Research Proposal for "Research on Research on Research: Analyzing historical trends in statistical and computational research from 1993 to 2024"

Naren Prakash
Department of Statistics and Data Science, University of California, Los Angeles

February 25, 2025

Abstract

This paper aims to analyze the changes in research paper output for different statistical and computational fields over the time period from 1993 to 2024. The research papers used for this analysis are sourced from a dataset of papers from the pre-print journal arXiv.

1 Introduction

In recent years, with the fields of artifical intelligence and machine learning becoming important parts of the public lexicon and increasingly becoming involved in our day to day lives, we've seen firsthand large changes in statistical and computational research. With statistical methods increasingly becoming intertwined with computational principles, such as its integration with aspects of computer science, the future of statistics and computation appear to be one and the same. How does this current research landscape compare with that of the landscape a mere 30 years ago? This paper aims to analyze historical trends in statistical and computational research, as tracked by papers submitted to the online pre-print journal arXiv, in order to visualize the dramatic changes we've seen over the years and find any subfields growing in the present that could yet transform the landscape of the future. This analysis of historical trends will be conducted using a specific dataset available on Kaggle Mishra (2025).

2 Research Questions

- What statistical and computational fields have seen the largest increase in publications?
- How have the most published statistical and computational fields changed over time?
- What statistical and computational fields are projected to grow the most in the coming years?

3 Literature Review

Looking at themes in statistical and computational research is nothing new. For instance, Gelman & Vehtari (2021) analyzed the dominant statistical ideas of the past 50 years, suggesting inferential methods, computational algorithms, and data nalysis have been the most impactful in the shifting of the research landscape. Smaller subsets of time have also been analyzed, with Jun et al. (2018) using Google Trends to track the growth of different subfields of research (with an emphasis on big data and application). This paper aims to look at a similar problem with a different lens, using publication outputs themselves as a way of analyzing changes in research focus and interest. In doing this, the paper aims to also obtain an indication of the subfields with increasing research interest in the short term that may lend itself to future publications. Evaluating the research trends of the future has often involved modeling itself, such as the hype cycle model Dedehayir & Steinert (2016). This model aims to track the life cycle of technological innovations. In a similar vein, this paper aims to use current and recent paper production output to indicate trends of the near future. As the methodology involves using the research paper pre-prints themselves, it may provide a clearer picture of specific publication interests and trends rather than topics and concepts in general.

4 Methods

4.1 Data Partioning

While acknowledging the connected nature of statistical and computational research in the present and future, this paper will partition the data into two halves. One half will be comprised of research deemed statistical in nature, and the other half will be comprised

of research deemed as computational. This split in the data is done to narrow down the problem and allow for ease of analysis and interpretation of the results. Along with this, the data will also be subset in terms of time. In order to keep each yearly subset of papers as a representative sample of all research output for that year, the time range will be limited to exclude publication in 2025. This is done so that the resulting analysis will focus on comparison with full yearly samples of research data, rather than extrapolating from the research output in the year 2025 as of now.

4.2 Quantifying growth

For the purposes of this paper, growth will be represented by two metrics. Firstly, the simple percentage change from year to year for each subfield will be considered. In addition to this, the proportion of overall research represented by each subfield over time will also be used to evaluate growth in research interest and output.

4.3 Trend analysis and short-term prediction

Lastly, the metrics of growth as well as the original data and additional engineered features will be used to create a prediction model for the short term growth of research subfields. In correspondence with the data partitioning, separate models will be constructed for the statistical research and the computational research. The objective here is to produce a time series model for short term projections of growth in research interests and outputs.

4.4 Data Description

The arXiv paper dataset consists of 136,238 observations and 10 columns. The 10 columns present in the data are: id, title, category, category code, published date, updated date,

authors, first author, summary, and summary word count. This data is scraped directly from arXiv, aiming to provide a representative sample of research published on the platform.

4.5 Limitations

Focusing on primarily numerical data as a sign of growth indicates a relatively simple way of quantifying growth. In reality, growth is a more complex idea and could benefit from the use of paper content for text data processing to supplement the numerical figures of growth. This is a potential avenue of further exploration and work.

5 Potential Impact

References

Dedehayir, O. & Steinert, M. (2016), 'The hype cycle model: A review and future directions', *Technological Forecasting and Social Change* **108**, 28–41.

URL: *https://www.sciencedirect.com/science/article/pii/S0040162516300270*

Gelman, A. & Vehtari, A. (2021), 'What are the Most Important Statistical Ideas of the Past 50 Years?', *Journal of the American Statistical Association* **116**(536), 2087–2097.

URL: *https://doi.org/10.1080/01621459.2021.1938081*

Jun, S.-P., Yoo, H. S. & Choi, S. (2018), 'Ten years of research change using google trends: From the perspective of big data utilizations and applications', *Technological Forecasting and Social Change* **130**, 69–87.

URL: *https://www.sciencedirect.com/science/article/pii/S0040162517315536*

Mishra, S. (2025), 'arxiv scientific research papers dataset', Kaggle.

URL: *https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset*