



ARTICLE

Regression model applied to rhizosphere data: A bibliometric review

 Aline Martineli Batista¹ and  Fábio Pratavia^{*,2}

¹Facultad de Ciencias Agrarias, Universidad Nacional del Litoral, and ICiAgro Litoral (UNL-CONICET), Esperanza, Prov. de Santa Fe, Argentina

²Luiz de Queiroz College of Agriculture, Department of Exact Sciences, University of São Paulo, Piracicaba, São Paulo, Brazil

*Corresponding author. Email: fabioprataviera828@gmail.com

(Received: September 29, 2023; Revised: December 21, 2023; Accepted: January 24, 2024; Published: August 30, 2024)

Abstract

The interaction of soil with plant roots in the rhizosphere plays an important role in various ecosystem services and food production, and it has been the focus of numerous studies. In turn, statistical modeling can aid in a more comprehensive understanding of this interaction, such as the application of regression models to rhizosphere data. Thus, the main objective of this work was to develop the first bibliometric analysis on regression models applied to rhizosphere data. Bibliometric data were obtained from Web of Science and Scopus databases. We use the topic retrieval as (“Rhizosphere”) AND (“Regression models” OR “Regression model” OR “Generalized Linear Models” OR “Generalized Linear Model”) to search for scientific articles that contain these keywords in their title, abstract, or keywords. The search encompassed articles published from 1900 to 2022, resulting in 34 articles, with the earliest record dating back to 1985. While studies of the rhizosphere are increasing, few studies apply regression models to their data. The use of more advanced techniques, such as Generalized Linear Models (MLG), Artificial Neural Network (ANN), Random Forest Model (RFM), Support Vector Machines (SVM), and Generalized Boosted Regression Modeling (GBM), became evident from 2016 onwards, which was associated with the computational advances and the development of artificial intelligence. Some articles demonstrated that the use of more robust models can provide more meaningful results to the researcher. Only one article was published in a journal dedicated to statistics, highlighting the diffusion of regression models into various fields. Collaborations involving co-authorship between researchers from different countries have led to higher citation rates, increasing the importance of the research to the scientific community. Perhaps one of the most notable limitations to increasing research using regression models is the absence of a statistician or researcher within the research groups who is well versed in statistical models and procedures.

Keywords: Bibliometrix; Bibliometric analysis; R software; Statistical modeling; Soil-root interaction.

1. Introduction

Soil, and in particular its interaction with plant roots, which has been the focus of many studies, plays a fundamental role in diverse ecosystem services and food production (van Veelen *et al.*, 2018). The region around the roots, called the rhizosphere, is the most active portion of the soil, where biogeochemical processes influence a range of landscape and global scale processes (McNear Jr., 2013). The rhizosphere was first defined as the soil volume adjacent to the plant roots (Hiltner, 1904). Currently, it is known that the rhizosphere is a dynamic system in which several processes occur that change the properties of the soil around the roots. Therefore, the changes promoted by plants in the rhizosphere are due to the root-soil interaction and occur mainly due to the organic compounds released by the root (Echer *et al.*, 2020).

Addressing the global challenges of climate change and population growth with a better understanding and control of rhizosphere processes is one of the most important scientific challenges today (McNear Jr., 2013). Rhizosphere management is useful for numerous processes that promote plant growth and health, such as improving nutrient efficiency and water uptake, mitigation drought stress, and suppressing disease (Ahmed *et al.*, 2014; Ayangbenro *et al.*, 2022; Fasusi *et al.*, 2021; Raaijmakers & Mazzola, 2016; Zia *et al.*, 2021). Thus, understanding interactions and processes that occur in the rhizosphere is fundamental to increasing our capacity address with relevant global challenges related to food production.

In this sense, statistical models can help to better understand these interactions and processes, such as regression models applied to rhizosphere data. Regression models can be used to investigate the relationship between a dependent variable and one or more independent (or explanatory) variables. It is often used to predict values of the dependent variable based on known or predicted values of the independent variables. The choice of model to be used depends on the nature of the data and the purpose of the analysis.

In general, analysis of variance (ANOVA) and simple or multiple linear regression are used, assuming normality and homogeneity of variance (based on adjusted model residuals), assumptions that are sometimes biased. In some situations where the assumptions are not met, transformations are used, such as the Box-Cox transformation. However, a simple interpretation transformation is not always obtained.

Identifying the statistical models applied to rhizosphere data is fundamental to support the development and improvement of current methodologies for modeling rhizosphere data. In this context, bibliometric analyses play an important role in generating knowledge for future practical applications.

Bibliometric reviews allow quantitative analysis of numerous criteria of a research topic or field (Mokhnacheva & Tsvetkova, 2020). This type of study is a good methodological tool for providing answers and evaluating or validating a given theory (Snyder, 2019). It consists of examining the production of articles on a given subject, mapping research institutions and identifying networks of researchers and their motivations (Chueke & Amatucci, 2015). This technique systematizes the bibliometric data to present the intellectual state of the research already carried out and the emerging trends of a subject, identifying problems to be investigated in future research (Chueke & Amatucci, 2015; Donthu *et al.*, 2021).

Considering the current relevance of rhizosphere studies and the importance of statistical modeling to identify relationships between variables, the main objective of this work was developed the first bibliometric analysis on statistical modeling used for rhizosphere data. Thus, a bibliometric analysis was carried out to investigate different types of regression models applied to the rhizosphere considering the period of 1900 - 2022. The results of this work will be useful for future research, pointing the out most used and new possibilities of statistical approaches for this study area, helping the development of important future studies.

The article is organized as follows: Section 2 presents material and methods. The results and

discussion on the main findings are presented in Section 3. Finally, Section 4 addresses some concluding remarks.

2. Material and methods

The methodological procedure adopted in this article was a systematic literature review, which analyzed international literature research related to studies of regression models applied to data from the rhizosphere, whose objective was to study or compare the modeling of the variables involved for this type of data.

Data were obtained from the database of Web of Science (WoS) and Scopus, as both provide lists of high-quality peer-reviewed articles. We use the topic retrieval as *(("Rhizosphere") AND ("Regression models" OR "Regression model" OR "Generalized Linear Models" OR "Generalized Linear Model"))* to search scientific articles with these words in their title, abstract or keywords, published since 1900 to 2022. We also applied a data cleaning filter that limited the sample to articles published in English. This search resulted in 81 articles, with the first register published in 1985.

These 81 articles were reviewed by the authors and some exclusion criteria were applied: (i) articles that did not focus on the topic under discussion (42 articles); (ii) exclusion of all languages other than English (3 articles); and (iii) duplicate articles (2 articles). Thus, our final sample included 34 articles. Table 1 provides a description of the final sample selected.

Table 1. Descriptive analysis: Main information regarding the collection

Description	
Articles	34
Period	1985-2022
Annual percentage growth rate	8.77
Average citations per article	13.59
Authors	155
Author appearances	170
Authors of single authored articles	0
Authors of multi authored articles	160
Articles per author	0.219
Authors per article	4.56
Co-authors per articles	5.00
Collaboration index	4.56

Data from the 34 articles were further processed using of the Bibliometrix package of the R software, which includes the graphical interface Biblioshiny (Aria & Cuccurullo, 2017). The following analyses were performed: (i) bibliometric analysis of citations, (ii) bibliometric analysis of co-citations by journals, (iii) analysis of the most cited articles, (iv) analysis of authors and countries of the articles, and (v) analysis of publications over time. In addition, the 34 articles were completely revised to analyze the types of regression models used in them.

3. Results and discussion

This section presents the main results of our analysis on regression models applied to rhizosphere data in the considered period.

3.1 Temporal distribution of publications

The evolution of the distribution of the number of publications over the period considered in the search was evaluated (Figure 1). A gradual increase in the number of articles was observed over the years, starting around the year 2000. The temporal distribution of publications reflects scientific trends and advances. As new techniques and technologies are developed, it is common to observe an increase in the number of publications related to specific areas. In particular, the use of more advanced techniques such as generalized linear models (MLG), artificial neural networks (ANN), random forest models (RFM), support vector machines (SVM), and generalized boosted regression modeling (GBM) became evident from 2016.

The emergence and popularization of high-performance computing and the advent of artificial intelligence (AI) have spurred research on more complex mathematical and statistical modeling (Basili *et al.*, 2008; Chi *et al.*, 2016; Jordan & Mitchell, 2015; Sarker, 2021). Nowadays, with the advancement and popularization of AI and its application to science, further use of advanced statistical methods is expected, as has been noted in some fields (Davenport & Kalakota, 2019; Raza *et al.*, 2022).

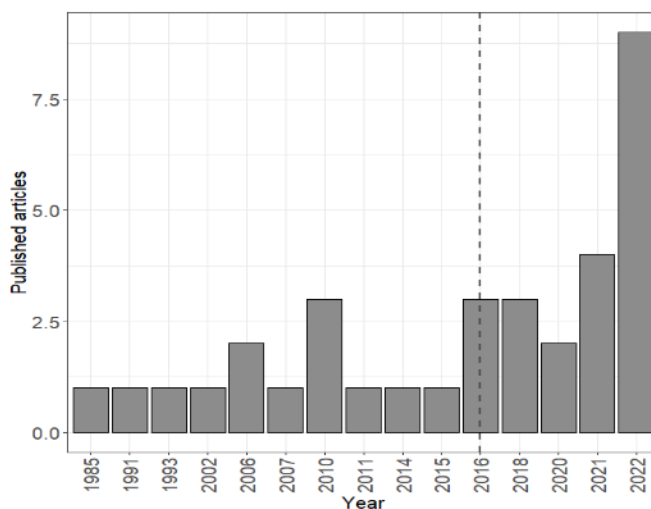


Figure 1. Publications per year 1985-2022. The line in the year 2016 indicates the moment when the use of more advanced techniques such as MLG, ANN, RFM, SVM, and GBM becomes evident.

The computational advancement plays a fundamental role in this context, because with the increase in processing power and the development of more sophisticated algorithms, researchers have been able to explore more complex mathematical and statistical models to analyze and interpret data in a more precise and comprehensive manner. This computational evolution has enabled the use of more advanced approaches, such as the utilization of machine learning algorithms and ANN, which allow for the processing of large amounts of data and the discovery of subtle and non-linear patterns. In summary, the temporal distribution of scientific publications reflects not only the growth of academic production but also scientific trends and advancements.

3.2 Types of regression models used

Univariate regression models are statistical techniques that can be used to investigate the relationship between a dependent variable, (Y), and one or more independent (or explanatory) variables,

(X). In general, the functional form of the regression is given by:

$$Y = f(\mathbf{X}). \quad (1)$$

The choice of the model to be used depends on the nature of the data and the objective of the analysis. In this regard, we conducted a detailed study of the types of regression models that were employed in the selected articles.

It is worth noting that a single article may employ more than one data analysis technique, such as multivariate analysis (principal components, cluster analysis, among others), ANOVA, multiple comparison tests, correlation analysis, etc. However, our focus was solely on the types of regression models.

Table 2 shows the authors, the abbreviation for the type of regression model used in the study, an indicator for the presence of data transformations, the journals in which the articles were published, the Journal's Impact Factor (IF) – based on 2022 Journal Impact Factor, *Journal Citation Reports* (Clarivate, 2023) –, and the field and quartile of the International Classification in which the publication is located. A brief description of the types of regression models used in the articles is given below.

Simple Linear Regression (SLR): is used when there is only one independent variable and the relationship with the dependent variable is linear.

Multiple Linear Regression (MLR): is used when there are several independent variables and the relationship with the dependent variable is linear.

Stepwise Multiple Linear Regression (SMLR): is a statistical method used to select the most relevant variables in a predictive model. It involves building a model by adding or removing variables one at a time, based on a set of predetermined criteria.

Polynomial Regression Model (PRM): is a type of regression analysis used to model the relationship between a dependent variable and one or more independent variables. It is an extension of simple linear regression, in which the relationship between the variables is modeled as an n th degree polynomial function.

Ridge Regression Model (RRM): is a type of linear regression used to handle multicollinearity, which occurs when independent variables in a regression model are highly correlated with each other.

Generalized Linear Models (GLMs): are a class of statistical models that extend the linear regression model to handle non-normally distributed response variables. They are used to model relationships between a dependent variable (response) and one or more independent variables (predictors) by specifying a probability distribution for the response variable.

Logistic Regression Model (LRM): is used when the dependent variable is categorical (for example, binary or nominal) and the independent variables are continuous or categorical.

Location-Scale Regressions (LSRs): are models in which the regression structure can be considered in both position and scale parameter, such as the gamma distribution in which the regressors can be considered in the parameter that represents the mean and also in the parameter that represents the dispersion.

Generalized Boosted Regressions Modeling (GBMs): are a popular machine learning technique used for regression problems. GBMs are an extension of the boosting algorithm, which iteratively

combines weak learners to create a strong learner.

Bayesian Dirichlet–Multinomial Regression Model (DMBVs): is a probabilistic model used to analyze count data in the context of multiple predictor variables. It is an extension of the standard multinomial logistic regression model, which assumes that the response variable follows a multinomial distribution.

Artificial Neural Network (ANN): is a machine learning model inspired by the structure and function of the human brain. It consists of a large number of interconnected processing nodes or neurons that work together to solve complex problems. Each neuron is connected to other neurons through pathways known as synapses, and these connections allow information to be transmitted throughout the network.

Support Vector Machines (SVMs): are types of machine learning algorithms that can be used for classification, regression, and outlier detection.

Random Forest Model (RFM): is a powerful machine learning algorithm used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to make predictions.

Based on the results presented in Table 2, we find that 34.3% of the articles employed some form of transformation, with the logarithmic transformation being the most common. Transformations were primarily applied in cases where the models used were SLR, MLR, and SMLR, which are part of the classical regression models. While it is not necessary to assume a probability distribution to obtain estimates in these models, it is observed that in many studies, data transformations were employed to meet the assumptions of the residuals. Furthermore, these models assume that the relationship between the response variable and the explanatory variable is linear, which is not always the case, as demonstrated in a previous study (Batista *et al.*, 2022).

We used the terms ("*Generalized Linear Models*" OR "*Generalized Linear Model*") in our search because we were interested in determining the extent to which research has used more robust techniques compared to classical regression models. From this perspective, it is evident that only one article used exclusively GLMs (Moreira *et al.*, 2020), while two others used GLMs in combination with other techniques, such as SVMs (Srivastava *et al.*, 2016), as well as SLR and RFM (Table 2).

The application of predictive machine learning models, such as RFM, ANN, and SVMs, to rhizosphere data has increased (Table 2). The use of machine learning models has gained popularity in recent years (Sarker, 2021). This can be explained by computational advances, the acquisition of large databases, and improvements in algorithms (Schmidt *et al.*, 2019).

3.3 Cited sources

The main sources cited by the fully reviewed articles were organized into three main clusters, where they are basically related to microbiology, soil science, and global challenges topics, as shown in Figure 2. These journals are consistent with the fact that the majority of articles resulting from our search were not published in journals with a focus on publishing statistical models or statistical advances (Table 2). Only one of the articles was published in a journal focused on statistical models (Pratavia *et al.*, 2022), as shown in Table 2. The application of these regression models to different topics is very interesting for their diffusion. For example, the first article published shared evidence that the ridge regression model was the best option for their data analysis (Leath & Carroll, 1985). It was published in the Canadian Journal of Plant Pathology because of its application to pathological data, but it basically consisted of comparing different statistical models to predict yield reduction by

Table 2. Descriptive analysis of the 35 selected articles, including authors (in citation format), models used, whether there was any transformation applied to the study variables, publication journal, impact factor (IF), and quantile-based area

Authors	Models used	Transformation	Sources	IF	SJR Subject Area (JCR Quartile)
DeWolf <i>et al.</i> , 2022	DMBVS	No	Mol. Ecol.	6.622	Ecol., Evol., Behav. and System. (Q1)
Schmidt <i>et al.</i> , 2022	SMLR	Yes	Agronomy-Basel	3.949	Agronomy and Crop Science (Q1)
Wu <i>et al.</i> , 2022	SLR	No	J. Trop. Ecol.	1.800	Ecol., Evol., Behav. and System. (Q2)
Prataviera <i>et al.</i> , 2022	LSRs	No	J. Appl. Stat.	1.416	Statistics and Probability (Q2)
Rezakhani <i>et al.</i> , 2022	SLR	No	Plant Soil	4.993	Plant and Science (Q1)
Li <i>et al.</i> , 2022b	MLR	No	Eur. J. Soil Sci.	4.178	Soil Science (Q1)
Beschoren da Costa <i>et al.</i> , 2022	GLMs, SLR and RFM	Yes	MBIO	7.786	Microbiology (Q1)
Ma <i>et al.</i> , 2022	MLR and ANN	Yes	Chemosphere	8.943	Chemistry (Q1)
Li <i>et al.</i> , 2022a	SMLR and ANN	Yes	Environ. Pollut.	9.988	Health, Toxic. and Mutagenesis (Q1)
Xia <i>et al.</i> , 2021	SMLR	No	Biol. Fertil. Soils	6.605	Agronomy and Crop Science (Q1)
Tang <i>et al.</i> , 2021	PRM and LRM	No	Plant Soil Environ.	2.328	Soil Science (Q2)
Wang <i>et al.</i> , 2021	MLR	No	Sci. Total Environ.	10.754	Environmental Chemistry (Q1)
Shimamura <i>et al.</i> , 2021	MLR	No	Soil Sci. Plant Nutr.	1.929	Plant Science (Q2)
Yao <i>et al.</i> , 2020	SLR and GBMs	Yes	Front. Microbiol.	6.064	Microbiology (Q1)
Moreira <i>et al.</i> , 2020	GLMs	No	J. Environ. Manage.	8.910	Environmental Engineering (Q1)
Monreal <i>et al.</i> , 2018	SLR and PRM	Yes	Rhizosphere	3.437	Agronomy and Crop Science (Q1)
Tian <i>et al.</i> , 2018	SLR	No	Environ. Sci. Pollut. Res.	5.190	Health, Toxic. and Mutagenesis (Q1)
Pan <i>et al.</i> , 2018	SLR and MLR	No	Forests	3.282	Forestry (Q1)
Chen <i>et al.</i> , 2016	SLR and SMLR	No	J. Plant Nutr. Soil Sci.	2.566	Plant Science (Q2)
Srivastava <i>et al.</i> , 2016	GLMs and SVMs	No	Environ. Earth Sci.	3.119	Earth-Surface Processes (Q2)
Fox <i>et al.</i> , 2016	SMLR	Yes	Pedobiologia	2.128	Ecol., Evol., Behav. and System. (Q2)
Mohanty <i>et al.</i> , 2015	SLR	Yes	Environ. Earth Sci.	3.119	Earth-Surface Processes (Q2)
Cloutier-Hurteau <i>et al.</i> , 2014	SLR and MLR	Yes	Environ. Sci. Pollut. Res.	5.190	Health, Toxic. and Mutagenesis (Q1)
Cloutier-Hurteau <i>et al.</i> , 2011	SLR and MLR	No	Can. J. For. Res.	2.331	Forestry (Q1)
Guo <i>et al.</i> , 2010	MLR	No	J. Environ. Radioact.	2.655	Environmental Chemistry (Q2)
Ma <i>et al.</i> , 2010b	MLR	No	Environ. Pollut.	9.988	Health, Toxic. and Mutagenesis (Q1)
Ma <i>et al.</i> , 2010a	MLR	No	J. Soils Sed.	3.536	Earth-Surface Processes (Q1)
Ibekwe <i>et al.</i> , 2007	SLR	Yes	Can. J. Microbiol.	3.226	Appl. Microbiol. and Biotechnol. (Q2)
Wang <i>et al.</i> , 2006	SLR and PRM	Yes	Soil Biol. Biochem.	8.546	Microbiology (Q1)
Guzmán-Plazola <i>et al.</i> , 2006	LRM	No	Nematropica	0.347	Agronomy and Crop Science (Q4)
Edge & Wyndham, 2002	SMLR	Yes	Can. J. Microbiol.	3.226	Appl. Microbiol. and Biotechnol. (Q2)
Duncan <i>et al.</i> , 1993	MLR	No	Phytopathology	4.010	Agronomy and Crop Science (Q1)
Manjunath & Habte, 1991	SMLR	No	Can. J. Bot.-Rev. Can. Bot.	1.397	NA
Leath & Carroll, 1985	RRM	No	Can. J. Plant Pathol.	2.074	Agronomy and Crop Science (Q2)

Fusarium sp. in soybean, including infection by *Fusarium* spp. from the rhizosphere.

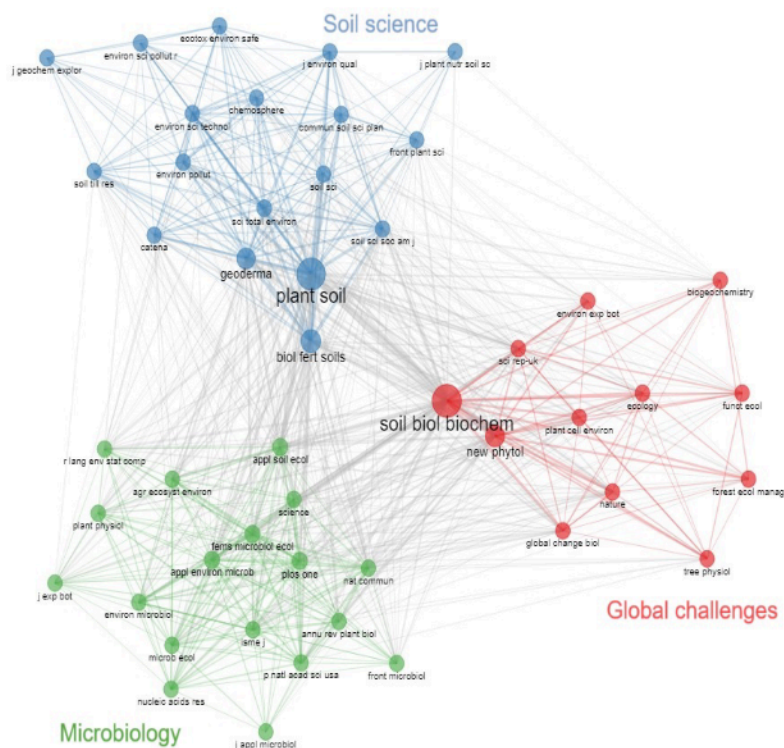


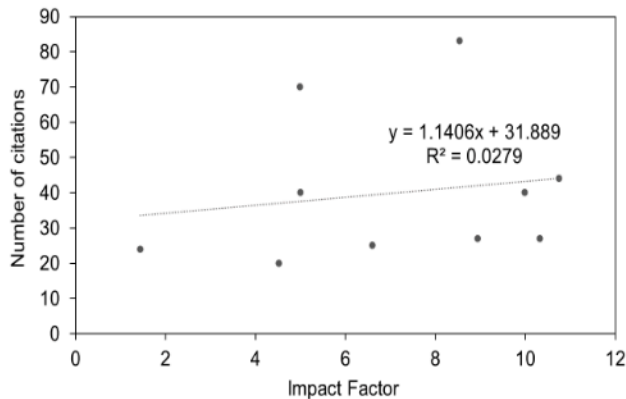
Figure 2. Co-citation network. Size of circles indicates the number of citations, links between the circles represent citations of different academic journals in the same publication, and a shorter distance between two circles indicates a higher number of co-occurrences between two citations. Clusters were labeled using different colors.

Table 3 shows the top 10 academic journals (sources) cited by fully revised articles. These journals aim to publish studies related to agriculture and/or the environmental sciences, i.e., relevant studies aimed at understanding the soil–plant relationship (rhizosphere). None of the journals have their scope focused on the publication of statistical models or statistical advances. However, there are records in the literature of articles published in journals with this scope that used regression analysis applied to soil science (Ibrahim *et al.*, 2023; Prativiera *et al.*, 2022; Prativiera *et al.*, 2021). Therefore, as the number of studies on the rhizosphere using regression models increases, it is expected that studies will be published in such journals.

Soil Biology and Biochemistry stands out with 81 citations, followed by Plant and Soil with 66 citations, while several other journals received less than 50 citations (Table 3). The number of times that a source was cited was not related to its IF (Figure 3). For instance, although New Phytologist has the second highest IF, it was cited only 27 times (Table 3).

Table 3. Descriptive analysis: Top 10-Most cited journal

2[0]*Sources	N° of Citation	2[0]*IF	SJR Subject Area (JCR Quartile)
Soil Biol. Biochem.	81	8.546	Microbiology (Q1)
Plant Soil	66	4.993	Plant Science (Q1)
Sci. Total Environ.	44	10.754	Environmental Chemistry (Q1)
Environ. Pollut.	40	9.988	Health, Toxic. and Mutagenesis (Q1)
Appl. Environ. Microb.	35	5.005	Appl. Microb. and Biotech. (Q1)
Chemosphere	27	8.943	Chemistry (Q1)
New Phytologist	27	10.323	Physiology (Q1)
Biol. Fert. Soils	24	6.605	Agronomy and Crop Science (Q1)
Environ Sci. Technol.	24	1.440	Chemistry (Q1)
FEMS Microbiol. Ecol.	20	4.519	Appl. Microbiol. and Biotechn. (Q1)

**Figure 3.** Number of citations in function of the Impact Factor (IF) of the academic journals on the considering the top-10 most cited journals.

3.4 Most cited articles

Table 4 shows the 10–most cited articles in absolute and relative terms (citations per year). Wang *et al.*, 2006 was the most cited article in absolute terms. They characterized the relationship between pH and biological activities of rhizosphere and non-rhizosphere soils under reduced pH during *Thlaspi caerulescens* phytoextraction using SLR and PRM. On the other hand, Moreira *et al.*, 2020 was the most recent paper among them and also the most cited in relative terms. They used GLMs to predict the effect of increasing salinity on bioinocula and rhizosphere bacterial communities.

The only two articles published in the 90's (Figure 1) are among the most cited. A SMLR was used to determine the extent to which rhizosphere acid production and root morphological characteristics could explain differences in mycorrhizal dependency of host plants (Manjunath & Habte, 1991). Two years later, Duncan *et al.*, 1993 used MLR and a stepwise procedure to regress the amount of *Tylenchulus semipenetrans* (a nematode) or *Phytophthora parasitica* (a oomycete) in the soil against a parameter of root quality and environmental variables in the citrus rhizosphere to detect seasonal patterns.

To investigate the survival of *Escherichia coli* O157:H7 in the non-rhizosphere and rhizosphere of lettuce after soil fumigation, Ibekwe *et al.*, 2007 plotted the population size of *E. coli* against time after inoculation using SLR. Ma *et al.*, 2010b investigated the role of molecular structure in determining the effect of the rhizosphere on polycyclic aromatic hydrocarbon (PAH) dissipation, i.e.,

Table 4. Descriptive analysis: Top 10-Most cited articles

Papers	Total Citations (TC)	TC per Year
Wang A.S., 2006, Soil Biol. Biochem.	74	4.11
Manjunath A., 1991, Can. J Bot.	52	1.57
Moreira H., 2020, J. Environ. Manage.	49	12.25
Cloutier-Hurteau B., 2014, Environ. Sci. Pollut. R.	24	2.40
Duncan L.W., 1993, Phytopathology	24	0.77
Fox A., 2016, Pedobiologia	20	2.50
Ma B., 2010, Environ. Pollut.	20	1.43
Chen L.X., 2016, J. Plant Nutr. Soil Sc.	19	2.37
Ibekwe A.M., 2007, Can. J. Microbiol.	19	1.12
Tian K, 2018, Environ. Sci. Pollut. R.	18	3.00

bioremediation effect. They used MLR in combination with an R package for genetic algorithms (genalg) to build datasets for *Poaceae* and *Fabaceae* plants. Another article related to PAH contamination is also one of the most cited (Tian *et al.*, 2018). The authors used SLR to investigate potential sources and levels of PAHs in rhizosphere soils of wheat fields affected by coal combustion.

Cloutier-Hurteau *et al.*, 2014 compared the predictions of Al, Cu, and Zn concentrations in tree fine roots using non-rhizosphere and rhizosphere soil properties to find the best approach for assessing the ecological risks of metals to trees. They used a MLR model with soil Al, Cu, and Zn speciation data and chemical and microbial properties as explanatory variables. Predictions of root metal concentrations were evaluated using both SLR and MLR. Fox *et al.*, 2016 used a SMLR to predict the effect of environmental variables on plant growth promotion in a soil with biochar. Rhizosphere bacteria were among the predictors used in the regression analyses. Chen *et al.*, 2016 used both SLR and SMLR to evaluate temporal variation of phosphorus fractions and phosphatase activities in rhizosphere and non-rhizosphere soils during the tree planting. SLR was used to identify the relationships between P fractions and phosphatase activities and stand age and SMLR were applied to correlate different P fractions and phosphatase activities with the tree growth.

None of these papers were published in a journal that focuses on statistical modeling, and presumably they were cited because of the area of research, not because of the statistical technique performed. However, it is interesting to observe the use of regression models applied to a wide range of topics.

3.5 Authors and countries

China was the country with the highest number of published articles (Table 5) and authors (Figure 4), followed by the USA (Table 5, Figure 4). Relevant information from our bibliometric data were observed: (i) the increase in the number of articles by country was positively reflected in the number of citations by country (Figure 5a), (ii) the co-authoring between authors from different countries reflected in a higher number of articles by country (Figure 5b), and (iii) the co-authoring between authors from different countries also reflected in a higher number of citations by country (Figure 5c). These observations are important because a higher number of citations of an article increases the importance of the developed research to the scientific community (Aksnes *et al.*, 2019).

Table 5. Descriptive analysis: Numbers of articles and citations by countries (based on first author's affiliation)

Country	N° Articles	% of Articles	N° Citations
China	14	41.18	143
USA	9	26.47	183
Canada	4	11.76	43
Belgium	1	2.94	2
Brazil	1	2.94	0
India	1	2.94	12
Iran	1	2.94	4
Ireland	1	2.94	20
Mexico	1	2.94	6
Portugal	1	2.94	49

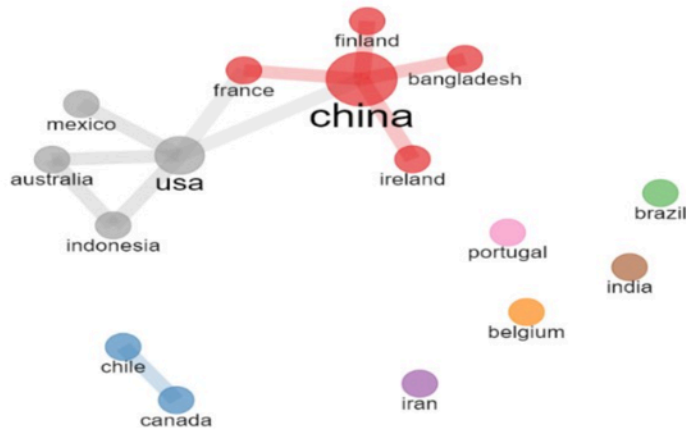


Figure 4. Collaboration network. Size of circles indicates the number of authors (including authors and co-authors) by country, links between the circles represent co-authoring between authors from different countries in the same publication, and a shorter distance between two circles indicates a higher number of co-occurrences between two co-authoring. Clusters were labeled using different colors.

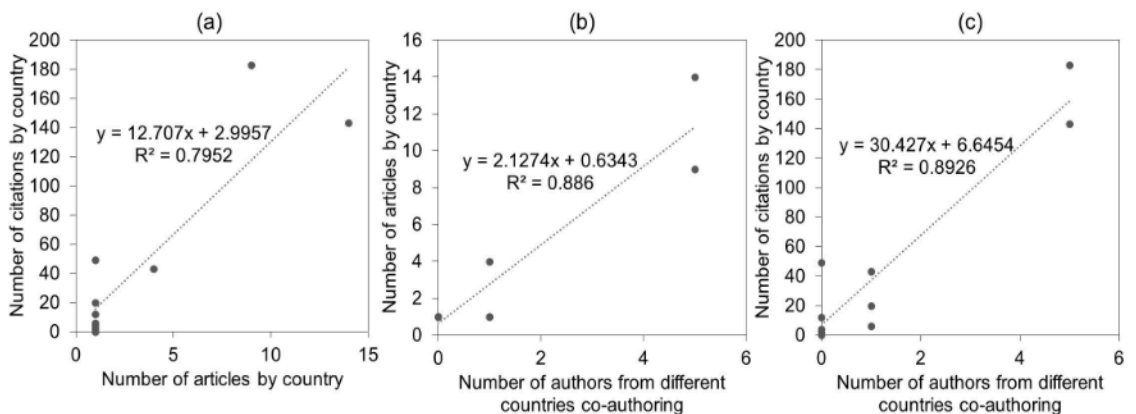


Figure 5. Relationship between the number of articles by country and the number of citations by country (a), the number of authors from different countries co-authoring and the number of articles by country (b), and the number of authors from different countries co-authoring and the number of citations by country.

4. Conclusion

Computational advances and the development of AI have enabled the utilization of more robust models that go beyond ANOVA and simple regression, thereby yielding more meaningful results for researchers. The reviewed articles spanned various domains, illustrating the widespread adoption of regression models across different fields. It is anticipated that as computational power continues to increase, regression analysis will become even more pervasive outside the realm of traditional statistics. Our observations underscore the significance of international collaboration among authors to enhance the relevance of their research within the scientific community. Our study was confined to applications within the rhizosphere due to the current prominence of such research. However, a more comprehensive investigation, encompassing diverse topics in soil science and potentially extending into other areas of agricultural science, could broaden researchers' perspectives and encourage the adoption of more robust statistical methods in their work. One of the most critical limitations hindering the utilization of the models presented in our article, as well as other advanced statistical techniques, is the absence of a statistician or researcher with expertise in statistical modeling within our research groups.

Acknowledgments

Authors would like to thank reviewers and editors for their comments.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Ahmed, M. A., Kroener, E., Holz, M., Zarebanadkouki, M. & Carminati, A. Mucilage exudation facilitates root water uptake in dry soils. *Functional Plant Biology* **41**, 1129–1137 (2014).
2. Aksnes, D. W., Langfeldt, L. & Wouters, P. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open* **9**, 2158244019829575 (2019).
3. Aria, M. & Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* **11**, 959–975 (2017).
4. Ayangbenro, A. S., Chukwuneme, C. F., Ayilara, M. S., Kutu, F. R., Khantsi, M., Adeleke, B. S., Glick, B. R. & Babalola, O. O. Harnessing the Rhizosphere Soil Microbiome of Organically Amended Soil for Plant Productivity. *Agronomy* **12**, 3179 (2022).
5. Basili, V. R., Carver, J. C., Cruzes, D., Hochstein, L. M., Hollingsworth, J. K., Shull, F. & Zelkowitz, M. V. Understanding the high-performance-computing community: A software engineer's perspective. *IEEE software* **25**, 29 (2008).
6. Batista, A. M., Libardi, P. L., Alves, M. E., Prativiera, F. & Giarola, N. F. B. Electrochemical Effects on Clay Dispersion in Rhizo- and Non-rhizospheric Soils. *Journal of Soil Science and Plant Nutrition* **22**, 3518–3526 (2022).
7. Beschoren da Costa, P., Benucci, G. M. N., Chou, M.-Y., Van Wyk, J., Chretien, M. & Bonito, G. Soil Origin and plant genotype modulate switchgrass aboveground productivity and root microbiome assembly. *Mbio* **13**, e00079–22 (2022).
8. Chen, L., Zhang, C. & Duan, W. Temporal variations in phosphorus fractions and phosphatase activities in rhizosphere and bulk soil during the development of *Larix olgensis* plantations. *Journal of plant nutrition and soil science* **179**, 67–77 (2016).

9. Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J. & Zhu, Y. Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE* **104**, 2207–2219 (2016).
10. Chueke, G. V. & Amatuucci, M. O que é bibliometria? Uma introdução ao Fórum. *Internext* **10**, 1–5 (2015).
11. Clarivate. Journal Impact Factor, Journal Citation Reports. *Accessed in 24 nov 2023*. **13** (2023).
12. Cloutier-Hurteau, B., Sauve, S. & Courchesne, F. Predicting Al, Cu, and Zn concentrations in the fine roots of trembling aspen (*Populus tremuloides*) using bulk and rhizosphere soil properties. *Canadian journal of forest research* **41**, 1267–1279 (2011).
13. Cloutier-Hurteau, B., Turmel, M.-C., Mercier, C. & Courchesne, F. The sequestration of trace elements by willow (*Salix purpurea*)—which soil properties favor uptake and accumulation? *Environmental Science and Pollution Research* **21**, 4759–4771 (2014).
14. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future healthcare journal* **6**, 94 (2019).
15. DeWolf, E., Brock, M. T., Calder, W. J., Kliebenstein, D. J., Katz, E., Li, B., Morrison, H. G., Maïgnien, L. & Weinig, C. The rhizosphere microbiome and host plant glucosinolates exhibit feedback cycles in *Brassica rapa*. *Molecular Ecology* (2022).
16. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. & Lim, W. M. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research* **133**, 285–296 (2021).
17. Duncan, L., Graham, J., Timmer, L., *et al.* Seasonal patterns associated with *Tylenchulus semipenetrans* and *Phytophthora parasitica* in the citrus rhizosphere. *Phytopathology* **83**, 573–581 (1993).
18. Echer, F. R., Peres, V. J. S. & Rosolem, C. A. Potassium application to the cover crop prior to cotton planting as a fertilization strategy in sandy soils. *Scientific Reports* **10**, 1–10 (2020).
19. Edge, T. A. & Wyndham, R. C. Predicting survival of a genetically engineered microorganism, *Pseudomonas chlororaphis* 3732RN-L11, in soil and wheat rhizosphere across Canada with linear multiple regression models. *Canadian journal of microbiology* **48**, 717–727 (2002).
20. Fasusi, O. A., Cruz, C. & Babalola, O. O. Agricultural sustainability: microbial biofertilizers in rhizosphere management. *Agriculture* **11**, 163 (2021).
21. Fox, A., Gahan, J., Ikoyi, I., Kwapinski, W., O'sullivan, O., Cotter, P. D. & Schmalenberger, A. *Miscanthus* biochar promotes growth of spring barley and shifts bacterial community structures including phosphorus and sulfur mobilizing bacteria. *Pedobiologia* **59**, 195–202 (2016).
22. Guo, P., Jia, X., Duan, T., Xu, J. & Chen, H. Influence of plant activity and phosphates on thorium bioavailability in soils from Baotou area, Inner Mongolia. *Journal of environmental radioactivity* **101**, 767–772 (2010).
23. Guzmán-Plazola, R., Navas, J. d. D. J., Caswell-Chen, E., Zavaleta-Mejía, E. & del Prado-Vera, I. C. Spatial distribution of *Meloidogyne* species and races in the tomato (*Lycopersicon esculentum* Mill.) producing region of Morelos, Mexico. *Nematropica*, 215–230 (2006).
24. Hiltner, L. Ober neuter erfahrungen und probleme auf dem gebiete der bodenbakteriologie unter besonderer berucksichtigung der grundung und brache. *Arbeiten der Deutschen Landwirtschaftlichen Gesellschaft* **98**, 59–78 (1904).
25. Ibekwe, A. M., Grieve, C. M. & Yang, C.-H. Survival of *Escherichia coli* O157: H7 in soil and on lettuce after soil fumigation. *Canadian Journal of Microbiology* **53**, 623–635 (2007).
26. Ibrahim, A. S., Musa, A. A., Abdulfatah, A. Y. & Idris, A. Developing soft-computing regression model for predicting soil bearing capacity using soil index properties. *Modeling Earth Systems and Environment* **9**, 1223–1232 (2023).

27. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
28. Leath, S & Carroll, R. Use of ridge regression to predict yield reduction by *Fusarium* sp. in selected soybean cultivars. *Canadian journal of plant pathology* **7**, 58–66 (1985).
29. Li, C., Zhang, C., Yu, T., Liu, X., Yang, Y., Hou, Q., Yang, Z., Ma, X. & Wang, L. Use of artificial neural network to evaluate cadmium contamination in farmland soils in a karst area with naturally high background values. *Environmental Pollution* **304**, 119234 (2022).
30. Li, P., Yin, R., Zhou, H., Yuan, X. & Feng, Z. Soil pH drives poplar rhizosphere soil microbial community responses to ozone pollution and nitrogen addition. *European Journal of Soil Science* **73**, e13186 (2022).
31. Ma, B., Chen, H., He, Y., Wang, H. & Xu, J. Evaluation of toxicity risk of polycyclic aromatic hydrocarbons (PAHs) in crops rhizosphere of contaminated field with sequential extraction. *Journal of Soils and Sediments* **10**, 955–963 (2010).
32. Ma, B., Chen, H., Xu, M., Hayat, T., He, Y. & Xu, J. Quantitative structure–activity relationship (QSAR) models for polycyclic aromatic hydrocarbons (PAHs) dissipation in rhizosphere based on molecular structure and effect size. *Environmental Pollution* **158**, 2773–2777 (2010).
33. Ma, X., Yang, Z., Yu, T. & Guan, D.-X. Probability of cultivating Se-rich maize in Se-poor farmland based on intensive field sampling and artificial neural network modelling. *Chemosphere* **309**, 136690 (2022).
34. Manjunath, A & Habte, M. Root morphological characteristics of host species having distinct mycorrhizal dependency. *Canadian Journal of Botany* **69**, 671–676 (1991).
35. McNear Jr., D. H. The Rhizosphere–Roots, Soil and Everything In Between. *Nature Education Knowledge* **4**, 1–15 (2013).
36. Mohanty, S., Kollah, B., Chaudhary, R. S., Singh, A. B. & Singh, M. Methane uptake in tropical soybean–wheat agroecosystem under different fertilizer regimes. *Environmental Earth Sciences* **74**, 5049–5061 (2015).
37. Mokhnacheva, Y. V. & Tsvetkova, V. Bibliometric analysis of soil science as a scientific area. *Eurasian Soil Science* **53**, 838–844 (2020).
38. Monreal, C. *et al.* Bacterial community structure associated with the addition of nitrogen and the dynamics of soluble carbon in the rhizosphere of canola (*Brassica napus*) grown in a Podzol. *Rhizosphere* **5**, 16–25 (2018).
39. Moreira, H., Pereira, S. I., Vega, A., Castro, P. M. & Marques, A. P. Synergistic effects of arbuscular mycorrhizal fungi and plant growth-promoting bacteria benefit maize growth under increasing soil salinity. *Journal of Environmental Management* **257**, 109982 (2020).
40. Pan, F., Liang, Y., Wang, K. & Zhang, W. Responses of fine root functional traits to soil nutrient limitations in a karst ecosystem of Southwest China. *Forests* **9**, 743 (2018).
41. Pratavia, F., Batista, A. M., Libardi, P. L., Cordeiro, G. & Ortega, E. M. M. Joint regression modeling of location and scale parameters of the skew t distribution with application in soil chemistry data. *Journal of Applied Statistics* **49**, 195–213 (2022).
42. Pratavia, F., Batista, A. M., Ortega, E. M., Cordeiro, G. M. & Silva, B. M. The Logit exponentiated power exponential regression with applications. *Annals of Data Science*, 1–23 (2021).
43. Raaijmakers, J. M. & Mazzola, M. Soil immune responses. *Science* **352**, 1392–1393 (2016).
44. Raza, M. A., Aziz, S., Noreen, M., Saeed, A., Anjum, I., Ahmed, M. & Raza, S. M. Artificial Intelligence (AI) in Pharmacy: An Overview of Innovations. *INNOVATIONS in pharmacy* **13** (2022).

45. Rezakhani, L., Motesharezadeh, B., Tehrani, M. M., Etesami, H. & Hosseini, H. M. The effect of silicon fertilization and phosphate-solubilizing bacteria on chemical forms of silicon and phosphorus uptake by wheat plant in a calcareous soil. *Plant and Soil*, 1–22 (2022).
46. Sarker, I. H. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science* **2**, 377 (2021).
47. Schmidt, J. E., Firl, A., Hamran, H., Imaniar, N. I., Crow, T. M. & Forbes, S. J. Impacts of Shade Trees on the Adjacent Cacao Rhizosphere in a Young Diversified Agroforestry System. *Agronomy* **12**, 195 (2022).
48. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 83 (2019).
49. Shimamura, E., Merckx, R. & Smolders, E. Limited effects of the soluble organic phosphorus fraction on the root phosphorus uptake efficiency of upland rice genotypes grown in acid soil. *Soil Science and Plant Nutrition* **67**, 120–129 (2021).
50. Snyder, H. Literature review as a research methodology: An overview and guidelines. *Journal of business research* **104**, 333–339 (2019).
51. Srivastava, P. K., Yaduvanshi, A., Singh, S. K., Islam, T. & Gupta, M. Support vector machines and generalized linear models for quantifying soil dehydrogenase activity in agro-forestry system of mid altitude central Himalaya. *Environmental Earth Sciences* **75**, 1–15 (2016).
52. Tang, L., Zhan, M., Shang, C., Yuan, J., Wan, Y. & Qin, M. Dynamics of root exuded carbon and its relationships with root traits of rapeseed and wheat. *Plant, Soil and Environment* **67**, 317–323 (2021).
53. Tian, K., Bao, H., Liu, X. & Wu, F. Accumulation and distribution of PAHs in winter wheat from areas influenced by coal combustion in China. *Environmental Science and Pollution Research* **25**, 23780–23790 (2018).
54. Van Veelen, A., Tourell, M. C., Koebernick, N., Pileio, G. & Roose, T. Correlative visualization of root mucilage degradation using X-ray CT and MRI. *Frontiers in Environmental Science* **6**, 32 (2018).
55. Wang, A. S., Angle, J. S., Chaney, R. L., Delorme, T. A. & McIntosh, M. Changes in soil biological activities under reduced soil pH during *Thlaspi caerulescens* phytoextraction. *Soil Biology and Biochemistry* **38**, 1451–1461 (2006).
56. Wang, Y., Hu, Z., Shen, L., Liu, C., Islam, A. T., Wu, Z., Dang, H. & Chen, S. The process of methanogenesis in paddy fields under different elevated CO₂ concentrations. *Science of The Total Environment* **773**, 145629 (2021).
57. Wu, X., Huang, C., Sha, L. & Wu, C. Influence of rhizosphere activity on litter decomposition in subtropical forest: implications of estimating soil organic matter contributions to soil respiration. *Journal of Tropical Ecology* **38**, 151–157 (2022).
58. Xia, Z., He, Y., Yu, L., Li, Z., Korpelainen, H. & Li, C. Revealing interactions between root phenolic metabolomes and rhizosphere bacterial communities in *Populus euphratica* plantations. *Biology and Fertility of Soils* **57**, 421–434 (2021).
59. Yao, Y., Yao, X., An, L., Bai, Y., Xie, D. & Wu, K. Rhizosphere bacterial community response to continuous cropping of Tibetan Barley. *Frontiers in microbiology* **11**, 551444 (2020).
60. Zia, R., Nawaz, M. S., Siddique, M. J., Hakim, S. & Imran, A. Plant survival under drought stress: Implications, adaptive responses, and integrated rhizosphere management strategy for stress mitigation. *Microbiological research* **242**, 126626 (2021).