Black Sea Journal of Statistics

doi: 10.5281/zenodo.23124738

Open Access Journal
E-ISSN: 3108 – 4265

Research Article
Volume 2- Issue 1: 28-43 / June 2026

BIBLIOMETRIC MAPPING OF ZERO-INFLATED STATISTICAL MODELS

Memduh AK[1]*, Cem TIRINK[1]

[1]*Iğdır University, Faculty of Agriculture, Department of Animal Science, 76000, Iğdır, Türkiye*

Abstract: In this study, the evolution of scientific literature published on zero-inflated models was examined using bibliometric methods. Records obtained from the Web of Science Core Collection database were analyzed in the R environment using bibliometrix and related packages. The main analysis period was set as 1997–2025, and a total of 1,067 publications were evaluated. Annual scientific output, citation trends, leading sources, and authors were examined, along with keyword co-occurrence networks, temporal trends, emerging research topics, thematic structure, and reference co-citation networks. The findings indicate that the literature on zero-inflated models has grown rapidly, particularly after 2010, reaching its highest annual publication count in 2024. While zero inflation, overdispersion, zero-inflated Poisson, hurdle models, and the negative binomial model form the core conceptual framework of the field, Bayesian inference, marginalized models, INLA, and deep learning have emerged as topics gaining increasing prominence in recent years. Citation network analysis highlighted the central role of seminal works by Lambert, Mullahy, Hall, and Ridout in the field's intellectual structure. The results indicate that zero-inflated models constitute a dynamic, interdisciplinary field of statistical research that continues to diversify methodologically.

Keywords: Bibliometric analysis, Zero-inflated models, Zero-inflated Poisson, Overdispersion, Count data, Science mapping, Thematic analysis

***Corresponding author:** Iğdır University, Faculty of Agriculture, Department of Animal Science, 76000, Iğdır, Türkiye
E mail: memduh.ak.33@gmail.com (M. AK)
Memduh AK https://orcid.org/0000-0001-5894-8986
CemTIRINK https://orcid.org/0000-0001-6902-5837

Received: April 03, 2026
Accepted: May 21, 2026
Published: June 15, 2026

Cite as: Ak, M., & Tırınk, C. (2026). Bibliometric mapping of zero-inflated statistical models. *Black Sea Journal of Statistics, 2*(1), 28-43.

1. Introduction

Count data is a type of data commonly encountered in many scientific disciplines that indicates how many times an observed event occurs within a specific time frame, area, or experimental unit. Examples of this type of data include the number of diseases or symptoms, the frequency of traffic accidents, insurance claims, species abundance, the number of harmful organisms, production defects, the frequency of healthcare utilization, and the number of technological applications. One of the most fundamental models in the analysis of count data is Poisson regression. However, the fundamental assumption of the Poisson distribution—that the mean equals the variance—is often not met in real-world data structures. In particular, overdispersion—defined as the observed variance being significantly larger than the mean—poses a significant modeling problem in count data analysis. While negative binomial regression is one of the primary alternatives developed to address this issue, in some datasets, a significant portion of the overdispersion stems from the fact that the number of observed zero values is much higher than predicted by the standard Poisson or negative binomial distributions (Cameron & Trivedi, 2013; Hilbe, 2011). This phenomenon is referred to in the literature as "excess zeros" or "zero inflation".

When zeros occur at a higher rate than expected in the normal counting process, this can affect not only the goodness-of-fit of classical counting models but also parameter estimates and the scientific conclusions drawn from those estimates. The fundamental idea behind zero-inflated models is that not all observed zeros may stem from the same data-generation mechanism. In this approach, a portion of the zero observations is treated as "structural zeros", arising from a separate condition where the occurrence of the event in question is impossible or extremely unlikely in practice. In contrast, the remaining zeros are considered "sampling zeros", which occur randomly from the normal counting distribution. Consequently, a zero-inflated model is generally expressed as a mixture of a component that produces zeros and a counting component that can produce both positive counts and some zeros. This approach allows zeros to be modeled not merely as a distributional issue but as an indicator of different data-generation mechanisms that may underlie the observed process (Lambert, 1992; Ridout et al., 1998; Young et al., 2022a). Lambert's (1992) Zero-Inflated Poisson (ZIP) regression model, developed for production errors, has become one of the most fundamental studies of this approach in modern statistical literature and has served as an important starting point for subsequent

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

methodological developments.

However, the conceptual foundation for modeling census data with excessive zeros had begun to take shape even before Lambert's (1992) work. Mullahy (1986) emphasized that standard census models might be insufficient to explain heterogeneity in data-generation processes and made significant contributions to modified census models, particularly through hurdle models. Although hurdle and zero-inflated models can be applied to similar data structures, they are not conceptually the same. In hurdle models, zero and positive outcomes are generated by two separate processes, and the positive count distribution is truncated at zero; therefore, zeros cannot be generated from the count component. In zero-inflated models, however, zeros can arise both from a specific zero-generation mechanism and from the normal process of a Poisson, negative binomial, or other count distribution. This distinction is not merely a difference in mathematical parameterization but is important for the interpretation of model parameters and the scientific justification of model selection (Mullahy, 1986; Feng, 2021).

Although the ZIP model is the most common zero-inflated approach, the fact that the mean-variance equality assumption of the Poisson component is not satisfied for nonzero observations has necessitated the use of more flexible distributions. For this reason, model families based on the zero-inflated negative binomial (ZINB), the zero-inflated generalized Poisson, and other discrete distributions have been developed. In particular, using the ZIP model when there is additional excess dispersion among nonzero observations can lead to significant biases in parameter estimates. Ridout, Hinde, and Demétrio (2001) demonstrated the significance of this problem through a score test they developed to compare the ZIP model with the ZINB alternative. Therefore, a high proportion of zeros in a dataset does not, by itself, indicate that the ZIP model is sufficient; it is necessary to evaluate both the structure of the zeros and the distributional characteristics of the positive counts (Ridout et al., 2001; Hilbe, 2011; Cameron & Trivedi, 2013).

Another noteworthy aspect of the development of models inflated with zeros is the growing importance of "the relationship between model selection and the scientific explanation of the data-generation mechanism". The fact that a dataset contains a large number of zeros does not always imply the existence of two distinct biological, economic, or physical states. Unobserved heterogeneity, sampling design, low event rates, missing explanatory variables, or differences in the tails of the distribution can also lead to high zero rates. Therefore, selecting a zero-inflated model based solely on information criteria or goodness-of-fit results may lead to the assumption of an incorrect data generation process. For example, in the traffic safety literature, the theoretical validity of zero-inflated models implying a binary structure, such as "completely safe" and "unsafe" road segments, has long been debated (Lord et al., 2005). Similarly, it has been shown that it is inappropriate to automatically use the Vuong test as a zero-inflated test in model comparisons (Wilson, 2015). These discussions highlight that the family of zero-inflated models should be evaluated not merely as statistical tools that provide a better fit, but as models that carry explicit assumptions about the data-generating mechanism.

In parallel with methodological advances, zero-inflated models have spread to a wide variety of application areas. Following Lambert's (1992) classic application to production errors, these models have been extensively used, particularly in biometrics, epidemiology, dentistry, ecology, agricultural sciences, transportation, insurance, and the social sciences. Böhning et al. (1999) addressed the high zero rates in the epidemiology of dental caries using the ZIP model; subsequent studies have shown that zero-inflated regressions have become increasingly common in dental epidemiological research (Preisser et al., 2012). However, Preisser et al. (2012) pointed out that interpreting parameters related to "sensitive" and "non-sensitive" groups, which are defined as latent in traditional zero-inflated models, as general population effects could be misleading. This issue became one of the key justifications for the subsequent development of marginalized zero-inflated models, leading to approaches that enable direct estimation of population-level effects via the marginal mean (Preisser et al., 2016). Thus, the literature has progressed from merely modeling zero inflation to more accurately linking model parameters to the scientific questions in applied research.

Ecology and environmental sciences are another important area of application for zero-inflated models. Since it is quite common for a species not to be observed at certain sampling points in studies of species presence and abundance, the proportion of zero values can be much higher than expected under the classical Poisson distribution. Chiogna and Gaetan (2007) extended the zero-inflated Poisson structure in animal abundance data using semiparametric methods, thereby enabling the modeling of nonlinear relationships with environmental variables. Arab et al. (2008), on the other hand, analyzed data with high zero rates obtained with different fishing gears in fish abundance studies on the Missouri River, using a Bayesian ZIP model that incorporates random effects. These examples demonstrate that zero inflation is not merely a theoretical distribution problem but a modeling issue directly related to the sampling process and ecological mechanisms. Today, methodological and computational research continues addressing zero inflation and over-dispersion together in ecological survey data (Sidumo et al., 2024).

Similarly, in agricultural sciences, there may be production, disease, pest, adaptation, or technology-use data that contain many zeros. Hall (2000) modeled within-subject correlation and between-subject heterogeneity in repeated measurements by adding random effects to Poisson and binomial models inflated

for zeros and applied this approach to experimental data on garden crops. Isgin et al. (2008), on the other hand, used ZIP and ZINB models to model the intensity of farmers' adoption of precision agriculture technologies. These studies demonstrate that the zero-inflated approach is important not only for correcting for excess zeros in event counts but also for examining, within the same model, the distinct mechanisms underlying "participation/non-participation in a process" and "the magnitude of the event count when participation occurs."

The use of zero-inflated models in insurance and transportation sciences also clearly demonstrates the field's interdisciplinary nature. The fact that many policyholders report no claims during the observation period has created a natural setting for applying zero inflation. Boucher et al. (2007) examined the use of these models in risk classification by comparing zero-inflated, hurdle, and various hybrid Poisson models in auto insurance portfolios. In the transportation and traffic safety literature, the fact that many intersections or road segments record no accidents during certain periods has led to the widespread use of ZIP and ZINB models (Lord et al., 2005). Subsequent studies have addressed unobserved heterogeneity and complex dependency structures in greater detail through stochastic-parameter, multivariate, and Bayesian zero-inflated models. This broad scope of application demonstrates that zero-inflated models have evolved from being methods specific to a single discipline into a general-purpose family of statistical models.

The expansion of the literature on zero-inflated models is not limited to their application in different fields. It has also led to the incorporation of features such as increasingly complex data structures, repeated measurements, clustered observations, spatial and temporal dependencies, multivariate counts, and random effects into these models. While studies such as those by Hall (2000) and Min and Agresti (2005) focused on random effects and repeated measurements, subsequent research has expanded into spatial models, count time series, multivariate models, and Bayesian hierarchical structures. Young, Roemmele, and Shi (2022b), by reviewing the literature beyond classical univariate zero-inflated regressions, demonstrated that significant progress has been made, particularly regarding correlated counts, discrete time series, spatial models, and multivariate zero-inflated structures. This progress demonstrates that, over time, the literature on zero-inflated models has evolved from a simple Poisson mixture model into a much broader family of models.

However, the growing popularity of zero-inflated methods has also sparked new debates about which model to prefer in which situations. Although ZIP, ZINB, and hurdle models are often compared as alternatives in practice, it should not be overlooked that they do not represent the same scientific mechanism. Feng (2021) emphasized that although zero-inflated and hurdle models can be used on similar datasets with a high proportion of zeros, their mechanisms for generating zeros differ. Similarly, findings in the dental literature regarding the misinterpretation of model parameters have shown that appropriate model selection cannot be reduced solely to fit criteria such as AIC, BIC, or log-likelihood (Preisser et al., 2012). For this reason, in the contemporary literature, model selection is evaluated in conjunction with the source of zeros, outliers, the structure of dependence, the population for which the effects of explanatory variables are to be interpreted, and the model's compatibility with the scientific question. Such a development suggests that the field is becoming increasingly diverse not only in the number of publications but also in its conceptual and methodological dimensions.

As the literature has grown more extensive and interdisciplinary, it has become difficult for traditional narrative-style compilations alone to reveal the field's full structure. While comprehensive reviews such as those by Young et al. (2022a, 2022b) explain in detail the theoretical and computational development of zero-inflated model families, bibliometric methods are needed to simultaneously and quantitatively reveal a research field's patterns of scientific output over time, its pioneering sources, influential authors, collaborations between countries and institutions, conceptual clusters, intellectual foundations, and current research frontiers. Bibliometric analysis enables the evaluation of a research field's development through quantitative indicators and relationship networks by systematically examining large volumes of scientific publication data (Donthu et al., 2021). The science mapping approach, on the other hand, focuses not only on the volume of output in a field but also on its "conceptual, intellectual, and social structure" (Cobo et al., 2011).

In bibliometric science mapping, keyword co-occurrence networks help identify the conceptual structure and thematic clusters of a research field; co-citation analyses reveal the intellectual knowledge base underpinning the field; bibliographic coupling analyses identify closely related current research fronts; while country and author collaboration networks enable an examination of the field's social organization. In addition, thematic maps, thematic evolution analyses, and temporal keyword studies can be used to assess how research topics have become more central, diversified, or lost significance over time. While tools such as VOSviewer have popularized the visual mapping of bibliometric networks (van Eck & Waltman, 2010), the `bibliometrix` R package has brought together bibliometric performance analysis and science mapping methods within a programmable and integrated working environment (Aria & Cuccurullo, 2017). The capabilities of these tools provide a suitable analytical framework for examining the scientific evolution of statistical methods, particularly those with long historical development and broad interdisciplinary reach.

It would not be accurate to claim that the bibliometric perspective on zero-inflated models has not been fully explored. Indeed, Tajuddin and Ismail (2025) conducted a three-decade bibliometric study examining "inflated" count distributions. While this study provides an important macro-level overview of the scientific development of inflated-count models, it also leaves room for a more focused bibliometric framework, specifically for zero-inflated models. In particular, a more detailed science mapping approach could examine how zero-inflated Poisson, zero-inflated negative binomial, and their numerous extensions have spread across different disciplines; how the field's conceptual clusters, key references, and current research frontiers are structured; which themes have emerged over time; and how international collaboration has taken shape.

At this point, the scope of the present study differs from the previous bibliometric research in several respects. First, the study focuses directly on "zero-inflated models" rather than the broader category of "inflated count distributions," thereby examining the literature on statistical methods related to zero inflation as a more homogeneous field of research. Second, the "Web of Science Core Collection" is used as the data source, and the analysis period is updated through 2025 for publication years. Third, rather than relying solely on performance indicators such as publication, source, author, and country frequencies, the study aims to evaluate complementary science mapping techniques collectively, including keyword co-occurrence networks, temporal keyword structures, thematic maps and thematic evolution, co-citation networks, bibliographic matching, and international collaboration networks. In this way, it will be possible to reveal not only "how much" the zero-inflated model literature has grown, but also "the statistical ideas around which it has developed, the studies upon which it is built, the research frontiers toward which it is directed, and the disciplines and research communities through which this development has taken place".

In this context, the primary objective of this study is to systematically examine the evolution of the scientific literature published on zero-inflated models using bibliometric and science mapping methods. To this end, the study aims to: (i) identify trends in scientific output and citations by year; (ii) identify the most productive and influential journals, authors, and countries; (iii) examine international research collaborations; (iv) determine the conceptual structure and thematic clusters of the field through author keywords; (v) assess changes in themes over time; (vi) to uncover the field's intellectual foundation through co-citation analysis, (vii) to identify current research frontiers through bibliographic matching, and (viii) to identify emerging research topics. In this regard, the study aims to evaluate the development of zero-inflated models in the statistical literature, spanning more than thirty years, within a comprehensive framework and to provide a systematic reference on the field's current structure and future potential research directions for both researchers developing methods and scientists applying these models in practical fields.

2. Material and Methods

2.1. Research Design

This study was conducted using bibliometric analysis and the science mapping approach to examine the evolution over time of scientific studies published on zero-inflated models, as well as their structure of scientific output, conceptual orientations, intellectual foundations, and the social structure among researchers. While bibliometric analysis enables the quantitative examination of large volumes of scientific publication data, the science mapping approach reveals the conceptual, intellectual, and social structures of research fields through networks (Cobo et al., 2011a; Donthu et al., 2021).

The study did not limit itself to examining only basic performance indicators; instead, it combined author-keyword co-occurrence analysis, thematic mapping, thematic evolution, co-citation analysis, bibliographic matching, and cross-national scientific collaboration analyses. Thus, the "zero-inflated" model literature was evaluated under different, complementary bibliometric dimensions: performance structure, conceptual structure, intellectual structure, and social structure. In designing the bibliometric analysis process, we drew upon the general science mapping approaches proposed by Aria and Cuccurullo (2017), Cobo et al. (2011a, 2011b), and Donthu et al. (2021).

2.2. Data Source and Literature Review Strategy

Bibliographic data were obtained from the Web of Science Core Collection (WoSCC) database. A key factor in selecting Web of Science was its provision of detailed bibliographic information on publications, as well as essential fields required for scientific mapping analyses—such as citation records, authors, institutions, countries, keywords, source journals, and cited references.

The literature review was conducted in August 2026. The search was performed in the Topic (TS) field of the Web of Science database. Since the topic field encompasses terms found in publication titles, abstracts, and keywords, the aim was to capture as broad a range as possible of both methodological and applied studies on zero-inflated models.

The final Web of Science query used in the study was constructed as follows:

TS=("zero-inflated model" OR "zero inflated model*" OR "zero-inflated regression" OR "zero inflated regression*" OR "zero-inflated distribution" OR "zero inflated distribution*" OR "zero-inflated count" OR "zero inflated count*")"

When determining the search terms, both hyphenated and non-hyphenated forms were included separately in the query. By adding an asterisk (*) at the end of the terms, the study aimed to capture singular and plural

forms as well as different word forms derived from the same root. For example, the phrase "zero-inflated model" covers both "zero-inflated model" and "zero-inflated models."

Since one of the primary objectives of the study was to reveal the interdisciplinary spread of zero-inflated models, no field restrictions were applied based on Web of Science Categories. Similarly, no restrictions were imposed regarding document type or publication language during the initial phase. Thus, the dataset was designed to include studies that use zero-inflated models across statistics, biometrics, health sciences, ecology, agriculture, economics, insurance, transportation, and other applied fields.

The records obtained from WoSCC were exported as a Plain Text File (.txt) using the "Full Record and Cited References" option. Due to export limitations, the records were downloaded as multiple files and merged into a single bibliographic dataset during the analysis phase.

2.3. Determining the Study Period

No restriction on the starting year was applied in the literature review, and all studies were retrieved starting from the earliest record on the subject found in Web of Science. It was determined that the earliest relevant publications in the dataset date back to 1997.

Since the year 2026 had not yet ended when the search was conducted, the number of publications for 2026 does not represent a full calendar year. To prevent the change in the number of publications for an incomplete year from causing an artificial decrease in keyword frequencies and other temporal indicators, the period 1997–2025 was designated as the primary bibliometric analysis period.

Records for the year 2026 were not deleted during the data cleaning phase but were preserved as a separate dataset. While these records can be used as indicators to show current publication activity, when necessary, they have not been included in annual scientific production trends, thematic evolution, emerging-topic analyses, or other temporal comparisons involving complete years. This was done to ensure comparability over the years.

2.4. Data Preprocessing and Data Cleaning

Text files obtained from Web of Science were converted into a bibliographic data frame in the R environment using the `convert2df()` function of the `bibliometrix` package (Aria & Cuccurullo, 2017). To prevent the analysis results of identical publications, which might appear in different export files, from being artificially inflated, automatic record deletion was not applied during the initial data import; instead, duplicate records were identified through a controlled cleaning process.

In the data cleaning process, the UT (Unique Article Identifier) field, assigned uniquely to each record by Web of Science, was used first. If multiple records had the same UT value, only a single record was retained. To prevent publications lacking UT information from being mistakenly identified as duplicates, these records were tracked using temporary unique sequence numbers.

In the second stage, Digital Object Identifier (DOI) values were checked. DOI values were converted to lowercase, unnecessary leading and trailing spaces were removed, and records with the same normalized DOI value were considered duplicates. Records lacking DOI information were not automatically deleted.

In addition, the publication title was converted to lowercase, punctuation was removed, and the title and publication year were evaluated together to report potential title-year duplicates. However, because identical or highly similar titles may represent different scientific records, automatic record deletion was not performed solely on the basis of title similarity. This was done to reduce the risk of false-positive duplicate deletions during data cleaning.

The publication year (PY) and total citation count (TC) fields were converted to numerical data types; the dataset was also checked to verify the presence of basic bibliographic fields such as author, title, source, document type, author keywords, Keywords Plus, abstract, address, cited references, WoS categories, DOI, and UT. Missing metadata rates were calculated separately for each bibliographic field.

2.5. Standardization of Author Keywords

In the analyses examining the conceptual structure of the research field, the author's keywords (Author Keywords; DE) were used as the basis. Since the same concept may appear under different spellings, resulting in multiple separate nodes in co-occurrence networks, the keywords were standardized prior to analysis.

During this process, keywords were converted to lowercase, hyphen characters were standardized, unnecessary punctuation was removed, and excess spaces were eliminated. Additionally, some terms that clearly represent the same scientific concept were combined under a common form. For example:

- "ZIP," "ZIP model," "zero-inflated Poisson model," and similar usages were grouped under "zero-inflated Poisson."
- "ZINB," "zero-inflated negative binomial model," and similar expressions were grouped under "zero-inflated negative binomial."
- "ZIB" and similar terms have been grouped under "zero-inflated binomial."
- terms such as "overdispersion" and "over-dispersed" have been grouped under "overdispersion,"
- Moreover, terms such as "excess zero" and "excess zeros" have been grouped under "excess zeros."

All original and standardized keywords have also been stored in a conversion table, thereby ensuring the traceability of the standardization processes.

Since there is a high likelihood that general terms such as "zero-inflated models," "zero-inflated regression," "zero-inflated count data," and "zero-inflated distribution", as a direct result of the search query, would excessively dominate the co-occurrence network, these terms were retained in the descriptive keyword frequency analyses. However, they were excluded from the detailed scientific

mapping networks created to identify the subthemes of the research field better. The aim was to reduce the artificial centralization effect that could arise from the structure of the search query itself.

2.6. Bibliometric Performance Analysis

A bibliometric performance analysis was conducted to quantitatively assess the scientific development of the literature on zero-inflated models. The performance analysis is based on an evaluation of the productivity and impact levels of key bibliometric units, such as publications, authors, sources, and countries, that contribute to a specific research field. While science mapping techniques reveal the conceptual, intellectual, and social relationships within a research field, performance analysis enables determination of the field's overall size, rate of development, and key scientific actors. For this reason, performance analysis and science mapping are considered two complementary core approaches in bibliometric research (Aria & Cuccurullo, 2017; Donthu et al., 2021). Furthermore, the systematic evaluation of bibliometric indicators helps to more objectively identify prominent researchers, publication sources, and scientific contributions within a broad and fragmented literature (Zupic & Čater, 2015).

In this context, the study first examined changes in scientific output over time using models adjusted for zero-inflated data. In addition, annual scientific output was assessed using yearly publication counts, and temporal trends were visualized using LOESS smoothing. The total number of documents was calculated for each publication year to determine the field's historical development trend, and annual scientific output was visualized. In bibliometric studies, changes in the annual number of publications are used as a key indicator to assess the development of a research field within the scientific community, researchers' changing interests over time, and the field's maturation process (Donthu et al., 2021). However, since the year 2026 had not yet concluded at the time of data collection for this study, a direct comparison of the number of publications for that year with previous years could be misleading. For this reason, records for 2026 have been retained in the dataset and shown separately in the annual output graph; however, only the period from 1997 to 2025, comprising completed publication years, was considered in assessing long-term trends. To visually illustrate the general publication trend in the field, a LOESS curve was fitted to the annual publication counts for 1997–2025. The LOESS curve was used solely for descriptive visualization of the publication trend and was not evaluated as a predictive model for future outcomes.

In addition to the volume of scientific output, the citation performance of the literature was also examined. For this purpose, the total number of citations, the average number of citations per publication, and the median number of citations were calculated for each publication year. Citation analysis is widely used in bibliometric research to assess the extent to which scientific studies are utilized by subsequent research and the degree to which they have become visible in the bibliographic landscape (Zupic & Čater, 2015; Donthu et al., 2021). However, it was noted that the total number of citations can be significantly influenced by the publication's age. Since older studies have had more time to accumulate citations, a direct comparison of absolute citation counts across different publication years is not sufficient on its own. Therefore, in addition to total citation counts, annual average and median citation values were also reported to enable a more balanced assessment of citation distribution. Furthermore, citation metrics were interpreted not as direct measures of scientific quality, but as indicators of the usage and bibliometric visibility of publications within the relevant literature.

To identify the primary publication channels in the research field, the performance of source journals was also evaluated. For each source, the number of published documents, the total number of citations for these documents, and the average citations per publication were calculated; the top 20 sources with the highest output were examined in detail. While the number of publications indicates a specific journal's quantitative contribution to a model literature inflated by zero-citation papers, the total and average citation values provide complementary information regarding the bibliographic visibility of studies published in these sources within the literature. Since evaluating sources based solely on the number of publications could overlook journals that publish a small number of high-impact studies, productivity and citation indicators were considered together. The combined evaluation of multiple indicators in bibliometric performance analysis helps mitigate the limitations that may arise from classifying key scientific actors in a research field based on a single metric (Donthu et al., 2021).

Similarly, productivity and citation metrics were calculated at the author level to identify the field's leading researchers. For each author, the total number of publications, the total number of citations for those publications, the year of the first publication, and the year of the most recent publication were determined. The scientific activities of the top 20 most productive authors were additionally visualized, taking the time dimension into account; thus, not only the total number of publications but also the periods during which the researchers were active in the model literature, which is inflated by zero-citation papers, were evaluated. While bibliometric analyses enable the identification of key researchers influencing the development of a research field, the inclusion of citation metrics provides complementary information regarding the visibility of this productivity within the literature (Aria & Cuccurullo, 2017; Donthu et al., 2021). However, because citations for a study in multi-authored publications are attributed to all authors, the author-level citations were interpreted as the total bibliographic visibility of the publications in

which the relevant authors appeared, rather than as independent or fractional citation shares.

Country-based performance indicators were also calculated to assess the study's geographic distribution and level of internationalization. Country information was derived from author addresses found in Web of Science records and standardized using Bibliometrix (Aria & Cuccurullo, 2017). In addition to the total number of publications and the total citation counts of relevant publications for each country, the Multiple Country Publication (MCP) indicator was calculated to assess the level of international research collaboration. Publications involving researchers from more than one country were classified as international co-publications, and the MCP ratio for each country was obtained by dividing the number of international co-publications by that country's total number of publications. This made it possible to distinguish between countries with high publication output and those with high international collaboration intensity. Co-authorship and country-based collaboration indicators are key components of bibliometric studies in assessing which researcher communities and geographic centers produce scientific knowledge (Zupic & Čater, 2015; Donthu et al., 2021).

Consequently, within the scope of bibliometric performance analysis, annual scientific output, annual citation dynamics, leading publication sources, the most productive authors, and countries' scientific output and international collaboration indicators were evaluated collectively. These indicators were used to determine the quantitative development of the "zero-inflated models" literature and its key scientific actors; inferences regarding the conceptual and intellectual structure of the research field were based on keyword co-occurrence, thematic mapping, co-citation, and bibliographic matching analyses conducted in subsequent stages. Thus, by evaluating performance indicators in conjunction with science mapping techniques, the study aimed to comprehensively examine not only the production and impact levels of the literature but also its scientific structure and transformations over time (Aria & Cuccurullo, 2017; Donthu et al., 2021).

All analyses were conducted as part of a bibliometric analysis using the "bibliometrix" package (v 5.4.1) in the R software (v. 2026.09.0+174) (Aria and Cuccurullo, 2017).

3. Results and Discussion

A literature search conducted in the Web of Science Core Collection database yielded 1,110 bibliographic records related to models inflated with zeros. After checking for duplicate records using the Web of Science unique record number (UT), DOI, and title-year combinations, no records were removed from the dataset, and all 1,110 unique records were retained. Since the year 2026 had not yet concluded at the time of the search, 43 publications from that year were excluded from the main temporal analyses; thus, the 1,067 publications spanning the 1997–2025 period, which covers completed publication years, formed the core bibliometric analysis dataset.

The dataset covering the 1997–2025 period comprises 1,067 documents published in 577 scientific sources. A total of 3,306 different authors were listed in these publications, and 31,943 cited references were identified. The annual compound growth rate of scientific output was calculated to be 17.1%, while the average age of the documents was 8.39 years. The average number of citations per document across the publications examined was 33.37, and the average annual number of citations per document was 3.097. In terms of author collaboration, there were an average of 3.70 co-authors per document, while 66 studies were published by a single author.

An analysis of document types revealed that the literature consists largely of research articles. During the main analysis period, 980 records were classified directly as "article," while 13 studies fell under the "article; proceedings paper" category, six under "article; book chapter," and two under "article; early access." Additionally, the dataset includes 40 proceedings papers, 16 reviews, 3 letters, 3 meeting abstracts, 2 corrections, and 2 editorial materials. These results indicate that scientific output in the literature on zero-inflated models has primarily taken the form of original research articles.

The most productive sources in terms of the number of publications in the literature on zero-inflated models are presented in Table 2 and Figure 3. With a total of 35 publications, "Statistics in Medicine" ranked first. This journal was followed by "Communications in Statistics–Theory and Methods" with 23 publications, the "Journal of Applied Statistics" with 20, "Statistical Methods in Medical Research" with 19, and the "Korean Journal of Applied Statistics" with 16. "Biometrics" and the "Journal of Statistical Computation and Simulation" were also among the leading sources, each with 15 publications.

When the number of publications and citation metrics were evaluated together, significant differences were identified between publication productivity and citation visibility. The 35 publications in "Statistics in Medicine", the most productive journal, received a total of 703 citations, averaging 20.09 citations per publication. In contrast, "Accident Analysis and Prevention", ranked eighth with only 13 publications, demonstrated the most striking citation performance, with a total of 2,526 citations and an average of 194.31 citations per publication.

Other sources with high average citation counts per publication included "Statistical Modeling" (49.91), "Ecological Modeling" (48.13), "Environmetrics" (39.78), "Computational Statistics & Data Analysis" (35.42), and "Caries Research" (34.50). It was observed that studies on zero-inflated models were published both in pure statistics and biometrics journals and in sources focused on applied fields such as health, traffic safety, ecology, and forestry.

The distribution of scientific publications on zero-inflated models over the years shows that the field produced relatively little in its early years but has grown significantly, especially since the late 2000s (Figure 1). While there was only one publication in 1997, no records were found in 1998, and one publication was published in each of 1999 and 2000. The annual number of publications increased to 13 in 2005, 18 in 2008, and 28 in 2010.

Table 1. Basic bibliometric characteristics of literature on zero-inflated models.

Bibliometric indicator	Values
Analysis period	1997–2025
Number of publications in the main analysis	1067
Number of excluded publications from 2026	43
Total retrieved records	1.110
Number of sources (journals, books, etc.)	577
Total number of authors	3.306
Author appearances	3.948
Number of author keywords	3.208
Number of author keywords plus	2.435
Number of cited references	31.943
Annual scientific production growth rate (%)	17.10
Average document age (years)	8.39
Average citations per document	33.37
Average annual citations per document	3.097
Average co-authors per document	3,70
Number of single-authored documents	66
Number of unique authors of single-authored documents	64

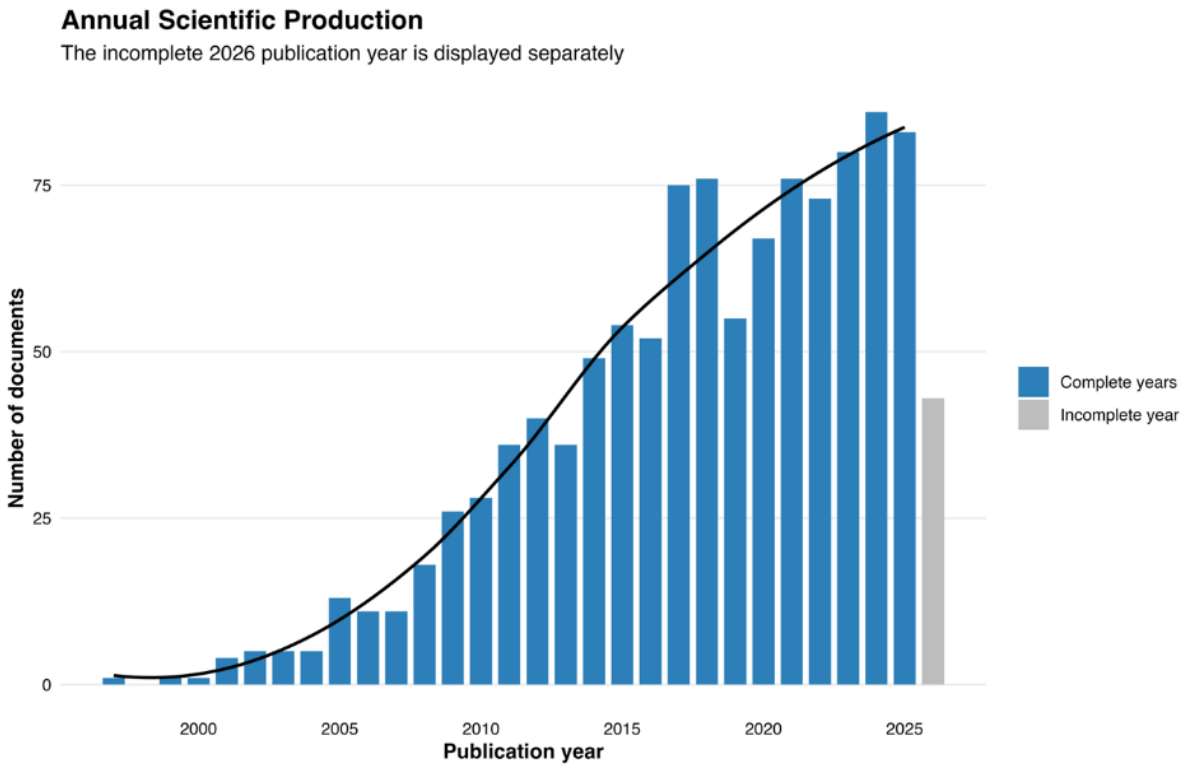


Figure 1. Annual scientific output based on zero-inflated models.

The increase became more pronounced after 2010. The annual number of publications was 36 in 2011, 40 in 2012, 54 in 2015, 75 in 2017, and 76 in 2018. Although a temporary decrease in the number of publications was observed in 2019, production increased again, reaching its highest level in the analysis period with 76 publications in 2021, 80 in 2023, and 86 in 2024. The 83 studies published in 2025 showed that this high publication volume continued in the last completed year as well.

Citation dynamics have shown a more heterogeneous structure compared to the growth in publication output (Figure 2). Publications in the main analysis period received a total of 35,606 citations based on their current citation status. While some older publication years have higher average citation levels, as might be expected, the average values are lower due to the shorter citation accumulation period for those publications.

The year 2017, in particular, has a remarkable citation profile. 75 studies published this year received a total of 13,217 citations, with an average citation count of 176.23. In contrast, the median citation counts of only 11 indicates a strongly right-skewed distribution and that the annual average is significantly influenced by a small number of publications with very high citation counts. Similarly, in 2008, the average citation count was 159.56 while the median was 28.5; and in 2005, the average was 141.15 while the median was 71. Therefore, it seems necessary to consider both the mean and median values when evaluating the citation effect.

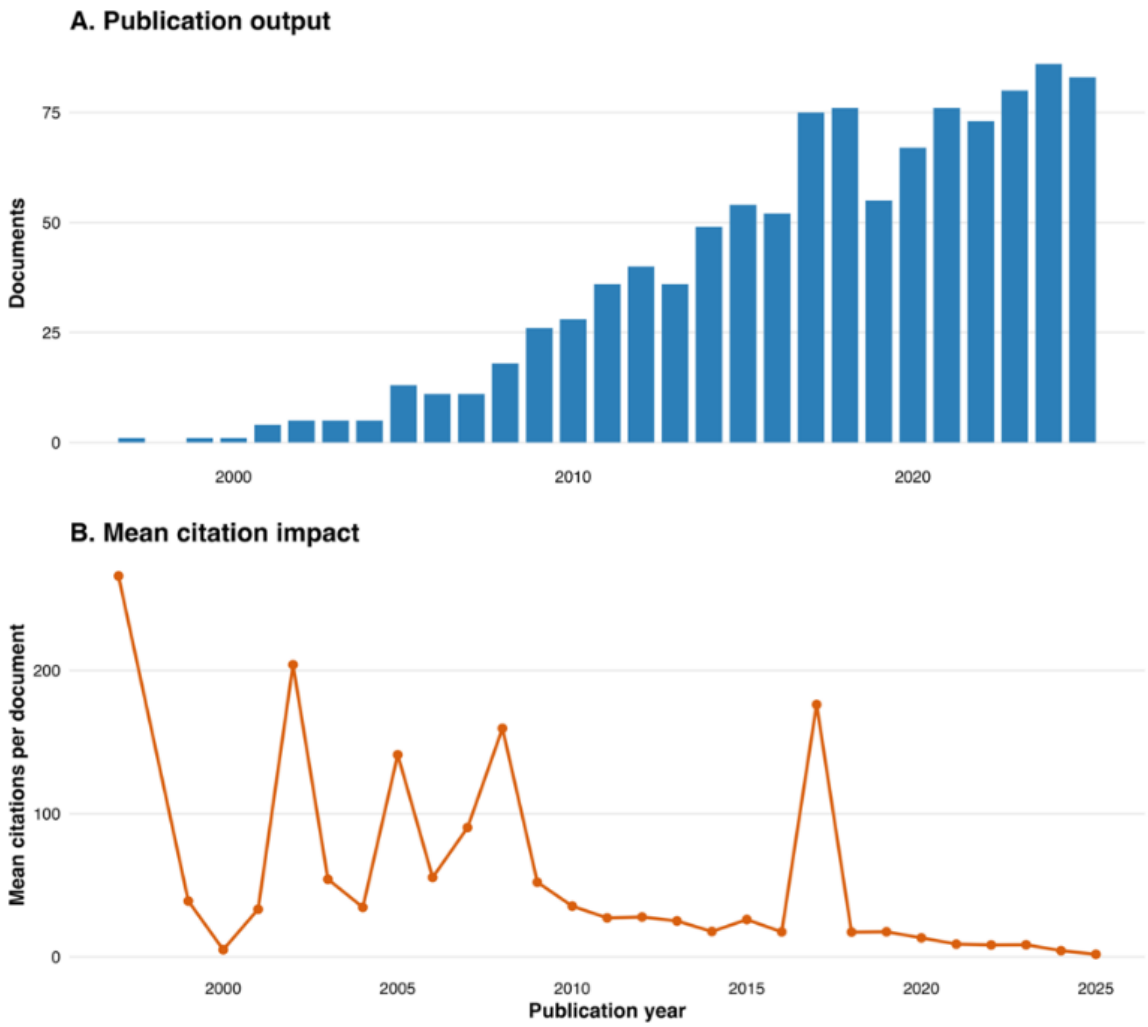


Figure 2. The sources that have produced the most publications in the literature on zero-inflated models

In terms of publication sources, it was determined that the literature on zero-inflated models is not concentrated in a single journal but is distributed across statistics, biometrics, and applied sciences. The highest number of publications was in the journal Statistics in Medicine, with 35 studies. This was followed by Communications in Statistics–Theory and Methods (n=23), Journal of Applied Statistics (n=20), Statistical Methods in Medical Research (n=19), and Korean Journal of Applied Statistics (n=16) (Table 2).

Significant differences were found between publication productivity and citation visibility. For example, Accident Analysis and Prevention, despite ranking eighth with only 13 publications, received 2,526 citations, averaging

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

194.31 citations per publication. In contrast, Statistics in Medicine, which ranked first in terms of publication volume, had an average of 20.09 citations per publication. Statistical Modeling (49.91), Ecological Modeling (48.13), and Environmetrics (39.78) were also among the sources with high average citation counts per publication.

In terms of author-based research, Hsu WW was identified as the most prolific researcher with 10 publications. Ganjali M and Todem D followed with nine each, Yang J with eight, and Aldirawi H and Faes C with seven each. However, a strong correlation between publication count and citation visibility was not found at the author level. For example, Mannering F, despite being represented by five publications, received a total of 1,434 citations for these publications. Similarly, Lee J's five publications received 365 citations, and Gelfand AE's six publications received 258 citations (Table 3).

Table 2. Top 10 most productive scientific sources

Journals	Documents	Total Citations	Mean Citations
Statistics in Medicine	35	703	20.09
Communications in Statistics-Theory and Methods	23	176	7.65
Journal of Applied Statistics	20	229	11.45
Statistical Methods in Medical Research	19	361	19.00
Korean Journal of Applied Statistics	16	21	1.31
Biometrics	15	190	12.67
Journal of Statistical Computation and Simulation	15	417	27.80
Accident Analysis and Prevention	13	2526	194.31
Biometrical Journal	13	366	28.15
Plos One	13	415	31.92

Table 3. The top 10 authors and their bibliometric indicators

Author	Documents	Total Citations	First Year	Last Year
Hsu WW	10	53	2012	2024
Ganjali M	9	46	2012	2025
Todem D	9	50	2012	2023
Yang J	8	57	2018	2023
Aldirawi H	7	55	2018	2023
Faes C	7	83	2014	2018
Chen T	6	41	2015	2022
Gelfand AE	6	258	2002	2025
Hardin JW	6	135	2009	2019
Kim K	6	39	2012	2023

An examination of standardized author keywords revealed that the core concepts in the research area centered on zero inflation, overdispersion, and alternative counting models (Figure 3). The most common keyword was “zero-inflated models,” used in 266 publications. This was followed by “zero inflation” (n=96), “overdispersion” (n=85), “count data” (n=64), “zero-inflated Poisson” (n=58), “hurdle model” (n=52), “negative binomial” (n=45), and “zero-inflated distribution” (n=45). Additionally, the terms zero-inflated negative binomial (n=28), Bayesian inference (n=27), excess zeros (n=25), Poisson regression (n=23), and EM algorithm (n=22) also stood out.

This distribution shows that the field is not solely limited to the general concept of “zero inflation” and also encompasses methodological axes such as overdispersion accompanying zero inflation, alternative distributions, hurdle structures, estimation algorithms, and Bayesian inference.

The co-occurrence network revealed the connections between concepts in greater detail and identified six communities (Figure 4). At the center of the network were, in particular, zero inflation, overdispersion, count data, zero-inflated Poisson, hurdle model, and negative binomial. The fact that these terms have large node sizes and numerous connections indicates that they form the fundamental conceptual backbone of the research field.

While one cluster includes concepts such as zero inflation, overdispersion, count data, excess zeros, and distribution-related topics, a separate cluster encompasses zero-inflated Poisson, zero-inflated negative binomial, variable selection, and generalized estimating equations. Another distinct cluster has formed around Bayesian inference, Gibbs sampling, Markov chain Monte Carlo, hierarchical models, and data augmentation. Thus, it has been observed that Bayesian computational approaches form a relatively independent methodological framework within the literature on zero-inflated models.

A temporal overlay network, based on the average publication years of keywords, revealed that methodological interest in the zero-inflated model literature has changed over time (Figure 5). While the average publication years of some classical terms such as score test, Poisson regression, negative binomial distribution, and random effects are concentrated in earlier periods, terms such as simulation, under dispersion, variable selection, model selection, and logistic regression are associated with more recent publications. The fact that high-frequency core concepts such as "zero inflation" and "count data" have remained central even in recent years indicates that these topics are not merely historical but are still actively used in the current literature.

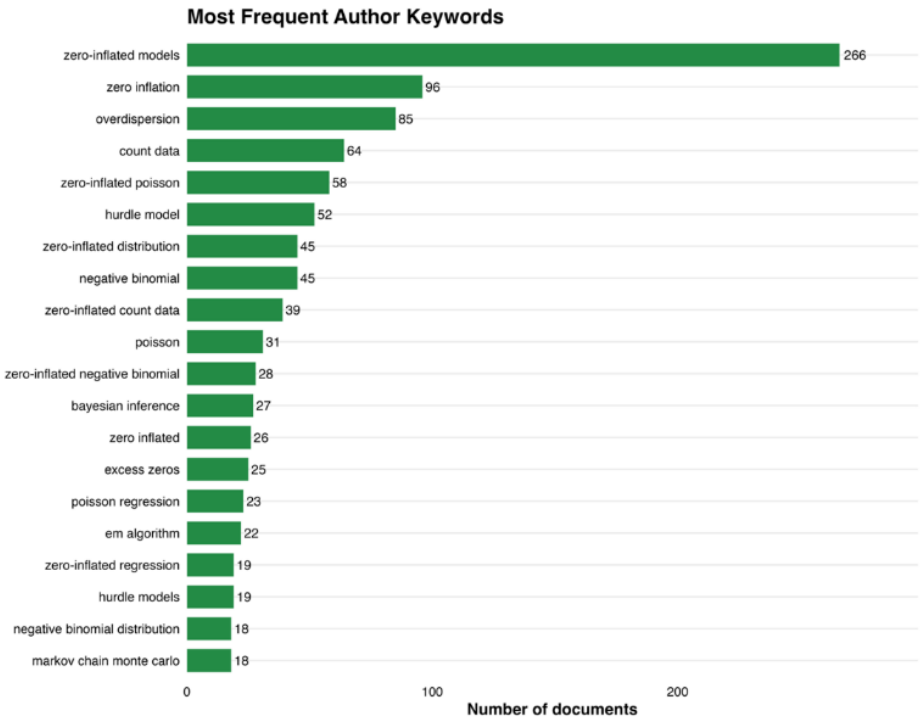


Figure 3. Most frequently used author keywords.

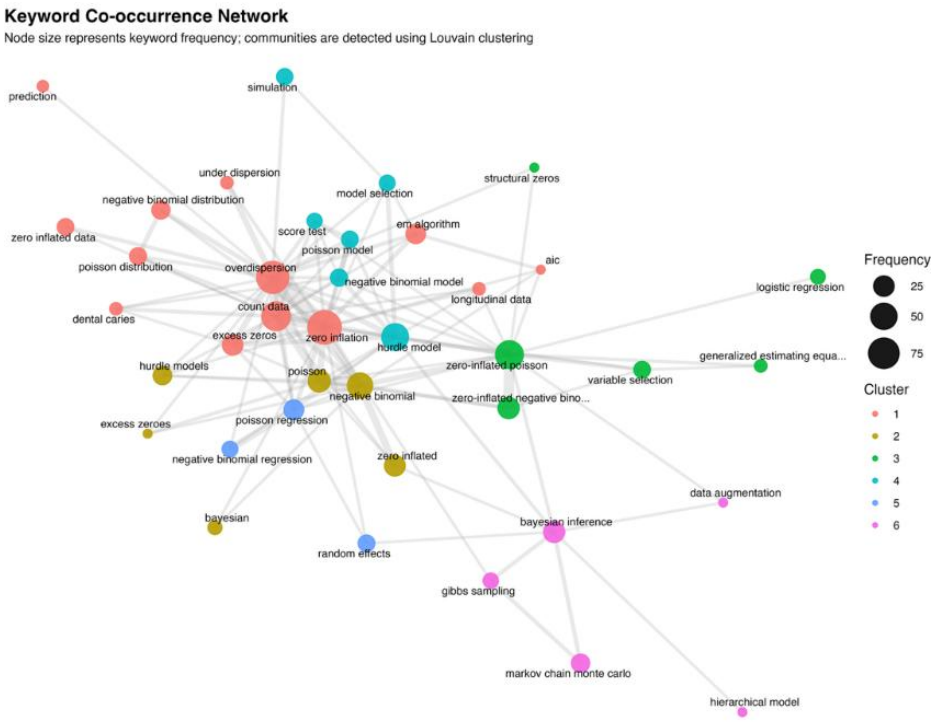


Figure 4. Keyword co-occurrence network.

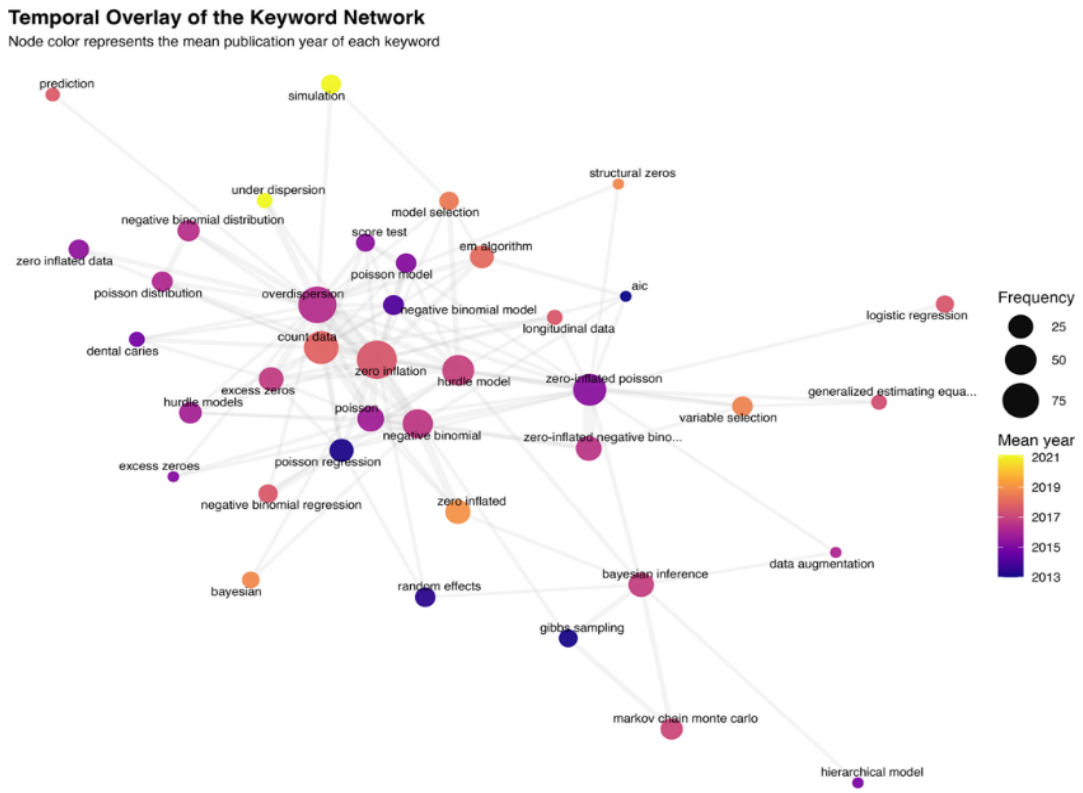


Figure 5. Temporal overlay of the keyword network.

Thematic map analysis revealed that the literature is organized around nine distinct thematic clusters (Table 4; Figure 6). The cluster with the highest total frequency was the zero-inflated/hurdle models cluster, which also exhibited the highest centrality value (0.430) with a frequency of 454. This cluster included concepts such as the EM algorithm, simulation, model selection, auto insurance, the microbiome, longitudinal data, and mixed models. The theme's high centrality indicates that it is strongly connected to other research themes.

The second-largest theme was zero inflation/overdispersion, with a frequency of 343 and a centrality of 0.289. Count data, excess zeros, the negative binomial model, the Poisson model, and the score test constituted the prominent sub-concepts of this cluster. On the thematic map, this cluster was located in the Basic Themes region, demonstrating that zero inflation and overdispersion are fundamental concepts linking the broader field. The "zero-inflated Poisson/zero-inflated count data" cluster, with 202 frequencies, stood out for its high density (16,932) and was located in the "core themes" region of the thematic map. This cluster was observed alongside more specific methodological concepts such as ZINB, variable selection, generalized estimating equations, maximum likelihood, and estimation. The negative binomial/Poisson cluster was another advanced theme characterized by a high combination of centrality and density. The Bayesian inference/Markov chain Monte Carlo cluster formed a distinct thematic structure with 99 occurrences; it was positioned alongside random effects, Gibbs sampling, hierarchical models, Bayesian analysis, and data augmentation.

Table 4. Clusters identified as a result of the thematic analysis

Cluster	Main theme	Frequency	Centrality	Density
1	Zero-inflated model / Hurdle model	454	0.430	13.629
2	Zero inflation / Overdispersion	343	0.289	11.175
3	Zero-inflated Poisson / Zero-inflated count data	202	0.188	16.932
4	Negative binomial / Poisson	147	0.196	14.546
5	Zero-inflated distribution / Negative binomial distribution	104	0.168	13.003
6	Bayesian inference / Markov chain Monte Carlo	99	0.186	13.123
7	Poisson regression / Zero-inflated regression	72	0.159	9.477
8	Zero inflated data / Count regression	27	0.037	13.776
9	Logistic regression / Maximum likelihood estimation	37	0.090	16.051

As a result of the thematic analysis, the clusters "logistic regression/maximum likelihood estimation" and "zero-inflated data/count regression" were located in the niche themes region on the thematic map due to their relatively low centrality but high density. In contrast, the Poisson regression/zero-inflated regression and zero-inflated distribution/negative binomial distribution clusters were positioned in the "emerging/declining" region due to their low centrality and low-to-moderate density. Since it cannot be determined solely from the thematic map whether the themes in this latter region are truly emerging or declining, interpretations should be made in conjunction with temporal keyword and growth analyses.

The literature on reference co-citation networks with zero-inflated models has shown that its methodological and theoretical foundation developed around three main co-citation clusters that are closely interconnected (Figure 7). Among the most prominent and influential nodes in the network are Lambert (1992) and Mullahy (1986), while Hall (2000), Böhning et al. (1999), Ridout et al. (2001), and Min and Agresti (2005) are also among the central references.

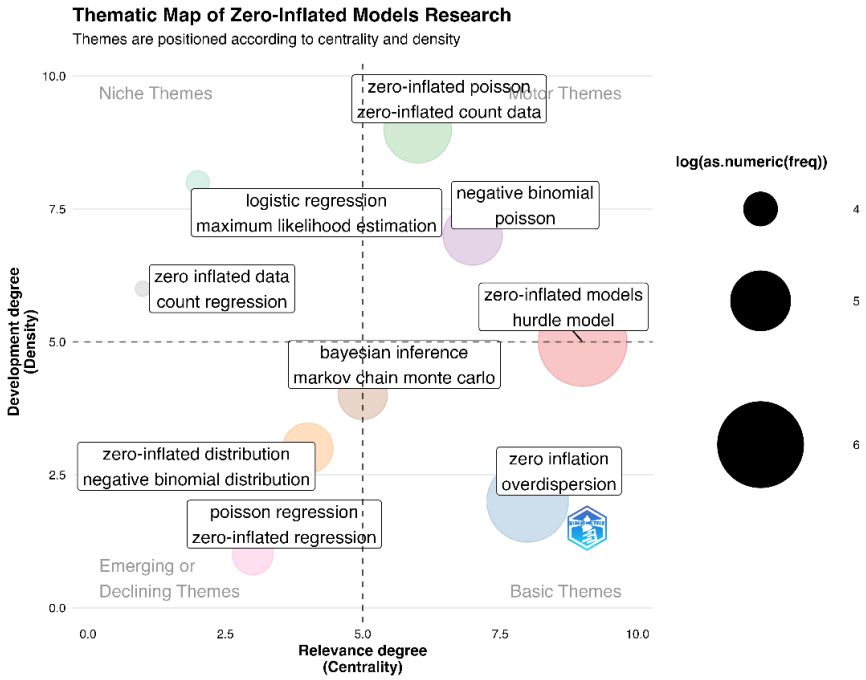


Figure 6. A thematic map of the research field using zero-inflated models.

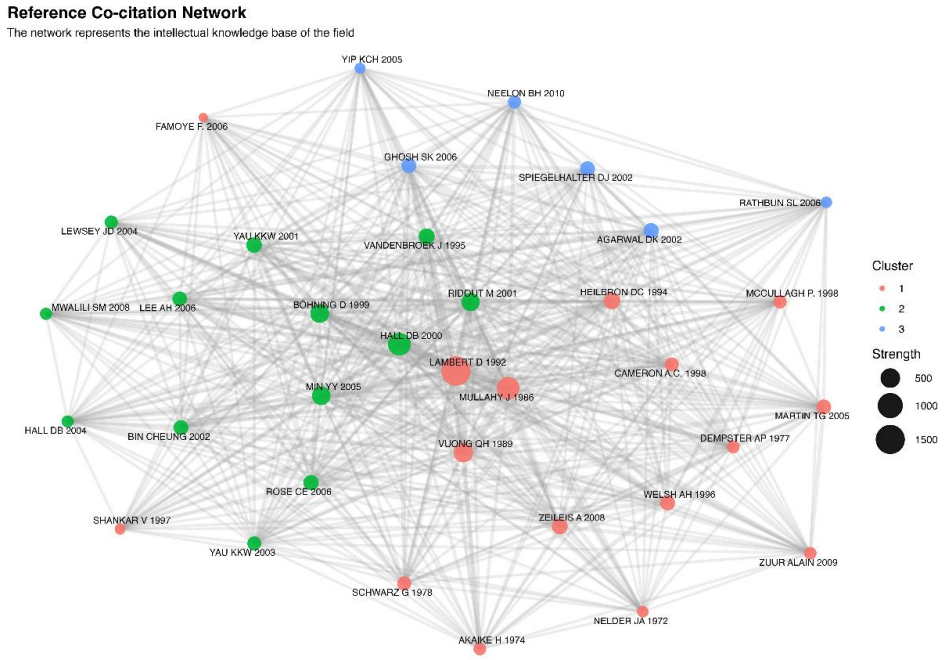


Figure 7. A knowledge based on literature regarding reference-citation networks and zero-inflated models.

The network also shows strong connections to Vuong (1989), Akaike (1974), Zeileis et al. (2008), Cameron (1998), McCullagh (1998), Nelder (1972), and other key sources on count models, model selection, and generalized linear models. This structure demonstrates that the literature on zero-inflated models has developed on an intellectual foundation grounded not only in ZIP/ZINB studies but also in the broader literature on count regression, model comparison, maximum likelihood, and generalized linear models.

4. Conclusion

This study comprehensively evaluated the development of scientific literature on zero-inflated models using data from the Web of Science Core Collection, through bibliometric performance analysis and science mapping techniques. An analysis of 1,067 publications from the main analysis period (1997–2025) revealed that the literature on zero-inflated models, which was initially a limited research area, has grown significantly over time. In particular, the rapid increase in publication output after 2010 and the fact that a significant portion of total publications was produced during the 2021–2025 period demonstrate that the zero-inflated model family continues to be important in current statistical research and applied sciences. The fact that the highest publication output occurred in 2024 and continued at a similar level in 2025 indicates that the field has not yet stabilized and that scientific interest persists.

Citation results show that the increase in publication output does not progress in a perfectly parallel manner with scientific visibility. It was determined that in some publication years, a small number of studies with very high citation counts significantly influenced the annual average citation values. This situation became particularly evident in 2017, and the high average citation value that year stemmed largely from a few highly influential methodological studies. This finding indicates that, when evaluating bibliometric performance, not only average citations count but also median values and the distribution of citations across publications must be taken into account.

The distribution of scientific sources has revealed that "zero-inflated" models exhibit a strong interdisciplinary structure that extends beyond statistical journals. While statistical and biometric journals such as "Statistics in Medicine", "Communications in Statistics–Theory and Methods", "Journal of Applied Statistics", and "Statistical Methods in Medical Research" demonstrate high publication productivity, it was determined that some studies published in applied fields such as traffic safety, ecology, and environmental sciences possess a very high citation impact. Similarly, author analyses have shown that there is no direct equivalence between the number of publications and bibliometric impact. While some researchers achieve high citation visibility with a limited number of publications, others have contributed to the development of literature through sustained publication output spread over longer periods.

Keyword analyses have shown that the conceptual core of the zero-inflated model literature centers on fundamental concepts such as zero inflation, overdispersion, count data, zero-inflated Poisson, hurdle model, negative binomial, and zero-inflated negative binomial. The co-occurrence network has revealed that these fundamental concepts are strongly interconnected and that different methodological frameworks have emerged within the field.

In particular, the clustering of terms such as Bayesian inference, Markov chain Monte Carlo, Gibbs sampling, hierarchical models, and data augmentation within the same conceptual framework indicates that Bayesian computation and hierarchical modeling approaches have evolved into a distinct research strand within the zero-inflated model literature.

Temporal analyses have shown that the research field has not only grown but has also undergone a methodological transformation. Alongside earlier concepts such as the score test, classical Poisson regression, and the negative binomial distribution, it has been observed that topics such as deep learning, marginalized models, count regression, INLA, Bayesian inference, variable selection, and under dispersion have gained prominence in recent years. In particular, the recent emergence of terms like deep learning and marginalized models indicates that zero-inflated data structures are now being addressed using more flexible and computationally intensive methods. However, the relative decline observed in certain keywords should not be interpreted as these topics having completely disappeared from the literature; rather, it should be viewed as an indication that the research agenda is diversifying toward new methodological problems and application areas.

The results of the thematic map demonstrate that the field possesses both a strong core of fundamental concepts and advanced methodological subthemes. The inclusion of the concepts of zero inflation and overdispersion among the core themes highlights that these two problems form the conceptual backbone of the research field; meanwhile, the presence of clusters related to zero-inflated Poisson, zero-inflated count data, and negative binomial models among the driving themes indicates that classical and extended count models continue to play central roles in the field's current development. The fact that Bayesian methods form a distinct thematic structure is another indication of the field's methodological diversity.

The co-citation analysis, meanwhile, has revealed that the literature on zero-inflated models possesses a strong and identifiable intellectual foundation. Pioneering works by researchers such as Lambert, Mullahy, Hall, Böhning, and Ridout have been at the center of the co-citation network; furthermore, it has been observed that classical methodological sources on model selection,

generalized linear models, over-dispersion, and count regression also constitute important components of the literature's knowledge base. Thus, it has been determined that research on zero-inflated models has not developed solely around a specific regression model; rather, it has progressed in an integrated manner with the broader literature on count data modeling, distribution theory, model comparison, and computational statistics.

When the study's overall findings are evaluated collectively, it is concluded that, over the course of approximately thirty years, the literature on zero-inflated models has evolved into a broader, more flexible, and interdisciplinary field of research than the classical ZIP and hurdle approaches. While the field's fundamental problems—excess zeros and excess dispersion—remain significant, current research is shifting toward more advanced methods such as Bayesian modeling, marginalized approaches, variable selection, complex dependency structures, and machine learning. This transformation suggests that future research will focus not only on achieving better model fit but also on more accurately defining the data-generation mechanisms underlying zeros, enhancing model interpretability, and developing new methods suitable for high-dimensional or complex data structures.

Unlike previous bibliometric evaluations in the literature, which were conducted within the broader framework of "inflated count distributions," this study has focused directly on zero-inflated models and assessed the field's scientific output, conceptual structure, thematic development, and intellectual knowledge base using an integrated science-mapping approach. Therefore, it is believed that the results obtained could serve as a useful reference for both statisticians developing new methods for zero-inflated models and researchers applying these methods in fields such as bibliometrics, health sciences, ecology, agriculture, traffic safety, and other applied disciplines, in assessing the current state of the field and its potential future directions.

However, the findings should be evaluated within the framework of certain limitations. The fact that the study is based solely on Web of Science Core Collection records may have resulted in studies indexed in other bibliographic databases but not included in the WoS being excluded from the analysis. Furthermore, bibliometric indicators reflect publication and citation behavior but do not directly measure the methodological quality of the studies. Since keyword-based analyses are sensitive to the terminology used by authors, the standardization of terms that are conceptually similar but expressed in different ways may influence the results to a certain extent. In future research, comparing different bibliographic databases, using more detailed topic modeling techniques, and combining bibliometric findings with systematic content analysis could contribute to a more detailed understanding of the methodological evolution of zero-inflated model literature.

In conclusion, zero-inflated models can be regarded as a mature yet still evolving dynamic field of statistical research. While classical ZIP, ZINB, and hurdle models continue to form the foundation of the field, emerging new methodological trends indicate that the problem of zero-inflated data will continue to be reexamined in the context of different data structures and scientific disciplines.

Author Contributions

The percentage of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	M.A	C.T
C	50	50
D	50	50
S	50	50
DCP	50	50
DAI	50	50
L	50	50
W	50	50
CR	50	50
SR	50	50
PM	50	50
FA	50	50

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

Arab, A., Wildhaber, M. L., Wikle, C. K., & Gentry, C. N. (2008). Zero-inflated modeling of fish catch per unit area resulting from multiple gears: Application to channel catfish and shovelnose sturgeon in the Missouri River. *North American Journal of Fisheries Management*, 28(4), 1044–1058. https://doi.org/10.1577/M06-250.1

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Bekalo, D. B., & Kebede, D. T. (2021). Zero-inflated models for count data: An application to number of antenatal care service visits. *Annals of Data Science*, 8, 683–708. https://doi.org/10.1007/s40745-021-00328-x

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L., & Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,

162(2), 195–209. https://doi.org/10.1111/1467-985X.00130

Boucher, J.-P., Denuit, M., & Guillén, M. (2007). Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, 11(4), 110–131. https://doi.org/10.1080/10920277.2007.10597487

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22, 155–205. https://doi.org/10.1007/BF02019280

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

Chiogna, M., & Gaetan, C. (2007). Semiparametric zero-inflated Poisson models with application to animal abundance studies. *Environmetrics*, 18(3), 303–314. https://doi.org/10.1002/env.830

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402. https://doi.org/10.1002/asi.21525

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, 5(1), 146–166. https://doi.org/10.1016/j.joi.2010.10.002

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, 8, Article 8. https://doi.org/10.1186/s40488-021-00121-4

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4), 1030–1039. https://doi.org/10.1111/j.0006-341X.2000.01030.x

Hilbe, J. M. (2012). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Isgin, T., Bilgic, A., Forster, D. L., & Batte, M. T. (2008). Using count data models to determine the factors affecting farmers' quantity decisions of precision farming technology adoption. *Computers and Electronics in Agriculture*, 62(2), 231–242. https://doi.org/10.1016/j.compag.2008.01.004

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. https://doi.org/10.1002/asi.5090140103

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14. https://doi.org/10.1080/00401706.1992.10485228

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1), 35–46. https://doi.org/10.1016/j.aap.2004.02.004

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1), 1–19. https://doi.org/10.1191/1471082X05st084oa

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Preisser, J. S., Das, K., Long, D. L., & Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine*, 35(10), 1722–1735. https://doi.org/10.1002/sim.6804

Preisser, J. S., Stamm, J. W., Long, D. L., & Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*, 46(4), 413–423. https://doi.org/10.1159/000338992

R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the 19th International Biometric Conference*. Cape Town, South Africa.

Ridout, M., Hinde, J., & Demétrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1), 219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Sidumo, B., Sonono, E., & Takaidza, I. (2024). Count regression and machine learning techniques for zero-inflated overdispersed count data: Application to ecological data. *Annals of Data Science*, 11, 803–817. https://doi.org/10.1007/s40745-023-00464-6

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. https://doi.org/10.1002/asi.4630240406

Tajuddin, R. R. M., & Ismail, N. (2025). Inflation properties in count data distributions: A three-decade bibliometric analysis. *Sains Malaysiana*, 54(1), 303–312. https://doi.org/10.17576/jsm-2025-5401-24

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84, 523–538. https://doi.org/10.1007/s11192-009-0146-3

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127, 51–53. https://doi.org/10.1016/j.econlet.2015.01.002

Young, D. S., Roemmele, E. S., & Shi, X. (2022b). Zero-inflated modeling part II: Zero-inflated models for complex data structures. *WIREs Computational Statistics*, 14(2), Article e1540. https://doi.org/10.1002/wics.1540

Young, D. S., Roemmele, E. S., & Yeh, P. (2022a). Zero-inflated modeling part I: Traditional zero-inflated count regression models, their applications, and computational tools. *WIREs Computational Statistics*, 14(1), Article e1541. https://doi.org/10.1002/wics.1541

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8), 1–25. https://doi.org/10.18637/jss.v027.i08

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, 18(3), 429–472. https://doi.org/10.1177/1094428114562629