# A Bibliometric Analysis of Astronomy Publications (2000–2025): Country Contributions, Citation Dynamics, and Journal Referencing Biases

Author name[a]

[a]*University of the Moon, , Earth, , ,*

**Abstract**

*Keywords:*

## 1. Introduction

## 2. Data Retrieval from the ADS API

To analyze trends in astronomy-related research, we retrieved bibliographic data from the SAO/NASA Astrophysics Data System (ADS)[1] using its public API. We focused on publications from 2000 to 2025 and selected the following keywords: *Asteroseismology, Exoplanet, Machine Learning, Deep Learning, Artificial Intelligence, Astrobiology, Star, Cosmology, Galaxy,* and *AGN*.

For each keyword, we queried all refereed journal articles, extracting metadata including publication venue, year, citations, abstracts, author affiliations, and Digital Object Identifiers (DOIs). We additionally retrieved reference lists for each article via batch requests, minimizing API calls while capturing citation networks.

Due to ADS's daily API request limits, data collection was performed over multiple days, with some articles retrieved more than once under different keywords or on different dates. These duplicate records occasionally contained differing metadata, especially for time-sensitive fields such as citation and readership counts. We combined all records by unique article identifier and retained the most recent information when merging such fields. Keyword labels were aggregated to preserve all relevant thematic associations.

A summary of the number of articles retrieved per keyword is given in Table 1. In total, we retrieved **1,354,544** records, of which **931,105** were unique by DOI. Of these, **928,619** had at least one listed author.

The code for data retrieval was written in Python using the `requests` and `pandas` libraries. A custom class, `ADSQuery`, handles API requests, batching, and rate-limiting. The results were saved and later consolidated using a deduplication script. The full pipeline is available on GitHub[2].

[1] `https://ui.adsabs.harvard.edu/`

[2] `https://github.com/YOUR-GITHUB-REPO`

Table 1: Number of articles retrieved from ADS by keyword (2000–2025).

| Keyword | Articles |
|---|---|
| AGN | 18,651 |
| Artificial Intelligence | 30,982 |
| Asteroseismology | 2,809 |
| Astrobiology | 356,469 |
| Cosmology | 146,547 |
| Deep Learning | 226,480 |
| Exoplanet | 12,571 |
| Galaxy | 154,130 |
| Machine Learning | 220,535 |
| Star | 185,370 |
| **Total retrieved** | **1,354,544** |
| **Total unique** | **931,105** |
| **With author metadata** | **928,619** |

## 3. First author affiliation and country

To enable geographic analysis of authorship patterns, we extracted and standardized country information from the first author's affiliation string in each article. These affiliation fields are often inconsistently formatted and may include full institutional addresses, abbreviations, or postal information, which makes automated extraction challenging.

The extraction process was implemented in a custom Python class, `AdsAffCountry`, which applies a multi-step procedure. First, we parsed the raw affiliation field—typically a stringified list—and extracted the final comma-separated segment of the first listed affiliation, assuming this was most likely to represent a country or regional identifier. In many cases, however, this segment contained ambiguous information, such as U.S. state abbreviations or postal codes. These cases were resolved using a set of heuristic rules that mapped known abbreviations and formats (e.g., "CA 93727" or "KS 66502") to "USA", along with handling common variants such as "U.S.A." or "United States".

After this initial suffix cleaning, many affiliations still could not be confidently matched to a recognized country. To improve coverage, we applied a normalization procedure using the

`country_converter` package, which maps free-form names to standardized country labels. Despite this, a large number of entries remained unresolved due to the presence of institution names, acronyms, or non-standard address elements (e.g., "Observatoire de Paris", "Max Planck Institut"). To address this, we sorted the unresolved suffixes by frequency of occurrence and manually reviewed the 200 most common ambiguous cases. With the assistance of ChatGPT, we assigned a valid country name to each of these based on institutional context and typical usage in the literature.

Finally, the output of each step was validated against a curated list of accepted short-form country names, and all successfully resolved affiliations were saved for further analysis.

Out of the **589,903** articles with at least one listed author, **81,140** (**13.8%**) had no identifiable country suffix in the first affiliation field. After the initial extraction, **244,161** articles (**41.4%**) lacked a valid country name. This number was reduced to **172,371** (**29.2%**) after suffix cleaning and heuristic interpretation, recovering **71,790** entries (**12.2%**). Following normalization and manual mapping, the number of unresolved entries dropped further to **92,838** (**15.7%**), yielding a total recovery of **151,323** articles (**25.7%**) beyond what was possible with strict parsing alone. Using

The resulting dataset, with standardized and validated country labels for the first author affiliations, was saved in a compressed format for subsequent statistical analyses.

## 4. Journal metadata and SJR classification

To analyze the publication venues of the articles in our dataset, we used the journal information provided by ADS and cross-referenced it with the SCImago Journal Rank (SJR) database[3]. While ADS includes the name of the publication outlet, it does not include journal-level metrics such as quartile rankings. These are provided by SJR, which publishes annual bibliometric data for journals indexed in Scopus.

The most recent SJR database available is for the year 2023. However, some journals in our dataset are no longer active or were not indexed in recent years. To improve matching coverage, we downloaded the SJR journal lists for the years 2023, 2020, 2015, 2010, 2005, and 2000. These databases were combined into a single reference table, ensuring that each journal was included only once. The integration was performed in reverse chronological order, starting from 2023 and progressively adding journals from earlier years only if they were not already present in the merged database.

We then developed a matching algorithm to associate each article's journal name with its corresponding entry in the SJR database. Exact string matches were insufficient due to minor naming inconsistencies (e.g., "The Astrophysical Journal" vs. "Astrophysical Journal"). To address this, we used fuzzy string matching, which compares two strings and assigns a similarity score based on their resemblance, even when they are not identical. This was implemented using Python's `difflib` library, which uses the Ratcliff/Obershelp algorithm to identify the longest contiguous matching subsequences between two strings. For example, "The Astrophysical Journal" and "Astrophysical Journal" received a similarity score of 0.91. We accepted matches with a similarity score of 0.9 or higher, which provided a good balance between robustness and precision.

To assess the reliability of this method, we manually inspected the 100 most frequently occurring journal names in the ADS dataset. Approximately 80% of them had an exact match in the SJR database, while the remaining 20% were all correctly identified through the fuzzy matching procedure. This validation supports the effectiveness of the algorithm in resolving small differences in journal naming conventions.

Initially, **178,194** articles could not be matched directly to any journal in the combined SJR database. After applying the fuzzy matching algorithm, the number of unmatched entries was reduced to **43,525**. These remaining unmatched records include articles from journals that are likely too new to be indexed by SJR, non-traditional publication sources, or venues outside the scope of the Scopus index.

The final output of this matching procedure included each journal's best available SJR quartile and the most recent year it was indexed. However, some journals—particularly newer or more specialized ones—remain unmatched. For example, *npj Science of Learning* and *npj Sustainable Agriculture* were not found in the SJR database. These journals are part of the Nature Partner Journals (npj) series and may be either too recent to be included in SJR or not indexed in Scopus at the time of data collection. In contrast, other npj titles such as *npj Quantum Information* are listed in SJR, indicating that inclusion varies across the series depending on indexing status and age. These limitations should be kept in mind when interpreting journal-level analyses.

## Acknowledgements

Thanks to ...

## Appendix A. Appendix title 1

## References

[3] `https://www.scimagojr.com/`