ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΔΙΟΙΚΗΣΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Analysis of differences about the bibliometric data of the Journal of Financial and Quantitative Analysis, from Scopus and Web of Science databases.

Ιωάννης Λαμπράκης

Επιβλέποντες:

Δημοσθένης Δριβαλιάρης

(Τμήμα Μηχανικών Οικονομίας και Διοίκησης, Πανεπιστήμιο Αιγαίου)

Ευάγγελος Βασιλείου

(Τμήμα Μηχανικών Οικονομίας και Διοίκησης, Πανεπιστήμιο Αιγαίου)

Χίος, 1/10/2020

Έχω διαβάσει και κατανοήσει τους κανόνες για τη λογοκλοπή και τον τρόπο σωστής αναφοράς των πηγών που περιέχονται στον Οδηγό συγγραφής διπλωματικών εργασιών του ΤΜΟΔ. Δηλώνω ότι, από όσα γνωρίζω, το περιεχόμενο της παρούσας διπλωματικής εργασίας είναι προϊόν δικής μου δουλειάς και υπάρχουν αναφορές σε όλες τις πηγές που χρησιμοποίησα.

Abstract

The present dissertation involves the analysis of the bibliometric differences, for the Journal of financial and quantitative analysis, using the databases of Scopus and Web of Science. It begins by describing some basic information about bibliometrics and some basic methods involving in bibliometric analysis, while the journal and the databases are introduced. Afterwards, the main analysis of this dissertation is described, including the process of getting, cleaning, and merging the data provided by the databases. In each procedure followed, the ramifications and differences are commented, to understand if it is possible for the two databases to cooperate.

Περίληψη

Η παρούσα διπλωματική εργασία αφορά τις βιβλιομετρικές διαφορές για το Journal of financial and quantitative analysis, χρησιμοποιώντας τις βάσεις δεδομένων του Scopus και του Web of Science. Ξεκινώντας περιγράφονται βασικές πληροφορίες για την ίδια την βιβλιομετρία και μέθοδοι οι οποίοι χρησιμοποιούνται σε μία, καθώς και πληροφορίες σχετικά με το περιοδικό και τις βάσεις δεδομένων. Έπειτα, στο κύριο μέρος της ανάλυσης περιγράφονται οι διαδικασίες συλλογής, καθαρισμού και ένωσης των δεδομένων που αποκτήθηκαν από τις δύο βάσεις δεδομένων. Σε κάθε στάδιο τις ανάλυσης, τα αποτελέσματα αλλά και η διαφορές μεταξύ των δύο βάσεων σχολιάζονται, έτσι ώστε να διαπιστωθεί το κατά πόσο οι δύο βάσεις μπορούν να χρησιμοποιηθούν σε συνδυασμό.

Table of Contents

Chapter 1

Aim of the analysis

The aim of this dissertation is to analyze the bibliometric data provided from online databases. It will try to identify the bibliometric differences of two databases, Scopus and Web of Science, concerning a common factor, the Journal of Financial and Quantitative Analysis. Identifying, measuring and interpreting their variations, will show if using multiple sources of information instead of one can make a bibliometric analysis more accurate. In addition, the procedure followed can be used for undertaking an analysis relevant to bibliometrics or for evaluating its results.

Chapter 2

Basic information

2.1 Bibliometrics

Living in the era of information, analyzing data while interpreting theirs values have become a major field of interest and inducement of creating new methods around them. Bibliometrics, is one of the tools for doing data analysis, having a large amount of applications. The following paragraphs will try to illustrate the progression of bibliometrics through time, the goals and applications of them while describing the different analyses which can be performed. One sources of information used to acquire the necessary data, was the Wikipedia page about bibliometrics [1]. The biggest assistance and guidance used for describing bibliometrics, was a book from Ana Andres [2], which provided some basic background concerning the topic and description about the methods and uses of bibliometrics.

2.1.1 Background

According to [1], the term bibliométrie was first used by Paul Otlet in 1934 and defined as "the measure of all aspects related to the publication and reading of books and documents.", while the anglicized version bibliometrics was first used by Alan Pritchard in a paper published in 1969, titled "Statistical Bibliography or Bibliometrics?", defining the term as "the application of mathematics and statistical methods to books and other media of communication". At almost the same time, Nalimov and Mulchenko in 1969, used the term scientometrics to refer to "the application of quantitative methods which are dealing with the analysis of science viewed as an information process". Because of the use of this term and for avoiding misunderstanding, scientometrics was restricted to the measurement of science communication, whereas bibliometrics was designed to deal with more general information process. However, at present, bibliometrics and scientometrics are used as synonyms [2]. Later, the term informatics introduced by Gorkova [2], which was a general sub-field of information science used for statistical analysis of communication processes in science. Lastly, the terms of webometrics or cybermetrics also appeared to designate the scientific literature's study from electronic resources.

The breakthrough in bibliometrics, according to [1] and [2], came with the work of Garfield (1955) and Price (1963). Garfield developed a multidisciplinary database in which authors could find articles from across many fields, called "Science Citation Index", being a tool for facilitating the researcher's task. On the other hand, Price's book "Little Science, Big Science" (1963), represented the first systematic approach to the structure of modern science that was applied to its whole. He explained in his book how science has progressed from "little science", which was carried out by a small group of educated scholars, to "big science", investment of developed

countries in science for prosperity. Considering the work done in the field until then, the development of new technologies during the 1990s and the availability of online data bases replacing the traditional indexation systems, made Bibliometric analysis much easier and accessible. Today there are online databases providing indexation information for vast numbers of papers, journals, books, and proceedings.

2.1.2 Before the bibliometric analysis

In order to acquire useful and accurate data about the analysis, there are some features which should be taken into consideration [2].

The first of them is to start with a clear topic definition. Gathering representative documents about the research field of interest needs to be clear for describing productivity in it. Documents gathered are also important about the outcome of the analysis, so they need to cover the whole domain of interest. The next step is to perform a bibliometric search for collecting those documents concerning our field of interest. The source of the information can be obtained by any database concerning the wanted area of study, but it is suggested to be gathered from a multidisciplinary database which include a range of disciplines concerning the publications available. Examples of those multidisciplinary databases are Web of Science, Scopus and Google Scholar.

For evaluating a database like those, Neuhaus and Daniel (2008) [2] have discussed the characteristics of them in order to be appropriate for a given bibliometric study. They pointed out that databases contribute in two ways to the development of a bibliometric study, as data source and as a platform that provides the analytical tools necessary for the bibliometric calculations. The first characteristic they indicate is coverage, which is referred to the extent which the sources processed by the database cover the written scholarly literature, the journal literature in a field. To ensure that the coverage is not biased towards certain countries, languages, publishers, or types of documents, it is essential that the documents on which a bibliometric study is based, to cover the content aimed to be analyzed. Another characteristic identified by these authors is the consistency and accuracy of the data. While it is inevitable that a database will have erroneous data, it is important to access the one that minimizes these errors. As a result, the manual checking required to evaluate the reliability of the data will be reduced. In addition to these characteristics, the data fields required while selecting a database should be taken into consideration. It is meant that to carry out a bibliometric study, it is prior to know about the analysis which will be performed. For example, to study the contribution of an individual will be different from that of an institution. The data fields which may be selected, can be author names, institutions, number of citations received, year of publication, year of the citations received, number of authors contributing to the publication and subject category of the journal. Moreover, certain types of documents can be selected, such as research articles, reviews, letters or comments, or different types of publications like journal articles, books, or proceedings. Those data fields will also influence the database which will be chosen for the bibliographic search, since not all databases may provide the desired information. Furthermore, the browsing options of a database are another characteristic mentioned by Neuhaus

and Daniel, closely related to data accuracy. One type of analyses that can be influenced by data errors is citation analysis since it is based on the citations received by an author or document. The database provides a match of those publications that have cited a given author or paper, but if the citing source has made a mistake the match will not be done correctly. So, to avoid possible inconsistences, it is important that the browsing options are included. Having mentioned the importance of knowing that the database provides the information needed, another characteristic of a database occurs, which is the search options. Each database has an interface, with different search, browsing and saving options giving the faculty of saving all the author names on a paper or the publications that cite a given author, or in ranking papers according to the number of times they have been cited, all of these depending on the objective of the analysis. The availability of these options will depend on the database chosen. Along with these, a database providing the characteristic of analytical tools, gives the option of analyzing the results obtained in a bibliographic search. These tools can supply information about the most productive authors among the publications identified or indicate the rate of productivity over time or include the option of analyzing collaboration maps between authors. Depending on the number of documents in a bibliographic search, it may not be possible that a given database will be able to analyze all records chosen at the same time. The final characteristic pointed out by Neuhaus and Daniel, is the saving and exporting options. The last step in a bibliographic search is to save the documents found to be used for the bibliometric analysis. The databases offer different formats for saving records or exporting them to bibliographic software. The appropriateness of a database will be decided based on the needs of the analysis.

Before beginning the main bibliometric analyses, it is important to provide a clear explanation of the bibliographic search conducted so that the target documents should be well defined. This means explicitly stating the period that has been taken into consideration, the database or databases used, and the keywords entered. It is also important to specify any other procedure used to select documents, like exclusion criteria or manual check of document content. This systematic procedure will ensure that the analysis will be reproducible.

2.1.3 Measuring productivity

2.1.3.1 Descriptive analysis

The easiest way to begin a bibliometric study is through a descriptive analysis, providing a quick and visual impression of certain aspects related to productivity [2].

One of the descriptive analyses which can be performed, is temporal evolution, the development of the documents collected over time. To conduct this analysis, it is necessary to have the year of publication of each study, the number of studies (frequency), percentage and cumulative percentage. These variables, along with a graphical representation, will make possible to identify the trend of scientific productivity in each area of study. The same analysis regarding temporal

evolution could, alternatively, focus on productivity over time in each journal. In this case, the number of publications will refer to those published in a specific journal.

Another descriptive analysis performed for a bibliometric study concerns the number of authors contributing to each publication. The count of the number of authors contributed to a report offers some indication of the degree of collaboration between authors. Another descriptive analysis also concerns the most productive authors, which draws up a list of the most productive authors are used, as well as the total number of publications. The percentage of publications the most productive authors have with respect to the total number of publications in the study or over the period in which they have been most productive can also be included in this descriptive analysis. The use of these analyses is not mandatory, so the choice of their use will depend on their suitability for describing the data and the features of each bibliometric study.

A common analysis along with these, entails the drawing up of a list with the most productive institutions or countries in each field, including a ranking of those authors who have collaborated in the greatest number of publications. For identifying the most productive institutions or countries, the affiliation of the authors who have taken part in writing the documents needs to be obtained. This procedure provides a better description of author participation as well as the collaboration between institutions and countries, if the listed affiliations include all the authors rather than only that of the first author. One technical problem which may occur with the analysis of scientific productivity by institution, as van Raan (2005) pointed out, is how the affiliation information of authors in obtained and arranged. This means that the same institution may be referred to by a different wording, resulting in a difficulty to distinguish an institute. In addition to that, due to the language differences the same institution could be mentioned in more than one language, resulting in the same problem as before.

There are also other indicators used in a descriptive analysis applied in a bibliometric study. These may concern the language of the documents, for counting how many of the publications included are written in each one of the languages found. The type of document is also a topic of study, by counting how many documents belong to each one of the literature types. Finally, some studies include a distribution of articles across different subject categories to which the document belongs, to categorize each.

2.1.3.2 Author production

Even though a basic descriptive analysis referring to the most productive authors can be carried out to identify the most highly productive people in each area of research, the data can also be analyzed differently. The author productivity can be examined using a widely used bibliometric law, Lotka's law, providing a better understanding of author productivity [2].

Alfred J. Lotka (1926) studied the patterns of author productivity while he developed one of the main laws in bibliometrics. He interpreted that, there are a lot of authors who publish only one study, in a given area of science, while there is a small group of authors who contribute with many publications. This proposition entails the basis of Lotka's law, referred to as the inverse square law on author productivity. The law takes the number of authors who have contributed with a single study and then predicts how many authors would have published "x" studies, according to this inverse square law.

The number of authors who produce "x" studies are proportional to "$1/x^2$", or as most commonly used $y_x = c/x^2$. In this equation, "y" is the number of authors with "x" publications, "c" is the number of authors with a single publication and "x" is the number of publications itself. The distribution about author production shows that the more articles produced by an author the more likely he or she is to produce others, so the productivity is related to the algorithm of the number of articles published by an author. The value to the n exponent ($n = 2$), can be calculated for each author productivity distribution, being higher or lower than 2, using the least squares method. To calculate c, the formula is $c = \frac{1}{\sum 1/x^n}$.

The productivity of the authors measured by Lotka's law does not depend on the science field it is applied, but the period. For the authors to have the opportunity to publish more than once or twice, a long period of time is necessary to be considered, set at ten years or more. For applying Lotka's Law it is important to know the author's credit for publication, as Wolfram (2003) [2] stated, publications with multiple authors can present a dilemma in productivity studies because there are several ways for authors to receive credit for publications they have co-authored. The credit having received the greatest support from the literature is the complete count, referring to the recognition and reception of equal treatment regardless of the number of authors associated with the article.

2.1.3.3 Journal productivity

Another aspect of analysis in a bibliometric study concerns journals in which the articles gathered are published. Like in the author productivity case, descriptive analysis about the journals can be carried out, although a more detailed analysis about the journal productivity may be applied for research purposes. In this case, Bradford's law is the main tool of the analysis [2].

Bradford's law was formulated in 1934 by Samuel C. Bradford, for studying the distribution of scientific literature. Through his work, the gathering of all the articles he could find related to geophysics between 1931 and 1933. He found a regulation called the inverse relationship between the number of articles published in a subject area and the number of journals in which the articles appear. The meaning of that, is that in each subject area, a small amount of journals account for a sizeable portion of the total publications in that area, whereas increasing numbers of journals

publish fewer articles in that area. The graphical representation of this law, is circles one inside the other for ranking and dividing into groups or categories the journals, depending on the number of articles they account for. These zones are called Bradford zones. While the number of articles in each zone must be the same, the number of journals publishing these articles will not be, since some journals will be more productive than other. In the first central zone, the smallest one, a small group of articles will be located, and the subsequent zones will have an increasing number of journals. The ratio between the number of journals in subsequent zones will be approximately $1: k: k^2$ and so forth. Using this law, most productive journals in a subject area can be identified.

Plotting the cumulative number of papers of each journal against the logarithm of their rank and ordering the journals from the most to the least productive, can provide the S-shape curve at the Bradford-Zipf plot. This S-shape presents a phenomenon known as the Gross droop (Gross,1967) [2], indicating that fewer articles than expected are being contributed by the least productive journals. An explanation given for this effect by Heine (1998), is that the observed droop is a result of an excess of low-productivity journals which contain lower than the expected number of articles. In this plot, the core journals are those that lie along the initial curved part of it, before it becomes a straight line. Although, in real data distributions the curve will not be exactly S-shaped.

In order to apply Bradford's law, all the articles in a given period of time and research field should be gathered and listed to the journal they have been published. Firstly, the Bradford's constant k should be calculated, explaining how the number of journals grows from one zone to the next. The formula for calculation k was formulated by Egghe (1986, 1990) [2] and Egghe and Rousseau (1990) [2], being $k = (e^{\gamma} \times Y_m)^{1/P}$, where γ is the Euler's number (γ = 0.5772), Y_m is the maximal productivity of the journal or rank one, and P is the number of zones or Bradford groups. After counting the number of Bradford zones, it is possible to calculate the number of journals belonging to Bradford's first group, using the formula $r_0 = \frac{T(k-1)}{(k^P - 1)}$, where T represents the total number of journals publishing articles in a given subject area, k is Bradford's constant and P is the number of Bradford groups. By calculating the number of journals belong to the first zone, it is then possible to find the expected number of journals in the next zones, using the ratio between them as mentioned before $1: k: k^2$ ... Confirming that the data fit Bradford's law, by comparing the exact number of multiplier k with the real number of journals, the equation for Bradford curve is calculated. To explain journal productivity, it is commonly used the Leimkuhler's formulation, $R(r) = a \log_e(1 + br)$, where $R(r)$ is the cumulative number of articles published by the journals of rank $1,2,3 \ldots r$. The value a being a constant, calculated as $a = y_0 / \log_e k$ where, y_0 is the number of articles found in each group (considering that each zone will include the same number of articles) and is calculated by $y_0 = A/P$, where A is the total number of articles found in the literature and P is the number of Bradford zones.

2.1.3.4 Journal ranks

Another analysis involving journal productivity in bibliometric research, is ranking journals according to the frequency of the documents they publish. Creating a list of journals with decreasing productivity order, provides important information and makes the analysis more comprehensive, while it is a prior step in applying Bradford's law [2].

The practical implications of identifying journal ranks have been pointed out by Lascar and Mendelsohn (2001), by analyzing journal productivity in structural biology. By this way they identified a group of key researchers in their field and gathered the studies they had published. Ranking of the most productive journals in the area, they were able to identify the most productive among them, for making proposals to their research center regarding the subscription to a given online journal package.

2.1.4 Author citation analysis

Another tool for identifying authors' relationships, is citation analysis. An author study which involves references to other authors' studies, which are related. The existence of those reveal connection between authors, groups of researchers, topics of study or countries. In addition, the impact and relevance that authors or studies have on a scientific community can be measured but citation analysis should not be the single criterion for judging their importance.

The number of citations received by an author during a given period, can be used for comparing research productivity and impact between authors, institutions, or countries. The citations are treated as uniformly positive recognitions of the contribution made by the author or work being cited, so the greater the number, the bigger the recognition.

2.1.4.1 Citation life

Price [2], found a regularity in the behavior of science, the obsolescence of literature. It is referred to the decline in the use of documents over time, where a document that is no longer cited will become obsolete. A proportion of documents may not be cited, while other may receive immediately citations after publication prior to becoming obsolete. It is also possible that some documents will remain uncited or rarely cited in the early years of their publication, but then become recognized.

Alongside with the obsolescence of literature, it can be also studied in the citation analysis the age of citations. This is assessed by knowing the year of the citation in comparison to the year of publication of the document. This analysis can show how citations made to articles are distributed over time.

Finally, the lifetime distribution of citations can be studied, in combination with an analysis of the type of citation. Among the citations that a study may receive, an important contribution makes the self-citations, aging more quickly than foreign citations do [2].

2.1.4.2 Author self-citations

Self-citation is a practice whereby authors cite their own previous work occurring whenever the set of co-authors of the citing paper and that of the cited one share at least one author.

The analysis of self-citation can be affected by homonyms, spelling variances or misspellings of an author's name.

2.2 Journal of Financial and Quantitative Analysis

The following chapter is referred to the Journal of Financial and Quantitative analysis, including information about the journal and its publication history. Sources used for undertaking this, were the journal's official website [3], the Cambridge university press site for the journal [4], which is the publisher of it, and the archive provided from this site for the journal [5]. It was also used from a digital library of academic journals, the Jstor, the journal's online archive [6] and the Wikipedia page for the journal [7].

2.2.1 JFQA background

The journal of Financial and Quantitative Analysis (JFQA), according to the front covers of the publications that Cambridge university press archive [5] provide, was first published in 1966 and exists until now. Its publications, according to Cambridge university press [4], involves theoretical and empirical research in financial economics, including topics of corporate finance, investments, capital and security markets, and quantitative methods of particular relevance to financial researchers. The Cambridge university press [4] also points out that, JFQA prints less than 10% of the more than 600 unsolicited manuscripts submitted annually and according to its official site [3] and its Wikipedia page [7], its acceptance rate is 8% and yearly submissions are 1106. It is also stated from the Cambridge University [4], that 3000 libraries, firms and individuals in 70 nations publish in the journal of financial and quantitative analysis while it serves an international community of sophisticated finance scholars, academics and practitioners. Moreover, the journal offers the annual William F. Sharpe Award [3], [7] for scholarship in Financial Research, recognizing researchers whose articles published have most contributed to the understanding of important areas about financial economics.

2.2.2 Publisher history

For investigating the publishers of the journal of financial and quantitative analysis, the archives of the journal from the Jstor [6] along with the Cambridge university press [5] were used. From those archives, the front covers of the journal for each issue of every volume were observed,

in order to identify the publisher of each issue that it was mentioned. There was a difference between the two archives [5], [6] where Cambridge university press mentioned them as “cover and front matter” while Jstor named them “front matter”. The Cambridge university press archive was missing some front matters but the Jstor had them covered. Those front matters missing, were from volume 29 (1994) until volume 35 (2000) in every issue, except volume 31 and the issues 4 from volume 29 and 33. Considering the above, it was found that the first publisher of the journal was the University of Washington graduate school of business administration jointly with the western finance association. Later, the two of them stop publishing the journal and the place of western finance association took the New York University Leonard N. Stern school of business. Afterwards, the Cambridge University press took charge of the Journal of financial and quantitative analysis publication, for the renamed University of Washington after the September of 2007 [8], Michael G. Foster school of business in cooperation with the university of Utah David Eccles school of business and the New York university Leonard N. Stern school of business. Finally, the Cambridge University press was publishing the journal for the University of Washington Michael G. Foster school of business cooperated with the Arizona state University W. P. Carey school of business and the University of North Carolina Kenan-Flagler business school. Below there is a matrix showing the historical background of the JFQA publishers.

Table 1: Journal of financial and quantitative analysis publishers' progression

Period	Publisher
1966 – 1987	University of Washington graduate school of business administration jointly with the western finance association
1988 – 1996	University of Washington graduate school of business administration
1997 – September of 2003	University of Washington school of business administration in cooperation with the New York university Leonard N. Stern school of business
December of 2003 – 2008	University of Washington school of business administration in cooperation with the University of Utah David Eccles school of business and the New York University Leonard N. Stern school of business
2009 – June of 2015	Cambridge university press for the University of Washington Michael G. Foster school of business in cooperation with the University of Utah David Eccles school of business and the New York University Leonard N. Stern school of business
August of 2015	Cambridge university press for the University of Washington Michael G. Foster school of business in cooperation with the Arizona state University W. P. Carey school of business and the New York University Leonard N. Stern school of business
October of 2015 – present	Cambridge university press for the University of Washington Michael G. Foster school of business in cooperation with the Arizona state University W. P. Carey school of business and the University of North Carolina Kenan-Flagler business school

2.2.3 Publication history

For tracing the number of publications the Journal of financial and quantitative analysis had each year, it was used once more the front covers from the archives of it, [6] and [5], and the information given from the sites as it were. By visiting each site and expanding the volumes' information, the number of issues were observed, while in cases where the year was not clear the covers of each issue were used. The journal has 54 volumes until today and the number of issues per volume published, differed from year to year. The first 3 volumes of the journal

consisted of 4 issues while the consecutive 17th volumes included 5 issues in each, with an exception of the 9th volume having 6 issues. Later, from the 18th until the 43rd volume there were 4 issues, while the rest of the volumes consisted of 6 issues each. Below there is matrix containing a more detailed description of the journal's volumes.

Table 2: Journal of financial and quantitative analysis volumes' progression

Volume	Number of issues	Period
1 – 3	4	1966 – 1968
4	5	1969 – 1970
5	5	1970
6	5	1971
7	6	1972
8	5	1973
9	6	1974
10 – 17	5	1975 – 1982
18 – 43	4	1983 – 2008
44 - 48	6	2009 – 2013
49	6	2014
50	6	2015
51 – 54	6	2016 - present

Respectively to the changes of the journal's volumes per year, the months of its issues publications has altered, from being every three months to bimonthly. Using again the archives, [6] and [5], as mentioned before about tracing the issues of the journal, the information needed were acquired, in this case by observing the months of publication. The first 3 volumes (1966-1968) consisted of 4 issues published in March, June, September and December. The fourth volume (1969-70) had 5 issues published respectively in March, June, September, December and January while the fifth volume (1970) had 5 issues, the 4th and 5th issue were published both in December of 1970. The next 3 volumes 6,7 and 8 (1971-73) published 5 issues in January, March, June, September and December respectively with an exception of the seventh volume (1972), which published one extra special issue in March. The Ninth volume (1974) consisted of 6 volumes published in January, March, June, September, November and December. The next 8 volumes, 10 to 17 (1975-1982), had 5 issues published respectively in March, June, September, November and December. The volumes 18 to 43 (1983-2008), consisted of 4 issues published on March, June, September and December respectively. The rest of the volumes up until today consisted of 6 issues which are published bimonthly in February, April, June, August, October and December, with two exceptions on volume 49 (2014) and volume 50 (2015). On volume 49 instead of October, issue 5 and 6 were both published on December while on volume 50 instead of February, issue 1 and 2 were both published on April.

2.3 Databases

Having described the journal itself, the sources of information required for the analysis needs to be discussed too. The databases used for this analysis, was Scopus and Web of science as they provide a large amount of data and variables about the journal having chronologically covered the archive. The websites used for this section, are the official website of Scopus [9], the official website of Web of science [10], their Wikipedia pages, [11] and [12], and the official web of Elsevier.

2.3.1 Scopus

Scopus, according to [9] and [11], is an abstract and citation database launched in 2004. It is owned by Elsevier [13], a Dutch information and analytics company which is one of the world's major providers of scientific, technical and medical information, established in 1880. Scopus requires a subscription in order to be accessed. The types of sources Scopus covers are book series, journals and trade journals concerning subject fields of life sciences, social sciences, physical sciences and health sciences dated from 1966 to present. Scopus consulting [9] and [11], also offers author profiles covering affiliations, number of publications and their bibliographic data, references and details on the number of citations each published document has received. It provides alerting features that allow registered users to track changes to a profile and a facility to calculate authors' h-index. The total number of its records as shown in [9] exceed the 69 million, containing nearly 36,377 titles and approximately 11,678 publishers, while it has 1.4 billion cited references.

2.3.2 Web of Science

Web of Science or as previously known Web of Knowledge, according to [10] and [12], is a website which provides subscription-based access to multiple databases that contain comprehensive citation data for many different academic disciplines. It was produced by the Institute for Scientific Information (ISI) [10], but is currently maintained by Clarivate Analytics. Due to Thomson Reuters work of Century of Social Sciences [12], a service containing files which trace social science research back to the beginning of the 20th century, Web of Science now has indexing coverage from the year 1900 to the present. Consulting [10] and [12], Web of science contains Full text articles, reviews, editorials, chronologies, abstracts, proceedings about journals and books, technical papers while its subject fields include sciences, social sciences, arts and humanities supported by 256 disciplines. The Web of Science core collection consists of six online databases [12]. These are the Science Citation Index Expanded, which covers journals encompassing 150 disciplines from 1900 to the present day, the Social Sciences Citation Index which covers journals in social science disciplines from 1900 to the present day, the Arts & Humanities Citation Index covering arts and humanities journals starting from 1975 while it contains scientific and social sciences journals, the Emerging Sources Citation Index covering sciences, social science, and humanities, the Book Citation Index which covers editorially selected books starting from 2005 and the Conference Proceedings Citation Index which covers conference

titles in the Sciences starting from 1990 to the present day. The total number of its records [10], are 161 million, while multidisciplinary coverage of it, encompasses over 50,000 scholarly books, 12,000 journals and 160,000 conference proceedings while the selection of those is made based on the impact evaluation of its. There are also 34,000 Journals indexed and 1,7 billion cited references.

Chapter 3

The data

3.1 Data acquisition

For conducting the analysis about the bibliometric differences, R studio was used. R studio, according to Wikipedia [14], is an integrated development environment for R, a programming language for statistical computing and graphics. It is a free and worldwide used program containing a large variety of packages, which can make an analysis much more convenient, descriptive and easier.

For obtaining the data of those two databases about the journal, a virtual private network (vpn) connection with the university was made, to gain access to the contents of those. Following the instructions of using each site [9], [10], the Journal of Financial and Quantitative Analysis was searched by its name, using at the search options all fields available. In both cases the results included other journals too, so from the source title field of both the databases the name of the journal was selected. This action returned all the documents available for the JFQA. Later, it was specified that the archive of each database, would be obtained from the day it was recorded, until the end of 2018. This limitation took place because the records of 2019 would be incomplete and, in some cases, inaccurate. Scopus dataset provided 2629 observations, dated from 1966 to 2018, while Web of science had 2431, dated from 1970 to 2018. Having confined the information needed, data were downloaded in 2/5/2019 using every variable available in each database, in bibtex format. This format was mandatory for using a specific package of R studio called “bibliometrix”, accountable for importing and normalizing the data into R studio. Additionally, both databases had a limit in downloading content, 2000 for Scopus and 500 for Web of science. Considering that, two bibtex files were downloaded for Scopus dataset and five for Web of science.

3.2 Data importation

Importing the data into R studio using the package “bibliometrix”, like mentioned before, Scopus dataset had 2629 observations and 27 variables while Web of science had 2431 observations and 36 variables. From those variables, 5 were created due to the use of the package. Consulting the field tag information of each database and the two datasets acquired, a first interpretation of the data was made. Below, there are two matrices depicting the variables included in each database, the meaning of each, a comment for each and the groups of the tag if existed.

Table 3: Scopus variables

Variable	Meaning	Comment	Groups
AU	Authors	The number of Authors in each observation varies from 1 to 5. In 5 cases there were not given author	
TI	Article Title		
SO	Publication Name	The name of the publisher	"JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS" in every observation
JI	ISO Source Abbreviation	Represents a standard abbreviation for the source title	"J. FINANC. QUANT. ANAL." In every observation
AB	Abstract	Contains the summary of the document	In 1,161 cases, it is not given
DE	Author Keywords	Words to make the article easier to be found	In 2,591 cases, it is not given
LA	Language	Specifies the language the document is written	"ENGLISH" in every observation
DT	Document Type	Specifies the type of the document	"ARTICLE", "ARTICLE IN PRESS", "CONFERENCE PAPER", "EDITORIAL", "ERRATUM", "LETTER", "NOTE", "REVIEW",
DT2	Document Type 2	Irrelevant variable because of the existence of the variable DT	"ARTICLE" in every case
TC	Times Cited	The Scopus count of times the observations has been cited	

Variable	Meaning	Comment	Groups
CR	Cited References	Refers the authors and the article the document has been cited	In 320 cases, it is not given
C1	Author Address	The address of the university the author is involved	In 134 cases, it is not given
DI	DOI (Digital Object Identifier)	Interoperable code for identifying the document	In 138 cases, it is not given
RP	Reprint Address	Contains the author and the university for Reprinting the document	In 816 cases, it is not given
FU	Funding Agency and Grant Number	Contains the name of the organization funded the document and the grant number	In 2318 cases, it is not given
SN	International Standard Serial Number (ISSN)	An international standardized code which identifies journals	"00221090" in every observation
PN	Part Number	The issue that document is included	
PP	Pages	The start and end page of the document	
PU	Publisher	Publisher of the document	In 1991 cases, it is not given and in 663 cases "CAMBRIDGE UNIVERSITY PRESS"
DB	Database	The name of the database	"SCOPUS" in every case
VL	Volume	The volume the document is	From 1 to 53
PY	Year	The year of publication	From 1966 to 2018
AU_UN	Author University	A variable created from the package used. Contains the universities of all authors	In 170 cases, it is not given
AU1_UN	First Author University	A variable created from the package used. Contains the university of the first author	

Variable	Meaning	Comment	Groups
AU_UN_NR	Author University cited		Not given in every case
SR_FULL		Discrete description for each observation. Contains the name of the first author, the year and the name of the journal	
SR		Abbreviation of the discrete description	

Table 4: Web of science variables

Variable	Meaning	Comment	Groups
AU	Authors	The number of Authors in each observation varies from 1 to 5.	
TI	Article Title		
SO	Publication Name	The name of the publisher	"JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS" in every observation
JI	ISO Source Abbreviation	Represents a standard abbreviation for the source title	"J. FINANC. QUANT. ANAL." In every observation
AB	Abstract	Contains the summary of the document	In 1249 cases, it is not given
ID	Keywords Plus	Words given for finding the document	In 1228 cases, it is not given
LA	Language	Specifies the language the document is written	"ENGLISH" in every observation
DT	Document Type	Specifies the type of the document	"ARTICLE", "ARTICLE, PROCEEDINGS PAPER", "CORRECTION", "CORRECTION, ADDITION", "DISCUSSION", "EDITORIAL MATERIAL",

Variable	Meaning	Comment	Groups
			"FICTION, CREATIVE PROSE", "LETTER", "MEETING ABSTRACT", "NOTE", "LETTER", "NOTE", "REVIEW", "BOOK REVIEW"
DT2	Document Type 2	Irrelevant variable because of the existence of the variable DT	"ARTICLE" in every case
TC	Times Cited	Web of Science Core Collection Times Cited Count	
CR	Cited References	Refers the authors and the article the document has been cited	In 137 cases, it is not given
C1	Author Address	The address of the university the author is involved	In 37 cases, it is not given
DI	DOI (Digital Object Identifier)	Interoperable code for identifying the document	In 53 cases, it is not given
PA	Publisher Address	It describes the address of the publisher	"32 AVENUE OF THE AMERICAS, NEW YORK, NY 10013-2473 USA" in every case
FU	Funding Agency and Grant Number	Contains the name of the organization funded the document and the grant number	In 2294 cases, it is not given
FX	Funding Text	Refers to the funding provider	In 2294 cases, it is not given
SN	International Standard Serial Number (ISSN)	An international standardized code which identifies journals	"0022-1090" in every case
PN	Part Number	The issue that document is included	

Variable	Meaning	Comment	Groups
PP	Pages	The start and end page of the document	
PU	Publisher	Publisher of the document	"CAMBRIDGE UNIV PRESS" in every case
VL	Volume	The volume the document is	From 10 to 53
PY	Year	The year of publication	From 1970 to 2018
UT	Accession Number	The accession number is a unique identifying number associated with each record in the product	
NR	Cited Reference Count		
SC	Research Areas	The research field the document covers	"BUSINESS \& ECONOMICS" in every case
U2	Usage Count (Since 2013)		
WC	Web of Science Categories	The research field the document covers	"BUSINESS, FINANCE, ECONOMICS" in every case
EM	E-mail Address	Provides the authors' email address	In 1613 cases, it is not given
GA	Document Delivery Number	A five to six-digit code, to identify the document's delivery	
RP	Reprint Address	Contains the author and the university for Reprinting the document	In 37 cases, it is not given
DB	Database	The name of the database	"ISI" in every observation
AU_UN	Author University	A variable created from the package used. Contains the universities of all authors	In 584 cases, it is not given
AU1_UN	First Author University	A variable created from the package used. Contains the university of the first author	

Variable	Meaning	Comment	Groups
AU_UN_NR	Author University cited		Empty in every case
SR_FULL		Discrete description for each observation. Contains the name of the first author, the year and the name of the journal	
SR		Abbreviation of the discrete description	

Chapter 4

Preprocessing the data

4.1 Duplicates

Having the data imported into R and their different variables observed, the attempt to normalize them was made. For doing that, duplicated entries searched in the fields of “Title”, “DOI”, “Page start”, “Page end”, “Volume” and “Issue”. This procedure was important for identifying mistaken entries and preparing the analysis of the data.

4.1.1 Title

The first variable used in each dataset for finding duplicates, was “Title”. Scopus dataset had 13 Titles, each one mentioned once or more times, making a total of 44 duplicated Titles. On the other hand, Web of science dataset had 16 Titles, each one mentioned once or more times, making a total of 37 duplicated Titles.

Table 5: Duplicated Titles

Scopus	Title	Number of repetitions	Web of Science	Title	Number of repetitions
	COMMENT: “AN AUTOREGRESSIVE FORECAST OF THE WORLD SUGAR FUTURE OPTION MARKET”	2		AUTOREGRESSIVE FORECAST OF WORLD SUGAR FUTURE OPTION MARKET – COMMENT	2

Scopus	Title	Number of repetitions	Web of Science	Title	Number of repetitions
	COMMENT: A TEST OF STONE'S TWO-INDEX MODEL OF RETURNS	2		CORPORATE DIVIDEND-SAVING DECISION	2
	CONVERTIBLE DEBT FINANCING	2		CORPORATE INTERNATIONAL DIVERSIFICATION AND MARKET ASSIGNED MEASURES OF RISK AND DIVERSIFICATION	2
	DISCUSSION	15		CORRECTION	7
	DIVIDEND PREDICTABILITY AROUND THE WORLD	2		EVIDENCE ON PRESENCE AND CAUSES OF SERIAL-CORRELATION IN MARKET MODEL RESIDUALS	2
	DOES THE DISPOSITION EFFECT MATTER IN CORPORATE TAKEOVERS? EVIDENCE FROM INSTITUTIONAL	2		FINITE-DIFFERENCE METHODS AND JUMP PROCESSES ARISING IN PRICING OF CONTINGENT	2

Scopus	Title	Number of repetitions	Web of Science	Title	Number of repetitions
	INVESTORS OF TARGET COMPANIES			CLAIMS – SYNTHESIS	
	ERRATA	6		FORECASTING AND ANALYSIS OF CORPORATE FINANCIAL PERFORMANCE WITH AN ECONOMETRIC MODEL OF FIRM	2
	FINITE DIFFERENCE METHODS AND JUMP PROCESSES ARISING IN THE PRICING OF CONTINGENT CLAIMS: A SYNTHESIS	2		FUNCTIONAL FORM, SKEWNESS EFFECT, AND RISK-RETURN RELATIONSHIP	2
	HINDSIGHT EFFECTS IN DOLLAR-WEIGHTED RETURNS	2		INFORMATION EFFECTS AND STOCK-MARKET RESPONSE TO SIGNS OF FIRM DETERIORATION	2

Scopus	Title	Number of repetitions	Web of Science	Title	Number of repetitions
	JFQA STYLE REQUIREMENTS	3		MODEL OF CAPITAL ASSET RISK	2
	SPILLOVER EFFECTS AMONG FINANCIAL INSTITUTIONS: A STATE-DEPENDENT SENSITIVITY VALUE-AT-RISK APPROACH	2		OPTIMAL EQUITY FINANCING OF CORPORATION	2
	THE ROLE OF GROWTH OPTIONS IN EXPLAINING STOCK RETURNS	2		OPTIMAL INVESTMENT FINANCING DECISIONS AND THE VALUE OF CONFIDENTIALITY	2
	WESTERN FINANCE ASSOCIATION	2		PORTFOLIO PERFORMANCE OF PROPERTY-LIABILITY INSURANCE COMPANIES	2
				SHOULD LARGE BANKS BE ALLOWED TO FAIL	2

Scopus	Title	Number of repetitions	Web of Science	Title	Number of repetitions
				STRANGE JOURNEY OF MONETARY INDICATORS	2
				SUFFICIENT CONDITION FOR A UNIQUE NONNEGATIVE INTERNAL RATE OF RETURN – COMMENT	2
sum	13	44		16	37

For further investigation, the observations with the same title were compared with the online archives of the journal [5], [6], to test their validity and discreteness. This inquiry made by searching each duplicated Title, based on its volume and issue. By doing that, some of the observations were found while others did not, leading to the discovery of possible duplicates. Below, there are two matrices depicting the problems observed for every possible duplicated entry, for each of the datasets.

Table 6: Scopus dataset

Title	Problem observed
DIVIDEND PREDICTABILITY AROUND THE WORLD	In the first case the issue was 56, not 5-6. The second case could not be found in the archive. So, the second one could be a **Duplicate.**
DOES THE DISPOSITION EFFECT MATTER IN CORPORATE TAKEOVERS? EVIDENCE FROM INSTITUTIONAL INVESTORS OF TARGET COMPANIES	In the first case the volume number was 96 (max 54) and it could not be found in the archive. The second one was found. So, it could be assumed that this may was a **Duplicate.**
FINITE DIFFERENCE METHODS AND JUMP PROCESSES ARISING IN THE PRICING OF CONTINGENT CLAIMS: A SYNTHESIS	In the first case we can see only the beginning of the second one. In the second case we can see a full article. Since the author was the same and the second published one year later it could be assumed that the second one was the continuum of his paper.
HINDSIGHT EFFECTS IN DOLLAR-WEIGHTED RETURNS	In the first case the volume number was 58 (max 54) and it couldn't be found in the archive. The second on was found. So, it could be assumed that this may was a **Duplicate**.
JFQA STYLE REQUIREMENTS	The three of them was identical
SPILLOVER EFFECTS AMONG FINANCIAL INSTITUTIONS: A STATE-DEPENDENT SENSITIVITY VALUE-AT-RISK APPROACH	In the first case the volume number was 96 (max 54) and it could not be found in the archive. The second one was found. So, it could be assumed that this may was a **Duplicate.** (Also, the name of Roland Fuss was shown as F<U+00BC>SS R in the dataset)
THE ROLE OF GROWTH OPTIONS IN EXPLAINING STOCK RETURNS	In the first case the volume number was 96 (max 54) and it could not be found in the archive. The second on was found. So, it could be assumed that this may was a **Duplicate**.
WESTERN FINANCE ASSOCIATION	Both of their names were different in the archive. The one with hot pages: 999-1000, was named 'Minutes of the executive committee meeting', and the one with hot pages: 997-998, was named 'Minutes of the Annual Meeting'.

Table 7: Web of Science dataset

Title	Problem observed
CORRECTION	In one case, the hot page was shown as U481 and in the archive was 481. In another case, the hot page was shown as U145 - \& and in the archive was 1009-1010. In another case which was shown as correction the full title found on the archive was 'A note on Modeling simple Dynamic cash Balance Problem: Errata. Finally, there was a case in the data with hot page 2155, but in the archive, it was referred as ERRATA and had no page.
FORECASTING AND ANALYSIS OF CORPORATE FINANCIAL PERFORMANCE WITH AN ECONOMETRIC MODEL OF FIRM	One of them was a comment to the other, but it was not mentioned in the dataset
MODEL OF CAPITAL ASSET RISK	One of them was a comment to the other, but it was not mentioned in the dataset
PORTFOLIO PERFORMANCE OF PROPERTY-LIABILITY INSURANCE COMPANIES	The hot page given in the dataset was wrong. The correct was 1595 – 1611.

After comparing the Titles with the journal's archives [5], [6], the Scopus archive [9] was visited in order to examine some cases. Those were 3 entries with Volume 96, 1 entry with Volume 58 and 1 entry with Volume, Issue and Page that could not match the journal's archive, all dated on 2014. So, it could be assumed that in the year 2014, those entries were wrongly accumulated or typed in the Scopus dataset and they may be duplicates.

4.1.2 DOI

For further investigation on duplicated entries, the variable of DOI was used for searching matches in each dataset. Scopus dataset had 5 observations with DOI mentioned twice and 138 observations with DOI equal to "not available". Web of Science dataset had no matching DOI. Those 5 observations found from the Scopus dataset using DOI, were matching 5 of the entries found using the Title, showing that those are possibly the duplicates.

4.1.3 Page start, Page end, Volume and Issue

To make sure that the entries found so far are duplicates or examine the cases of finding others, the variables of Page start, Page end, Volume and Issue were used as criteria, with different combinations of those.

The combination of Page start, Page end, Volume and Issue, returned from the Scopus dataset 3 observations, matching all those fields with "not available" except volume, which was 96, confirming the already found possible duplicates. In addition, 56 observations matched, because of having "not available" value in each of these fields mentioned. On the contrary, Web of science dataset had no observation matching all these parameters.

The next criterions used were Page start, Volume and Issue. Scopus dataset returned the same 3 observations that mentioned before, 9 cases where Page start was the same, while Issue and Volume were "not available" and 56 observations having "not available" values in all the fields mentioned. Web of science dataset once again had no observation matched using these parameters.

The last combination of those variables used, was Page end, Volume and Issue. In Scopus dataset were found the same 3 observations mentioned before, 70 observations matched by Issue and Volume while having Page end "not available" and 56 entries having "not available" values in all the fields mentioned. Web of science dataset had 68 cases where Issue and Volume were matching while Page end was "not available". From those cases none had the same Page start variable. Additionally, there were 2 cases where Page end, Issue and Volume were the same while Page start was not. Comparing those two observations with the Scopus dataset and the archives of the journal [5], [6], it was found that the Page end of one of them was wrong.

4.1.4 Results

Having performed those 5 different examinations, 5 duplicated entries were found in Scopus dataset while none in Web of Science. 5 of them were found using as criterion the Title or DOI, while 3 of them were found due to the other combinations of Page start/end, Volume and Issue. The Title of those duplicated observations and the different variables they possessed, are depicted below.

Table 8: Duplicated Scopus observations

Title	Different variables
THE ROLE OF GROWTH OPTIONS IN EXPLAINING STOCK RETURNS	"Citation", "Issue", "Page start", "Page end", "Volume"
DOES THE DISPOSITION EFFECT MATTER IN CORPORATE TAKEOVERS? EVIDENCE FROM INSTITUTIONAL INVESTORS OF TARGET COMPANIES	"Citation", "Issue", "Page start", "Page end", "Volume"
SPILLOVER EFFECTS AMONG FINANCIAL INSTITUTIONS: A STATE-DEPENDENT SENSITIVITY VALUE-AT-RISK APPROACH	"Reprint Address", "Citation", "Issue", "Page start", "Page end", "Volume"
DIVIDEND PREDICTABILITY AROUND THE WORLD	"Reprint Address", "Citation", "Issue", "Page start", "Page end", "Volume"
HINDSIGHT EFFECTS IN DOLLAR-WEIGHTED RETURNS	"Citation", "Page start", "Page end", "Volume"

Removing the duplicated entries, Scopus dataset had 2624 observations and Web of Science 2431. In addition, the Page mistake found at the Web of Science dataset comparing Page end, Volume and Issue was corrected.

4.2 Missing values

Having checked and removed the duplicated entries from the datasets, they were examined for missing values. All the variables available were observed for missing values, although the field of interest were Title, DOI, Author, Year, Volume, Issue, Page start, Page end. The number of those observations was recorded along with combinations of relevant variables. Combinations of those relevant variables that are not described, had no match. The results are depicted below in the matrices, separately for Scopus and Web of science datasets.

Table 9: Scopus missing values

Variables	Count
Author	14
Title	0
Publication Name	0
ISO source Abbreviation	0
Abstract	1161
Authors' Keywords	2591
language	0
Document type	0
Document type 2	0
Times cited	0
Cited references	320
Author Address	134
DOI	138
Reprint Address	816
Funding Agency & Grant Number	2318
ISSN	0
Part Number	65
Page start	56
Page end	144
Publisher	1991
Bibliographic database	0
Volume	65
Year	0
Author university	170
First Author university	0
AU_UN_NR	2624
SR_FULL	0
SR	0

Table 10: Missing values among combinations of relevant variables in Scopus dataset

Variables	Count
DOI - Page end	1
Author - Issue - Page start - Page end - Volume	2
Author - Page end	10
Issue - Page start - Page end - Volume	56
Issue - Volume	65

Contemplating those variable combinations of missing values for Scopus dataset, some things were recorded. The entry of the first combination, DOI – Page end, was an introduction to a special issue dated on 2003. The entries of the second combination, Author – Issue – Page start – Page end – Volume, were journal covers of Volume 53 dated on 2018. The entries of the third combination, Author – Page end, contained again the 2 of the covers mentioned before and 3 style requirements dated from 1983 to 1985. 3 of them were ERRATA dated on 1988, 2002 and 2018, 1 was a DISCUSSANT dated on 1973 and the last of them was named after a university dated on 1980. The entries of the fourth combination, Issue – Page start – Page end – Volume, were all articles dated on 2018 with document type equal to “Article in press”. Lastly, the entries of the combination, Issue – Volume, were the observations mentioned before with document type equal to “Article in press” in addition to 9 more entries dated in 2018 with the same document type.

Table 11: Web of science missing values

Variables	Count
Author	0
Title	0
Publication name	0
ISO source Abbreviation	0
Abstract	1249
Keywords associated by ISI database	1228
Language	0
Document type	0
Document type 2	0
Times cited	0
Cited references	137
Author address	37
DOI	53
Publisher address	0
Funding Agency & Grant Number	2294
Funding text	2294
ISSN	0
ISSUE	0
Page start	3
Page end	88
Publisher	0
Volume	0
Year	0
Unique article identifier	0
Cited reference count	0
Research areas	0
Usage count	0
Web of science categories	0
E-mail address	1613
Document delivery number	0
Reprint address	37
Bibliographic database	0
Author university	584
First Author university	0
AU_UN_NR	2431
SR_FULL	0
SR	0

Table 12: Missing values among combinations of relevant variables in Web of science dataset

Variables	Count
DOI - Page start - Page end	3
DOI - Page end	31

Contemplating those variable combinations of missing values for Web of science dataset, some things were recorded. The entries of the first combination, DOI – Page start – Page end, were 2 having document type equal to “editorial material” dated on 2003 and 2006 while the other 1 had document type equal to “correction” and dated on 2002. The entries of the second combination, DOI – Page end, were the entries mentioned before with 3 others having document type equal to “editorial material” dated on 1971, 1996 and 2000, 2 having document type equal to “correction, addition” dated on 1972 and 1988. The rest 23 of them had document type equal to “meeting abstract”, with 22 of them dated on 1977 and 1 on 1976.

Chapter 5

Connecting the datasets

Having processed the two datasets individually for finding duplicates and missing values, created the foundation for merging them. By doing that, information about the nature of the data were provided, along with similarities and differences. Succeeding the process, would correspond to a merged dataset which could be used as a tool for a more precise bibliometric analysis.

For merging 2624 Scopus observations with 2431 Web of science observations, common variables of the two datasets were used, with different combinations of those. Variables used, were DOI, Title, Volume, Issue, Page start and Page end.

The first variable used was DOI. Before merging the two datasets, observations having missing values at DOI were removed, because those would match each other resulting in multiple wrongly connected entries. The result was a matrix of 2632 observations, having 2232 matched entries, while those that did not included in the procedure were 138 from Scopus and 53 from Web of science. The rest of the data that did not match, 392 from Scopus and 199 from Web of science, including those entries with missing DOI that excluded before, merged using the variable of Title. In this case titles that repeated in each dataset were excluded for avoiding non-equilibrium observations from merging. Those observations were 26 from Scopus dataset and 3 from Web of science dataset. So, merging those 366 Scopus with 196 Web of science observations, gave a matrix of 417 observations, having 145 matched entries.

The process of merging continued with the same pattern, using different combinations of the last variables. The first combination used was Volume, Issue, Page start and Page end for the observations that did not match with the previous methods, 221 from Scopus and 51 from Web of science. There were 56 observation with missing values in each of the variables, so they excluded from each process followed. Adding those titles that did not included in the previous procedure, formed a Scopus dataset of 191 observations and 54 of Web of science. Merging those, gave a matrix of 209 observations, having 36 matched entries. The second combination of variables used, was Volume, Issue and Page start. Scopus dataset observations that did not match with the previous method, were 155 and 18 from Web of science. Merging those, gave a matrix of 172 observations, having 1 matched entry. Finally, the last combination used was Volume, Issue and Page end, merging the rest of the data that did not match. Those were 210 form Scopus and 17 from Web of science. The result was matrix of 224 observations, having 3 matched entries and 221 unmatched.

Concluding with the procedure described, five matrices were created, containing matched entries, each one based on different variables. Connecting those, gave a matrix of 2638 observations, having 2417 entries matched. Below is a matrix containing the results of each method used.

Table 13: Merge of the two datasets

Variables used for merge	Observations included	Observations excluded	results
DOI	Scopus: 2486 WoS: 2378	Scopus: 138 WoS: 53	Total: 2632 Matched: 2232
Title	Scopus: 366 WoS: 196	Scopus: 26 WoS: 3	Total: 417 Matched: 145
Volume, Issue, Page start, Page end	Scopus: 191 WoS: 54	Scopus: 56	Total: 209 Matched: 36
Volume, Issue, Page Start	Scopus: 155 WoS: 18	Scopus: 56	Total: 172 Matched: 1
Volume, Issue, Page end	Scopus: 210 WoS: 17	Scopus: 56	Total: 168 Matched: 3
Merging the above			Total: 2638 Matched: 2417

Chapter 6

Evaluation of the merged dataset

Having a file containing both the matched and unmatched observations from the two datasets, the adequacy of those were checked. To do that, the results from the merged dataset evaluated using two factors, the matched and unmatched observations. The procedure used to assess those, is described in this chapter.

6.1 Unmatched observations

The unmatched observations from the merged dataset were 221, 207 from Scopus and 14 from Web of Science. To check if there were any possible combinations of entries that could be matched, the dataset with the least observations were chosen. By doing that, 2 observations were found, that in both datasets had symbols instead of numbers at the fields of “Page start” and “Page end”. After checking that these two observations were the same using the archives of the journal [5], [6], the variables of page were changed at the initial datasets of Scopus and Web of science and the procedure mentioned above for merging the two datasets was repeated all over.

The result of merging anew the datasets, was a matrix with 2636 observations. From those, 2419 entries matched while distinct observations of Scopus were 205 and Web of Science 12.

6.2 Matched observations

Having matched the two datasets with every possible combination that could be found and having evaluated that there are no more entries that could be matched, the matched observations could be evaluated for checking the result of the procedure followed.

As far as this evaluation is concerned, the variables used to match the datasets were “DOI”, “Title”, “Volume”, “Issue”, “Page start” and “Page end”. In some cases, instead of having different values in each dataset, there were missing values. The results of comparing those variables and combinations of those were depicted in the matrix below.

Table 14: Differences in matched observations

Scopus variable	Web of science variable	Different values	Scopus missing values	Web of science missing values	Both missing values
Page start	Page start	8	0	2	0
Page end	Page end	44	0	2	81
Page start/end	Page start/end	4	0	2	0
DOI	DOI	6	135	43	3
Issue	Issue	0	4	0	0
Volume	Volume	0	4	0	0
Page start/end, Volume, Issue	Page start/end, Volume, Issue	0	0	0	0
Author	Author	368	5	0	0
Year	Year	21	0	0	0
Title	Title	1000	0	0	0

Using those variables mentioned above, the total number of observations that were different was 1229, while 241 observations had missing values so they could not be matched.

6.3 Differences between matched observations

Having found the inconsistences between the matched observations, the archives of the journals [5], [6] was used in order to check which of the given information was correct. Below there is the analysis done about the differences that were spotted, at the variables of interest “Title”, “DOI”, “Page start”, “Page end”, “Volume” and “Issue”.

6.3.1 Page start

At the *Page start* variable, Scopus dataset had 5 wrong values while the corresponding Web of science values were correct. From those, *4* were on volume 53, *2* from issue 2 and *2* from issue 4. All these *4* had Page start equal to 1. The other one was on volume 49 issue 3 and differed for 10 pages.

On the other hand, Web of science had *3* wrong values, with no consistency about the nature of the value, while those Scopus entries were correct. *2* of them differed for a few pages and the other one had a symbol in front of the page.

They were also *2* cases were Web of science dataset had missing values, at one introduction on volume 38 issue 1 and one errata on volume 37 issue 4.

6.3.2 Page end

At the *Page end* variable, Scopus dataset had *30* wrong entries while the corresponding Web of science values were correct. *4* of them were the same as mentioned before on volume 53 where the pages completely differed from the correct ones. There was an entry which differed for 20 pages on volume 38 issue 1. The rest of the entries had a difference for 1 to 2 pages from the correct ones.

Web of science had *14* wrong entries while the matched ones from Scopus dataset were correct. *1* entry differed for 54 pages on volume 11 issue 2. *3* entries for 10 pages, on volume 10 issue 2, volume 8 issue 1 and volume 7 issue 5. 3 entries had symbols instead of numbers, like U8, U7 and \& and those were on volume 44 issue 6, volume 34 issue 4 and volume 7 issue 2. The rest of them differed for 1 to 2 pages from the correct ones.

Web of science also had *2* cases where there were missing values on Page end, at one correction and one note, on volume 25 issue 3 and on volume 9 issue 4 respectively.

Moreover, there were *81* cases where both Scopus and Web of science had missing values at the Page end. From those 4 were Corrections, 3 Articles, 1 Editorial material, 2 Letters, 4 Notes and the rest of them were Meeting abstracts. Those entries were correct, since they were papers with one page only.

6.3.3 DOI

At the *DOI* variable, in Scopus dataset there were *3* wrong entries while the corresponding entries of Web of science were correct. The *3* of them were completely different from the correct *DOI*. *2* of them were on volume 41 issue 2 and the other one on volume 40 issue 4.

In Web of science there were also *3* wrong entries, while those of Scopus were correct. In the first one there were a missing “0”, in the other one a “G” was placed wrongly and the last of them

was completely wrong. Those were respectively on volume 53 issue 1, volume 44 issue 3 and volume 11 issue 1.

In addition, Scopus had *135* missing values at the *DOI* variable. There was no observable distinctive pattern for why these were missing. One of them dated on 1998 and the others from 2004 until 2008. Web of science on the other hand, had *43* entries with missing values at *DOI*. Most of them dated from 1970 until 1995 while two of them in 2002 and 2005. There were also *3* cases where both Scopus and Web of science had missing *DOI*. Those were dated on 1998, 2002 and 2003. The reason that these missing values were appeared may be that they were old entries and the information was not been saved.

6.3.4 Issue and Volume

As far as the *issue* and *volume* is concerned, Scopus dataset had 4 missing values, which was the cases mentioned before, *2* on volume 53 issue 4 and *2* on issue 2. Those corresponding entries on Web of science dataset were correct.

6.3.5 Year

In the *year* variable, Scopus dataset had *21* wrong values, where Web of science was correct. In *11* cases the year instead of 2015 was 2016 and in 10 cases the year instead of 2011 was 2010.

6.3.6 Author

The next step of the analysis was about the *Author names*. Comparing the common observations of Scopus and Web of science, found 368 differences. In general, Scopus *Author names* were more inaccurate that those of Web of science, having *273* against *93* while in 2 cases both were wrong. For correcting those values, the archives were advised once more to see which of the two datasets had the correct *Author name.* The two datasets abbreviated the names, using the last name of the author and then the first letters of each of his/her names.

There was a problem with some names, like those of Chinese authors that they were abbreviated differently than the others which made the process of correcting and unifying those difficult. Scopus dataset abbreviated the names using the first part of their name, while Web of science was using both parts of it. The way Web of science noted them made those Authors' names distinct and more accurate to the archive, it was taken as correct.

At the Scopus dataset, there were *79* spelling mistakes. At same cases where there was a letter from a different alphabet, Scopus dataset had no letter on its behalf, while Web of science dataset use a corresponding letter from the English alphabet. *105* of those Author names were wrongly abbreviated. At those cases the names didn't matched the ones the archive had or the letters for abbreviating the Author's name was wrong. Those Chinese Authors' names mentioned before are included here. There were also *6* cases where in the middle of the abbreviated name it was the characters "JR" without being part of the Author's name. *12* cases had wrong order of the Authors, *1* case had one more Author, 3 cases had an Author's name twice and there were *67* cases missing 1 or 2 Author names.

At the Scopus dataset, there were also *5* cases with Author name "NA NA", where Web of science dataset given one. Checking those from the archive it was found that 1 of them didn't have a name on the archive, at 3 of them the name of the author couldn't be deduced and in 1 case the name was given at the archive.

At the Web of science dataset, there were *26 cases* with spelling mistakes and *35* wrongly abbreviated names. 3 cases had a missing Author and at 29 cases there was an Author name duplicated. Duplicated Author names had cases only at the 6,7,8,9 volumes of the Journal.

Finally, there were 2 cases where both datasets had to be corrected. In the first case Scopus dataset had a spelling mistake, while Web of science dataset had wrongly abbreviated the name. In the second case Scopus dataset had wrongly abbreviated the name, while Web of science missed a space between the two parts of the Author's name.

6.3.7 Title

Between merged articles, there were *1000* differences. Web of science dataset had *961* wrong titles, Scopus dataset had *15* and *24* of them were common mistakes. In all cases when the one dataset was wrong the other one was correct, except those *24* cases where both were wrong.

The Scopus dataset wrong titles had some categories of mistakes. *3* of them had different word placement or differed by one word from the correct one. *1* of the missed a space between words. *6* of them missed a symbol, like comma or question mark. *5* of them had spelling mistakes.

The Web of science dataset also had some categories of mistakes. *13* of them had spelling mistakes or spacing between words. *836* of them had missing or different symbols or words from the correct one, which was the most common mistake in the dataset. Those were dash instead of colon, colon in cases where it was not, use of word "and" instead of "&", missing notations like question mark or comma, using numbers instead of letters, missing a word or letter like "A" or

“The” and miss dots between country abbreviations like “US” instead of “U.S.”. Moreover, there were *112* cases having different title or wrong order of the words used.

Finally, the *24* common mistakes of both the datasets, had missing colon between word or missing quotation marks including a word.

Chapter 7

Conclusion

The outcomes of the analysis offered some inferences about the process undertaken, while indicating the different peculiarities each database had.

The process of the analysis did not used much time, because the corrections needed to be made were much less than the whole sample. In addition, the sample itself increased, since Scopus databases provided originally, 2629 observation and Web of Science database 2431, while the final sample had 2636 observations. Evaluating the databases became easier since they were correcting each other, having correct values where the other had not, which provided a higher precision value to the outcome.

So, merging the databases of Scopus and Web of Science for the Journal of Financial and Quantitative Analysis made possible, increasing the sample's total number, accuracy, consistency and facilitated the evaluation of observations.

Chapter 8

Further Research

Having analyzed the bibliometric differences between the two databases, ideas about how the analysis could continue were made.

The same procedure used in this analysis, can be reiterated at more than two databases. Doing that, would show if the accuracy and volume of the records about the same journal increase, while understanding if the process adds any more significance to the results.

Undergoing a bibliometric analysis with two merged databases would be a next step, after having finished with this analysis. This would provide information about accuracy of the process done in this analysis while finding probable obstacles, in relation to the names of the authors, the document types and titles of the articles.

Considering the above, recognizing the pattern that each data base records their ‘easy to mistake variables’, like authors, title and document type, will provide the base for homogenizing those before starting a bibliometric analysis.

References

[1] "Wikipedia page: Bibliometrics," [Online]. Available: https://en.wikipedia.org/wiki/Bibliometrics. [Accessed 5 11 2019].

[2] A. Andres, Measuring Academic Research. How to undertake a Bibliometric Study, Chandos Publishing, 2010.

[3] "Official website: Journal of financial and quantitative analysis," [Online]. Available: https://jfqa.org/. [Accessed 6 11 2019].

[4] "Cambridge unviersity press: Journal of finacial and quantitative analysis," [Online]. Available: https://www.cambridge.org/core/journals/journal-of-financial-and-quantitative-analysis. [Accessed 6 11 2019].

[5] "Cambridge university press: Journal of financial and quantitative analysis archive," [Online]. Available: https://www.cambridge.org/core/journals/journal-of-financial-and-quantitative-analysis/all-issues. [Accessed 6 11 2019].

[6] "Jstor: Journal of financial and quantitative analysis archive," [Online]. Available: https://www.jstor.org/journal/jfinaquananal. [Accessed 6 11 2019].

[7] "Wikipedia: Journal of financial and quantitative analysis," [Online]. Available: https://en.wikipedia.org/wiki/Journal_of_Financial_and_Quantitative_Analysis. [Accessed 6 11 2019].

[8] "Wikipedia: Foster School of Business," [Online]. Available: https://en.wikipedia.org/wiki/Foster_School_of_Business. [Accessed 6 11 2019].

[9] "Official website: Scopus," [Online]. Available: https://www.scopus.com/. [Accessed 6 11 2019].

[10] "Official website: Web of science," [Online]. Available: http://wokinfo.com/. [Accessed 6 11 2019].

[11] "Wikipedia: Scopus," [Online]. Available: https://en.wikipedia.org/wiki/Scopus. [Accessed 6 11 2019].

[12] "Wikipedia: Web of science," [Online]. Available: https://en.wikipedia.org/wiki/Web_of_Science. [Accessed 6 11 2019].

[13] "Official website: Elsevier," [Online]. Available: https://www.elsevier.com/. [Accessed 6 11 2019].

[14] "Wikipedia: RStudio," [Online]. Available: https://en.wikipedia.org/wiki/RStudio. [Accessed 6 11 2019].