SJR Best Quartile	...	Ref. / Doc.	%Female	Overton	SDG
0	1	4700152228	Molecular Systems Biology	journal	17444292	Springer Science and Business Media Deutschlan...	Yes	No	8,164	Q1	...	53,70	32,77	5	0
1	2	19169	Journal of Operations Management	journal	18731317, 02726963	John Wiley & Sons Inc.	No	No	5,439	Q1	...	89,71	21,84	17	0
2	3	17945	Bioinformatics	journal	13674811, 13674803	Oxford University Press	Yes	No	5,275	Q1	...	24,10	22,66	46	0
3	4	12402	MIS Quarterly: Management Information Systems	journal	02767783, 21629730	Management Information Systems Research Center	No	No	5,234	Q1	...	77,60	23,72	22	0
4	5	19300156903	Foundations and Trends in Machine Learning	journal	19358245, 19358237	Now Publishers Inc	No	No	4,637	Q1	...	139,33	0,00	1	0

Figure 3.3: Computer Science DataFrame for 2012

After processing and initial analysis, the datasets were grouped by year and stored in separate dictionaries (one for each year and each area, each subject category). In these dictionaries, the original names of the corresponding Areas and Categories were additionally recorded for each dataset. This approach allows us to preserve the correct labels from the source data, improve the readability of visualisations, and simplify further analysis and comparison of results between years and areas.

Table 3.1 reports the final descriptive statistics of the constructed datasets after the data collection and preprocessing stages using SCImago data for the years 2012, 2019, 2022, and 2024. For each subject area, two indicators are provided: the dataset size (number of journals)

and the maximum SJR (SCImago Journal Rank) value observed in the corresponding year. The table is organized into three panels: (a) Computer Science, (b) Mathematics, and (c) Decision Sciences. Values shown in **bold** indicate the highest maximum SJR within each subject area across all considered years

Year	Computer Science	
	Number of journals	**Max SJR**
2012	5960	8.164
2019	4920	13.637
2022	2378	13.775
2024	2438	**22.797**

(a) - Area Computer Science

Year	Mathematics	
	Number of journals	**Max SJR**
2012	1925	8.901
2019	2553	9.444
2022	1897	6.760
2024	1914	**12.168**

(b) - Area Mathematics

Year	Decision Sciences	
	Number of journals	**Max SJR**
2012	513	7.171
2019	798	9.444
2022	548	7.126
2024	564	**10.181**

(c) - Area Decision Sciences

Table 3.1: Dataset size (number of journals) and maximum SJR by year for each subject area. Panels (a)–(c) correspond to Computer Science, Mathematics, and Decision Sciences, respectively.

The Area of Computer Science includes 12 subject categories. Mathematics contains 14 categories, while the Decision Science Area comprises only 4 categories and is the smallest category. The number of categories within each category remained unchanged throughout the years under review. The complete list of categories can be found in Appendix A, Figures A.1–A.3.

Figure 3.4 shows the SJR distribution for the three subject areas across four years. In all cases, the distribution is clearly right skewed: most journals have low SJR values, while only a small number reach very high scores, forming a long right tail. A logarithmic scale is applied to the SJR axis to improve readability and to better visualize the tail.

The trend in the first two years differs significantly from the trend in the last two years. This appears to be related to changes in the number of journals around 2020, as highlighted in Figure 3.1. This is due to the superliner growth in publications caused by the COVID-19

pandemic in 2020, which led to the launch of new platforms for the rapid dissemination of research and accelerated review processes.

When comparing subject areas, Computer Science exhibits the largest spread and the most pronounced high-SJR tail. Mathematics is more concentrated, with fewer extreme values. Decision Sciences has the smallest sample size and a relatively compact distribution. Overall, the general shape remains similar across years, while sample sizes and the presence of top-ranked journals vary. Colors correspond to the different years, and the legend also reports the number of journals (N) in each dataset for the corresponding area–year combination.

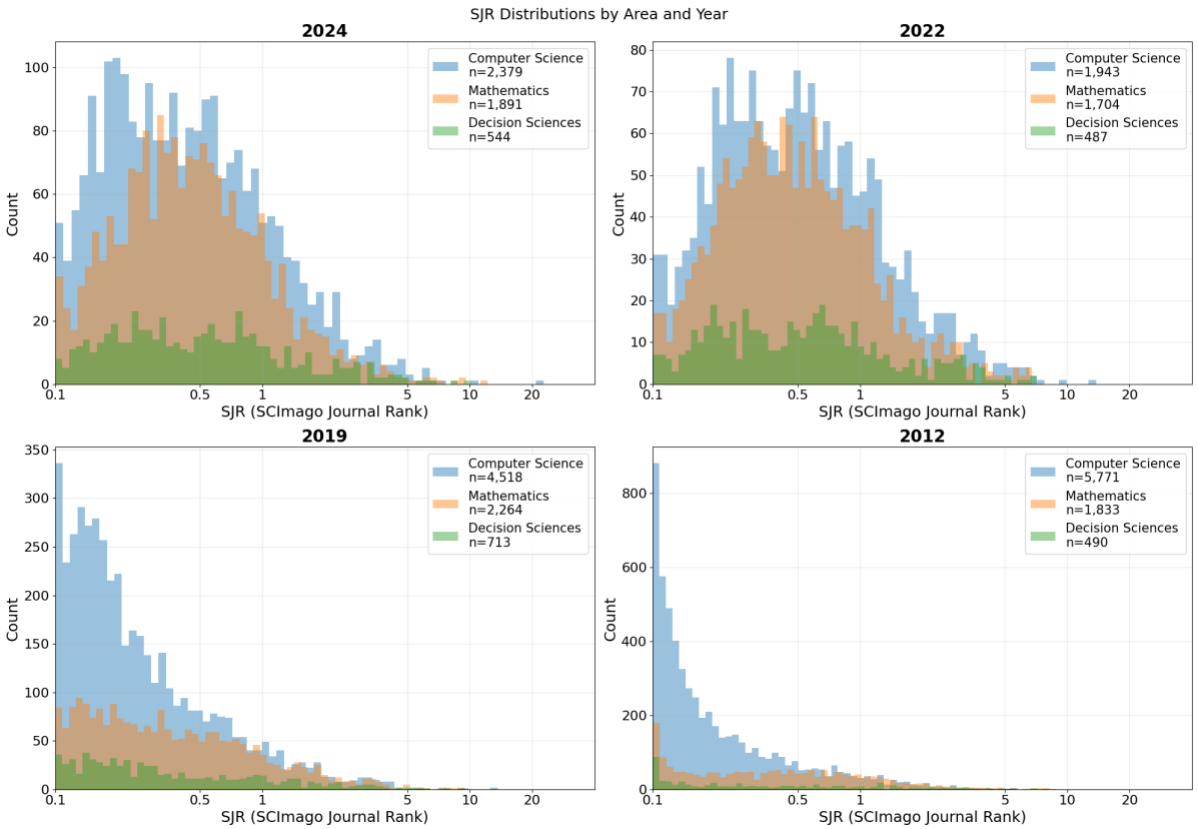

Figure 3.4: SJR Distributions by Subject Area and Year

For further work, *SourceId* (unique numerical identifier of the journal) values were used as the main unique journal identifier. This choice reduces potential matching errors caused by variations in journal titles (e.g., abbreviations, alternative spellings, language-specific naming conventions, or minor title updates over time). In practice, relying on SourceId also simplifies the detection of overlaps between categories and supports reproducible computations of coverage and similarity metrics based on set operations.

Chapter 4

Methodology

This chapter describes the methodology for selecting dataset sizes and conducting the analysis across multiple sampling thresholds within each subject area. It then computes coverage by subject categories within an area, enabling a comparison of category representation across different cutoffs of the top-ranked journals. Next, Jaccard similarity is introduced as a measure of set overlap and is used to examine how intersections change as the area subset expands. In addition, graphs are constructed between subject categories, where edges represent the magnitude of overlap (Jaccard coefficient), and signature tables of graph structural metrics are generated. These steps are repeated for each selected year (2012, 2019, 2022, and 2024). Finally, cross-area similarity is assessed using cosine similarity on standardized signatures and PCA to identify the main directions of variation and clusters among Area – Year structural profiles.

4.1 Sampling strategy and methodology review

For preliminary analysis and coverage calculation, it was decided to truncate the area and subject category datasets by 5%, 10%, 15% and 20% percentiles. These values were selected as a balance that would produce smoother results without sharp fluctuations, but without overfilling the sample with values close to the complete dataset.

While the selection of percentages for the sample as a whole was successful, the approach in which truncation occurs for both datasets (area subject and subject category) showed unstable results, where the coverage for many journals reached 100%, thus the truncated list of Subject journals was completely contained in the truncated Area list (Figure 4).

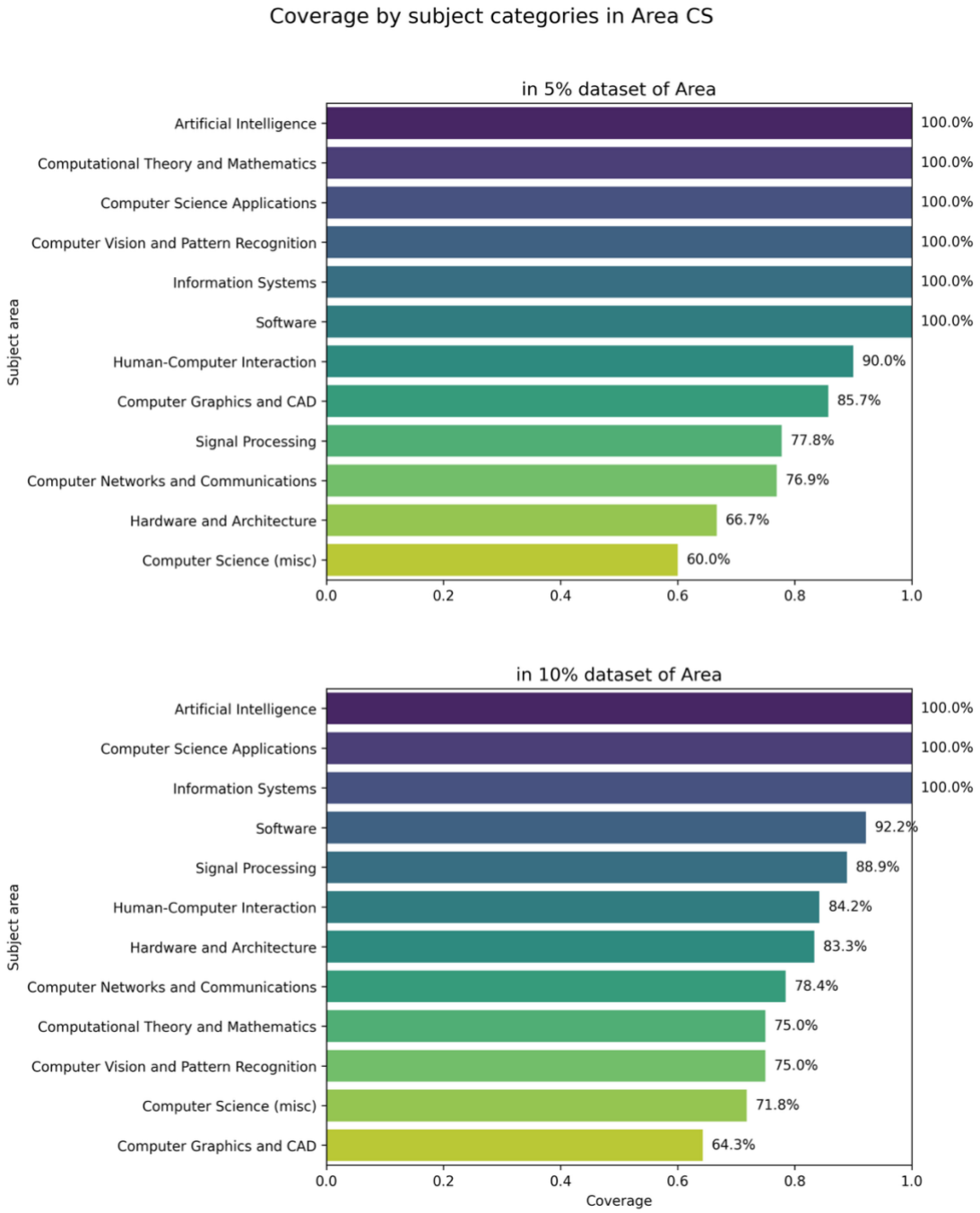

Figure 4.1: Comparison of coverage percentage across different both dataset sizes

This result can be interpreted as an artefact in the sampling method. The problem is that truncation on both sides increases the concentration on the highest-ranked journals, which simultaneously appear in both the top Area cut-off and the top Subject Category cut-off. As a result, the intersection becomes almost equal to the smaller set, and the coverage metric loses its discriminatory power.

4.2 Coverage by percent

The study uses the coverage metric to assess the representation of subject categories within an Area. Coverage determines the proportion of journals on a specific Subject that are present in a selected subset of journals in the Area. As part of the coverage calculation, only the Area set is truncated: for each percentage (5%, 10%, 15%, 20%), the top percentage of Area journals is selected to create a set. For each subject category, the full set of journals is used without truncation. Coverage is defined as the proportion of journals in the category present in the selected Area subset. The coverage is calculated as *Coverage = Intersection / Total_Subject*, where:

- Intersection is the number of common Sourceids (unique journal identifiers) between the truncated Area dataset and the full set of subject category data.
- Total_Subject is the number of unique Sourceids in the subject category dataset.

The idea behind this analysis was to demonstrate whether there are differences in SJR ratings between filtering all journals within the entire subject area without selecting a specific category and filtering by rating directly within the subject category. It was also useful to understand which subject category is most represented at the top of the field, as this information is not displayed on the SCImago portal when viewing within an Area.

Computer Science will be used as an illustrative example, but the other two Subject Areas (Mathematics and Decision Sciences) showed almost similar results. The results of the analysis showed that, overall, the ranking within a subject category correlates closely with the selected percentage of the Area dataset size. Figure 4.2 shows two bar plots comparing two sizes of the Computer Science Area dataset: the left bar plot is 5% Area dataset size and the right bar plot is 10% Area dataset size, respectively. The x-axis corresponds to the category names, and the y-axis shows the coverage within the Area.

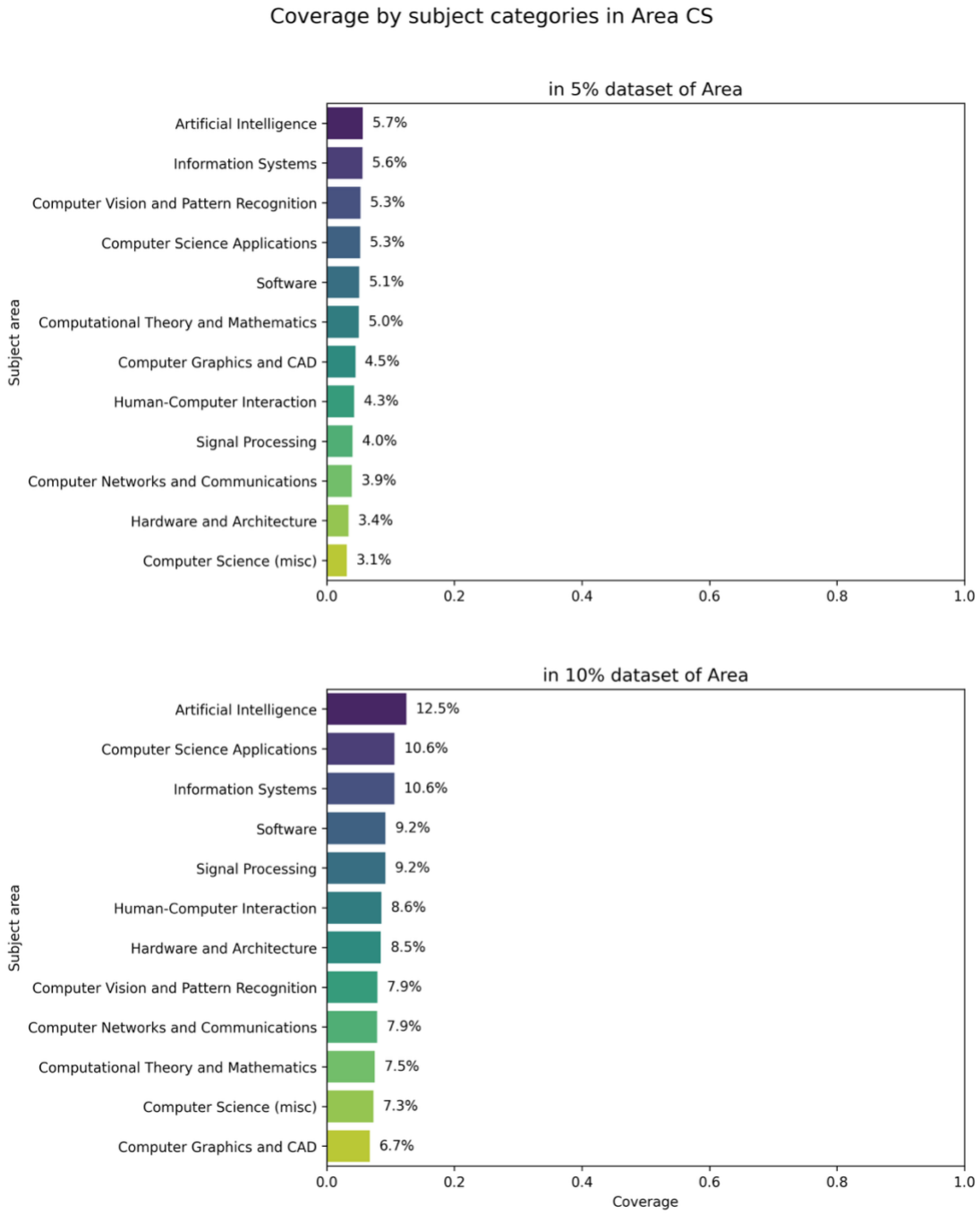

Figure 4.2: Comparison of coverage by subjects in two sizes of the Area dataset

4.3 Jaccard Similarity between Subject Categories and Area

Since the coverage metric only shows the coverage ratio, and the results may be unstable due to differences in dataset sizes, this study uses the Jaccard coefficient for a more detailed and accurate analysis, which shows the overall similarity of two sets, taking into account their size.

The Jaccard similarity metric assesses the extent to which journals of the selected Subject Category are represented in the top X% of Area journals, sorted by SJR rank.

Additionally, it allows to analyse the effect of expanding the Area subset (5%, 10%, 15%, 20%) on the degree of overlap between the sets, thus revealing the saturation effect: at what moment the addition of journals to Area begins to bring less and less growth in journals from a given Subject Category. The Jaccard coefficient is calculated as the number of common journals in two sets divided by the number of unique journals that appear in at least one of these sets. A coefficient closer to 1 indicates a stronger overlap, while values closer to 0 indicate a weaker overlap.

For each Subject category, the Jaccard similarity coefficient is calculated separately between the set of journals in this category and the set of journals in the selected Area.

Figure 4.3 shows 12 bar plots (1 for each Subject category of the CS area) that represent the Jaccard similarity results for each subject, where the x-axis corresponds to the percentage of the CS area dataset, and the y-axis corresponds to the Jaccard coefficient value.

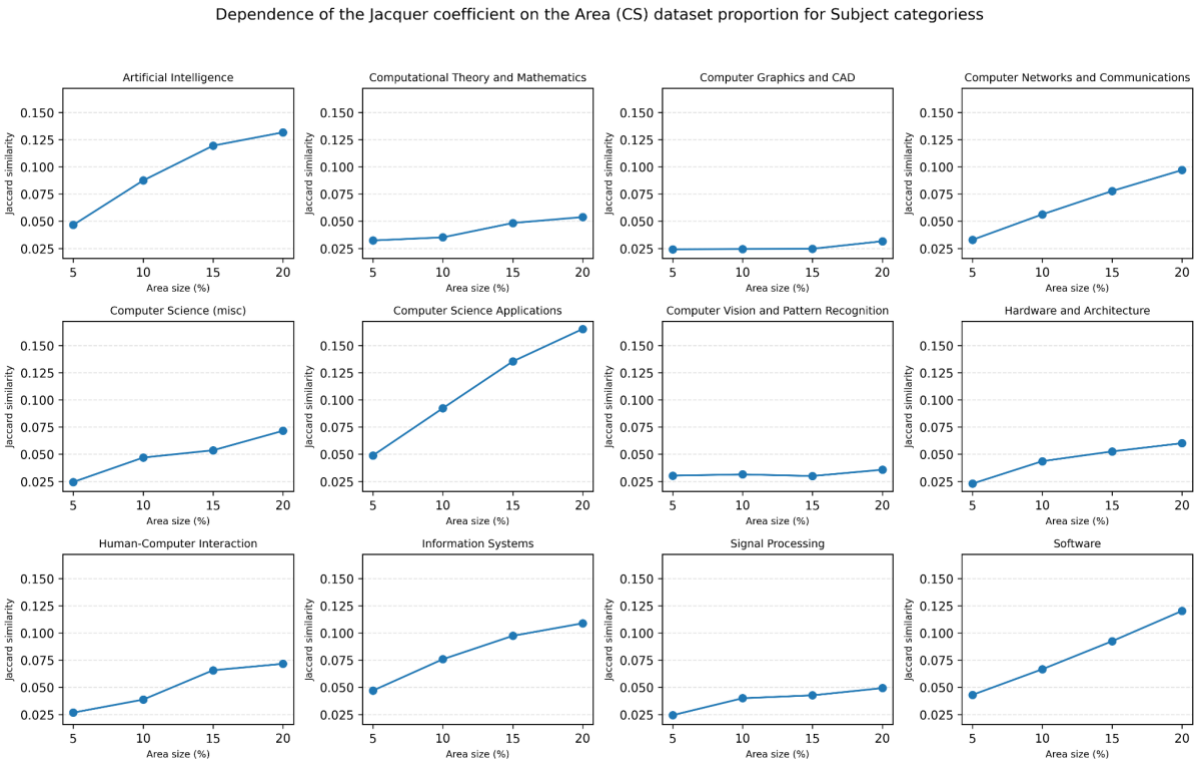

Figure 4.3: Comparison of CS subject categories by Jaccard coefficient

4.4 Subject Similarity Network (Jaccard Overlap)

After analysing Subject Categories within an Area using coverage metrics and the Jaccard coefficient, the next step is to study the relationships between the Subject Categories themselves. To do this, each Subject is considered as a set of journals, and the similarity between pairs of Subjects is measured by the Jaccard coefficient. A weighted graph is constructed based on pairwise values: the vertices (nodes) correspond to Subject Categories, and the edges reflect the degree of overlap between them. This method allows identifying clusters of related topics and determining categories that serve as connectors between several subject categories.

For this phase, a weighted undirected similarity graph between subject categories is formed based on SourceId journal identifiers. For each Subject Category from the list of DataFrames, journals with unique SourceId identifiers are extracted, after which the Jaccard coefficient is calculated for all Subject Category. Nodes corresponding to Subject Category names are added to the graph with a size attribute reflecting the number of unique journals in the category. A value of 0.05 (Jaccard coefficient ≥ 0.05) is selected as the minimum threshold for establishing a connection. This means that an edge between two subject categories was added only if their Jaccard similarity was not less than 0.05. Its allows insignificant connections to be filtered out and makes the visualisation more structured and understandable.

The edge weight reflects the Jaccard value between nodes, and the absolute value of the intersection is reflected as overlap. Additionally, for each node, an attribute is accumulated with the total number of intersections, equal to the sum of intersections with other categories across all added edges, which helps to determine the characteristic of the degree of cross-category overlap.

After creating the graph (Figure 4.4), it is possible to identify a central group of components: where large categories (Information Systems, Software, Computer Networks and Communications, Computer Science Applications, Artificial Intelligence) are connected by a large number of edges with a higher Jaccard coefficient. They can be identified by red edges (Jaccard coefficient ≥ 0.20) and orange edges (Jaccard coefficient ≥ 0.15). This indicates that

a significant proportion of journals belong to several such categories at the same time, forming an interconnected thematic space within CS. It is also possible to identify nodes with the highest degree (with the maximum number of connections to other nodes). Such categories act as network connectors: they have a wide intersection with different directions and actually serve as connectors between groups of topics. This may mean that these categories are either more universal within the CS Area or reflect interdisciplinary journals that are classified in several Subject Categories at the same time.

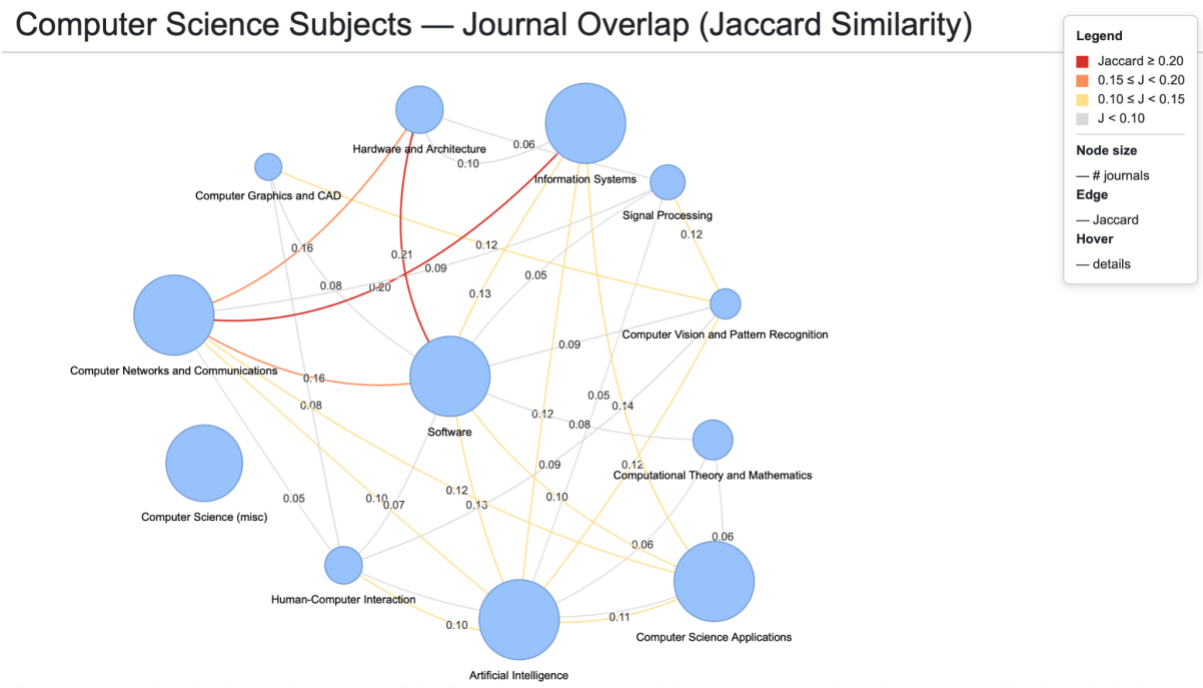

Figure 4.4: Graph between subjects in the Subject Area CS

For quantitative comparison of the obtained Subject - Subject graphs (constructed using Jaccard similarity), the graph signatures approach is used. For each graph, a set of aggregated metrics is calculated: the number of nodes, the number of edges, density, average and maximum degree, edge weight characteristics (average and maximum Jaccard value), and network connectivity (number of components). The resulting feature vector serves as a compact representation of the area structure and allows graphs to be compared between different areas and years, as well as other methods of multidimensional analysis to be applied.

Analysis of the results of the structural profile of the graph between subjects (Table 4.1) showed that the graph contains 31 edges with a density of 0.47, which indicates that almost half of all possible connections are present. The average degree of a node is 5.17, which shows that, on average, each category is connected to approximately five other categories, and *max_degree* = 10 indicates the presence of a pronounced connector node. On average, the overlap between related Subject Categories is moderate (*avg_jaccard* = 0.10), but there are pairs of categories with significantly stronger overlap (*max_jaccard* = 0.205), indicating locally closely related subject areas within CS.

Metric name	**Value**
num_nodes	12
num_edges	31
density	0.4696969696969697
avg_degree	5.166666666666667
max_degree	10
avg_jaccard	0.10411229348445501
max_jaccard	0.20516962843295639
avg_overlap	77.38709677419355
num_components	2

Table 4.1 Result of the structural profile of the graph of subjects in the CS Area

4.5 Year-by-Year Dynamics

To analyse the dynamics of the above metrics and calculations, year-by-year analysis was used to determine which patterns are stable over time and which change from year to year. All calculations described in the previous sections were repeated for the data sets for the years specified in Chapter 3 (2012, 2016, 2019 and 2024). For each year, the corresponding Subject Category datasets were downloaded and updated data frames were formed, after which the metrics were calculated for the Area subsets, formed as the top X% by SJR rank.

For the percentage coverage metric, a calculation procedure was created, as in section 4.2, but for the data set for each year. After the calculation was completed, the results were aggregated into a single data frame consisting of columns: *year*, *area_percent*, *subject*, *coverage*, *intersection*, and *total_subject*, which contain data for each subject by year, its coverage number, intersections, and the total number of subjects, as well as the percentage of the sample size for Subject Area. In the next step, for comparative analysis by year based on the general data frame, a composite bar plot with 4 panels was constructed, where each panel corresponded to the year of metric calculation. On each bar plot, the x-axis corresponded to the coverage number, and the y-axis displayed the names of the subjects.

To calculate the Jaccard coefficient for each year, the process was similar to that described in section 4.2, but it was recalculated for all years, and the results were stored in a single data frame. After the calculations and saving, a panel line graph was constructed based on the data frame, in which a separate graph was constructed for each subject. The graph shows the change in the Jaccard coefficient depending on the size of the Area sample. For each Subject Category, the dependence of Jaccard similarity on the size of the Area subset is displayed, and different years correspond to separate lines of different colours.

The procedure for constructing the inter-subject graph (Subject – Subject) by year repeats the methodology described in Section 4.3. For each year, sets of journals were formed by SourceId for all Subject Categories, the pairwise Jaccard coefficient was calculated, and a similarity network was constructed with threshold filtering of weak links. The main difference was in compiling a general table of signature profiles across all years. For each year, the count's signature was calculated, and then a summary table of signatures by year was formed. This made it possible to compare years quantitatively and move on to the next stage of analysis.

4.6 Cross-area comparison

A quantitative analysis of the similarity of structural signatures of graphs between Area -Year combinations was performed. For this purpose, key network features (density, average degree, average and maximum Jaccard coefficient, number of edges, average overlap) were selected from the summary table of signatures. Next, a feature matrix was formed, where the rows correspond to areas and years, and the columns correspond to the selected network metrics (see Appendix B, Figure B.1) . The features were then brought to a comparable scale using standardisation to exclude the dominance of metrics with large numerical values.

Cosine similarity was calculated for all Area–Year pairs on the standardised signature vectors, which made it possible to obtain a matrix of pairwise similarity of structural profiles. The resulting matrix was visualised as a heat map, where higher values correspond to a more similar network structure of graphs between the areas and years under consideration.

In addition, Principal Component Analysis (PCA) was used as a dimension reduction method for further inter-regional and inter-annual comparison of structural signatures. A standardised feature matrix X containing selected network metrics for each Area – Year combination was fed into the PCA. The data was then projected into a two-dimensional space with two principal components. PC1 captures the largest amount of variance in the data, while PC2 captures the largest amount of the remaining variance and is independent of PC1. The application of PCA allows identifying groups of graphs that are similar in structure, assessing differences between Areas, and observing changes in the structural profile over time.

Chapter 5

Results and Analysis

This chapter presents the main results of the comparison between Subject Areas and Subject Categories. First, the coverage between each area and its subject categories is analyses using several threshold values (5-20%) to show how the representation of categories changes as the sample size increases. Then, the dynamics are analysed by year (2012, 2019, 2022 and 2024) to assess the stability of coverage patterns and identify shifts in the leading categories. Next, intersections are evaluated using the Jaccard similarity coefficient by year and dataset size between each subject area and the top segments of its Area. Finally, the relationships between categories are characterised using graph-based structural signatures and compared across fields using cosine similarity and PCA.

5.1 Analysis of coverage between Area and subjects

For a clear comparison of how the representation of categories changes as the Area sample size increases, the results are visualized using a four-panel bar plot. Each panel corresponds to a specific percentage slice of the Area dataset, allowing for a direct comparison of coverage values between categories at different thresholds (Figure 5.1).

The plots show a comparison of coverage by subjects for the Area of Computer Science for 2024. It can be seen that the general trend of growth in coverage percentage is closely related to an increase in the percentage of the selected dataset slice: the larger the share of Area (5%, 10%, 15%, 20%), the higher the representation of journals from the corresponding categories. The graphs also show that the most “leading” category for any sample size remains Artificial Intelligence, i.e., it consistently provides the largest share of intersections with the top of the overall field ranking.

Overall, the fluctuations between the top categories for 5% and 10% are not very pronounced – the distribution of coverage remains fairly similar. Starting at 15% and especially for 20% of the dataset size, changes become noticeable for the categories *Artificial Intelligence* and *Human-Computer Interaction*: their coverage grows more strongly, and the difference relative to smaller segments reaches approximately 5%. For *Artificial Intelligence*, high representation is to be expected, as this category remains one of the most prominent in the field at any sample size, and its saturation will only increase the influence of *Artificial Intelligence*. For *Human-Computer Interaction*, this growth looks more significant: it means that journals in this category are more likely to be included in the sample when the dataset sample threshold is expanded. Thus, a significant portion of the journals in the *Human-Computer Interaction* category are concentrated not at the very top of the field's ranking, but slightly below. Therefore, as the sample size increases, their presence in the field becomes noticeably higher.

At the same time, there's also a reverse trend: the *Computer Graphics and Computer-Aided Design (CAD)* and *Computer Vision and Pattern Recognition* categories are seeing a relative decline – as it moved to larger segments, their share of coverage is growing more slowly.

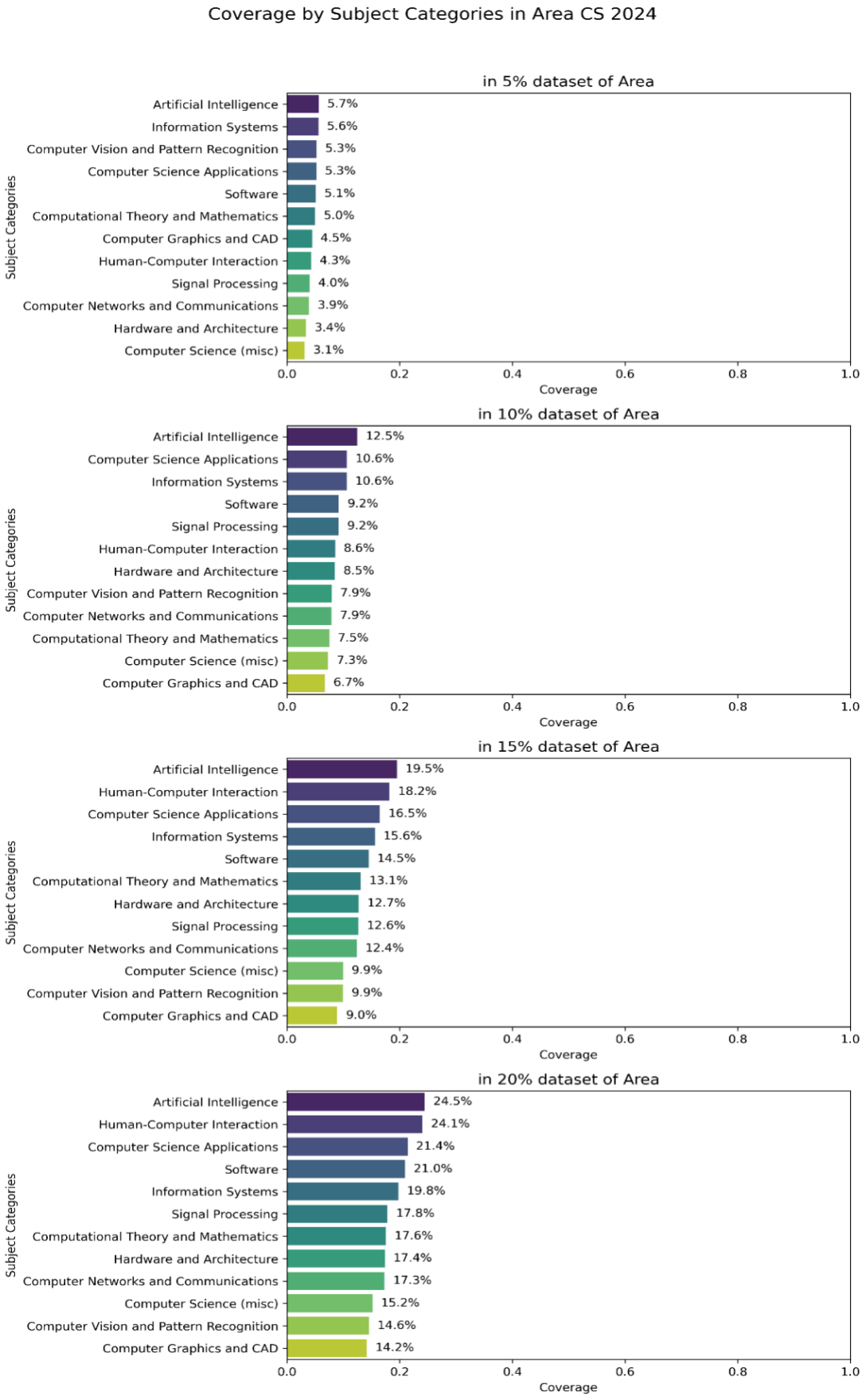

Figure 5.1: Comparison of coverage by Subjects in four sizes of the CS Area dataset

A coverage analysis was also performed for the Area of Mathematics. Figure 5.2 shows a comparison of coverage by year. Overall, the same logic applies as before: as the size of the Area cut-off increases (from 5% to 20%), the coverage values for most categories increase, i.e., expanding the threshold leads to the inclusion of a larger number of journals from different Subjects in the upper part of the Area ranking.

When considering the main changes, *Analysis* is leading in a small sample (about 7.5%). When the sample is expanded, *Theoretical Computer Science* takes the lead – its coverage grows significantly and reaches about 21.7% in a 15% sample of the dataset. In the largest sample, *Mathematical Physics* takes the lead (about 27.1%), with *Theoretical Computer Science* remaining very close behind (about 26.1%). *Geometry and Topology* and *Numerical Analysis* also strengthen significantly, rising from the bottom half of the list closer to the top. It can also be observed that journals in the *Mathematics (misc.)* category consistently rank in the middle of the top. Given that misc. includes journals that cannot be clearly assigned to a specific subject category, this may mean that such publications have fairly stable positions in the overall ranking of the field, but do not form the core of leaders, as more specialised areas do. In other words, *Mathematics (misc.)* reflects a broad and fairly significant interdisciplinary layer of journals that is consistently present at the top of the Area ranking but does not dominate among the leaders.

This changes in the top positions can be interpreted as a difference in the distribution of journals on the ranking scale: some categories (e.g., *Theoretical Computer Science*) are more strongly represented in the top range and are clearly visible even at low thresholds, while others (e.g., *Mathematical Physics*) begin to appear particularly noticeable when the threshold is expanded – that is, a significant proportion of journals in this category are not only at the top of the Area, but also in a wider range of high positions in the Area ranking.

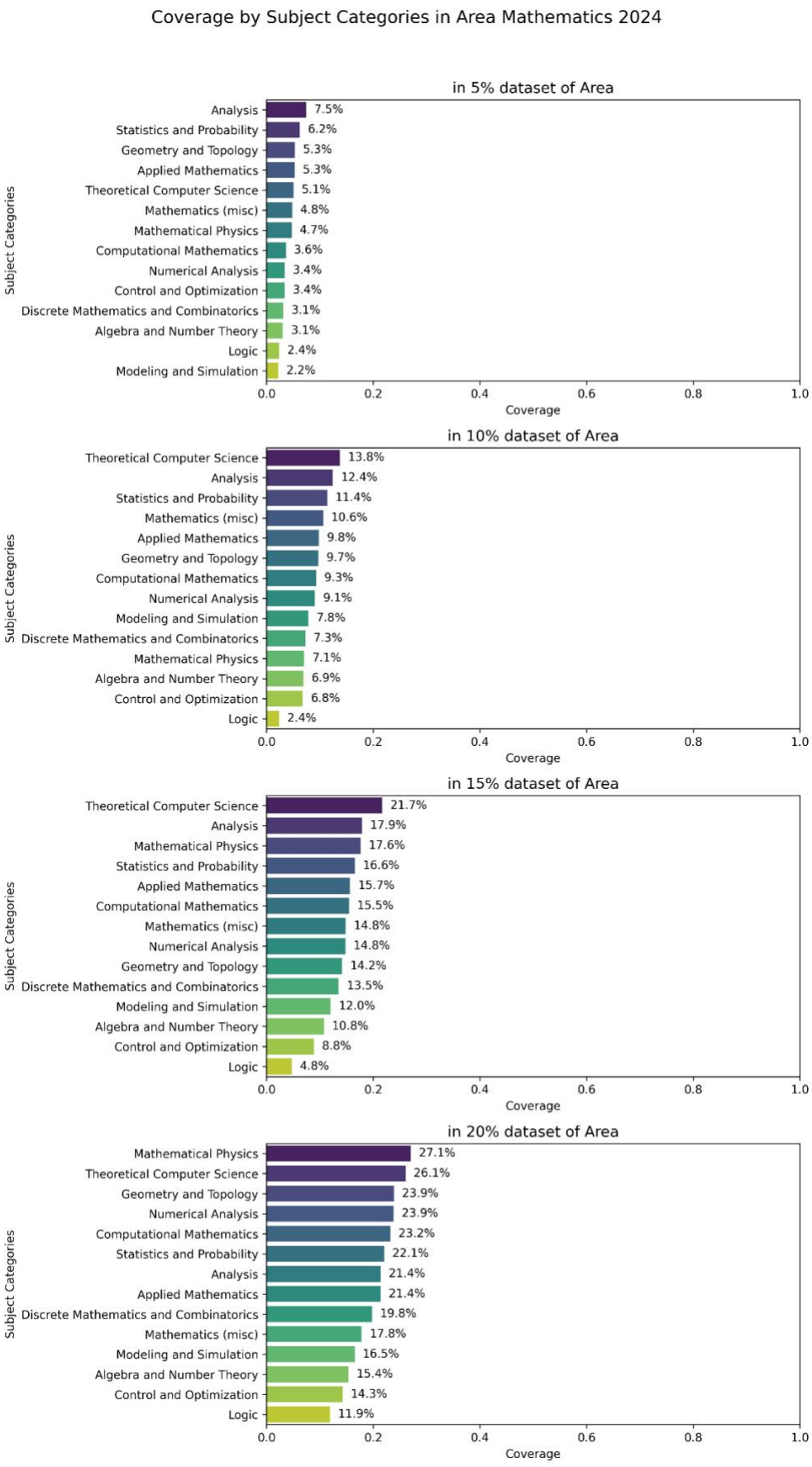

Figure 5.2: Comparison of coverage by Subjects in four sizes of the Area dataset

A similar trend can also be observed in the field of decision-making sciences (Figure 5.3). The category *Decision Sciences (misc.)* clearly stands out: it leads in all sample datasets, and its coverage steadily increases as the cutoff expands. In particular, at a cutoff of 20%, it reaches 29.6%, creating a noticeable gap compared to other subject categories. This suggests that *Decision Sciences (misc.)* forms the most stable core of overlap with the top of the *Decision Sciences Area* ranking, i.e, journals classified in this category are consistently represented among the highest-ranked journals in the broader field at different threshold levels.

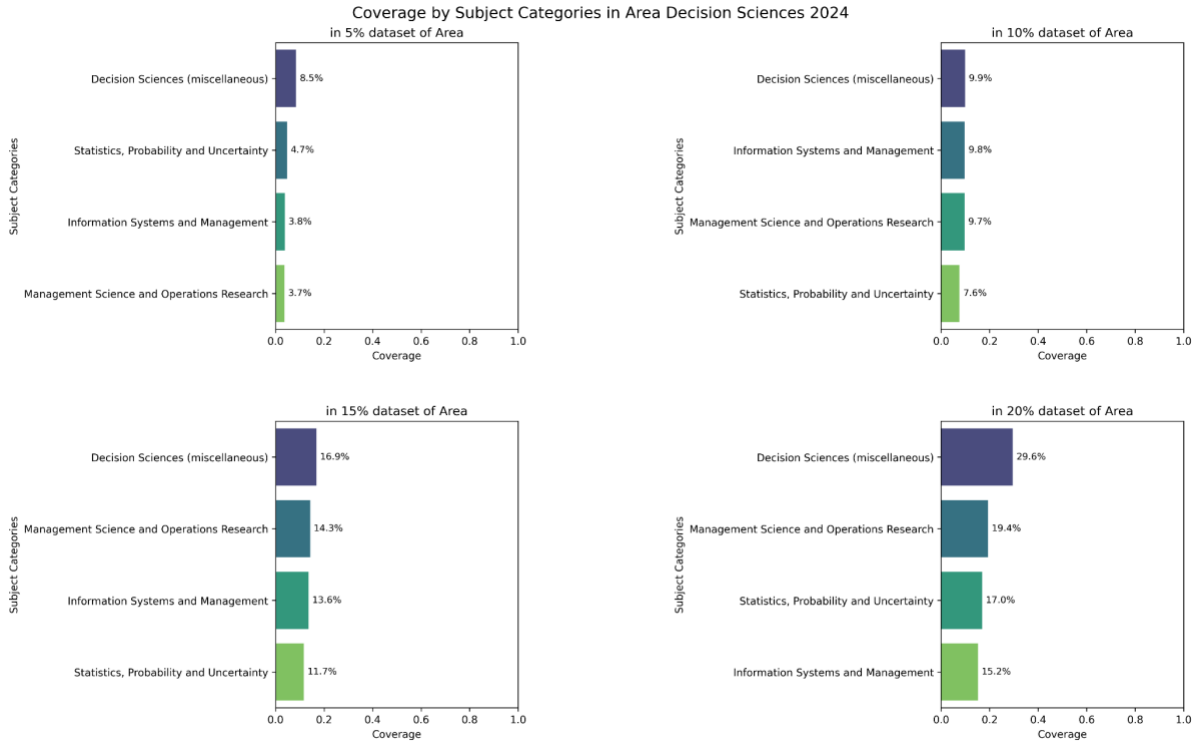

Figure 5.3: Comparison of coverage by Subjects in four sizes of the Area dataset

Overall, this analysis showed that individual Area datasets and datasets for each subject category generally correlate well with each other regardless of the selected cut- off size. This means that the ranking of journals within a specific subject category is in most cases comparable to the overall ranking within the corresponding Area, and when moving from Area to Subject, there is no fundamentally different structure of the ‘top’, but rather a refinement of

the representation of individual categories, since journals can be included in two categories at the same time.

It is important to note one limitation of this analysis. The comparison considered journals that may belong to several thematic categories at the same time, so there are duplicate entries between different groups in the data. This may lead to partial overlap of samples and, as a result, affect the calculation of the aggregate coverage percentage, slightly overstating the final values for some categories. At the same time, the study consciously adhered to a policy of using complete source data sets, without manually transferring or excluding duplicate journals from one category to another, as such intervention could significantly change the distribution of overlaps and, accordingly, significantly affect the results of the analysis and their interpretation.

5.2 Year-by-year analysis

Continuing the analysis of coverage, Figure 5.4 shows four bar charts comparing coverage by year (2012, 2019, 2022, 2024) for the Area of Computer Science with a fixed sample size of 15% (the remaining threshold values are given in Appendix C). Overall, the dynamics over the years appear to be even, without any sharp spikes, but there are noticeable changes in the distribution of categories at the top of the ranking. As shown in the plots, until 2022, the leading positions were mainly occupied by categories not directly related to AI, such as *Computer Science Applications*, *Software, Computer Science (misc.)*, and *Computational Theory and Mathematics*. Starting in 2022, the situation changes: *Artificial Intelligence* takes first place (18.7% in 2022 and 19.5% in 2024) and continues to maintain its leadership. This shift can be explained by the general growth in interest in generative models and practical AI applications, which has intensified since the widespread adoption of tools such as ChatGPT.

The results also clearly show a stable core area: categories such as *Software*, *Information Systems*, *Computer Science Applications*, and *Hardware and Architecture* are

present in the top rankings for all years considered and maintain comparable coverage values. This shows that the basic structure of the top of the CS ranking is generally stable, and the main changes are related to the redistribution of leadership and the strengthening of individual areas.

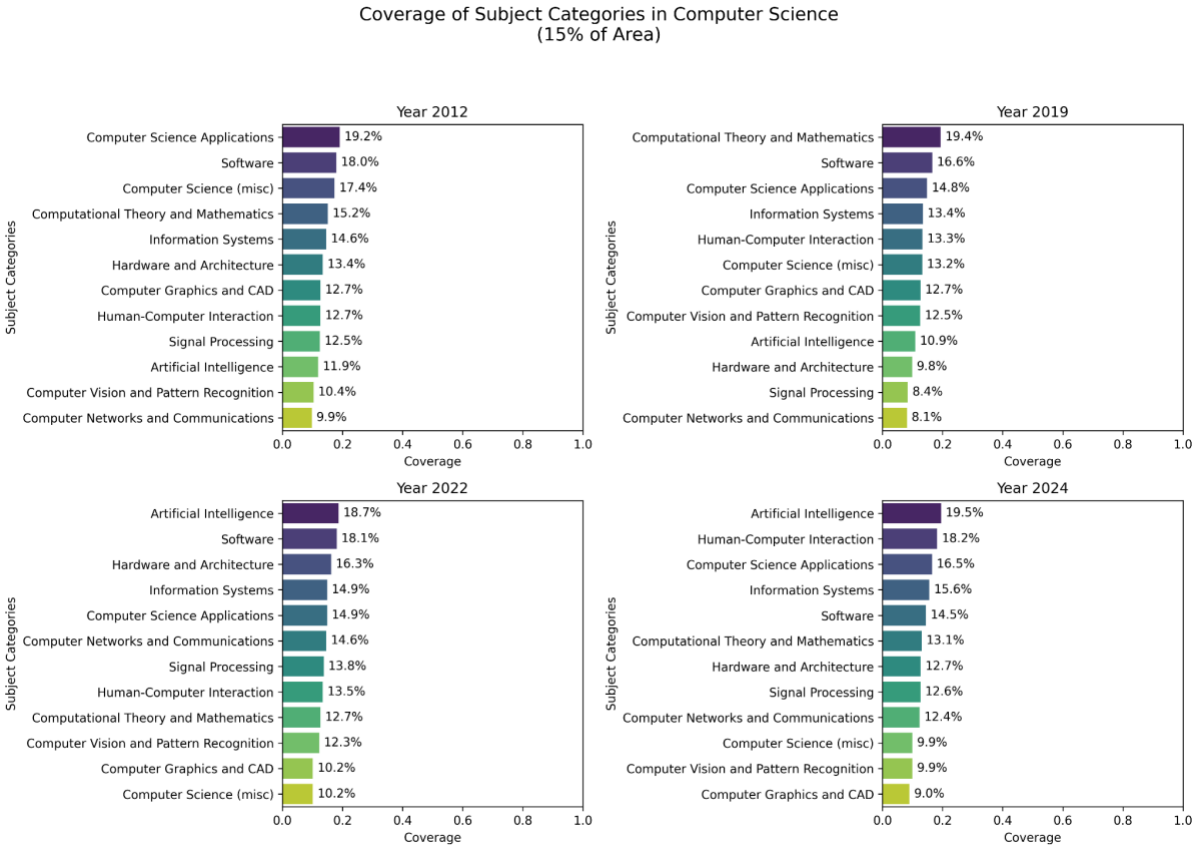

Figure 5.4: Comparison of coverage by subjects and years in four sizes of the CS Area dataset

Continuing the analysis, Figure 5.5 shows four bar plots with coverage by year (2012, 2019, 2022, 2024) for the Mathematics Area with a fixed sampling threshold of 15%. In general, it is noticeable that the distribution structure remains consistent, with the exception of the *Analysis* category, but the rest of the dynamics of the top remain similar

The main categories at the top of the ranking are repeated from year to year. In 2012 and 2019, the leaders are *Analysis*, *Statistics and Probability*, *Geometry and Topology*, and *Mathematical Physics* – that is, the classical areas of mathematics consistently form the top. In 2022 - 2024, Theoretical Computer Science stands out more prominently (about 19.4% in 2022 and 21.7%

in 2024), but the other key categories remain roughly in the same positions and with similar values.

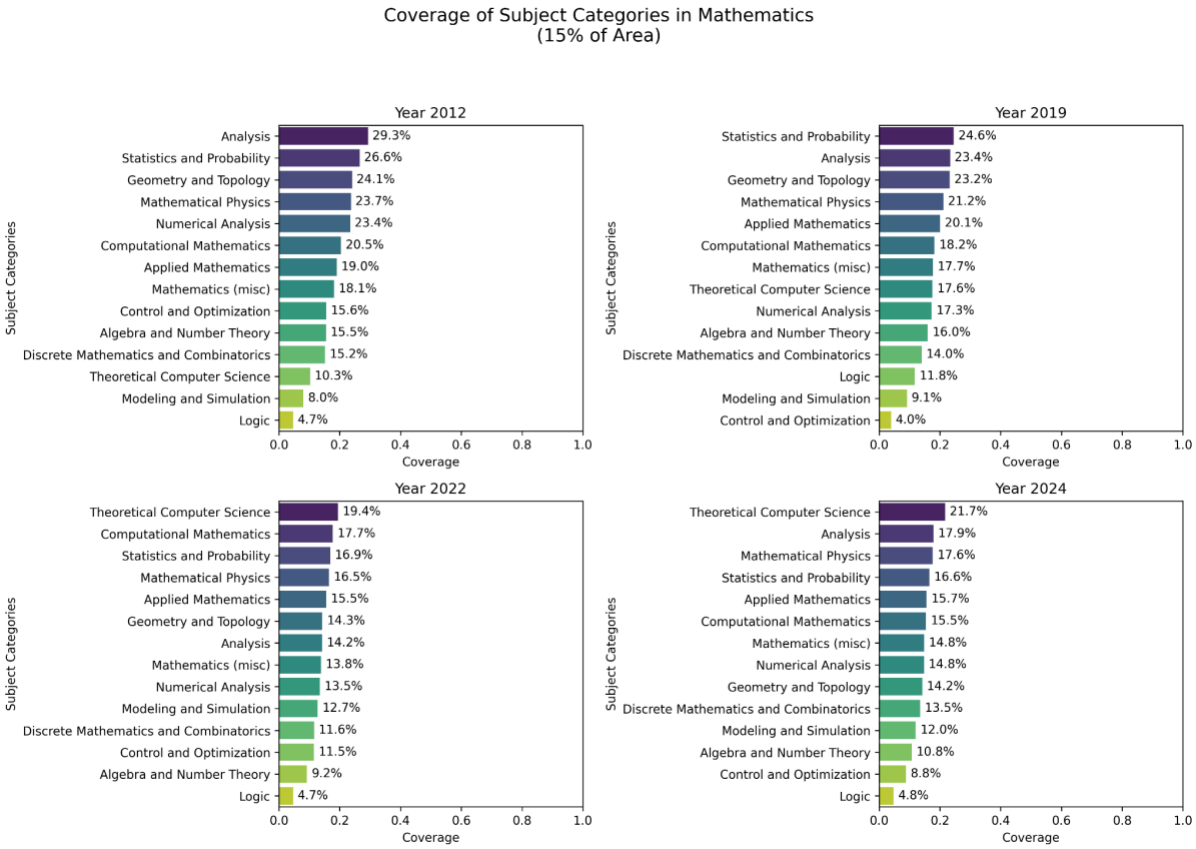

Figure 5.5: Comparison of coverage by subjects and years in four sizes of the Mathematics Area dataset

Analysis of the results for Area Decision Sciences also showed stable results among samples of different sizes from four years. The results can be found in Appendix C.

The annual analysis also performed a comparison of Jaccard similarity between each subject and Area Computer Science for different cut sizes (5%, 10%, 15%, 20%). Figure 5.6 shows separate graphs constructed for each subject, showing how the intersection of the subject's and Area's journal sets changes depending on the year and the selected threshold value.

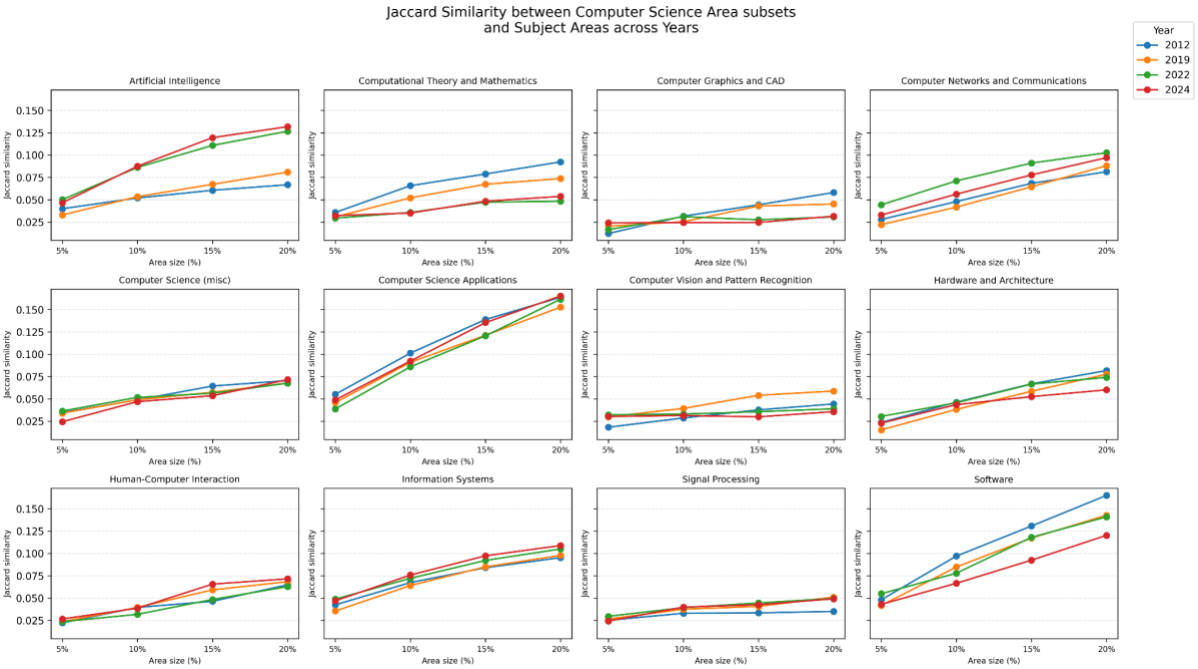

Figure 5.6: Jaccard Similarity Comparison Chart Across Years and Dataset Sizes for the CS Area

The results showed that the categories *Artificial Intelligence*, *Computer Networks and Communications*, *Computer Science Applications*, *Information Systems*, and showed a more significant increase in the Jaccard coefficient with an increase in the share of the Area CS dataset (from 5 to 20%), while for other categories the increase was less noticeable. This indicates that the listed categories overlap more with the upper subsets of Area CS.

Therefore, they are more represented in the upper segments of the ranking in the field under consideration. It is also important to note that this pattern remains consistent for all years considered (2012, 2019, 2022, 2024). There are no significant changes in the overall structure of intersections over the years, and the differences between years are only sporadic, with the exception of the growth of the metric for the *Artificial Intelligence* category for the years 2022 and 2024, which is associated with the growing popularity of the category in those years.

A similar analysis for the Area of Mathematics also showed stable results, with the exception of a sharp increase in the Analysis category in 2012 (Figure 5.7). In most categories, the expected growth of Jacquard is observed with an increase in the size of the Area cut.

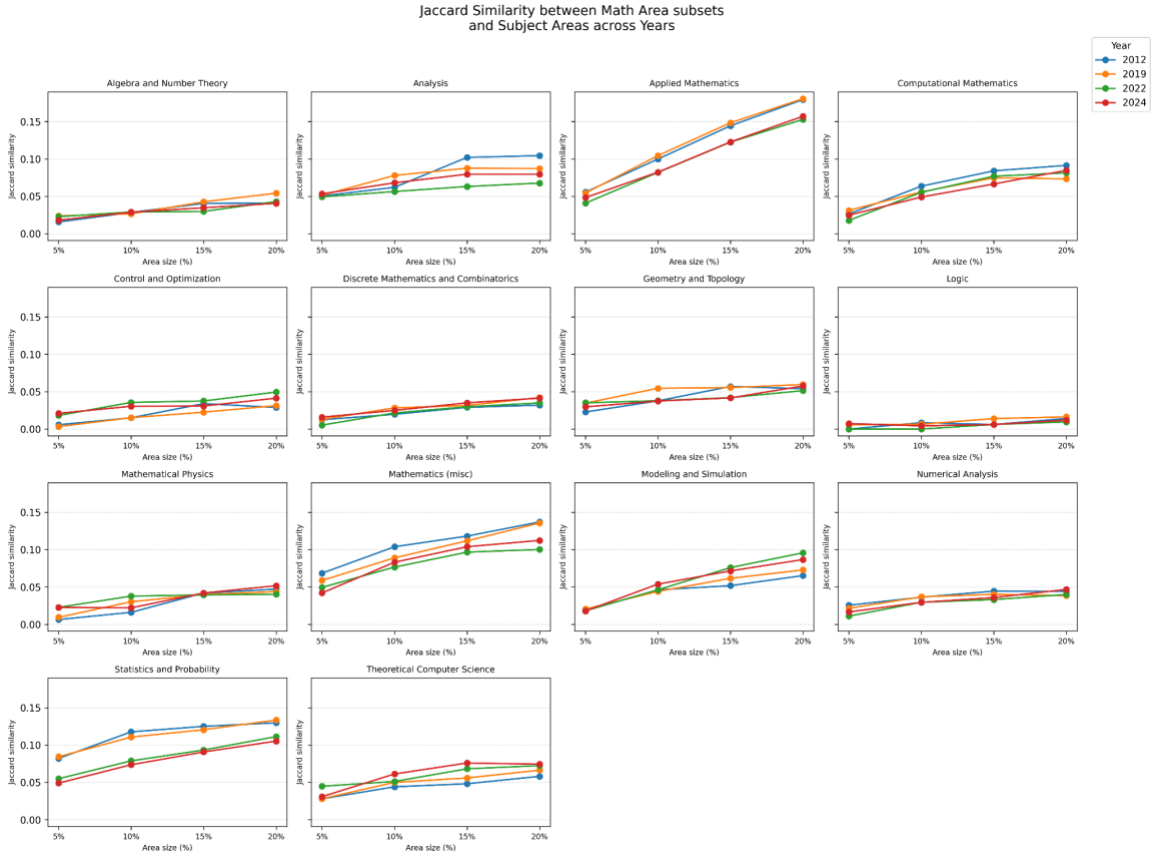

Figure 5.7: Jaccard Similarity Comparison Chart Across Years and Dataset Sizes for the Math Area

At the same time, the results of the sample for different years showed close similarities to each other, which indicates the stability of the results when crossing time intervals. The most pronounced values and growth of the Jaccard coefficient are seen in the categories *Applied Mathematics*, *Statistics and Probability*, and *Mathematics (misc.)*, while for several narrower categories, the values remain low and change less significantly.

An overall analysis of coverage dynamics by year for different areas showed fairly stable results. The analysis did not reveal any sharp spikes or other anomalies. However, the

results made it possible to identify patterns in the formation of cores within the top of the field, as well as trends among certain areas (for example, the growth of AI in the CS area), which become noticeable when comparing several years

Furthermore, a general analysis of the dynamics of the Jaccard coefficient by year for different areas showed a stable picture. When comparing 2012, 2019, 2022 and 2024, there are no sharp spikes or anomalies.

An annual analysis was also performed for the signatures of constructed graphs: connections between subject categories were compared using the Jaccard similarity measure. For the field of Computer Science (Table 5.1), the results show a moderate increase in the density of inter-category connections between 2012 and 2024.

num_nodes	num_edges	density	avg_degree	max_degree	avg_jaccard	max_jaccard	avg_overlap	num_components	year
12	24	0.363636	4.000000	9	0.106963	0.192308	196.333333	2	2012
12	43	0.651515	7.166667	10	0.125684	0.263110	249.744186	2	2019
12	32	0.484848	5.333333	10	0.101631	0.204698	71.875000	2	2022
12	31	0.469697	5.166667	10	0.104112	0.205170	77.387097	2	2024

Table 5.1: Signature table of Area CS subjects by year

A sharp increase in values can be observed in 2019, due to the large number of journals in the Area CS this year, which may partially increase the metrics. The parameters *num_edges*, *density*, *avg_degree*, and *avg_overlap* in 2012 show that the journals in that year were more concentrated and formed fewer connections between subject categories compared to 2024 (*num_edges* = 24 vs. *num_edges* = 32) .

The number of *num_components*, which shows the number of unconnected subgraphs, remains unchanged. This means that the structure of the division into components is preserved, and at least one category has no intersections with the others. In this case, it is *Computer Science*

(misc.), which is generally understandable, since misc includes journals that do not clearly belong to other subject categories.

In the Area of Mathematics (Table 5.2), the results show marked differences between 2012 and other periods in terms of the *num_components* indicator: four separate groups of subgraphs stand out, indicating a higher degree of category disconnection compared to other time periods. The categories *Mathematics (misc.)*, *Logic* and *Theoretical Computer Science* have no connections with other subject categories.

num_nodes	num_edges	density	avg_degree	max_degree	avg_jaccard	max_jaccard	avg_overlap	num_components	year
14	15	0.164835	2.142857	6	0.087047	0.146497	32.800000	4	2012
14	24	0.263736	3.428571	8	0.094714	0.215000	49.958333	2	2019
14	28	0.307692	4.000000	9	0.094984	0.216080	42.678571	2	2022
14	27	0.296703	3.857143	9	0.095336	0.215000	43.333333	2	2024

Table 5.2: Signature table of Area Mathematics subjects by year

The Decision Sciences Subject Area presents a stable dynamic between 2019 and 2024, whereas the 2012 sample shows a more pronounced sparsity of categories (*num_components* = 3). The Jaccard coefficient indicators have remained largely unchanged (Table 5.3).

num_nodes	num_edges	density	avg_degree	max_degree	avg_jaccard	max_jaccard	avg_overlap	num_components	year
4	1	0.166667	0.5	1	0.075209	0.075209	27.0	3	2012
4	2	0.333333	1.0	2	0.067819	0.073756	34.0	2	2019
4	2	0.333333	1.0	2	0.080324	0.088825	28.5	2	2022
4	2	0.333333	1.0	2	0.083726	0.089674	30.5	2	2024

Table 5.3: Signature table of Area Decision Sciences subjects by year

Comparing signature data graphs for different years allows for an assessment of the dynamics of changes in their structural characteristics, as well as the identification of stable trends in the formation of links between subject categories. The results show that over time there has been an increase in the number of connections between categories, which may

indicate a strengthening of their interconnections and an expansion of interdisciplinary intersections. Meanwhile, the other metrics remained mostly unchanged or showed only a minimal increase.

5.3 Cross-Area Analysis

For comparison between Subject Area characteristics of graphs by year, a summary table was created for all areas (Figure B.1). Based on this table, cosine similarity values were calculated, a heat map was constructed.In addition, a principal component analysis (PCA) was performed, which revealed the main directions of variation among graph metrics by subject area and year.

Figure 5.8 shows a heat map of cosine similarity between structural signatures of graphs (Area – Year). Overall, the similarity between signatures from different areas is low, indicating differences in their network structures. The most pronounced inter-area exception is observed for the pair *Mathematics(2012)* and *Decision Sciences* (all years), where the similarity values are significantly higher. Within the areas, there is a high consistency of signatures by year for *Decision Sciences*, while for *Mathematics* and *Computer Science*, the year 2012 stands out with less similarity to later periods. There is also a slight similarity between C*omputer Science* and *Mathematics* for 2019 - 2024, while in 2012 it is practically absent.

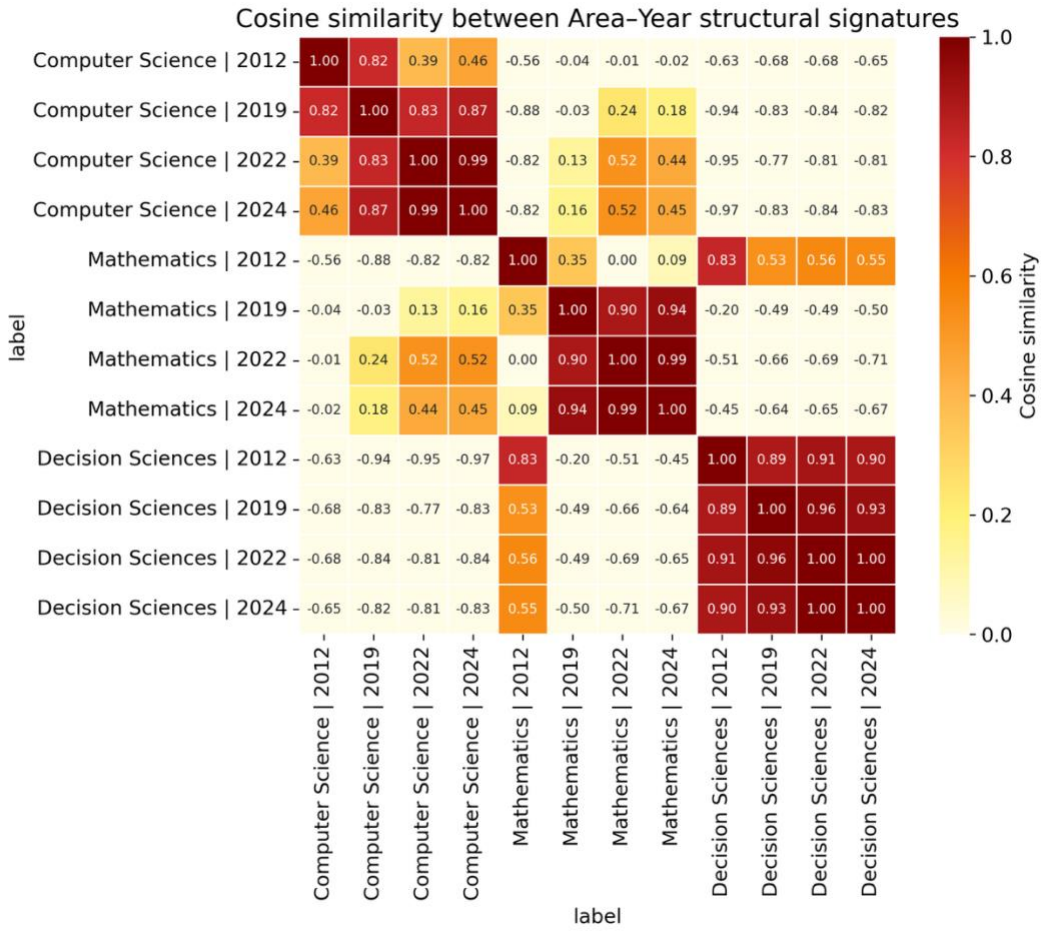

Figure 5.8: Heatmap of Cosine Similarity for Area – Year Graph Signatures

It is important to note that for the Area of Decision Sciences, the dataset is smaller and includes only four subject categories. With this number of vertices, the space of possible graph configurations is limited, and structural metrics take discrete values and change less significantly. Therefore, high cosine similarity between years may reflect both the actual stability of relationships and the effect of a small number of categories and data volume.

To represent the PCA results, the projection of structural signatures of graphs constructed for each region-year pair was used. On the scatterplot (see Figure 5.9), each point corresponds to one region in a specific year: the shape of the marker indicates the region, and the colour indicates the time slice. The graph shows that the points are mainly grouped by region. Thus, Computer Science forms a cluster at positive PC1 values, Decision Sciences at

negative PC1 values, and Mathematics shifts down the PC2 axis. It is also visible that the point for the Computer Science field for 2019 (marker ‘x’) is significantly shifted to the right along the PC1 axis (to 5). This may be since in 2019, the CS field included significantly more journals than in other years in this field, which potentially affected the calculated structural metrics (e.g., the number of intersections and related indicators).

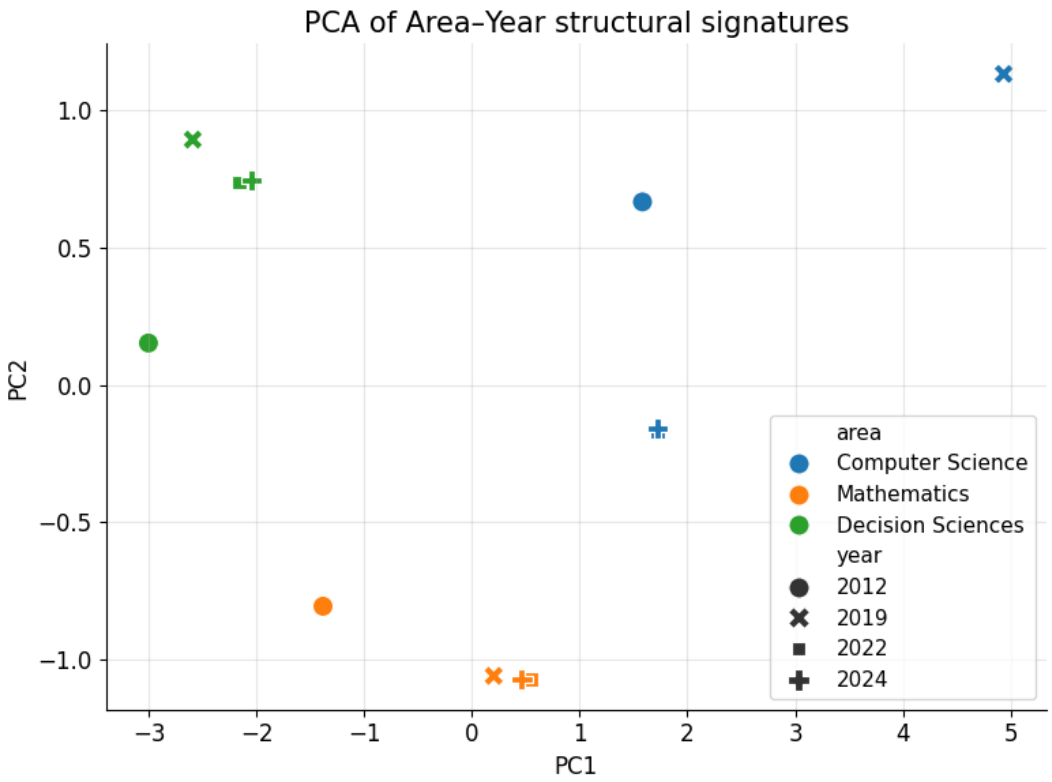

Figure 5.9: PCA of Structural Signatures by Scientific Area and Year

In summary, the cosine-similarity heat map shows generally low similarity between different subject areas, indicating distinct network structures. Both cosine similarity and PCA suggest that structural signatures are relatively stable within each area across years, while clear differences persist between areas. PCA additionally highlights a pronounced shift for Computer Science in 2019, pointing to an atypical year that may require further interpretation. Overall, these results indicate that such metrics can be used to characterize and compare subject areas over time, as well as to identify atypical years that may require additional interpretation.

Chapter 6

Conclusions

This study analyses scientific publications and journals based on data from the SCImago portal to compare and characterise scientific fields through their subject categories and mutual intersections. The study covers three fields: Computer Science, Mathematics, and Decision Sciences. The study also includes four time periods: 2012, 2019, 2022, and 2024. These periods were chosen as representative stages in the development of the scientific agenda – from the era associated with the AlexNet breakthrough to the current stage of growing interest in generative artificial intelligence.

For this work, several datasets were formed for each field and each year for that area. The data was collected manually from the SCImago portal—the download was performed iteratively for each field, for each subject category, and for each year, respectively. The data obtained was combined and structured in the Pandas environment, which ensured its uniform presentation and allowed for visualisation and comparative analysis while maintaining identical filtering rules for all time slices and fields. A total of 132 CSV files with logs for each field and each subject category for all 4 years were collected and processed.

The methodology and analysis showed how the composition and connectivity of subject categories change when moving from the highest-ranked journals to broader subsets. To this end, datasets were cut off at thresholds (5–20% of the top of the ranking), category coverage metrics within the area were used, and the Jaccard similarity coefficient was applied to assess intersections between sets of journals as the sample was successively expanded. Additionally, a graph representation of the relationships between categories was constructed, where the edges reflect the degree of overlap (according to the Jaccard coefficient), and signature tables of the structural characteristics of the resulting graphs were formed. A comparison of the structural profiles of the areas between the four years was also made.

The results demonstrate that the ratings of Area datasets and corresponding Subject Categories datasets in most cases show good agreement regardless of the selected top-% threshold. The dynamics of coverage and Jaccard coefficient indicators over the years (2012, 2019, 2022, 2024) remain generally stable, with no significant anomalies. At the same time, an increase in the number of links between categories has been observed over time, which may reflect the increase in interdisciplinary intersections. A comparison of fields by signatures and years (based on cosine similarity and PCA) indicates low similarity between different subject areas, with relative stability in the structure within each of them.

Analysis of the results showed that structural differences between the scientific fields under consideration persist throughout all the time periods studied, indicating the stability of their internal organisation and the specificity of their thematic areas. Thus, the work carried out demonstrated that the combination of various methods – dataset truncation (top-percent), coverage and intersection metrics, and graph structural features - forms a comprehensive and reproducible toolkit for characterising and comparatively analysing scientific fields.

As a further development to the work done, it would be advisable to develop a methodology for identifying duplicate journals within regions and integrate it into the existing workflow to automatically exclude such journals or reduce their weight in the ranking. Additionally, more detailed comparative studies between areas and time intervals should be conducted to refine and deepen the results obtained. It is also important to form a unified pipeline in which sets of fields and categories, as well as time intervals of interest, can be specified. This will simplify and speed up the entire task cycle: from data loading and pre-processing to structural analysis by year and calculation of the resulting metrics.

References

Glänzel, W., Moed, H.F. Journal impact measures in bibliometric research. *Scientometrics* 53, 171–193 (2002). ISSN: 0138-9130 , 1588-2861; DOI: 10.1023/A:1014848323806 URL https://doi.org/10.1023/A:1014848323806

Garfield, E. Citation Analysis as a Tool in Journal Evaluation. *Science*, *178*(4060), 471–479 (1972). ISSN: 0036-8075 , 1095-9203; doi: 10.1126/science.178.4060.471 URL http://www.jstor.org/stable/1735096

Pinski, G., Narin, F. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, Information Processing & Management, Volume 12, Issue 5, 1976, Pages 297-312, ISSN 0306-4573, doi: 10.1016/0306-4573(76)90048-0 URL https://doi.org/10.1016/0306-4573(76)90048-0

Hirsch, J.E. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics* 85, 741–754 (2010). doi: 10.1007/s11192-010-0193-9 URL https://doi.org/10.1007/s11192-010-0193-9

Bontis N, Serenko A. A follow-up ranking of academic journals. Journal of Knowledge Management, Vol. 13 No. 1 pp. 16–26 (2009), doi: 10.1108/13673270910931134
URL https://doi.org/10.1108/13673270910931134

Borja González-Pereira, Vicente P. Guerrero-Bote, Félix Moya-Anegón,
A new approach to the metric of journals' scientific prestige: The SJR indicator, Journal of Informetrics, Volume 4, Issue 3, 2010, Pages 379-391, ISSN 1751-1577, doi: 10.1016/j.joi.2010.03.002, URL https://doi.org/10.1016/j.joi.2010.03.002.

Google. Google Scholar. URL https://scholar.google.com

Harzing, A.-W. Publish or Perish. Harzing.com (updated 6 Feb 2026). URL https://harzing.com/resources/publish-or-perish

SCImago. SCImago Journal & Country Rank (SJR). URL https://www.scimagojr.com

Appendices

Appendix A. Lists of Subjects Category for each Area

Subject category name
Artificial Intelligence
Computational Theory and Mathematics
Computer Graphics and Computer-Aided Design
Computer Networks and Communications
Computer Science Applications
Computer Science (miscellaneous)
Computer Vision and Pattern Recognition
Hardware and Architecture
Human-Computer Interaction
Information Systems
Signal Processing
Software

Figure A.1: Table of Computer Science subject categories.

Subject category name
Algebra and Number Theory
Analysis
Applied Mathematics
Computational Mathematics
Control and Optimization
Discrete Mathematics and Combinatorics
Geometry and Topology
Logic
Mathematical Physics
Mathematics (misc)
Modeling and Simulation
Numerical Analysis
Statistics and Probability
Theoretical Computer Science

Figure A.2: Table of Mathematics subject categories.

Subject category name
Statistics, Probability and Uncertainty
Management Science and Operations Research
Information Systems and Management
Decision Sciences (miscellaneous)

Figure A.3: Table of Decision Sciences subject categories.

Appendix B. The feature matrix

num_nodes	num_edges	density	avg_degree	max_degree	avg_jaccard	max_jaccard	avg_overlap	num_compo	area	year	num_nodes
12	24	0.363636	4.000000	9	0.106963	0.192308	196.333333	2	Computer Science	2012	12
12	43	0.651515	7.166667	10	0.125684	0.263110	249.744186	2	Computer Science	2019	12
12	32	0.484848	5.333333	10	0.101631	0.204698	71.875000	2	Computer Science	2022	12
12	31	0.469697	5.166667	10	0.104112	0.205170	77.387097	2	Computer Science	2024	12
14	15	0.164835	2.142857	6	0.087047	0.146497	32.800000	4	Mathematics	2012	14
14	24	0.263736	3.428571	8	0.094714	0.215000	49.958333	2	Mathematics	2019	14
14	28	0.307692	4.000000	9	0.094984	0.216080	42.678571	2	Mathematics	2022	14
14	27	0.296703	3.857143	9	0.095336	0.215000	43.333333	2	Mathematics	2024	14
4	1	0.166667	0.500000	1	0.075209	0.075209	27.000000	3	Decision Sciences	2012	4
4	2	0.333333	1.000000	2	0.067819	0.073756	34.000000	2	Decision Sciences	2019	4
4	2	0.333333	1.000000	2	0.080324	0.088825	28.500000	2	Decision Sciences	2022	4
4	2	0.333333	1.000000	2	0.083726	0.089674	30.500000	2	Decision Sciences	2024	4

Figure B.1: Table of feature matrix.

Appendix C. Comparison of coverage by subjects

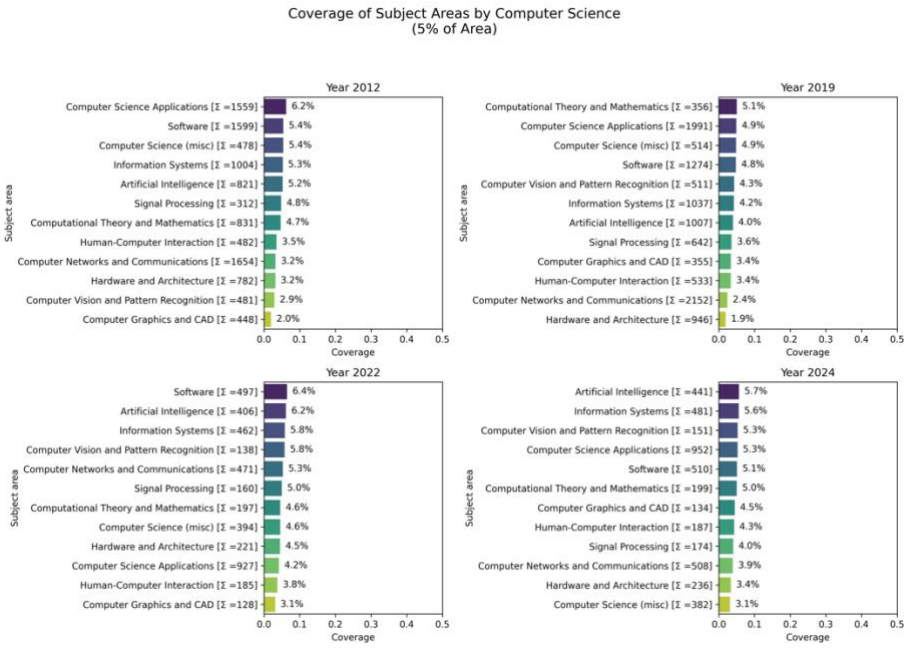

Table C.1: Comparison of coverage by subjects and years in four sizes of the CS Area (5% dataset)

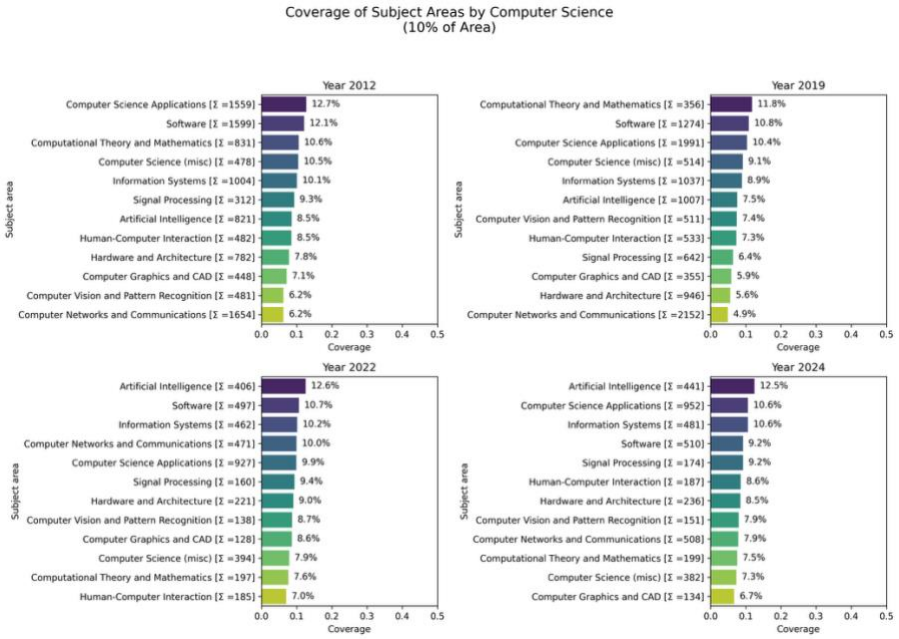

Table C.2: Comparison of coverage by subjects and years in four sizes of the CS Area (10% dataset)

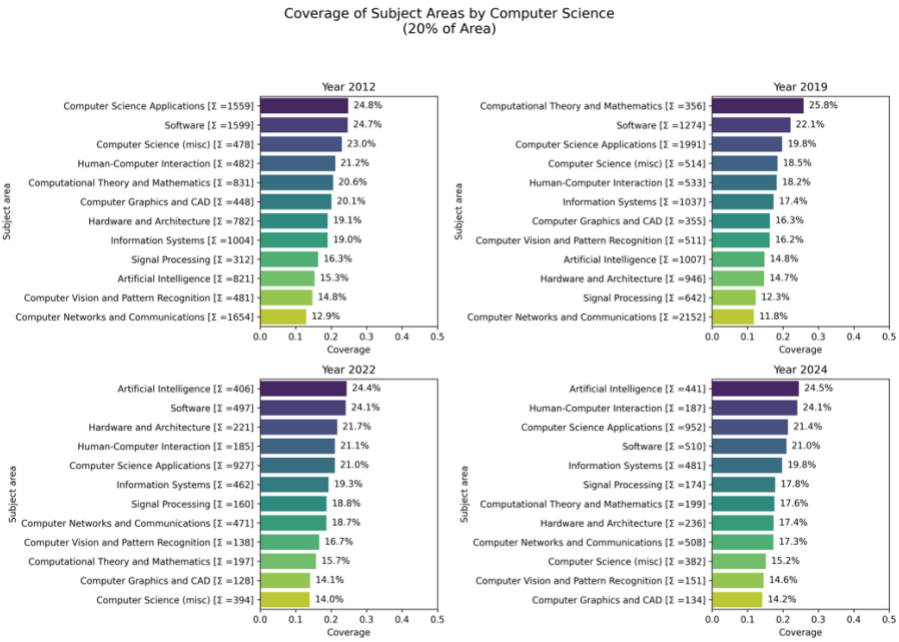

Table C.3: Comparison of coverage by subjects and years in four sizes of the CS Area (20% dataset)

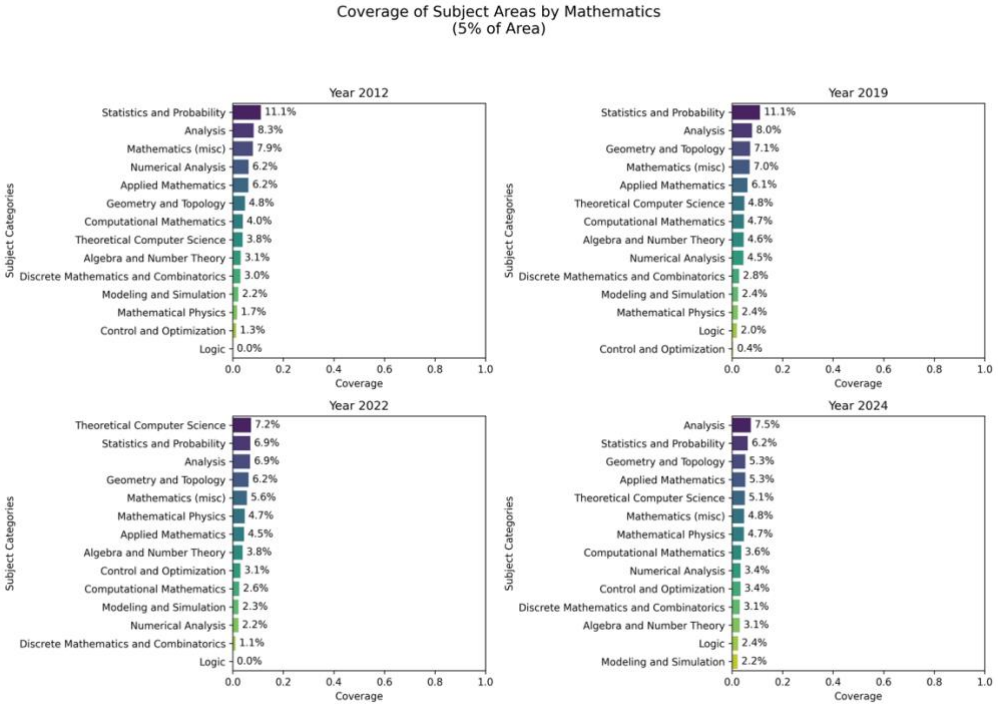

Table C.4: Comparison of coverage by subjects and years in four sizes of the Mathematics Area (5% dataset)

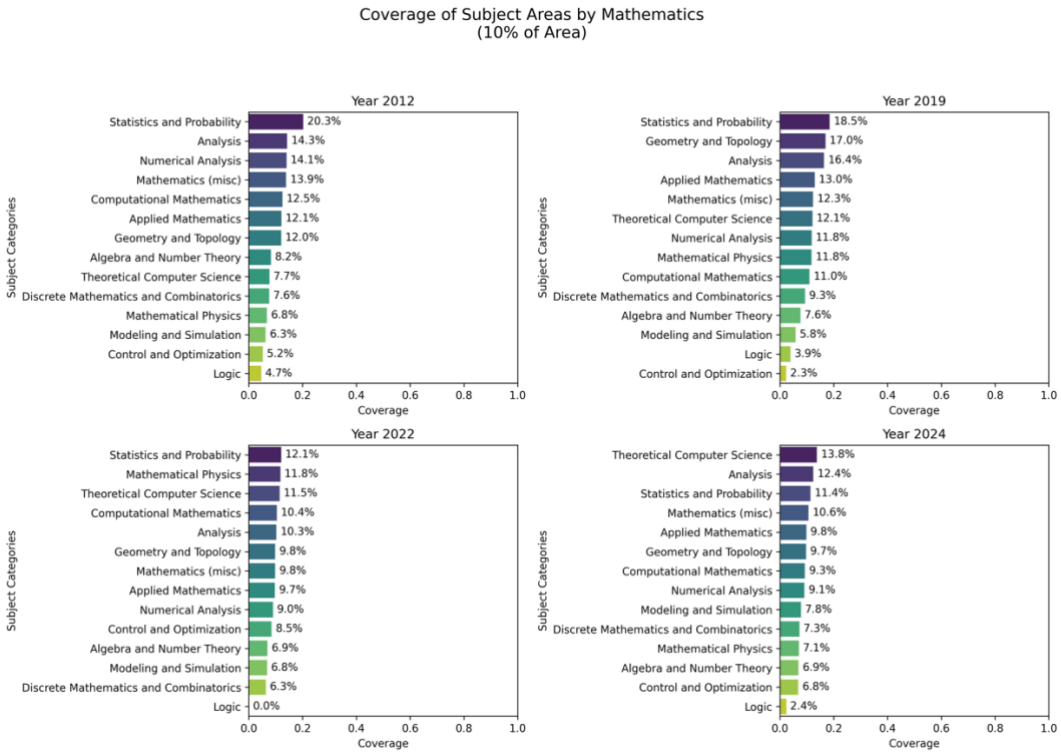

Table C.5: Comparison of coverage by subjects and years in four sizes of the Mathematics Area (10% dataset)

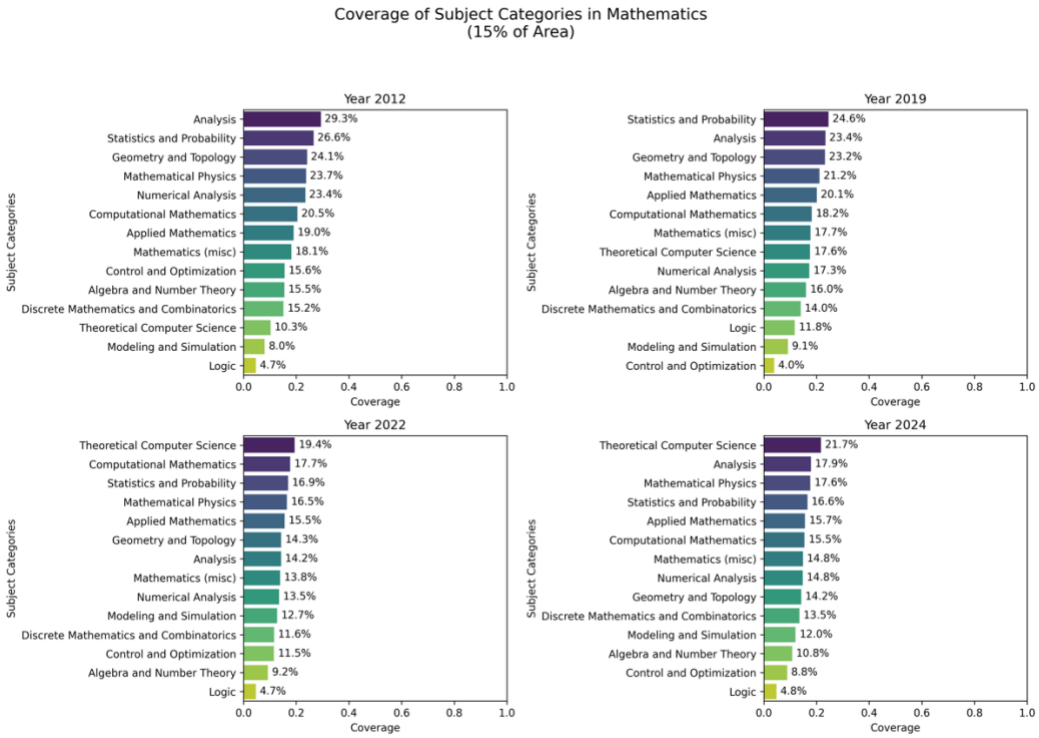

Table C.6: Comparison of coverage by subjects and years in four sizes of the Mathematics Area (15% dataset)

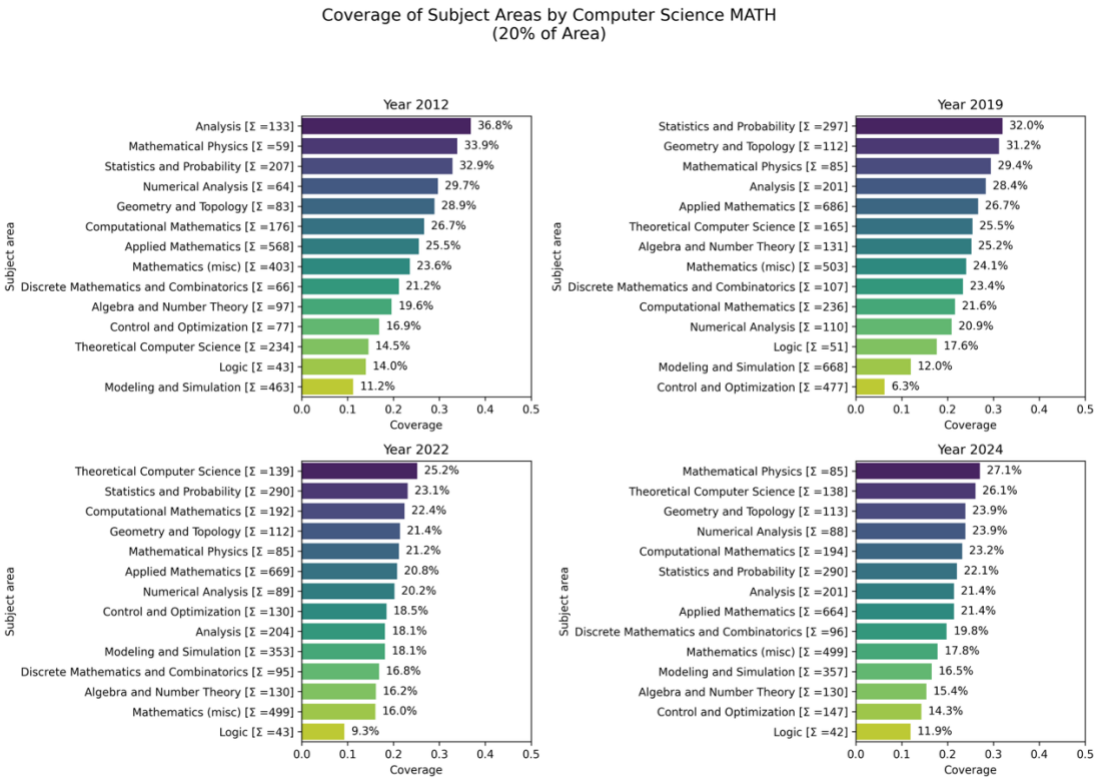

Table C.7: Comparison of coverage by subjects and years in four sizes of the Mathematics Area (20% dataset)

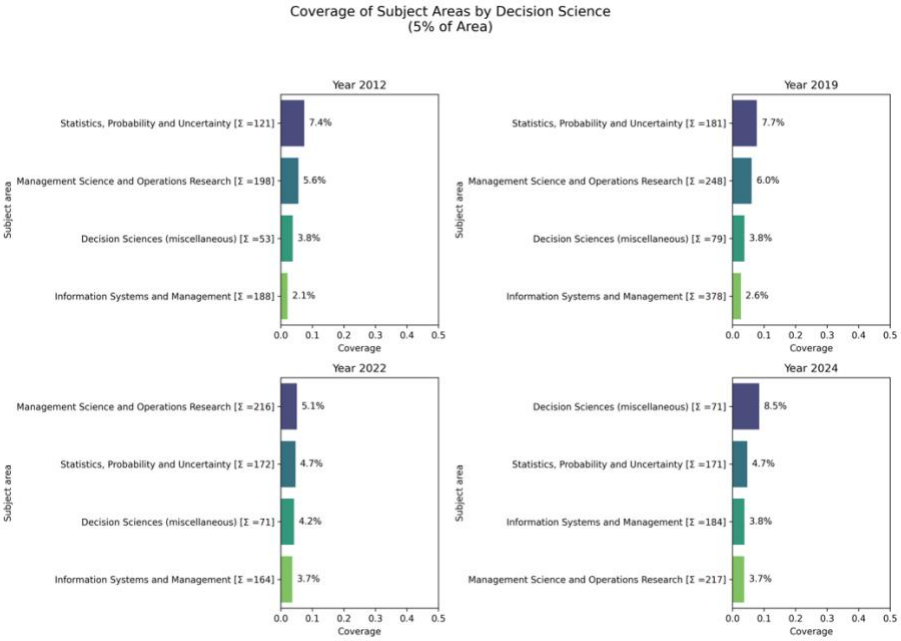

Table C.8: Comparison of coverage by subjects and years in four sizes of the Decision Sciences Area (5% dataset)

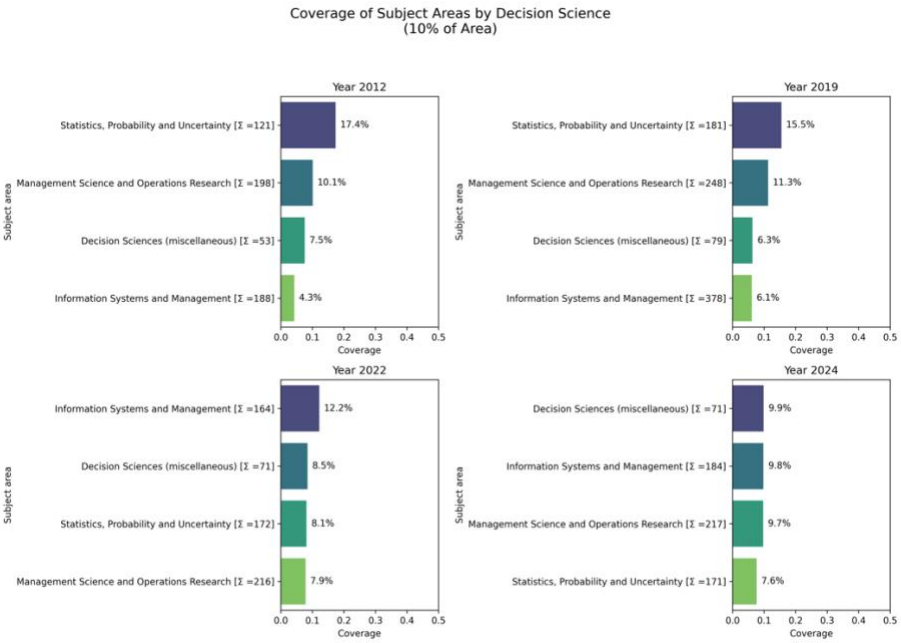

Table C.9: Comparison of coverage by subjects and years in four sizes of the Decision Sciences Area (10% dataset)

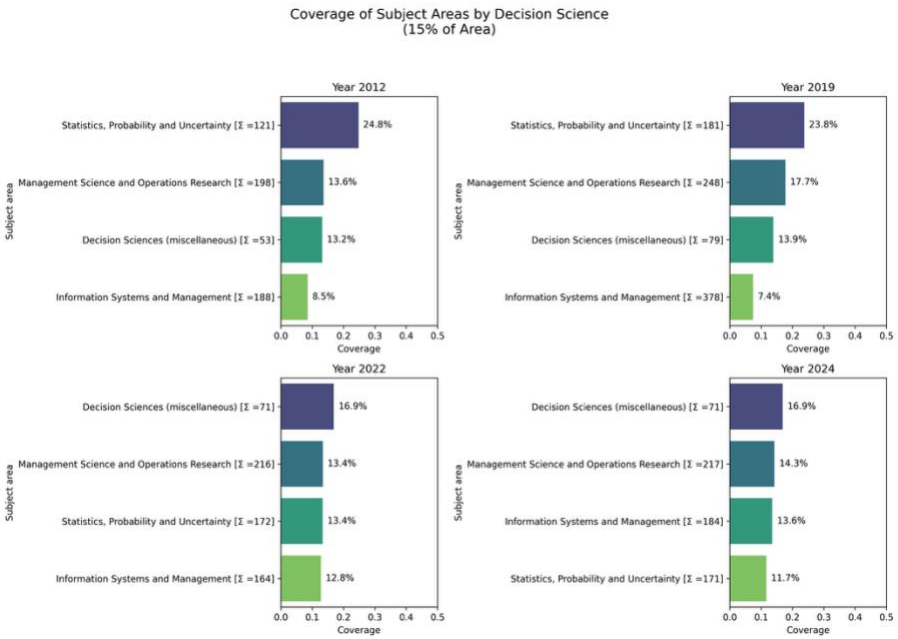

Table C.10: Comparison of coverage by subjects and years in four sizes of the Decision Sciences Area (15% dataset)

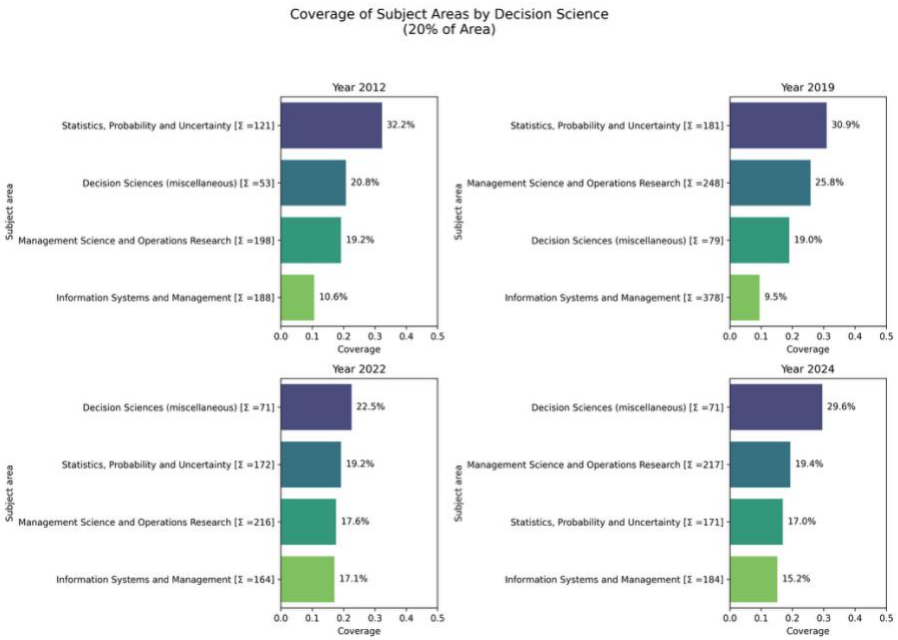

Table C.11: Comparison of coverage by subjects and years in four sizes of the Decision Sciences Area (20% dataset)