**ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ**

**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΔΙΟΙΚΗΣΗΣ**

**ΤΜΗΜΑΜΗΧΑΝΙΚΩΝΟΙΚΟΝΟΜΙΑΣΚΑΙΔΙΟΙΚΗΣΗΣ**

# Analysis of problems in the bibliometric data of the journal "Journal of Finance" from the Scopus database.

**ΕΛΕΝΗ ΜΑΥΡΙΔΗ ΠΡΙΝΤΕΖΗ**
**Α.Μ: 2312013070**

*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**

θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές στην παρούσα διπλωματική εργασία Βασιλείου Ευάγγελο και Δριβαλιάρη Δημοσθένη για την βοήθεια τους.

Επίσης θέλω να ευχαριστήσω τηνοικογένεια μου και όλους τους φίλους μου για την στήριξη και την υπομονή τους.

## ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια υπάρχει μεγάλο ενδιαφέρον στον κλάδο της ανάλυσης δεδομένων που αφορούν δεδομένα από επιστημονικά περιοδικά. Βασική πηγή των δεδομένων στην εργασία είναι το Scopus , όμως στην συνέχεια συμπεριλαμβάνονται και δεδομένα από μία δεύτερη πηγή το Web of Science . Η παρούσα διπλωματική εργασία αφορά δεδομένα από το επιστημονικό περιοδικό Journal of Finance. Για να διερευνηθούν τα προβλήματα και τα σφάλματα που υπάρχουν στα δεδομένα ,αρχικά επεξεργαζόμαστε τις διπλοεγγραφές και τα missingvalues στις πήγες των δεδομένων. Η ένωση των δεδομένων σε ένα κοινό όσο των δυνατών γίνεται πιο καθαρό αρχείο είναι ο απώτερος σκοπός. Η καταγραφή των βημάτων που ακολουθήσαμε και των προβλημάτων που αντιμετωπίσαμε αναλύονται διεξοδικά.

## ABSTRACT

In recent years there has been great interest in the field of data analysis involving data from scientific journals. The main source of data in the work is scopus, but then data from a second source, the web of science, is included. This dissertation deals with data from the scientific journal Journal of Finance. In order to investigate the problems and errors in the data, we first process the duplicates and missing values in the data sources. The ultimate goal is to merge the data into an audience as clean as possible. The recording of the steps we followed and the problems we encountered are analyzed in detail.

## Contents

## List of tables

# 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].

## 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 (i.e. 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 travellers” 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.[4, 5].

## 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 analyse 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, is 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) 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 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[8].

## 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 |
| 1949-1976 | 4 | 4-22 |
| 1968-1997 | 5 | 23-53 |
| 1998-2018 | 6 | 54-73 |

**Table 2.3.a**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. Data Source – Scopus

This chapter analyzes the source we have used to load the data, as well as the types of data that belong to it.

## 3.1.Overview

Scopus[11]is a bibliographic database, which is chronologically dated from November 2004 until today. More specifically, it belongs to Elsevier[12]. Elsevier is a public company, which was founded in 1880. This public company plays an integral role in progression of knowledge, as it is a modern-day distinguished, an academic publishing company which is related to scientific publications and a global analytics company specialized in science and health. Scopus is the largest database of abstract bibliographic references, including smart tools for monitoring, analysing and visualizing different types of research. In addition, it also includes abstracts and citations for academic journal articles. Scopus covers a wider range of journals and offers author's information profiles that cover collaborations, editions and bibliographic data, as well as reports and details on the number of reports issued by published documents.

Undoubtedly, with 22,800 titles from more than 5,000 international publishers, Scopus [13]offers the most comprehensive picture of the global research in many scientific fields including science, technology, medicine, social science, the arts and human science. To visualize the width of Scopus it is useful to present some numerical information, mentioning that it includes over 21,950 scientific journals, 280 trade publications, over 560 book series, Over 8 million conferences, more than 150,000 books with 20,000 added each year, over 69 million entries, 62.4+ million records. After 1969 the number of references was more than 6.6 million records in less than a year. Moreover, the oldest record dates back in 1788, which was a patent and since then more than 39 million patents were registrations from five patent offices.

Scopus supports researchers and librarians in three key areas:

1. Search (Search by document, author or affiliation)

2. Discover (Find related documents by shared references, authors and / or keywords)

3. Analyze (Track citations over time for a set of authors or documents with Citation Overview / Tracker)[14]

## 3.2.Document Types covered in Scopus

The following table shows the data types in detail[14]:

| Document types | Definition |
|---|---|
| Article | Original research or opinion. |

| | |
|---|---|
| | Articles in peer-reviewed journals are usually several pages in length, most often subdivided into sections: abstract, introduction, materials & methods, results, conclusions, discussion and references. However, case reports, technical and research notes and short communications are also considered to be articles and may be as short as one page in length. Articles in trade journals are typically shorter than in peer-reviewed journals, and may also be as brief as one page in length. |
| Book | A whole monograph or entire book<br>Book type is assigned to the whole. Additionally, for books with individual chapters,<br>each chapter, along with a general item summarizing the book, is also indexed with the source<br>type Book |
| Chapter | A book chapter.<br>Complete chapter in a book or book series volume where the item is identified as a chapter by a heading or section indicator |
| Conference paper | Original article reporting data presented at a conference or symposium.<br>Conference papers are of any length reporting data from a conference, with the exception of conference abstracts. Conference papers may range in length and content from full papers and published conference summaries to short items as short as one page in length |
| Editorial | Summary of several articles or provides editorial opinions or news.<br>Editorials are typically identified as editorial, introduction, leading article, preface or foreword, and are usually listed at the beginning of the table of contents |
| Erratum | Report of an error, correction or retraction of a previously published paper<br>Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided. |
| Letter | Letter to or correspondence with the editor. Letters are individual letters or replies. Each individual letter or reply is processed as a single item. |
| Note | Note, discussion or commentary.<br>Notes are short items that are not readily suited to other item types. They may or may not share characteristics of other item types, such as author, affiliation and references. Discussions and commentaries that follow an article are defined as notes and considered to be items in their own right. Notes also include questions and answers, as well as comments on other (often translated) articles. In trade journals, notes are generally shorter than half a page in length |
| Review | Significant review of original research also includes conference papers.<br>Reviews typically have an extensive bibliography. Educational items that review specific issues within the literature are also considered to be reviews. As non-original articles, reviews lack the most typical sections of original articles such as materials & methods and results |

| | |
|---|---|
| Short survey | Short or mini-review of original research.<br>Short surveys are similar to reviews, but usually are shorter (not more than a few pages) and with a less extensive bibliography |

**Table 3.2.a**Document types category

# 4. Data

The data comes from the economic journal which called “The Journal of Finance”. The process of obtaining them and the field types of the data concerning them are presented below.

## 4.1.Download Process

The data present in this study have been downloaded from www.scopus.com[11]. To find references in a magazine, we use the tool “search for documents” entering the journal’s name into the search box, namely “Journal of Finance”. The search will match all fields of the document record. The result will be sorted by Date. To limit the effects, we selected the refine results field in the source title field of the Journal of Finance and displayed 5,019 document results. The following step was to subtract from the field year "the year 2019,2020eport of an error, correction or retraction of a previously published paper

Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided. eport of an error, correction or retraction of a previously published paper

Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided. eport of an error, correction or retraction of a previously published paper

Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided. eport of an error, correction or retraction of a previously published paper

Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided. and we had finally 4,942 documents as a result. Before exporting the document, we chose to present all available records from 1946 to 2018 and then choose the method of export. We downloaded the data from Scopus in bibtexformat.Bibtex[15]is report management software for formatting report lists. Its purpose is to facilitate the reporting of sources by separating bibliographic information from the presentation of such information. The files are in .bib format and contain the database with the list of entries that the user wants to use. Each entry corresponds to a bibliographic record.

The data that was downloaded on the 28th of March 2019 contains 4942 entries which represent the rows and divided into 27 fields which are the columns and all the data covering the years from 1946 to 2018. Based on the export document setting according to the scopus website, the columns are divided in 5 separate groups depending on the information that they contain. The first group is the so-called "citation information" and consists of information about the authors name, document title, year published, source title, volume, issue and pages, citation count, source and document type and DOI (Digital Object Identifier). The second group is the "bibliographic information" which contains affiliations, serial identifiers (ISSN), PubMed ID, publisher, editors, language of the original document, corresponding

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 other important information divided by trade names ,accession numbers and chemicals , conference information .To all these groups the references was included .This document contain 27 columns which are Author names(AU), Title(TI), Publication Name (SO) , Abstract (AB) , Authors Keywords(DE) , Language (LA) , Document Type (DT) , Document Type 2(DT2), Scopus Collection Times Cited Count (TC) , Cited References (CT) , Author Address (C1) ,Digital Object Identifier (DOI) ,Reprint Address (RP) ,Funding Agency and Grant Number(FU) , International Standard Serial Number(SN), Part Number (PN) , Page Number (PP),Publisher (PU) ,Bibliographic Database (DB) , Volume (VL) , Year Published (PY) ,Author University(AU_UN) , Author1 University(AU1_UN) , AU_UN_NR , SR,SR_FULL.

## 4.2.Field Types

This table consists of three columns. In the first column with name ‘”symbol” all the abbreviations are present, in the second one their full names are illustrated and lastly, in the third column a more detail explanation for each name is given.

| SYMBOL | FULL 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 |
| AB | Abstract | summary of the subject of the publication |
| DE | Keywords | keywords used by authors |
| 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 count | Scopus core collection times cited count |
| CR | Cited references | Details about the cited references |
| C1 | Author address | Information about the University address |
| DI | DOI | Digital object identifier |
| RP | Reprint address | Authors name and university name |
| FU | Funding agency and grant number | Funding information for the search document |
| SN | ISSN | International standard serial number |
| PN | Part number | Issue |

| | | |
|---|---|---|
| PU | Publisher | The publisher's name |
| DB | Database | Bibliographic database |
| VL | Volume | The volume of a journal |
| PY | Publication year | The year of publication |
| AU_UN | Author university | Name of the university |
| AU1_UN | Author university | Name of the university |
| AU_UN_NR | Empty column | |
| SR_FULL | Au , year , ji | Author name , publication year , source |
| SR | Row names | Author name , publication year , source |
| PP | Pages | One or more page numbers or range of numbers separated by - |

**Table 4.2.a**Field Type

# 5. Duplicates

Duplicate data are entries that have been added by a system user more than ones. Duplicate records are in the same rows in a dataset. This means that for a pair of, duplicate records, the value in each row coincide at all levels. That means, we were talking about identical rows in the dataset. Databases do not allow duplicates, for this reason we used a variety of methods to eliminate them to the greatest extent, eventually resulting in a "clean" database. The “cleaning” of the data base may not be fully achieved, that means, there may be some duplicates that we need, depending on the data they contain.

## 5.1.Checking Types

The duplicates were checked in the database in 5 different stages. That means, we did 5 separate checks to obtain a final file without duplicates. More in detail, we used the following methodology: we initially checked for duplicates in our database and instantly we noticed that there were double values, based on some columns. The logic we follow was to maintain the original file and control the duplicates based on a specific column in the table to do corrections. Every duplicate that was found was kept in a different file and afterwards was checked for weather it is indeed duplicated or weather it has any values that are duplicated by mistake. We afterwards placed the remaining data in another archive. We continue with the control based on the next column and again create 2 archives, that means, we follow the same procedure for all 5 levels of controls. Finally we end up with a database, which is our pure archive, and we have the rest of the archives from each audit stage containing the duplicates from each audit.

The five controls are based on the columns:

- Document type
- Title
- Volume – Issue – Pages
- Volume – Issue – Start Page
- Volume – Issue – End Page.

### 5.1.1. DOI

The first stage of the checks begins with the DOI column. This column contains information about the Digital object identifier. A Digital Object Identifier is a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web. A DOI will help the reader easily locate a document from the citation. In the dataset we have two observations, the one was that the first record has a missing value and the other is displayed in 2 lines. Also only one duplicate appears at this stage twice. The reason that occurs twice is the different values in the TC, SR columns. The information in this TC column are different, as every database entry has its unique times cited count. So two different observations with two different TC are considered in this matrix. The SR differs only in the information

concerning the source of the article, the title of the journal. On the other hand, our content derives from the same journal, so we are not talking about different data.

In the next table is illustrated the observation, which appears more than one time:

| DOI |
|---|
| Na |
| 10.1111/J.1540-6261.1990.TB02426.X |

**Table 5.1.a**DOI column entry

Following are present the SR, TC column of the dataset:

| DOI | TC | SR |
|---|---|---|
| 10.1111/J.1540-6261.1990.TB02426.X | 1021 | CARTER R ,1990, J FINANC |
| 10.1111/J.1540-6261.1990.TB02426.X | 79 | CARTER R, 1990, J FINANC-a |

**Table 5.1.b**SR, TC entries

### 5.1.2. Title

The next control step concerns the TI column, which contains the titles from the document included in our journal. At this stage there are 30 titles featuring duplicates. This result is a bit complicated in explaining, that's why we will split it on the basis of each individual title, amounts of the times that appears each. This information is displayed in the table below. These 30 titles are analyzed in 350 rows. The first column contained the title of the article, the second the times the title appears in our archive and the third column the analysis of the similarities or differences between the titles.

| Document Title | No Rep | Difference |
|---|---|---|
| Discussion | 217 | Different articles |
| | | Author name missing : 23 |
| The effect of bond refunding of shareholder wealth : comment | 2 | Different articles |
| Reply | 58 | Different articles |
| Growth , consolidation and mergers in banking : comment | 2 | |
| Valuation, leverage and the cost of capital in the case of depreciable assets: comment | 2 | |
| Errata | 3 | |
| The weighted average cost of capital: some question on its definition, interpretation , and use : comment | 2 | |
| Optimal life insurance: comment | 2 | |
| An intertemporal approach to the optimization of dividend policy with predetermined investments : comment | 3 | |
| Analysis of the leaseorbuy decision : comment | 4 | |
| Financial disintermediation in a macroeconomic framework: comment | 2 | |
| Portfolio returns and the random walk theory: comment | 2 | |
| The effect of FHLB bond operation of | 3 | |

| | | |
|---|---|---|
| savings inflows at savings and loan associations: comment | | |
| Notice to the membership of the American finance association | 2 | Only the same Author |
| | | Different articles |
| A reply | 9 | |
| Test of portfolio building rules : comment | 2 | |
| Premiums on convertible bonds: comment | 4 | |
| Leverage, dividend policy and the cost of capital : a comment | 2 | |
| Erratum | 4 | |
| Is the federal reserve system really necessary: a comment | 2 | |
| Compensatory cyclical bank asset adjustments : comment | 2 | |
| The theoretical value of a stock right: a comment | 2 | |
| Report of the executive secretary and treasurer | 3 | Only the same author |
| | | Different articles |
| Minutes of the annual membership meeting | 4 | Two author names is missing |
| | | Different articles |
| A rejoinder | 2 | |
| Forward and futures prices: evidence from the foreign exchange markets | 2 | |
| Initial public offerings and underwriter reputation | 2 | Same articles |
| | | Difference between TC, CR columns |
| Editorial announcement | 2 | |
| Introduction | 2 | |
| Mean variance versus direct utility maximization : a comment | 2 | |

**Table 5.1.c**Duplicate entries

According to the entries in the table above, we draw some conclusions about duplicates at this stage of control. Only one archive has all the information itself except for the TC and SR columns, which, as we have described above, follow a different logic. In other words, we're talking about the only real double-entry. The rest refer to different articles.

### 5.1.3. Volume, Issue, Page

The combination of three different columns is the next step in testing. The PN column containing information about the issue of the articles and is combined with the VL column which containing the data volume information, and finally the two above are combined with the PP column, which contains the article page listing. From this combination we have as result two duplicates, which appear in four rows. The first duplicate is “real”, as we are talking about the same article, but the TC and SR columns are different so it appears as a different entry. We have also highlighted in a

previous section. The next duplicate refers to two different articles, which have many entries in common. However, according to the journal verification we concluded that they are two different articles.

The next table represents the first duplicate:

| PN | VL | PP | TI | TC | SR |
|---|---|---|---|---|---|
| 4 | 45 | 1045-1067 | Initial public offerings and underwriter reputation | 1021 | CarterR,1990,J-FINANC |
| 4 | 45 | 1045-1067 | Initial public offerings and underwriter reputation | 79 | Carter R, 1990, J FINANC-a |

**Table 5.1.d**First duplicate entry

Following, the second duplicate are illustrated:

| PN | VL | PP | AU | TI |
|---|---|---|---|---|
| 4 | 69 | 1845-1846 | Singleton KJ | REPORT OF THE EDITOR OF THE JOURNAL OF FINANCE FOR THE YEAR 2013 |
| 4 | 69 | 1845-1846 | Schalljeim | REPORT OF THE EXECUTIVE SECRETARY AND TREASURER: FOR THE FISCAL YEAR ENDING SEPTEMBER 30, 2013 REPORT OF THE EXECUTIVE SECRETARY AND TREASURER |

**Table 5.1.e**The second duplicate entry

Afterwards we divided the column containing the pages into two separate columns. By the division the first derived column included, the pages from which the article begins and the second the page where it ends.

### 5.1.4. Volume, Issue, Start Page

In this stage, two double articles appear. The common elements of which are many, but they differ in the following columns. The first entry is actually a duplicateentrie and the columns where the differences are displayed are few. In the second record the only common elements is the combination of these three columns.

| PN | VL | STARTPAGE | TC | SR |
|---|---|---|---|---|
| 4 | 45 | 1045 | 79 | CARTER R, 1990, J FINANC-a |
| 4 | 45 | 1045 | 1021 | CARTER R, 1990, J FINANC |

**Table 5.1.f**First duplicate entry and the different columns information

Following, the second duplicate are illustrated:

| PN | VL | STARTPAGE | TITLE | AUTHOR |
|---|---|---|---|---|
| 4 | 69 | 1845 | REPORT OF THE EDITOR OF THE JOURNAL OF FINANCE FOR THE YEAR 2013 | SINGLETON KJ |
| 4 | 69 | 1845 | REPORT OF THE EXECUTIVE SECRETARY AND TREASURER: FOR THE FISCAL YEAR ENDING SEPTEMBER 30, 2013 REPORT OF | SCHALLHEIM J |

| | | | | |
|---|---|---|---|---|
| | | | THE EXECUTIVE SECRETARY AND TREASURER | |

**Table 5.1.g**Second duplicate entry

### 5.1.5. Volume, Issue, End Page

The last stage of the control is based on the combination of the end page, the volume and the issue. From this stage, six duplicates emerged, appearing in 12 rows. Only the one duplicate is "real", as we are talking about the same article, but the TC and SR columns are different, so it appears as a different entry. All the others duplicate entries even if they have the same 3 columns, they are not the same article, as none of their other elements are the same. To sum up, we came to the conclusion that we have ten different articles and one real duplicate article.

The next table represents the "real" duplicate:

| VL | PN | ENDPAGE | TITLE | SR | TC |
|---|---|---|---|---|---|
| 45 | 4 | 1067 | INITIAL PUBLIC OFFERINGS AND UNDERWRITER REPUTATION | CARTER R, 1990, J FINANC | 1021 |
| 45 | 4 | 1067 | INITIAL PUBLIC OFFERINGS AND UNDERWRITER REPUTATION | CARTER R, 1990, J FINANC-a | 79 |

**Table 5.1.h**The real duplicate entry

Following, the next table presents the other duplication:

| VL | PN | ENDPAGE | TITLE | AUTHOR |
|---|---|---|---|---|
| 14 | 1 | 77 | REPLY | WEISS NC |
| 14 | 1 | 77 | A FURTHER NOTE ON TIME DEPOSIT INTEREST RATES | MORRISON GR |
| 16 | 1 | 51 | CONCENTRATION IN INSTITUTIONAL COMMONSTOCK PORTFOLIOS | MILLER NC |
| 16 | 1 | 51 | ERRATUM | NA NA |
| 40 | 3 | 756 | INDEX OPTIONS: THE EARLY EVIDENCE | ENVUNE J, RUDD A |
| 40 | 3 | 756 | DISCUSSION | MACBETH JD |
| 60 | 2 | 839 | LIFTING THE VEIL: AN ANALYSIS OF PRE-TRADE TRANSPARENCY AT THE NYSE | DVOK T |
| 60 | 2 | 839 | DO DOMESTIC INVESTORS HAVE AN INFORMATION ADVANTAGE? EVIDENCE FROM INDONESIA | BOEHMER E, SAAR G, YU L |
| 69 | 4 | 1846 | REPORT OF THE EDITOR OF THE JOURNAL OF FINANCE FOR THE YEAR 2013 | SCHALLHEIM J |

| | | | | |
|---|---|---|---|---|
| 69 | 4 | 1846 | REPORT OF THE EXECUTIVE SECRETARY AND TREASURER: FOR THE FISCAL YEAR ENDING SEPTEMBER 30, 2013 REPORT OF THE EXECUTIVE SECRETARY AND TREASURER | SINGLETON KJ |

**Table 5.1.i**The other duplicated entries

# 6. Missing Values

Once the duplicate records are found, we continue the analysis by finding the Missing values in the two different databases. A missing value is one whose value is unknown. Missing values are represented in R by the “NA” symbol as not availableor “NaN”(not a number) as impossible values (e.g. dividing by zero). NA is a special value whose properties are different from other values. In this chapter, we will first analyze the meaning and the types of Missing Data. Next we will presentfinally the role that NA play in our own databases[16].

## 6.1.The meaning of the missing data and the three types

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.

To find what to do with the missing values in our data set, we must first understand what kind of missing data we have. There are three kinds of missing data, Missing at Random (MAR), Missing Completely at Random (MCAR), Missing Not at Random (MNAR).“Missing Completely at Random” and “Missing at Random” are both considered ‘ignorable’ because we don’t have to include any information about the missing data itself when we deal with the missing data.MNAR is called “non-ignorable” because the missing data mechanism itself has to be modeled as you deal with the missing data. You have to include some model for why the data are missing and what the likely values are.

**Missing at Random (MAR):** means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data. What it means, is that the missingness of data can be predicted by other features in the dataset. The missing values in any feature are dependent on the values of other features.

**Missing Completely at Random (MCAR):** It is the highest level of randomness. MCARmeans there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others. Also with these data we can undertake analyses using only observations that have complete data (provided we have enough of such observations).The MCARassumption is rarely a good assumption. It is only likely to be true in situations where the data is missing due to some truly random phenomena.

**Missing not at Random (MNAR):**These kinds of data are the most complicated one both in terms of finding it and dealing with it. MNAR, means there is a relationship between the propensity of a value to be missing and its values the fact that the data is missing is related to the unobserved data, i.e. the data that we don’t

have, the missingness is related to factors that we didn't account for. The easiest way to understand why the data is missing is to understand the data collection process. Two possible reasons are that the missing value depends on the hypothetical value or the missing value depends on the value of another variable.

**MARvsMNAR:** The only true way to distinguish between MNAR and MAR is to measure some of that missing data. If their responses on the key items differ by very much, that's good evidence that the data are MNAR.However in most missing data situations, we don't have the luxury of getting a hold of the missing data. So while we can't test it directly, we can examine patterns in the data get an idea of what's the most likely mechanism[17].

## 6.2.Missing data in the datasets

The missing data are the "blank" entries in a file. More specifically, the missing values are displayed when no data value is stored for the variable in an observation. A missing data has a significant effect on the conclusions that can be drawn from the data. A Missing data may be due to a non-response: no information is provided on one or more objects or an entire unit ("subject").

After checking and clearing the data from duplicates, we examined the counting and recording of the missing values in all our data. In total, we have 24085 missing prices. We examined the NA in more detail in each column separately, as well as different columns in combination. From the previous, we noticed that from the AU_UN_NR column (information about the author's university) all the entries are missing, so we are talking about a blank column. Similarly, from the DE (author keywords) column 4939 entries were missing. By checking the missing values from FU (text funding) and PU (publisher), we realised that there are lacking 4924 and 4161 entries from each column, respectively. Continuing, we noticed that AU_UN (university editors) column lacks 3115, C1 column (address editors) 3032, AB (summary) 2103, and column CR (mentioned above) 1004 entries. Last but not least, by observing the data set it realised, that in RP (address) column there are missing 802 entries, as well as 5 registrations in DI (DOI).

More analytically, the missing prices for each column separately, as well as for their combinations, are illustrated in the table below:

| COLUMN | NUMBER OF NA'S VALUE |
|---|---|
| AU | 0 |
| TI | 0 |
| SO | 0 |
| JI | 0 |
| AB | 2103 |
| DE | 4393 |

| | |
|---|---|
| LA | 0 |
| DT | 0 |
| DT2 | 0 |
| TC | 0 |
| CR | 1004 |
| C1 | 3032 |
| DI | 5 |
| RP | 802 |
| FU | 4924 |
| SN | 0 |
| PN | 0 |
| PP | 0 |
| PU | 4161 |
| DB | 0 |
| VL | 0 |
| PY | 0 |
| AU_UN | 3115 |
| AU_UN_NR | 4942 |
| SR_FULL | 0 |
| SR | 0 |
| STARTPAGE | 7 |
| ENDPAGE | 9 |
| DOI–ENDPAGE | 0 |
| AU-PN-STARTPAGE-ENDPAGE-VL | 0 |
| AU-END-STARTPAGE-VL | 0 |
| AU-ENDPAGE | 0 |
| PN-VL | 0 |

**Table 6.1.a**Missing Values based on the specific columns

| SYMBOL | CATEGORY OF MISSING VALUE |
|---|---|
| AB | MCAR |
| CR | MCAR |
| DE | MCAR |
| C1 | MCAR |
| D1 | MCAR |
| RP | MCAR |
| AU_UN | MCAR |

| | |
|---|---|
| AU_UN_NR | MAR |
| PU | MCAR |
| FU | MAR |
| START PAGE | MCAR |
| END PAGE | MCAR |

**Table 6.1.b**The category of NA

# 7. Merging the datasets

In this chapter, we will deal with the junction of data from the two databases, since we have first excluded duplicates, as we saw in the previous chapter. After finding the "clean" datasets using the four criteria, we will proceed to find the errors or the differences in the assigned entries between the two databases.

The four criteria are:

- DOI
- Title
- Volume, Issue, Start Page
- Volume, Issue, End Page

## 7.1.Merged based on specific column and collaboration

### 7.1.1. DOI

The procedure we followed was to separate the duplicate records from the original databases based on some different criteria, in order to create files with the “unique” records at each step. Then we created different files for each of the four different criteria, containing the data from both databases, Scopus and Web of Science. The logic we followed was that those data which were not combined with the first criterion, after we find them first, we have to combine them with the second criterion and continue the same logic for the other two criteria. The first criterion we started the process is the DOI (Digital object identifier). The records that have a single DOI are in Web of Science 5006 observations with 38 columns and in Scopus there are 4935 observations with 28 columns. The total file we created consists of 1798 rows and 66 columns.

### 7.1.2. Title

The data that were not joined based on DOI, we have to combine them with the second criterion, the Title. First we found that the data where were not joined in the previous step based on the DI, were 3208 observations with 38 columns from the Web of Science and from Scopus we have as result 3137 observations with 28 columns. After that, we have joined these assignments based on the Title and we have as a result 3038 observations and 2798 for Web of Science, Scopus respectively. Finally, the final file we have created with the common entries from the two databases contains 527 entries with 66 columns.

### 7.1.3. Volume, Issue, Start Page

For the third criterion, we have to combine three variables for the union of the data that were not joined in the previous step. These variables are volume, issue, start page. After finding, 2556 for Web of Science and 2271 for Scopus, assignments that were not united based on the Title, we continue with their union with the above three criteria. As a result we have 2540 registrations and 2270 registrations for Web of Science and Scopus respectively. The total file at this stage consists of 1261 observations with 66 columns.

### 7.1.4. Volume, Issue, End Page

Last but not least, we have the last criterion for finding a cleaner file. The combination of columns Volume (VL), Issue (PN), End page (ENDPAGE). For the Web of Science, 1279 entries were not joined from the previous step, while in Scopus 1009observations. So we continued with their combination with the three criteria and we got as a result two files with 1235 entries from Web of Science and 1007 entries from Scopus. The total file shows us only 1 entry with 66 columns.

## 7.2.Total registry

In this section, all the above criteria will be merged as a total file, so that the differences or even the errors that exist between the two sources, Web of Science and Scopus, can be found and compared below. The combination of the above four separate files as one, gives us as a result a dataset with 3587 observation with 66 columns.

| CRITERIA | UNIQUE WOS | UNIQUE SCOPUS | TOTAL FILE | NOT JOIN WOS | NOT JOIN SCOPUS |
|---|---|---|---|---|---|
| DI | 5006 | 4935 | **1798** | 3208 | 3137 |
| TI | 3083 | 2798 | **527** | 2556 | 2271 |
| SP,VL, PN | 2540 | 2270 | **1261** | 1279 | 1009 |
| EP,VL, PN | 1235 | 1007 | **1** | 1234 | 1006 |

**Table 7.2.a**Total table

## 7.3.Comparisons and Differences between Sources

The two sources Scopus and Web of Science present some differences in terms of records in the assigned data. These differences are due either to erroneous entries or to some spelling errors or to a different wording of the same values. Therefore, the purpose is to identify these errors.In the following, regarding the incorrect recording of the names of the authors or the assignment of a different author, 335 different author names was presented. Volume and Issue have no difference. Start page has only 10 different entries in the two databases, while end page has 101different entries.these differences are analyzed below depending on each different variable.

### 7.3.1. Comparison between DI

In the DOI column the differences are 1789 and are mainly due to a differententity of the digital object identifier, the source Web of Science presents all these different assignments. The format you compile Doi follows the series 10.1111 / J.1540-6261.1981.TB01075.X based on scopus while the web of science in some observations has assigned Doi in the form 10.2307 / 2327299. These remarks concern the same paper, only Doi's syntax differs so he presents them as different. The following table lists some of these data as an example.

| DI.x | DI.y | TI | AU | V L | P N |
|---|---|---|---|---|---|
| 10.2307/2327067 | 10.1111/J.1540-6261.1979.TB03455.X | RISK AND RETURN | GEHRAK | 34 | 1 |
| 10.2307/2327524 | 10.1111/J.1540-6261.1981.TB00650.X | THE WEEKEND EURODOLLAR GAME | COATS WL | 36 | 3 |
| 10.1111/J.1540-6261.1996.TB04073.X | 10.2307/2329398 | EQUILIBRIUM ANALYSIS OF PORTFOLIO INSURANCE | GROSSMAN SJ;ZHOU Z | 51 | 4 |

**Table 7.3.a**Different entries between Doi

### 7.3.2. Comparison between TI

The next variable covered by the analysis is the title of the paper. Our final file has 2153 common titles but has 1434 different. These differences are due to typographical errors, which result from either spelling errors or unnecessary space or extra words.

Wos in moredetail:

- uses words instead of some symbols. A typical example is the% symbol described in wos as PERCENT.
- .Unexpected gaps between words and punctuation create differences between the two sources.
- Another difference is that scopus leaves spaces between some words as opposed to wos which joins them using the hyphen (-).
- To the titles used the symbol &wos mistakenly adds the / symbol.

More details for scopus:

- Does not separate some words with a space between them
- Skips the question marks at the end of some titles, this phenomenon has not been observed often
- For the abbreviation reference it uses the dot as a point of separation of the initial letters, while wos either writes the whole word or the two initials are stuck. For example, in scopus the initials for the United States are referred to as U.S. while in wos either in full or as US
- To separate the title, from notes or other features, the (:) is used while in wos the hyphen (-). This is often done in comments, note etc

The following table shows some examples with these errors from both sources.

| Error category | TI from Scopus | TI from Wos |
|---|---|---|
| punctuation | LUCAS IN THE LABORATORY | ``LUCAS'' IN THE LABORATORY |

| | | |
|---|---|---|
| Separation with (.) | WHY DO FOREIGN FIRMS LEAVE U.S. EQUITY MARKETS? | WHY DO FOREIGN FIRMS LEAVE US EQUITY MARKETS? |
| % | 100% MARGINS: COMBATING SPECULATION IN INDIVIDUAL SECURITY ISSUES | 100 PERCENT MARGINS - COMBATING SPECULATION IN INDIVIDUAL SECURITY ISSUES |
| Different wording | TWO NOTES ON THE UNIQUENESS OF COMMERCIAL BANKS | 2 NOTES ON UNIQUENESS OF COMMERCIAL BANKS |
| (: )instead (-) | A GENERAL DIVERSIFICATION THEOREM: A NOTE | A GENERAL DIVERSIFICATION THEOREM- A NOTE |
| /& | A NONLINEAR FACTOR ANALYSIS OF S&P 500 INDEX OPTION RETURNS | A NONLINEAR FACTOR ANALYSIS OF S\&P 500 INDEX OPTION RETURNS |
| Separation of words | NEWISSUE STOCK PRICE BEHAVIOR | NEW-ISSUE STOCK PRICE BEHAVIOR |
| Lack of articles, words | THE MEASUREMENT OF THE VOL<br>ATILITY OF COMMON STOCK PRICES | MEASUREMENT OF VOLATILITY OF COMMON STOCK PRICES |

**Table 7.3.b**Difference between TI

### 7.3.3. Comparison between Authors

The authors cover an important part of the study to compare the similarity of injections from the two bases. The common authors are 3252 while the different ones are only 335. Their differences are due to different factors, there are also typographical errors, omissions of the names of some authors, as well as a confusing order of their recording and some repetitions. Allthesedifferencesarelistedbelow..

- Typographical errors of names that mainly concern omissions of middle names or spaces between the names, or different typing of the same names
- Repeat the last name more than once
- Completely different assigned names
- Lack of a author
- The source scopus indicates authors who do not exist as NA, while the source wos as Anonymus, both sources express the same thing
- In wos there are some above assigned authors due to an error in their assignment
- The order of assignments of multiple authors differs
- The wos font includes special characters in vowels and letters, while in scopus they are omitted (Ä, Ş, Ľ)

The following table shows some of these errors

| Error Categor | C oun | AU scopus | AU wos | Wright |
|---|---|---|---|---|

| y | t | | | |
|---|---|---|---|---|
| Typographical errors | 2870 | VAN HORNE JC | VANHORNE JC | VANHORNE JC |
| Repeat the last name | 5 | MUKHERJEE RN | MUKHERJE RN MUKHERJERN | MUKHERJE RN |
| lack of the author name | 1 | SERDAR DINCI;EREL I | DINCIS;EREL I | SERDAR DINCI;EREL I |
| Lack of the author | 4 | SOLEDAD M;PERIAM;SCHMUKLER SL | PERIAMSM;SCHMUKLER SL | SOLEDAD M;PERIAM;SCHMUKLER SL |
| NA NA, Anomymus | 2 | NA NA | ANONYMOUS A | NA |
| Special characters | 32 | PSTORL;VERONESI P | PASTOR L;VERONESI P | PĂSTORL;VERONESI P |
| Error in author | 4 | NA NA | SINGLETON KJ | NANA |
| More authors | 1 | ELTON EJ;GRUBER MJ | ELTON EJ;GRUBERMJ;GUPTAMK;HAMADARS;PINCHES GE | ELTON EJ;GRUBER MJ; |

**Table 7.3.c**Difference between Authors

### 7.3.4. Comparison between Start Page

The differences between the home pages are only 10 and the similarities are 3579. 8 of these differences are simply due to incorrectly assigned values from one of the two sources. The 2 assigned values are NA. The following table shows these differences between the two sources and the 3rd column contains the source that has the correct pages based on the official site of the magazine.

| STARTPAGE.x | STARTPAGE.y | CORRECT |
|---|---|---|
| 1365 | 1364 | scopus |
| 1845 | 1844 | scopus |
| 639 | 637 | scopus |
| 599 | 598 | scopus |
| 1269 | 1268 | scopus |
| 802 | 803 | wos |
| 574 | 575 | scopus |
| 1037 | 1039 | wos |
| NA | NA | v |

| | | |
|---|---|---|
| 1307 | NA | scopus |

**Table 7.3.d**Difference between Start pages

### 7.3.5. Comparison between End pages

Last but not least, the comparison between end pages has as result 101 different pages and 3497 κοινεςεγγραφές. 90 of these differences are due to a typographical error in the assigned values from one of the two sources. The remaining 11 values are NA. The table below will present the assignments and the 3rd column will contain the correct value. We noticed that out of the 90 values, the 63 values have only one page difference and all the assignments of scopus are correct. The remaining 27 values are listed below along with the NA assignments.

| ENDPAGE.x | ENDPAGE.y | CORRECT |
|---|---|---|
| 2348 | NA | Scopus |
| NA | NA | 1862 |
| NA | NA | Viii |
| 1329 | NA | Scopus |
| NA | NA | 1305 |
| 1563 | NA | Scopus |
| 807 | NA | Scopus |
| 1461 | NA | Scopus |
| 1362 | NA | Scopus |
| 1367 | NA | Scopus |
| 471 | NA | Scopus |
| 638 | 636 | Scopus |
| 62 | 92 | Wos |
| 839 | 812 | Scopus |
| 1047 | 1074 | Wos |
| 187 | 178 | Wos |
| 1033 | 1034 | Scopus |
| 539 | 549 | Scopus |
| 729 | 735 | Scopus |
| 717 | 719 | Scopus |
| 733 | 735 | Scopus |
| 675 | 679 | Scopus |
| 775 | 777 | Wos |
| 630 | 639 | Scopus |
| 226 | 227 | Scopus |
| 483 | 487 | Scopus |
| 308 | 318 | Scopus |

| | | |
|---|---|---|
| 740 | 752 | Scopus |
| 602 | 609 | Scopus |
| 1222 | 1022 | Wos |
| 95 | 96 | Scopus |
| 1121 | 1142 | Scopus |
| 1066 | 1075 | Scopus |
| 315 | 327 | Scopus |
| 633 | 637 | Scopus |
| 656 | 660 | Scopus |
| 697 | 699 | Scopus |
| 346 | 364 | Scopus |

**Table 7.3.e**Differecne between End Pages

# 8. Conclusion

## 8.1.Differences and Errors in the merged file

The complete file containing the data from the two separate data sources Scopus and Web of Science consists of 3587 observations and 74 columns. Our goal is to find the differences in the records between the two sources. We used five columns as a basis for this comparison, the names of the authors, the Titles, the DOI, the start and end pages. The names of the authors differ either in spelling mistakes, or in a different way of writing the abbreviations of the first names, mainly and more rarely they differ in their last names. The titles in turn differ most often in the support points (for example -, /). Digital object identifier differs many times, because the way of listing in the two sources is different, while they are mentioned in the same paper. The pages are usually due to a typo to record their correct number, at this point it should be noted that the end pages have more differences than the start ones. We compared these comparisons in the following tables.

The tables have the comparisons of the five variables depending on the similarities and differences, always based on one variable.

The first table records for the 1789 different Doi we found comparing the two sources, how many differences they have in the other columns. We always consider as logic a second basic column where we find the differences and then we alternate the other 3 to find the other differences.

| Second given column | No of difference | TITLE | AUTHOR | START PAGE | END PAGE |
|---|---|---|---|---|---|
| AU | 135 | 102 | | 0 | 8 |
| TI | 1262 | | 102 | 1 | 33 |
| START | 3 | 1 | 0 | | 0 |
| END | 40 | 33 | 8 | 0 | |

**Table 8.1.a**The differences between the different Doi and the other four columns

The second table show us the 1798 similar Doi and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | TITLE | AUTHOR | START PAGE | END PAGE |
|---|---|---|---|---|---|
| AU | 200 | 36 | | 0 | 3 |
| TI | 172 | | 36 | 1 | 4 |
| START | 1 | 1 | 0 | | 1 |
| END | 50 | 4 | 3 | 2 | |

**Table 8.1.b**The difference between the similar Doi and the other four columns

The third table show us the 1434different Title and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | AUTHOR | START PAGE | END PAGE |
|---|---|---|---|---|---|
| AU | 138 | 102 | | 0 | 7 |
| DOI | 1262 | | 102 | 1 | 33 |
| START | 3 | 3 | 0 | | 2 |
| END | 90 | 40 | 11 | 2 | |

**Table 8.1.c**The difference between the different Title and the other four columns

The fourth table show us the 2153similar Titles and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | AUTHOR | START PAGE | END PAGE |
|---|---|---|---|---|---|
| AU | 197 | 33 | | 0 | 4 |
| DOI | 527 | | 33 | 2 | 7 |
| START | 6 | 2 | 0 | | 1 |
| END | 53 | 7 | 4 | 1 | |

**Table 8.1.d**The difference between the similar Titles and the other four columns

The next table show us the 335 different Authors names and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | START PAGE | END PAGE |
|---|---|---|---|---|---|
| DOI | 135 | | 102 | 0 | 8 |
| TI | 138 | 102 | | 0 | 7 |
| START | 0 | 0 | 0 | | 0 |
| END | 11 | 8 | 7 | 0 | |

**Table 8.1.e**The difference between the different AU and the other four columns

The sixth table show us the 3252 similar Authors name and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | START PAGE | END PAGE |
|---|---|---|---|---|---|
| DOI | 1654 | | 1160 | 3 | 32 |

| | | | | | |
|---|---|---|---|---|---|
| TI | 1296 | 1160 | | 2 | 30 |
| START | 8 | 3 | 2 | | 2 |
| END | 79 | 32 | 30 | 2 | |

**Table 8.1.f**The difference between the similar AU and the other four columns

The next table show us the 8 different Start pages and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | AUTHOR | END PAGE |
|---|---|---|---|---|---|
| DOI | 3 | | 1 | 0 | 0 |
| TI | 2 | 1 | | 0 | 1 |
| AU | 0 | | | | |
| END | 2 | 0 | 1 | 0 | |

**Table 8.1.g**The difference between the different Start pages and the other four columns

This table show us the 3577 similar Start page numbers and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | AUTHOR | END PAGE |
|---|---|---|---|---|---|
| DOI | 1786 | | 1261 | 135 | 40 |
| TI | 1432 | 1261 | | 138 | 36 |
| AU | 344 | 135 | 138 | | 11 |
| END | 88 | 40 | 36 | 11 | |

**Table 8.1.h**The difference between the similar Start pages and the other four columns

The following table show us the 90 different End page numbers and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | AUTHOR | START PAGE |
|---|---|---|---|---|---|
| DOI | 40 | | 33 | 8 | 0 |
| TI | 37 | 33 | | 7 | 1 |
| AU | 11 | 8 | 7 | | 0 |
| START | 2 | 0 | 1 | 0 | |

**Table 8.1.i**The difference between the different End pages and the other four columns

The last table show us the 3486 similar End page numbers and again the differences between the other columns. The logic we followed is the same as above.

| Second given column | No of difference | DOI | TITLE | AUTHOR | START PAGE |
|---|---|---|---|---|---|
| DOI | 1744 | | 1224 | 127 | 3 |
| TI | 1388 | 127 | | 128 | 0 |
| AU | 320 | 127 | 128 | | 0 |
| START | 6 | 3 | 1 | 0 | |

**Table 8.1.j**The difference between the similar End pages and the other four columns

## 8.2.Conclusion

The two scopus, wos databases have many errors. Having tried to remove from the data the duplicates and the NA based on some criteria, we end up with a common file. As clean a file as possible consists of 3587 comments. Most are due to typographical errors or a different recording of the same data. We checked the errors in the key variables and recorded them to see the exclusions from the two bases between them.

## 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] "shodhgangainflibnet," [Online]. Available: https://shodhganga.inflibnet.ac.in/bitstream/10603/5109/10/10_chapter%201.pdf. [Accessed 5 April 2020].

[4] "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," [Online]. Available: file:///C:/Users/elena/Downloads/ZupicCater-2015-Bibliometricmethodsinmanagementandorganization%20(2).pdf.

[6] Γ.Γεωργάκη, "Βιβλιομετρική ανάλυση του επιστημονικού περιεχομένου του προγραμμάτος σπουδών Περιβαλλοντικός Σχεδιασμός," 25 September 2018. [Online]. [Accessed 20 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] Elsevier, "Scopus," November 2004. [Online]. Available: https://www.scopus.com/search/form.uri?display=basic. [Accessed 08 December 2019].

[12] Elsevier, «Elsevier,» 1880. [Online]. Available: https://www.elsevier.com/. [Πρόσβαση 08 December 2019].

[13] Wikipedia, "Wikipedia," [Online]. Available: https://el.wikipedia.org/wiki/Scopus. [Accessed 08 December 2019].

[14] Elsevier, "Scopus Content Coverage Guide," 2017. [Online]. Available: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf. [Accessed 08 December 2019].

[15] Wikipedia, "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/BibTeX. [Accessed 18 12 2019].

[16] A. Swalin, «towards data science,» 31 Janury 2018. [Online]. Available: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4. [Accessed 13 September 2020].

[17] Y.Kinha, «kdnuggets,» [Online]. Available: https://www.kdnuggets.com/2020/06/missing-values-dataset.html?fbclid=IwAR2pAUktlE_zHeCTtbg4MIDxHdB9ZtopiezB9pgo0Lweso1r_rnE8sbn0Gc. [Accessed 15 September 2020].