ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΔΙΟΙΚΗΣΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Analysis of problems in the bibliometric data of the Journal of Finance from the Web of Science database.

ΣΟΥΧΛΑ ΣΤΑΜΑΤΙΝΑ
Α.Μ:2312014127

Επιβλέποντες
Βασιλείου Ευάγγελος
(Τμήμα Μηχανικών Οικονομίας και Διοίκησης, Πανεπιστήμιο Αιγαίου)
Δριβαλιάρης Δημοσθένης
(Τμήμα Μηχανικών Οικονομίας και Διοίκησης, Πανεπιστήμιο Αιγαίου)

Μέλη Επιτροπής
Κωνσταντέλου Αναστασία
(Τμήμα Μηχανικών Οικονομίας και Διοίκησης, Πανεπιστήμιο Αιγαίου)

ΧΙΟΣ, 2020

Στην οικογένειά μου
και τους φίλους μου

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές μου Βασιλείου Ευάγγελο και Δριβαλιάρη Δημοσθένη για την στήριξή και την βοήθεια τους καθώς και την κυριά Κωνσταντέλου Αναστασία ως μέλος επιτροπής.

Θέλω επίσης να ευχαριστήσω την οικογένεια μου και όλους τους φίλους μου για την συμπαράσταση και την υπομονή τους καθ όλη την διάρκεια της φοιτητικής ζωής.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή θα επικεντρωθούμε στην ανάλυση προβλημάτων που αντιμετωπίσαμε στην πηγή δεδομένων μας Web of Science αλλά και στην συνέχεια στην πηγή Scopus με βάση το περιοδικό μας "The Journal of Finance". Για να επιτευχθεί αυτή η ανάλυση προβλημάτων πρέπει να υπάρξει σύγκριση μεταξύ των δύο πηγών με διάφορα κριτήρια σε κάθε στάδιο με την βοήθεια της βιβλιομετρικής ανάλυσης. Τα στάδια που θα εξετάσουμε στην εργασία αυτή είναι οι διπλοεγγραφές κάθε πηγής όπου θα εντοπιστούν με συγκεκριμένα κριτήρια και στην συνέχεια θα εξαιρεθούν για να περάσουμε στο επόμενο βήμα που είναι η ένωση των δυο πηγών μας. Η ένωση θα μας βοηθήσει να βρούμε ποιες εγγραφές είναι κοινές από τις βάσεις και να καταλήξουμε με ένα όσον το δυνατόν καθαρό αρχείο.

Abstract

In this work we will focus on the analysis of problems we encountered in our Web of Science data source and then in the Scopus source based on our journal "The Journal of Finance". To achieve this problem analysis there must be a comparison between the two sources with different criteria at each stage with the help of bibliometric analysis. The stages that we will examine in this work are the duplicates of each source where they will be identified with specific criteria and then they will be excluded to move on to the next step which is the union of our two sources. The association will help us find which records are common to the databases and come up with a file that is as clean as possible.

CONTENTS

List of tables *8*

1. Introduction to Bibliometrics *9*

1.1 Historical Evolution of Bibliometrics **9**

1.2 Application of Bibliometrics **10**

1.3 List of methods based on bibliometrics **12**

1.4 Bibliometric analysis and its impact parameters **13**

2. Research *15*

2.1 American Finance Association (AFA) & the Journal Of Finance **15**

2.2 AFA – History **15**

2.3 Citation Information **15**

3. Web of Science *17*

3.1 Introduction **17**

3.2 History **17**

3.3 Data bases **17**

4. The data *19*

4.1 Data types **19**

4.2 Field types **19**

5. Duplicates *22*

5.1 Checking Types **22**

5.2 Digital Object Identifier (DOI) **22**

5.3 Title (TI) **23**

5.4 Volume (VL), Issue (PN), Pages (PP) **29**

5.5 Volume (VL), Issue (PN), Start Page **31**

5.6 Volume (VL), Issue (PN), End Page **32**

6. Missing Value *33*

6.1 Introduction **33**

6.2 Missing Data **33**

6.3 Methods for handling missing data **33**

6.4 Checking missing values **34**

7. Merging the data sets *37*

7.1 DOI(Digital Object Identifier) **37**

7.2 TI(Title) **37**

7.3 PN (Issue), VL(Volume), START PAGE **37**

7.4 PN (Issue), VL(Volume), END PAGE **37**

8. Total registry *39*

8.1 Comparisons and differences between Sources 39
8.2 Differences 39
8.3 Similarities 42
REFERENCES 44

List of tables

Table 2.3 1: Citation information per year 16
Table 4.2 1: Explain columns 21
Table 5.2 1: Duplicates in case DOI 23
Table 5.3 1: Duplicates in case TI 28
Table 5.4 1: Duplicates in case Volume, ISSUE, Pages 31
Table 6 1: Missing value for each column 31
Table 6 2 1: Missing Value for combine columns 32
Table 8.21:Differences in relation to DOI column 40
Table 8.22:Differences in relation to TI column 41
Table 8.23 : Differences in relation to AU column 41
Table 8.24: Differences in relation to START column 42
Table 8.25 : Differences in relation to END column 42
Table 8.3 1: The similarities between Scopus and Web of Science 42

1. Introduction to Bibliometrics

In this chapter we will review the history of the bibliometrics and their pioneers. In the following, we will mention the application of bibliometrics and a list of methods based on the bibliometrics. Finally, we will report a bibliometric analysis and its parameters.

1.1 Historical Evolution of Bibliometrics

The idea of conducting a research and examination of literature has its roots at the beginning of the century. In this section a historical overview of all the pioneer of bibliometric is represented, covering the period from 1917 until 20th century. Starting from 1917 and reaching the 80's and beyond. In 1917, the scientists FJ.Cole and Nellie Eales published a statistical analysis of the history of comparative anatomy. This date marked a milestone in the history of bibliometric analysis, as Cole and Eales were among the first to use the published research work to create a quantitative picture of the progress being made in a research field. Their work describes the contribution of Bibliometry as well as the problems it poses [1]. Otlet was then the one who used the term Bibliometry to describe the technique used to quantify science and scientists. Otlet (1920), a pioneer in the science of information and its theory, insists on the difference between Bibliometry and Statistical Bibliography, arguing that science from its inception is measured or quantified by applying statistical methods to information sources. Otlet's view is that Bibliography is established as a general science that systematically collects and classifies the totality of data, which relates to the production, maintenance, circulation, and use of all kinds of writing and documents. Otlet proposed a number of basic principles for the field of Librarianship, taking into account a number of factors that affect or surround the text. These include the language, the intervals contained and the factors mentioned among others, in the form, layout and price of the unit as well as in factors that belong to the statistics, such as comparison indicators. It also pays attention to the frequency at which a given author or work is read. From this data it implies that a "frequency of use" curve can be designed, taking into account the number of editions of a text depending on the author and its content or the context of the social extensions in which it appears [1].In 1926,when Alfred J. Lotka published his pioneering study on the frequency distribution of scientific productivity determined from a decennial index (1907- 1916) of Chemical Abstracts. Lotka concluded that "the number (of authors) making n contributions is about 1/n2 of those making one; and the proportion of all contributors, that makes a single contribution, is about 60 per cent." This result can be considered as a rule of thumb even today [2]. During almost the same period, in 1927, Gross and Gross (1927) published a study focusing on citation to help decide which chemistry journals would be best purchased from small college libraries. In particular, they examined 3633 citations from the 1926 volume of the Journal of the American Chemical Society. This study is considered to be the first citation analysis, although it was not a reference analysis in the current sense [1].Eight years after Lotka's article appeared, Bradford (1934) published his study on the frequency distribution of papers over journals. He found that "if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the numbers of periodicals in the nucleus and the succeeding zones will beas 1:b:b2...". An important consequence of the law is that in a search for a specific topic, a large number of related articles will be concentrated in a small number of journal titles (Nordstrom, 2005) [2]. These laws usually make estimates of reporting indicators, as well as of various library services. However, it was S.W. Fernberger of the University of Pennsylvania who developed the statistics on the publication. Fernberger (1936) studied the evolution of researchers and gave increasing emphasis to publication as a criterion for eligibility. Fernberger was the one who imposed the notions of productivity and the index for measuring the productivity of science [1].Then in 1949, Zipf (1949) formulated an interesting law in bibliometrics and quantitative linguistics that he derived from the study of word frequency in a text. It can be considered a generalisation of the laws by Lotka and Bradford. He

formulated the following underlying principle of his law although he has never shown how this principle applies to his equation. "The Principle of Least Effort means... that a person...will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems...." (Zipf, 1949) [2].The situation changed dramatically in the early 1960s, when historian Derek de Solla Price published his fundamental work in Bibliometry, which is analyzed in his two books, the first on "Science from the Babylonian Era" (1961) and the second for "Little Science, Great Science" (1963) [1]. In his book entitled “Little Science – Big Science” (1963), Derek de Solla Price analysed the recent system of science communication and thus presented the first systematic approach to the structure of modern science applied to the science as a whole. At the same time, he laid the foundation of modern research evaluation techniques [2]. In 1969, the term "Bibliometry" was proposed by Alan Pritchard (1969) as the most representative and was defined as "the application of mathematical and statistical methods to books and other publications" or, more specifically, to "the quantitative study of bibliographical references. As they appear in the bibliographies, with the aim of providing evolutionary models in science and technology. " Although Bibliometry was then used as a model to measure the output of scientists' publications nearly a century ago, the term was first introduced, as mentioned above, by Alan Pritchard in his work entitled "Statistical Bibliography or Bibliometry?" in 1969. But what greatly helped the quantitative analysis of scientific publications was the work of Eugene Garfield in the 1960s and the indexes he introduced under the name Social-Arts-and Humanities Science Citations Indexes, through the Institute for Scientific Information (ISI). Garfield's original idea and goal was to provide researchers with a fast and effective way of finding published articles that addressed the various areas of their research (Garfield, 1968). However, he soon expanded his study and work by evaluating the reports cited, thus: "The conclusion to be drawn is that as the scientific enterprise grows larger and more complex and its role in society becomes ever greater, and the more critical, the more difficult, costly, but also necessary will be to evaluate and clearly identify the largest and most important contributions” (Garfield, 1979b). Garfield attempted to portray the analysis of references as a legitimate and practical tool for the evaluation of scientific production. Price (1976) introduced an interest in the science of science, based on precise quantitative analysis, and on the one hand of the rates of scientific production, that is, the number of scientific books and journals per unit of time, and on the other the number of people employed in the field Science. In the 1970s and 1980s, Bibliometry saw a steep rise and a new orientation. Then at the beginning of the eighties, Bibliometrics evolved into a separate field with characteristic profiles, subfields and scientific communication structures. Institutionalization of the field began in 1978 with the release of Scientometrics, international conferences since 1983, and Evaluation Research since 1991 [1]. Later, from the early 1980s, bibliometrics could develop into a separate scientific discipline with a specific research profile, some subfields and corresponding scientific communication structures [2]. In the 1990s, bibliometrics became a research management tool with sophisticated techniques. The fact that bibliometric methods are already applied to the field “bibliometrics” itself also indicates the rapid development of the discipline. At that time, most basic models for scientific communication were developed. Among these are first models for essential concepts in scientific communication like growth and ageing of information. Literature and information was assumed to grow exponentially, but in individual research disciplines the growth can also be linear or logistic. Finally, the logistic model has been widely accepted since both exponential and linear growth can be considered special phases within the logistic model. The concept of ageing or obsolescence is intimately linked with the growth of science. In information science and bibliometrics, changing frequency of citations given or received over time is assumed to reflect ageing of scientific literature.

1.2 Application of Bibliometrics

Today, bibliometric analysis helps in a wide range of fields. the most important of these are [3]:

i. Bibliography for Librarians (Methodology)

This is the main research area of the library and is traditionally funded by the usual grants. The methodological research is carried out mainly in this field.

ii. Bibliography for scientific disciplines (scientific information)

In this field, we see that bibliography helps in many scientific fields, such as in the history of science, where it contributes to the clarification of the evolution and evolution of sciences, identifying the historical movements that are reflected in the results of researchers. An examination of the scientific literature supports and reinforces the analysis of the scientific community and its structure in a given society, as well as the motivations and networks of researchers. Researchers in the scientific field are the largest and most diverse group of interests in accounting. Due to their original scientific orientation, their interests are closely linked to their specialty. This field can be considered as an extension of scientific information by metric means. Here we also find a common issue with quantitative research in information retrieval.

iii. Bibliography on Scientific Policy and Administration (Scientific Policy)

The science policy it provides indicators for measuring productivity and scientific quality. At this point are the national, regional and institutional structures of science and their comparative presentation. In essence, it provides a basis for the evaluation and orientation of E&A.

Bibliometric techniques have evolved over time and continue in this direction and have the following:

- The counting of documents per country, institution / author and author.
- Measuring reports to assess the impact of published work on the scientific community.
- Counting coherent reports (ie how many times two reports are mentioned and referred to in a single document).

So we mentioned the areas in which the bibliometric study has been restored and every year it becomes necessary, but why is it necessary? Below we will see the ways that the bibliometric analysis uses and helps the branches that we mentioned above. So, according to our sources, we have that:

- To quantify research and development
- Determine the completeness of the secondary journals.
- Identify the uses and publishers of secondary journals
- Identify the main magazines in different industries to formulate a need based on market policy.
- Launch an effective multi-level network system.
- To regulate the inflow of information and their communication.
- Development of standardization standards.
- Predict the Productivity of Publishers, country or the whole discipline

Today, bibliometric analysis is applied to a wide range of fields:

In the history of science, where it helps to clarify the development and evolution of the disciplines, by identifying the historical movements that are reflected in the results produced by the researchers.

In the social sciences, where, by examining the scientific literature, it supports and enhances the analysis of the scientific community and its structure in a given society, as well as the motivations and networks of researchers

In the documentation, where it can calculate the number of journals available per library, as well as identify the journals that constitute the core, secondary sources and periphery of a discipline.

In science policy, where it provides indicators of measuring productivity and scientific quality. In essence, that is to say, it provides a basis on which to evaluate and orient R&D.

The question where bibliometric answers, have evolved over time and continue in this direction and are as follows:

- The counting of documents by country-by-institution, institution / author and author.
- Counting the reports in order to estimate the impact of the published work on the scientific community.
- Counting the co-reports (that is to say how many times two research papers are cited and cited together in a single document) [1, 4, 3].

1.3 List of methods based on bibliometrics

Derek J. de Solla Price (1965) proposed scientific methods of science for the study of science. Bibliometric methods use bibliographic data from databases to construct scientific images of building blocks. They introduce a measure of objectivity in the evaluation of scientific literature and can be used for the detection of informal research networks. References to research fields, compiled over time, reflect the authors 'assessments of the subject, methodology, and value of other authors' work. Bibliometric methods are concerned with performance analysis and science mapping. Performance analysis deals with the evaluation of research and publication performance of individuals and institutions. The mapping of science reveals the structure and dynamics of the scientific fields, where they are useful to the researcher when he wants to review a particular line of research. Bibliometric methods are quantitatively rigorous in evaluating literature, demonstrating emerging categories in review articles. The main bibliometric methods are five. The first three use citation data to construct measures of influence and similarity: citation analysis, pooling analysis, and bibliographic coupling. Co-author analysis uses co-author data to measure collaboration. Keyword analysis finds links between concepts that coexist in document titles, keywords, or summaries.

Co-citation analysis uses co-citation measurements to construct measures of similarity between documents, authors, or journals. Co-referencing is defined as the frequency with which two units are referred to together. A fundamental assumption of the interview is that the more two elements are mentioned together, the more likely it is that their content is relevant. Co-citation links documents, authors, or periodicals in the way authors use them.

Document co-citation analysis connects specific published documents. Author co-citation analysis (ACA) connects bodies of writings by a person and therefore the authors who produced them. ACA can identify important authors and connect them through citation records. What is mapped is an author's citation image. Journal co-citation analysis (JCA) aims to connect related scientific journals. A special form of co-citation is tri-citation analysis, which examines the "intellectual fellow travelers" of a particular author or publication by analyzing works which have been co-cited with them. It has the potential for researching the legacy of important authors or seminal studies. Tri-citation is a variant of co-citation analysis where the focal author or publication is always one of the cited publications and provides the context for co-citation analysis.

Bibliographic coupling uses the number of references shared by two documents as a measure of the similarity between them. The more the bibliographies of two articles overlap, the stronger their connection. The number of references shared between two documents is static over time as the number of references within the article is unchanged, while relatedness based on co-citation develops with citation patterns. As citation habits change, bibliographic coupling is best performed within a limited timeframe It is best to analyze publications from roughly the same period of time. A bibliographic coupling connection is established by the authors of the articles in focus, whereas a co-citation connection is established by the authors who are citing the examined works. When two documents are highly co-cited this means that each individual document is also highly individually cited. This indicates that documents selected through co-citation thresholds are deemed more important by the researchers who are citing them.

Co-author analysis examines the social networks scientists create by collaborating on scientific articles. A relationship between two authors is established when they co-publish a paper. Co-authoring scientific publications is presumed to be a measure of collaboration. Co-authorship reflects stronger social ties than other relatedness measures, which makes it

particularly suitable for examining social networks rather than intellectual structures of research fields. In addition, because bibliographic data contains information about authors' institutional affiliations and their geographical location, co-author analysis can examine the issues of collaboration on the level of institutions and countries. Co-authorship as a measure of collaboration assumes that authoring a publication is synonymous with being responsible for the work done. However, just because a person's name appears as a co-author of a scientific article it is not necessarily because they contributed a significant amount of work, but could be purely "honorary authorship" for social or other reasons. On the other hand, there might be scientists who contributed to the work but whose names do not appear on the author sheet.

Co-word analysis is a content analysis technique that uses the words in documents to establish relationships and build a conceptual structure of the domain. The idea underlying the method is that, when words frequently co-occur in documents, it means that the concepts behind those words are closely related. It is the only method that uses the actual content of the documents to construct a similarity measure, while the others connect documents indirectly through citations or co-authorships. The output of co- 7 word analysis is a network of themes and their relations which represent the conceptual space of a field. This semantic map helps to understand its cognitive structure. A series of such maps produced for different time periods can trace the changes in this conceptual space. Co-word analysis can be applied to document titles, keywords, abstracts or full texts. The unit of analysis is a concept, not a document, author or journal. The quality of results from co-word analysis depends on variety of factors – the quality of keywords, the scope of the database and the sophistication of statistical methods used for analysis. Solely using keywords for co-word analysis is a problem for two reasons. First, many journals' bibliographic data do not contain keywords. Second, relying just on keywords suffers from so-called "indexer effect" – where the validity of the map is dependent on whether the indexers captured all relevant aspects of the text. The solution is to use abstracts or full texts, but this introduces noise into the data as the algorithms have difficulty distinguishing the importance of words in large corpuses of text [5] [4].

1.4 Bibliometric analysis and its impact parameters

As mentioned above, bibliometric analysis aims to determine the trends of scientific research, through the processing of data derived from the literature, at the level of institution, country or set of countries, scientific field, author, etc. This project seeks to identify networks in the scientific community. The number of publications and the analysis of the reports in the publications are the main indicators for the recording of the scientific work. Bibliometric analysis is influenced by parameters, which are [6]:

i. Time reporting period: Obviously, the number of reports contained in a research paper is related to the amount of time that has elapsed since its original publication, which means that older works have a larger number of reports. It should be noted, however, that the large number of reports does not always ensure their impact on the scientific field, nor does it ensure the quality of the research work. In any case, to address issues related to the large number references in older publications specify specific time intervals for measuring references per publication.

ii. Scientific disciplines: There are several differences between the scientific disciplines associated with practical research publications, reporting practices, and the duration of the research project, which does not allow the comparison of bibliographic indicators between different research disciplines. In particular, we note that in the field of medicine there is a very large number of scientific articles in journals per year, which in a short period of time since their publication, reach a maximum number of reports. In contrast, the number of publications per year in the social sciences is much smaller with a large, however, period of publication, in which reports are recorded. On the other

hand, there are sciences, such as computer science, that are not used to publishing their scientific work in journals, but it is their permanent practice to publish their research results in conferences. All this makes it difficult and precarious to compare indicators between different scientific fields.

iii. Type of scientific publications: The type of scientific publications plays a decisive role in the number of references included in them. It has been observed that review articles contain a very large number of reports in relation to other types of research papers, with the result that the selection of the appropriate bibliometric indicator for the evaluation of researchers is of utmost importance.

2. Research

The study on the economic journal, which called “The Journal of Finance”, has its roots in the past, as in 1946 the first steps have been made. With the pass of the years this study gas been developed, giving to the researcher the opportunity to analyze this evolution over time. In the current study we followed the progress on the journal from 1946 until 2018. All the information about the journal, included in this paragraph, are received from the following three websites: journal's website [7], the American Finance Association website [8] , the Wikipedia website [9]and the jstor website [10].

2.1 American Finance Association (AFA) & the Journal Of Finance

According to the American Finance Association website [8] “The Journal of Finance” publishes leading research across all the major fields of finance. It is one of the most widely cited journals in academic finance, and in all of economics. Each of the six issues per year reaches over 8,000 academics, finance professionals, libraries, and government and financial institutions around the world. The journal is the official publication of The American Finance Association, the premier academic organization devoted to the study and promotion of knowledge about financial economics”.

2.2 AFA – History

The American Finance Association (AFA) [8] is the premiere academic organization devoted to the study and promotion of knowledge about financial economics. The purpose of the Association is to provide for the mutual association of persons with an interest in finance to improve public understanding of financial problems, and to provide for the exchange of financial ideas through the distribution of a periodical and other media; to encourage the study of finance in colleges and universities; to conduct such other activities as may be appropriate for a non-profit, professional society in the field of finance.

The American Finance Association [8] is an academic organization, which was established in December 1939 in Philadelphia. The first journal, published in 1942, was called “American Finance”. Afterwards, it was renamed as “The Journal of Finance” and became a regular serial journal in August 1946. The editor in chief is Stefan Nagel since 2016. The journal was being published from 1946-2015 by the American Association and from 2015 up to now by Wiley-Blackwell.

2.3 Citation Information

It is an academic journal that publishes leading-edge research in all areas of finance and is considered to be the leading journal for academic finance and economics. Each year are published up to 6 issues and in total 73 volumes of the journal. The number of published issues varies by year. In 1946 the journal published 1 issue in August, 1947 1 issue in April and more issue in October. In 1948 published 3 issues in the months February, June and October. From 1949 to 1967 he published the journal 4 issues per year in the months March, May, September, and December. From 1968 to 1983, the journal published 5 issues each year in the months March, May, June, September, and December. In the following 14 years from 1984 to 1997, the magazine continued to publish 5 issues but the months of publication differ as opposed to May published each July. From 1998 to now the months of publication are February, April, June, August, October, December and each year corresponds to 6 issues [7].

Year	Number of issues	Volume
1946	**1**	**1**
1947	**2**	**2**
1948	**3**	**3**
1949-1967	**4**	**4-22**
1968-1997	**5**	**23-53**
1998-2018	**6**	**54-73**

Table 2.3 1: Citation information per year

1946 is the only year in which the magazine has an extra issue that is listed as S2. This issue contains an introduction to pages 1-3 and follows 3 articles from pages 4-12, 13-22 and 23-43 respectively. The Articles are “II recommendations for further research: the capital market as a whole”,

“III recommendations for further research: particular sectors of the capital market” and “IV an inventory of recent and current research” [7].

3. Web of Science

In this chapter we will analyze the web of science database, we will see in detail which databases it consists of and when it was created.

3.1 Introduction

Web of science (previously known as Web of Knowledge) is a bibliographic basis of scientific articles consisting of 12,500 scientific journals worldwide. Is a website which provides subscription-based access to multiple databases that provide comprehensive citation data for many different academic disciplines. It was originally produced by the Institute for Scientific Information (ISI) and is currently maintained by Clarivate Analytics (previously the Intellectual Property and Science business of Thomson Reuters) [11], [12].

According to web of science website [11]''Web of Science is the world's most trusted publisher-independent global citation database. Guided by the legacy of Dr Eugene Garfield, inventor of the world's first citation index, Web of Science is the most powerful research engine, delivering your library with best-in-class publication and citation data for confident discovery, access and assessment''.

The Web of Science Core Collection contains 34.200 journals indexed cover-to-cover,12M peer-reviewed full-text open access versions,12.5M records with funding data, 80M patents for over 40M inventions, 4,900 publishing partners, Backfiles to 1900 with cover-to-cover indexing 1.6B cited references [11].

3.2 History

The main factors that made up the web of science are three. Originally the first player emerged in the 1950s with the influx of government dollars into research and development after World War II, where the need for a better information management system was created. The second factor was the growing dissatisfaction with the ability to classify topics to cover the active researcher. The third and last factor was the need to implement machines for easy creation and data collection.

In 1955 Dr. Eugene founder and President of ISI (now Clarivate Analytics) was deeply involved in research on mechanical indicators. In 1960, he and his associates founded a database that included 5,000 patent applications for chemicals owned by two private pharmaceutical companies. Then in 1962 they needed to create a base on genetics, which led to the creation of three databases to cover the literature over a period of five and 14 years with a varied number of source publications adapted to each. Finally, in 1963 Eugene Garfield proceeded with the private publication of his multi-disciplinary citation index as the first issue of the Science Citation Index (SCI), which now covers 5,600 journals in more than 150 scientific specialties [11].

3.3 Data bases

The Web of Science Core Collection consists of six online databases [11], [12].

i. Science Citation Index Expanded (SCIE): created in 1964 as SCI and includes over 9,200 of the world's most impactful journals across 178 scientific disciplines. More than 53 million records and 1.18 billion cited references date back from 1900 to present.

ii. Social Sciences Citation Index (SSCI): contains over 3,400 journals across 58 social sciences disciplines, as well as selected items from 3,500 of the world's leading

scientific and technical journals. More than 9.37 million records and 122 million cited references date back from 1900 to present.

iii. Arts & Humanities Citation Index (AHCI): contains over 1,800 journals across 28 arts & humanities disciplines. More than 4.9 million records and 33.4 million cited references date back from 1975 to present.

iv. Emerging Sources Citation Index (ESCI): contains more than 3 million records and 74.4 million cited references date back from 2005 to present.

v. Conference Proceedings Citation Index (CPCI): contains over 205,900 conference proceedings, with 70 million cited references dating back from 1990 to present.

vi. Book Citation Index (BKCI): includes over 104,500 editorially selected books, with 10,000 new books added each year. Containing more than 53.2 million cited references, coverage dates back from 2005 to present.

4. The data

In this chapter we will describe the process we followed to get our data as well as the data fields.

4.1 Data types

We originally visited www.apps.webofknowledge.com [13] so we could find information about our magazine. The search is done by the name of the journal "Journal of Finance" in all fields. As a result we get 82,522 document results which is incorrect because we are returning records that do not belong to our magazine. So we have to limit these results based on our title in the source title field. With this intervention we finally get 6114 records and we have removed them in 2019 and we have reached 6037 records. To export the document, we have chosen to present all available records from 1970 to 2018 and then choose the method of export. We downloaded the data from web of science in bibtex format. BibTeX [14] is reference management software for formatting lists of references. The purpose of BibTeX is to make it easy to cite sources in a consistent manner, by separating bibliographic information from the presentation of this information.

The data was downloaded on the 28th of March ,2019 and was composed by 6037 observations and 37 variables and all the data between the years 1970 and 2018 are included. The columns are divided in 5 groups depending on the information that they contain. .The first group is the so-called " citation information" and consists information about the authors name, document title , year published ,source title ,volume ,issue and pages ,citation count , source and document type and DOI. The second group is the "bibliographic information" which contains affiliation ,serial identifiers(ISSN), pubMed ID , publisher , editors , language , address , abbreviated source title columns respectively. Following , abstract and keywords are present , which include abstract, author keywords and index keywords . "Finding details" is another category ,that likewise contains , number, acronym ,sponsor and funding text columns .Last but not least the final group contains important information divided by trade name ,accession number and conference information .

4.2 Field types

This document consists of 37 columns which we will see in the following table who they are and what they mean.

SYMBOL	COLUMN NAME	ENTIRE CONTENTS
AU	Author	First or Last Name of the authors
TI	Title	The title of the document
SO	Source	Publication name or source
JI	ISO source abbreviation	Source abbreviation
AB	Abstract	Summary of the subject of the publication
DE	Authors Keywords	Keywords used by authors
ID	Keywords associated or Keywords Plus	Keywords Plus
LA	Language	e.g. English
DT	Document Type	Type of the document e.g. article , note , review
DT2	Document Type 2	Only article
TC	Times Cited	Web of science core collection times cited count
CR	Cited References	Details about the cited references
C1	Authors address	Information about the University address
DI	Digital Object Identifier	Digital Object Identifier
PA	Publisher	Publisher Address
FU	Funding	Funding Agency and Grant Number
FX	Funding Text	Funding Text
SN	ISSN	International Standard Serial Number
PN	Issue	Issue
PP	Pages	Page Number
PU	Publisher	The publisher's name
VL	Volume	The volume of a journal
PY	Year	The year of publication
UT	Unique article identified	Unique Article Identified
NR	Cited References	Cited References (in wos core collection)

SC	**Subject Categories**	**Subject Categories**
U2	**Usage Count**	**Usage Count**
WC	**Web of Science Categories**	**Web of Science Categories**
EM	**Email Address**	**The e-mail address of the author(s)**
GA	**IDS Number**	**Identifies an issue of a journal. Used to order copies of articles from a document delivery service**
RP	**Reprint Address**	**The address of the reprint author. It may include reprint author, organization, sub organization, street, city, state or province, zip or postal code**
DB	**Bibliographic Database**	**Which bibliographic database is used**
AU_UN	**Authors University**	**Which is the authors university**
AU1_UN	**Authors University**	**Which is the authors university**
AU_UN_NR	**Authors University**	**Which is the authors university**
SR_FULL	**Author name , publication year , source**	**Author name , publication year , source**
SR	**Software Review**	**Contain information which distinguishes each paper**

Table 4.2 1: Explain columns

5. Duplicates

After collecting our data from our source Web of Science [13] our priority is to identify the duplicates in the 6037 observations and 37 variables. Duplicates data are entries that have been added by a system user multiple times.

5.1 Checking Types

In this section we will analyze the process we followed to check the validity of our data regarding duplicates. Initially, as first step, we checked if there are duplicates in our data and we discovered that there are many that we will check methodically. The audits should be based on different columns of our file, as well as a combination of some columns. the process we followed is to create two files in each control. The first file contains the actual duplicates, ie in how many lines and how many times they appear in our file. Is to end with a "clean" file without duplicates. So we did 5 checks which were based on:

- Digital Object Identifier (DOI),
- Tittle (TI),
- Volume (VL), Pages(PP), Issue(PN),
- Volume (VL), Issue (PN), Start Page,
- Volume (VL), Issue (PN), End Page,

and below we will see their results.

5.2 Digital Object Identifier (DOI)

The first duplicate test was based on the column DOI (Digital Object Identifier). As a result of this test we got 13 duplicates in 25 lines. The first eleven as we see in the table below are all BOOK REVIEW, we also see differences in the columns TI (Title), TC (Times Cited Count), CR(Cited References), RP (Reprint Address), AU_UN (Author University) etc. The last document as we see in our table has all the columns mentioned above (TI, TC, CR, RP, AU_UN), but here we see a difference in the DT (document type) column because one is an ARTICLE and the other EDITORIAL MATERIAL. The 13th entry is NA and does not appear in these 25 lines. Below we see in detail the table for each doi the differences that exist and in which columns as well as how many times each DOI appears.

DOI (Digital Object Identifier)	Rep	Difference
10.2307/2325788 10.2307/2325462 10.2307/2326885 10.2307/2326748	2	TI(tittle), CR (Cited References), C1 (Author Address), RP (Reprint Address), UT(Accession Number), AU_UN (Author University), AU1_UN (Author University), SR(Software Review)
10.2307/2325790 10.2307/2978662	2	TI(tittle), CR (Cited References),
10.2307/2327567	2	TI (tittle), C1 (Author Address), RP (Reprint Address), UT(Accession Number), AU UN (Author University), AU1_UN (Author University), SR (Software Review)
10.2307/2978660 10.2307/2326585 10.2307/2327078	2	TI (tittle), CR (Cited References), C1 (Author Address), RP (Reprint Address), UT(Accession Number),

		AU1_UN (Author University), SR(Software Review)
10.2307/2978663	**3**	**TI (tittle), CR (Cited References), UT (Accession Number), SR (Software Review)**
10.1111/0022-1082.00153	**2**	**TI (tittle), DT(Document type), UT (Accession Number), SR (Software Review)**

Table 5.2 1: Duplicates in case DOI

5.3 Title (TI)

The second duplicate test was based on the column TI (Tittle). In this case we have 109 duplicates that appear in 249 lines. So we see that some titles appear over 2 times so we have 13 titles that appear 3 times each and 3 titles appear 4 times each and refer to different papers .The 39 titles, although have the same authors but are mentioned in different papers. The 40 columns have all the other columns different and are different papers, while 13 titles have similar names, but they are also different. Finally we have 13 corrections that appear in 14 lines and they are different papers.

TITTLE	REP	DIFFERENCE
TEST OF PORTFOLIO BUILDING RULES - COMMENT	**2**	**Different article**
DEPOSIT COMPOSITION AND COMMERCIAL BANK EARNINGS	**2**	**Different article**
IS THERE AN OPTIMAL MONEY SUPPLY	**2**	**Different article**
LEVERAGE, DIVIDEND POLICY AND COST OF CAPITAL - COMMENT	**3**	**Different article**
PREMIUMS ON CONVERTIBLE BONDS - COMMENT	**4**	**Different article**
EFFECT OF FHLB BOND OPERATIONS ON SAVINGS INFLOWS AT SAVINGS AND LOAN ASSOCIATIONS - COMMENT	**2**	**Different article**
SESSION TOPIC - CAUSES AND PREDICTIONS OF RATES OF RETURN ON STOCKS AND BONDS - DISCUSSION	**2**	**Different article**
SESSION TOPIC - CAPITAL MANAGEMENT AND CAPITAL THEORY - DISCUSSION	**2**	**Different article**
SESSION TOPIC - MULTINATIONAL FIRM - BANE OR BOON - DISCUSSION	**3**	**Different article**
SESSION TOPIC - FEDERAL RESERVE SYSTEM - DISCUSSION	**4**	**Different article**
SESSION TOPIC - FINANCE AND BANKING - REFEREED PAPERS .2. - DISCUSSION	**3**	**Different article**

ANALYSIS OF LEASE-OR-BUY DECISION - COMMENT	**4**	**Different article**
MICROECONOMIC APPROACH TO BANKING COMPETITION - REPLY	**2**	**Different article**
CORRECTION	**14**	**Different article**
PREMIUMS ON CONVERTIBLE BONDS - REPLY	**2**	**Different article** **Same author**
EFFECT OF FHLB BOND OPERATIONS ON SAVINGS INFLOWS AT SAVINGS AND LOAN ASSOCIATIONS - REPLY	**2**	**Different article**
OPTIMAL FINANCING AND CAPITAL STRUCTURE PROGRAMS FOR FIRM - REPLY	**2**	**Different article**
METHODOLOGY OF TESTING FOR INDEPENDENCE IN FUTURES PRICES - REPLY	**2**	**Different article**
COMPANY CONTRIBUTIONS TO DISCRETIONARY PROFIT-SHARING PLANS - REPLY	**2**	**Different article**
GROWTH, CONSOLIDATION AND MERGERS IN BANKING - COMMENT	**2**	**Different article**
PAYMENTS IMPACT OF FOREIGN-INVESTMENT CONTROLS - REPLY	**2**	**Different article**
VALUATION, LEVERAGE AND COST OF CAPITAL IN CASE OF DEPRECIABLE ASSETS - COMMENT	**2**	**Different article**
WEIGHTED AVERAGE COST OF CAPITAL - SOME QUESTIONS ON ITS DEFINITION, INTERPRETATION, AND USE - COMMENT	**3**	**Different article**
TREASURY BILL AUCTION PROCEDURES - EMPIRICAL-INVESTIGATION - COMMENT	**2**	**Different article**
OPTIMAL LIFE-INSURANCE - COMMENT	**2**	**Different article**
INTER-TEMPORAL APPROACH TO OPTIMIZATION OF DIVIDEND POLICY WITH PREDETERMINED INVESTMENTS - COMMENT	**2**	**Different article**
FINANCIAL DISINTERMEDIATION IN A MACROECONOMIC FRAMEWORK - COMMENT	**2**	**Different article**
FINANCIAL-MARKETS AND BUSINESS FINANCE - DISCUSSION	**2**	**Different article**
STATE-OF-THE-ART STUDIES IN FINANCIAL THEORY - DISCUSSION	**2**	**Different article**
EMPIRICAL-RESEARCH ON CAPITAL-MARKETS - DISCUSSION	**2**	**Different article**
MONETARY-POLICY - ASSESSING THE BURNS YEARS - DISCUSSION	**3**	**Different article**
MULTIPERIOD FINANCIAL MODELS - DISCUSSION	**2**	**Different article**
EFFECT OF BOND REFUNDING ON SHAREHOLDER WEALTH - COMMENT	**2**	**Different article**
INFLATION ACCOUNTING, A GUIDE FOR ACCOUNTANT AND FINANCIAL	**2**	**Same author** **Different article**

ANALYST - DAVIDSON,S, STICKNEY,CP, WEIL,RL		
COMBINING FINANCIAL AND ACTUARIAL RISK - SIMULATION ANALYSIS - DISCUSSION	**2**	**Different article**
THE USE OF VOLATILITY MEASURES IN ASSESSING MARKET-EFFICIENCY - DISCUSSION	**2**	**Different article**
TREASURY BILL FACTORS AND COMMON-STOCK RETURNS - DISCUSSION	**2**	**Different article**
RATE-OF-RETURN REGULATION AND UTILITY CAPITAL STRUCTURE DECISIONS - DISCUSSION	**3**	**Different article**
CORPORATE EXCHANGE RISK MANAGEMENT - THEME AND ABERRATIONS - DISCUSSION	**3**	**Different article**
A COMPARISON OF ALTERNATIVE MODELS FOR PRICING GNMA MORTGAGE-BACKED SECURITIES - DISCUSSION	**2**	**Different article**
AN AGGREGATE MODEL OF THE CREDIT UNION INDUSTRY - DISCUSSION	**3**	**Different article**
ESSENTIALS OF FINANCE - JONES,RG, DUDLEY,D	**2**	**Different article**
CAPITAL-BUDGETING TECHNIQUES - WILKES,FM	**2**	**Different article**
MULTINATIONAL BUSINESS FINANCE - EITEMAN,DK, STONEHILL,AI	**2**	**Different article**
TESTING AN AGGRESSIVE INVESTMENT STRATEGY USING VALUE LINE RANKS - A COMMENT	**2**	**Different article**
TENDER OFFERS AND MANAGEMENT RESISTANCE - DISCUSSION	**2**	**Different article**
THE DESIGN OF A COMPANYS BANKING SYSTEM - DISCUSSION	**2**	**Different article**
INFLATION RISK AND REGULATORY LAG - DISCUSSION	**2**	**Different article**
ON THE DETERMINANTS OF NET FOREIGN-INVESTMENT - DISCUSSION	**3**	**Different article**
VALUATION OF SAFE HARBOR TAX BENEFIT TRANSFER LEASES - DISCUSSION	**3**	**Different article**
STOCHASTIC-PROCESSES FOR INTEREST-RATES AND EQUILIBRIUM BOND PRICES - DISCUSSION	**3**	**Different article**
YIELD APPROXIMATIONS - A HISTORICAL-PERSPECTIVE	**2**	**Different article**
MEAN-VARIANCE VERSUS DIRECT UTILITY MAXIMIZATION - A COMMENT	**2**	**Different article**
INVESTMENTS - JACOB,NL, PETTIT,RR	**2**	**Different article**
FORWARD AND FUTURES PRICES - EVIDENCE FROM THE FOREIGN-EXCHANGE MARKETS	**2**	**Different article**

EXECUTIVE CAREERS AND COMPENSATION SURROUNDING TAKEOVER BIDS	**2**	**Same author Different article**
VOLUME AND AUTOCOVARIANCES IN SHORT-HORIZON INDIVIDUAL SECURITY RETURNS	**2**	**Different article**
ARBITRAGE CHAINS	**2**	**Same author Different article**
A NONPARAMETRIC APPROACH TO PRICING AND HEDGING DERIVATIVE SECURITIES VIA LEARNING NETWORKS	**2**	**Same author Different article**
RATIONAL PREPAYMENTS AND THE VALUATION OF COLLATERALIZED MORTGAGE OBLIGATIONS	**2**	**Different article**
THE IMPACT OF PUBLIC INFORMATION ON THE STOCK-MARKET	**2**	**Same author Different article**
FINANCIAL DISTRESS AND CORPORATE PERFORMANCE	**2**	**Same author Different article**
IMPLIED BINOMIAL TREES	**2**	**Same author Different article**
OPTION VALUATION WITH SYSTEMATIC STOCHASTIC VOLATILITY	**2**	**Same author Different article**
INFLUENCE COSTS AND CAPITAL STRUCTURE	**2**	**Same author Different article**
SECURITY DESIGN	**2**	**Same author Different article**
THE INVESTMENT PERFORMANCE OF UNITED-STATES EQUITY PENSION FUND MANAGERS - AN EMPIRICAL-INVESTIGATION	**2**	**Same author Different article**
INVISIBLE PARAMETERS IN OPTION PRICES	**2**	**Same author Different article**
TOP-MANAGEMENT COMPENSATION AND CAPITAL STRUCTURE	**2**	**Same author Different article**
THE EFFECT OF MARKET-SEGMENTATION AND ILLIQUIDITY ON ASSET PRICES - EVIDENCE FROM EXCHANGE LISTINGS	**2**	**Same author Different article**
MARKET INTEGRATION AND PRICE EXECUTION FOR NYSE-LISTED SECURITIES	**2**	**Same author Different article**
INTERACTIONS OF CORPORATE FINANCING AND INVESTMENT DECISIONS - A DYNAMIC FRAMEWORK	**2**	**Same author Different article**
MEASURING ASSET VALUES FOR CASH SETTLEMENT IN DERIVATIVE MARKETS - HEDONIC REPEATED-MEASURES INDEXES AND PERPETUAL FUTURES	**2**	**Same author Different article**

TESTING THE PREDICTIVE POWER OF DIVIDEND YIELDS	**2**	**Same author Different article**
A NEW METHOD OF PORTFOLIO PERFORMANCE MEASUREMENT.	**2**	**Different article**
SPECIAL REPO RATES	**2**	**Same author Different article**
BACKWARDATION IN OIL FUTURES MARKETS - THEORY AND EMPIRICAL-EVIDENCE	**2**	**Same author Different article**
THE VALUATION OF CASH FLOW FORECASTS - AN EMPIRICAL-ANALYSIS	**3**	**Same author Different article**
SURVIVAL	**2**	**Same author Different article**
PERFORMANCE CHANGES FOLLOWING TOP MANAGEMENT DISMISSALS	**2**	**Same author Different article**
DYNAMIC ASSET ALLOCATION AND THE INFORMATIONAL EFFICIENCY OF MARKETS	**2**	**Same author Different article**
ONE SECURITY, MANY MARKETS - DETERMINING THE CONTRIBUTIONS TO PRICE DISCOVERY	**2**	**Same author Different article**
OIL AND THE STOCK MARKETS	**2**	**Same author Different article**
MANAGERS OF FINANCIALLY DISTRESSED FIRMS - VILLAINS OR SCAPEGOATS	**2**	**Same author Different article**
EX-DAY BEHAVIOR - TAX OR SHORT-TERM TRADING EFFECTS	**2**	**Same author Different article**
A SIMPLE APPROACH TO VALUING RISKY FIXED AND FLOATING RATE DEBT	**2**	**Same author Different article**
TIME-VARYING WORLD MARKET INTEGRATION	**2**	**Same author Different article**
PRICE REACTIONS TO DIVIDEND INITIATIONS AND OMISSIONS - OVERREACTION OR DRIFT	**2**	**Same author Different article**
THE MATURITY STRUCTURE OF CORPORATE-DEBT	**2**	**Same author Different article**
LATTICE MODELS FOR PRICING AMERICAN INTEREST-RATE CLAIMS	**2**	**Same author Different article**
THE PERFORMANCE OF HEDGE FUNDS: RISK, RETURN, AND INCENTIVES	**2**	**Same author Different article**
HETEROGENEOUS INFORMATION ARRIVALS AND RETURN VOLATILITY DYNAMICS: UNCOVERING THE LONG-RUN IN HIGH FREQUENCY RETURNS	**2**	**Same author Different article**

HOW COSTLY IS FINANCIAL (NOT ECONOMIC) DISTRESS? EVIDENCE FROM HIGHLY LEVERAGED TRANSACTIONS THAT BECAME DISTRESSED	**2**	**Same author Different article**
MERGING MARKETS	**2**	**Different article**
DO CHANGES IN DIVIDENDS SIGNAL THE FUTURE OR THE PAST?	**2**	**Different article**
INTERNATIONAL PORTFOLIO INVESTMENT FLOWS	**2**	**Different article**
DIVIDENDS, ASYMMETRIC INFORMATION, AND AGENCY CONFLICTS: EVIDENCE FROM A COMPARISON OF THE DIVIDEND POLICIES OF JAPANESE AND US FIRMS	**2**	**Same author Different article**
MARKET SEGMENTATION AND STOCK PRICES: EVIDENCE FROM AN EMERGING MARKET	**2**	**Same author Different article**
STOCK MARKET EFFICIENCY AND ECONOMIC EFFICIENCY: IS THERE A CONNECTION?	**2**	**Same author Different article**
DEBT, LEASES, TAXES, AND THE ENDOGENEITY OF CORPORATE TAX STATUS	**2**	**Same author Different article**
TAX INCENTIVES TO HEDGE	**2**	**Same author Different article**
LEGAL DETERMINANTS OF EXTERNAL FINANCE	**2**	**Different article**
SYMPOSIUM ON PUBLIC POLICY ISSUES IN FINANCE	**2**	**Same author Different article**
ARE INVESTORS RELUCTANT TO REALIZE THEIR LOSSES?	**2**	**Same author Different article**
ASSESSING SPECIFICATION ERRORS IN STOCHASTIC DISCOUNT FACTOR MODELS	**2**	**Same author Different article**
AGENCY PROBLEMS, EQUITY OWNERSHIP, AND CORPORATE DIVERSIFICATION	**2**	**Same author Different article**
QUOTES, ORDER FLOW, AND PRICE DISCOVERY	**2**	**Same author Different article**
UNTITLED	**2**	**Different article**

Table 5.3 1: Duplicates in case TI

5.4 Volume (VL), Issue (PN), Pages (PP)

The third test we got the combination of 3 columns volume, issue, pages where we found 38 records in 79 rows. In this case we have 49 book review,28 meeting abstract and 2 corrections. In the table below we see that we have 21 duplicates that while they have same Volume, Issue, Pages and the same authors refer to different articles, as well as 17 duplicates while have in common the Volume, Issue, Pages are different articles.

PP/VL/PN	REP	DIFFERENCE
PP: 986-988 VL:25 PN:4	2	Same author, Different tittle But different articles
PP: 981-982 VL: 36 PN: 4	2	Same author, Different tittle But different articles
PP: 980 VL: 50 PN: 3	2	different articles
PP: 977 VL: 50 PN: 3	2	different articles
PP: 974 VL: 50 PN: 3	2	different articles
PP: 926-927 VL: 30 PN: 3	2	Same author, Different tittle But different articles
PP: 826-827 VL: 26 PN: 3	2	Same author, Different tittle But different articles
PP: 818-819 VL: 26 PN: 3	2	Same author, Different tittle But different articles
PP: 814-816 VL: 26 PN: 3	2	Same author, Different tittle But different articles
PP: 787-789 VL: 28 PN: 3	3	Same author, Different tittle But different articles
PP: 785-787 VL: 28 PN: 3	2	Same author, Different tittle But different articles
PP: 781-784 VL: 28 PN: 3	2	Same author, Different tittle But different articles
PP: 678-680 VL: 33 PN: 2	2	Same author, Different tittle But different articles
PP: 60-61 VL: 2 PN: 1	4	Same author, Different tittle But different articles

PP: 58-59 **VL: 2** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 56-57 **VL: 2** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 421-426 **VL: 54** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 289-290 **VL: 29** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 287-289 **VL: 29** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 238-240 **VL: 26** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 230-233 **VL: 26** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 182-188 **VL: 25** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 166-168 **VL: 27** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 1659 **VL: 47** **PN: 4**	**2**	**Different corrections**
PP: 1659 **VL: 49** **PN: 3**	**2**	**Different articles**
PP: 1063 **VL: 49** **PN: 3**	**2**	**Different articles**
PP: 1068 **VL: 49** **PN: 3**	**2**	**Different articles**
PP: 1068 **VL: 51** **PN: 3**	**2**	**Different articles**
PP: 1070-1071 **VL: 49** **PN: 3**	**2**	**Different articles**
PP: 1077 **VL: 49** **PN: 3**	**2**	**Different articles**
PP: 974 **VL: 50** **PN: 3**	**2**	**Different articles**
PP:977 **VL: 50** **PN: 3**	**2**	**Different articles**

PP: 980 **VL: 50** **PN: 3**	**2**	**Different articles**
PP:421-426 **VL: 54** **PN: 1**	**2**	**Same author,** **Different tittle** **But different articles**
PP: 1214 **VL: 52** **PN: 3**	**2**	**Different articles**
PP: 1226 **VL: 52** **PN: 3**	**2**	**Different articles**
PP: 1239 **VL: 52** **PN: 3**	**2**	**Different articles**
PP: 1245 **VL: 52** **PN: 3**	**2**	**Different articles**
PP: 1256 **VL: 52** **PN: 3**	**2**	**Different articles**

Table 5.4 1: Duplicates in case Volume, ISSUE, Pages

5.5 Volume (VL), Issue (PN), Start Page

We split the pages into 2 columns where we named them the startpage and the end page and we continued the same procedure with the duplicates.

In the first case with the VL(volume), PN(issue), START(start page), we got 166 records in 375 rows. We have 71 book reviews. The 21 of these 23 book reviews appear in 2 lines the remaining 2 in three and six lines respectively. Finally, we still have 49 Book reviews, 3 articles, 1 article progressing paper, 2 correction addition and 1 editorial material that according to our data but also with our magazine are different papers. In the following table we will see some examples and what are the differences in our duplicates.

PP/VL/START	REP	TITLE	DIFFERENCE
START: 60 **VL: 2** **PN: 1**	**4**	**APPRAISING CAPITAL WORKS - BROSTER,EJ** **STATISTICAL SAMPLING FOR ACCOUNTING INFORMATION - CYERT,RM AND DAVIDSON,HJ** **TRANSPORT FINANCE AND ACCOUNTING - LEE,GA** **HANDBOOK OF SAMPLING FOR AUDITING AND ACCOUNTING - ARKIN,H**	**Same author, Different title But different articles**
START: 907 **VL: 27** **PN: 4**	**2**	**STATE OF FINANCE FIELD - FURTHER COMMENT**	**Different Title Different Author**

		REINVESTMENT ASSUMPTIONS IN CHOOSING BETWEEN NET PRESENT VALUE AND INTERNAL RATE OF RETURN	
START: 907 VL: 27 PN: 4	**2**	**CORRECTION**	**Same Title, Different Author, But different articles**
START: 1501 VL: 54 PN: 4	**2**	**PANEL ON GLOBAL FINANCIAL MARKETS AND PUBLIC POLICY FED POLICY, FINANCIAL MARKET EFFICIENCY, AND CAPITAL FLOWS**	**Different Title Different Author**

Table 5.5.1 Duplicates in case Volume, ISSUE, Start Page

5.6 Volume (VL), Issue (PN), End Page

In the second case with the VL(volume), PN(issue), END(end page), we get as a result 60 records in 390 rows. We noticed that there were several mistakes in the dataset, as the 323 lines were empty in the pages column. And from these 60 observations we see that on 30 there are NA so we conclude that it is not a valid criterion to find our duplicates in data.

PP/VL/END	REP	TITLE	DIFFERENCE
END: 61 VL: 2 PN: 1	**4**	**APPRAISING CAPITAL WORKS - BROSTER,EJ TRANSPORT FINANCE AND ACCOUNTING - LEE,GA STATISTICAL SAMPLING FOR ACCOUNTING INFORMATION - CYERT,RM AND DAVIDSON,HJ HANDBOOK OF SAMPLING FOR AUDITING AND ACCOUNTING - ARKIN,H**	**Same author, Different title But different articles**
END: NA VL: 2 PN: 1	**3**	**QUANTITATIVE FRAMEWORK FOR FINANCIAL MANAGEMENT - PETERSON,DE CORPORATE FINANCIAL MANAGEMENT - KENT,RP MANAGEMENT PLANNING AND CONTROL - DEVERALL,CS**	**Different Title Different Author**

Table 5.6 1 Duplicates in case Volume, ISSUE, END Page

6. Missing Value

6.1 Introduction

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data [15] .Missing data are defined as values that are not available and that would be meaningful if they are observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. Most datasets in the real world contain missing data.

6.2 Missing Data

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions .The best way to prepare for missing values is to understand the data you have: understand how the missing values are represented, how the data was collected, where the values are not supposed to be missing, and which are used specifically to represent the absence of data.

First, we need to understand what types of data are missing. Missingness is generally categorized into 3 categories [16]:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

Missing Completely at Random (MCAR): If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. It is safe to ignore many of the complexities that arise because of the missing data, apart from the obvious loss of information. Most simple fixes only work under the restrictive and often unrealistic MCAR assumption [17].

Missing at Random (MAR): If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). It is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption [17].

Missing not at Random (MNAR): If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios [17].

6.3 Methods for handling missing data

Below we will analyze seven ways we can use for missing data [18]:

Listwise deletion (or complete case analysis): Delete all data from any participant with missing values. If a case has missing data for any of the variables, then simply exclude that case from the analysis.

Recover the Values: You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we've found having an additional check for missing values before the participant leaves helps.

Educated Guessing: It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix.

Average Imputation: Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.

Common-Point Imputation: use the middle point or most commonly chosen value. This is a bit more structured than guessing, but it's still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.

Regression Substitution: You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator.

Multiple Imputation: The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation , software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions.

6.4 Checking missing values

We initially measured how much is missing in each column as in some column combinations for both sets of data. We first see that in our data almost all the variables in the DE (Author Keywords) column are missing, as 6037 values are missing 6034. We also noticed that many values are missing from the FU columns (Funding Agency and Grant Number) and FX (Funding Text), as there are no entries 5905 and 5911 respectively. Continuing, we saw that the AU_UN column (Author University) has 108331 missing entries, the C1 column (Address Writers) 240, AB (Abstract) 3978, EM (Email Address) 5931 and CR (Reported Reference) 709. Summarizing, in the RP column (print address) there are no 240 entries, in DOI (Digital Object Identifier) 1006, in the total entries START (Start Page) 1 and END (End Page) 351.

MISSING VALUES	SUMMATION
AU	**0**
TI	**0**
SO	**0**
JI	**0**
AB	**3978**
DE	**6034**
ID	**3936**

LA	**0**
DT	**0**
DT2	**0**
TC	**0**
CR	**709**
C1	**240**
DI	**1006**
PA	**0**
RP	**240**
FU	**5905**
FX	**5911**
SN	**0**
PN	**0**
PP	**1**
UT	**0**
NR	**0**
PU	**0**
SC	**0**
U2	**0**
WC	**0**
EM	**5931**
GA	**0**
DB	**0**
VL	**0**
PY	**0**
AU_UN	**1831**
AU_UN1	**0**
SR_FULL	**0**
SR	**0**
START PAGE	**1**
END PAGE	**351**

Table 6 1: Missing value for each column

Finally we combine the columns DOI-END, AU-PN-START-END-VL, AU-END, AU-END-START-VL, PN-VL to thoroughly analyze which records are missing. In the DOI-END combination gave us an effect of 274 missing records and in the other combinations we did not have missing values.

MISSING VALUES	
DOI-END	**274**

AU-PN-START-END-VL	**0**
AU-END	**0**
AU-END-START-VL	**0**
PN-VL	**0**

Table 6 2 1: Missing Value for combine columns

7. Merging the data sets

In this section, the papers from two databases, Web of Science and Scopus, will be merged, where they have been cleared of duplicates. This union should be done with some criteria so that we can reach a final "clean" file. Initially, the compounds must be based on the following criteria:

- DOI
- TI (Title)
- PN (Issue) ,VL (Volume), START PAGE
- PN (Issue), VL (Volume), END PAGE

7.1 DOI(Digital Object Identifier)

The first criterion is DOI (Digital Object Identifier). In this case we have to combine the data sets from these two data sources Scopus and Web of Science which have common DOI column. In order to succeed this process, first we have to exclude from the original Scopus and Web of Science files the duplicate documents based on DOI. After removing the duplicates from original Scopus with 4942 observations and 27 variables and original Web of Science with 6037 observations and 37 variables, we dropped to files with 4935 lines and 27 columns for Scopus and 5006 lines and 37 columns for Web of Science. The union of these two files with DOI criteria gives us a final file with 1798 lines and 66 columns of common data from our two sources.

7.2 TI(Title)

The second criterion is TI (title). In this case the data which has not been merged with the DOI criterion, have to be merged based on TI. From Scopus based on DOI have not joined 3137 observations and 27 variables and from Web of Science 3208 observations and 37 variables. Before the union with the Title criterion, we have to exclude the duplicate based on the TI from those that did not unite. From the exclusion of duplicate documents we have 2798 rows and 27 columns from Scopus and 3083 rows and 37 columns for Web of Science. So taking these data sets from the 2 sources we proceeded to the union based on the title and we have as a final result with 527 lines and 66 columns common elements.

7.3 PN (Issue), VL(Volume), START PAGE

Our next criterion is PN/ VL/ START (issue, volume, start page). We follow the same procedure again. First of all we have to find how many papers were not joined with the previous criterion of the TI (title). Those that were not united based on the title from the source Scopus are 2271 observations and 27 variables and from the Web of Science source 2556 observations and 37 variables. So after the deduction of the duplicates from these files based on PN /VL /START, we continue with the union of the same criterion. Thus, excluding the duplicates documents, we have as a result for the source Scopus 2270 observations and 27 variables and from Web of Science source the 2540 observations and 37 variables. Finally, making their union based on PN /VL /START we end up with a final file with 1261 common papers and 66 columns.

7.4 PN (Issue), VL(Volume), END PAGE

Our last criterion is PN /VL /END (issue, volume, end page). We have found how many papers not joined by the previous step (PN /VL /START), in this case we joined them with PN /VL / END. Those which were not joined by the PN /VL /START, from the Scopus source are 1007 observations and 26 variables while from the Web of Science source are 1235 observations and 36 variables. We continue with the same procedure removing the duplicates based on the criterion PN /VL /END and then with the union based on this. After removing the duplicates from Scopus we have 1006 observations and 26 variables and from Web of Science source

1234 observations and 36 variables. Finally, making their union based on PN /VL /END we end up with a final file with 1 common papers and 63 columns.

8. Total registry

Having completed the whole merging process, our next activity is to merge all the above files (criteria) into one total file. This merging of the four data sets gives us a result of 3587 observations and 66 variables.

8.1 Comparisons and differences between Sources

This merging of the files above will help us to identify both the errors and the differences between our two different sources, Scopus and Web of Science.

The columns, which we will examine the differences of our sources are the following: DOI (Digital Object Identifier), TI (Title), AU (Author), PN (Issue), VL (Volume), START (Start page), END (End page). Starting from the DOI column we see that the differences are 1789 observations. We continued with the column TI , where we have to deal with 1434 differences (where either it may be due to a spelling mistake or in a different spelling). In the column AU we identified 335 differences. Then we noticed that in the columns PN , VL there are no differences as all the elements of the columns are the same from our two sources. Finally in the columns START we have 10 differences due to different entries in the column END we have 101 differences. In the following table we will see a summary table with the differences that exist in each criterion

DIFFERENCES	RESULTS
DOI	**1789**
TI	**1434**
AU	**335**
PN	**0**
VL	**0**
START	**10**
END	**101**

Table 8.11 : The differences between Scopus and Web of Science

8.2 Differences

Initially starting with DOI (Digital Object Identifier). As we have mentioned the differences of criterion according to our sources of Web of Science and Scopus is 1789. In the following table we present the differences that exist between the two sources. Where in the first column of the table named “DI.X” is the Doi according to the Scopus source, in the second column named “DI.Y” is the Doi according to the Web of Science source and the third column shows us the differences.

As we can see from our DI.X and DI.Y columns the doi column is assigned differently from the 2 databases. that's why we end up with such a large result of different Doi.

DI.X	DI.Y	Differences
10.1111/J.1540-6261.1980.TB02203.X	10.2307/2327093	different entries
10.1111/J.1540-6261.1990.TB03739.X	10.2307/2328761	different entries
10.1111/J.1540-6261.1982.TB03618.X	10.2307/2327850	different entries
10.1111/J.1540-6261.1980.TB03517.X	10.2307/2327217	different entries

Table 8.21:Differences in relation to DOI column

Then we will see the differences of our sources in relation to the column TI (title). In this case we have 1434 differences in our sources and in the table below we see where TI.X the titles from the source Scopus, where TI.Y the titles from the source Web of Science and where differences we will describe the differences that exist between our two sources.

In this case we see in the first line of our table that there are typographical errors as from the source scopus we see that the title is separated by ":" and from the source web of science with "-". In the second line we see that from the source scopus has normally been assigned a title while from Web of Science it has been assigned as Untitled. In the third line we see that we have two different entries, ie different titles. Finally we see that some titles are written differently as in the last example which from the source scopus is "the two" and from web of science "2".

TI.X	TI.Y	DIFFERENCES
EXPECTATIONS, TOBIN'S Q, AND INVESTMENT: A NOTE	EXPECTATIONS, TOBIN Q, AND INVESTMENT - A NOTE	Misprint
LETTER FROM THE NEW EDITOR	UNTITLED	UNTITLED
REPLIES	WEIGHTED AVERAGE COST OF CAPITAL - SOME QUESTIONS ON ITS DEFINITION, INTERPRETATION, AND USE - REPLY	different entries
THE TWO FACES OF BOND REFUNDING: REPLY	2 FACES OF BOND REFUNDING - REPLY	Written Differently

Table 8.22:Differences in relation to TI column

Then we will see the differences in relation to the AU(Author). The differences from our two sources are 335. In the third table we present in the first column with the name “AU.X” the authors from the source Scopus, in the second column with the name “AU.Y” the authors from Web of Science and the third column with the name “Differences” we see the differences about the two sources.

In the case of the authors (AU) we see that here too there are several differences between our sources. As we see in the first line of our table the Scopus source has been entered as the author's name “NA NA” ie the author's name is not available in relation to the Web of Science source as we see that it has been entered as “SINGLETON KJ”. In the second line we see that there are typographical errors, as the Scopus source seems to be missing a letter since in the author's name we see the name “CUSTDIO C” while in the Web of Science source we see the name “CUSTODIO C”. In the third line of the table we see that the Scopus source has a name “ELTON EJ;GRUBER MJ” while the Web of Science source has four authors' names “ELTON EJ;GRUBER MJ;GUPTA MK;HAMADA RS;PINCHES GE”. Finally we see the most common difference of our sources because we see different author entries.

AU.X	**AU.Y**	**Differences**
NA NA	**SINGLETON KJ**	**NA**
CUSTDIO C	**CUSTODIO C**	**Misprint**
ELTON EJ;GRUBER MJ	**ELTON EJ;GRUBER MJ;GUPTA MK;HAMADA RS;PINCHES GE**	**More authors**
ECONOMICS TECPFB	**COPELAND TE**	**different entries**

Table 8.23 : Differences in relation to AU column

We follow the same procedure examining the START(Start page) column. We have identified 10 differences in the START column. In the following table we follow the same philosophy where in the first column named “START PAGE.X” we have the initial pages from the Scopus source, in the second column named “START PAGE.Y” the initial pages from the Web of Science source and in the third column with the name “Differences” we have the differences from our two sources. We notice that all our differences are due to different page entries from our sources. However, according to our magazine, the correct pages are those of the Scopus source (START PAGE.X).

START PAGE.X	START PAGE.Y	Differences
1365	1364	different entries
639	637	different entries
599	598	different entries
1269	1268	different entries

Table 8.24: Differences in relation to START column

Finally, the table below concerns the differences between the two sources in relation to the END column (End page). In this case there are 101 differences. So we see in the first column named "END PAGE.X" the final pages based on scopus, "END PAGE.Y" the final pages of web of science and in the third column the various between them. The 101 differences concern different entries as as we see they have differences from 1 to 3 pages usually. Finally, we have 11 SE values that are missing and concern the source of Web of Science and 3 more SE where they exist in both of our sources.

END PAGE.X	END PAGE.Y	Differences
1846	1845	different entries
638	636	different entries
471	NA	NA
NA	NA	NA

Table 8.25 : Differences in relation to END column

8.3 Similarities

In the following table we will examine the similarities of our two sources Scopus and Web of Science.

SIMILARITIES	RESULTS
DOI	1789
TI	2153
AU	3252
PN	3587
VL	3587
START	3579
END	3497

Table 8.3 1: The similarities between Scopus and Web of Science

Conclusions

So we conclude that there are several problems in our Web of Science database as no matter how many criteria we have chosen as much as we have excluded our duplicate registrations there are still several registrations which have not been merged with any criteria from our two databases.

So we see that there are both typographical errors and different entries in all the criteria we examined from our bases. We also see differences in papers, comments, etc. where one database includes them (Web of Science) and the other not equal and the big difference of the data since in web of science we have 6037 registrations and in Scopus 4942.

Nevertheless, we end up with a file that was as clean as possible with 3587 observations, ie with a file where less than half are common to our two databases.

REFERENCES

[1] S.Papavlasopoulos, "kallipos.gr," 2015. [Online]. Available: https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/4755/1/00_master_document.pdf#page=37&zoom=100,52,618. [Accessed 12 1 2020].

[2] W. Glanzel , bibliometrics as a research field, US & UK: Course Handouts, 2003.

[3] "shodhganga.inflibnet," [Online]. Available: https://shodhganga.inflibnet.ac.in/bitstream/10603/5109/10/10_chapter%201.pdf. [Accessed 7 April 2020].

[4] Wikipedia, "Wikipedia" [Online]. Available: https://en.wikipedia.org/wiki/Bibliometrics?fbclid=IwAR3wst1LlO9p7mYcWGUWcO5Rsu_0L9M4aB8773NX1xDQsEmQgRs6yU2lJYw. [Accessed 18 12 2019].

[5] T. Č. Ivan Župič, "Bibliometric methods in management and organization," 2003.

[6] Γ. Γεωργάκη, "Βιβλιομετρική ανάλυση του επιστημονικού περιεχομένου του προγράμματος σπουδών Περιβαλλοντικός Σχεδιασμός," 25 September 2018. [Online]. [Accessed 5 February 2020].

[7] J. W. Sons, "Wiley online library," 1999. [Online]. Available: https://onlinelibrary.wiley.com/journal/15406261. [Accessed 02 December 2019].

[8] S. Nagel, "The American Finance Association," December 1939. [Online]. Available: https://afajof.org/journal-of-finance/. [Accessed 02 December 2019].

[9] I. Wikimedia Foundation, "Wikipedia," 23 November 2019. [Online]. Available: https://en.wikipedia.org/wiki/The_Journal_of_Finance. [Accessed December 12 2019].

[10] ITHAKA, "JSTOR," 2002. [Online]. Available: https://www.jstor.org/journal/jfinance. [Accessed 02 December 2019].

[11] W.O.Science, "Clarivate," [Online]. Available: https://clarivate.com/webofsciencegroup/solutions/web-of-science/. [Accessed 8 December 2019].

[12] Wikipedia, "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Web_of_Science. [Accessed 8 December 2019].

[13] C.Analysis, "WebOfKnowledge," [Online]. Available: www.apps.webofknowledge.com. [Accessed 2019 December 8].

[14] Wikipedia, "wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/BibTeX. [Accessed 2019 December 9].

[15] S.Statistics, "Statistics Solutions," [Online]. Available: https://www.statisticssolutions.com/missing-values-in-data/. [Accessed 13 July 2020].

[16] B.Roy, "towardsdatascience," 3 September 2019. [Online]. Available: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184. [Accessed 14 July 2020].

[17] V.Silaparasetty, "medium," [Online]. Available: https://medium.com/@vinitasilaparasetty/guide-to-handling-missing-values-in-data-science-37d62edbfdc1. [Accessed September 8 2020].

[18] J.Sauro, "measuringu," 2 June 2015. [Online]. Available: https://measuringu.com/handle-missing-data/. [Accessed 8 September 2020].