Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2019) 1547-1562

ECCOMAS
Proceedia

COMPDYN 2019
7th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Crete, Greece, 24–26 June 2019

LITERATURE REVIEW OF HISTORICAL MASONRY STRUCTURES WITH MACHINE LEARNING

Vagelis Plevris[1], **German Solorzano**[1], **Nikolaos Bakas**[2]

[1] OsloMet—Oslo Metropolitan University
Department of Civil Engineering and Energy Technology
Pilestredet 35, Oslo 0166, Norway
e-mail: vageli@oslomet.no, germanso@oslomet.no

[2] Neapolis University Pafos
School of Architecture, Land & Environmental Sciences
2 Danais Avenue, 8042 Paphos, Cyprus
n.bakas@nup.ac.cy

Abstract

The objective of this work is to use machine learning techniques to analyze a big database of papers published since 1990 regarding the conservation of masonry buildings considered as historical heritage. In this study, a database of nearly three thousand papers obtained from Scopus database [1] were investigated. In a first stage, the papers are analyzed using basic statistics in order to describe the evolution of the research during the past thirty years. In the second stage, Genetic Algorithms (GA) were implemented to create bibliometric maps for the visualization of co-word analysis that was performed using the provided database. The obtained diagrams constitute comprehensive maps of relevant characteristics among the investigated literature such as similarities between author keywords or author names. The maps are constructed using a rigorous methodology that involves the mapping of each item (for example, a keyword) to a two-dimensional point. The distances between the items represent their dissimilarity for a specific characteristic. The numerical procedure for the construction of the bibliometric map involves an optimization task solved by means of GA. The obtained result is a powerful tool for data analysis, which provides a deep insight of relevant characteristics of a large database in a very short time.

Keywords: Multidimensional Scaling, Bibliometric Mapping, Co-Word Analysis, Optimization, Knowledge Management, Masonry, Historical Structures.

doi: 10.7712/120119.7018.21053

1 INTRODUCTION

The number of research articles published in scientific journals or conference proceedings has shown an exponential growth during the last decades. Journal articles first appeared in 1665 and according to the work of Arif E. Jinha [2] there were around 50 million scholarly articles in existence already in 2010 and the number has grown significantly since then. Bornmann and Mutz [3] investigated the rate at which science has grown since the mid-1600s, identifying three essential growth phases in the development of science. A recent study by Van Noorden [4] showed that the evolution of global scientific output is equivalent to a doubling every nine years, on average.

Scopus [1] is Elsevier's abstract and citation database launched in 2004 and it is available online by subscription. It is the largest abstract and citation database of peer-reviewed literature, with bibliometrics tools to track, analyze and visualize research. Scopus contains abstracts and citations for academic journal articles, covering nearly 36,377 titles from approximately 11,678 publishers, of which 34,346 (as of May 2019) [5] are peer-reviewed journals in top-level subject fields: life sciences, social sciences, physical sciences and health sciences. According to Elsevier, Scopus has 55 million records dating back to 1823 where 84% of these contain references dating from 1996.

A simple Scopus search within the subject area of "Engineering" (query string "SUBJAREA (engi)") gave us 13,274,640 document results in a query made in June 2019, with the first paper being published in 1861 and the last to be published in 2020 . Figure 1 shows the evolution of these papers in time with the focus period being 1900-2018 (12,977,414 papers in this period). The blue line shows the number of papers published each year and the orange line shows the cumulative sum since 1900. The dotted red line shows the trend as a line in the graph, where of course the y-axis is in logarithmic scale.

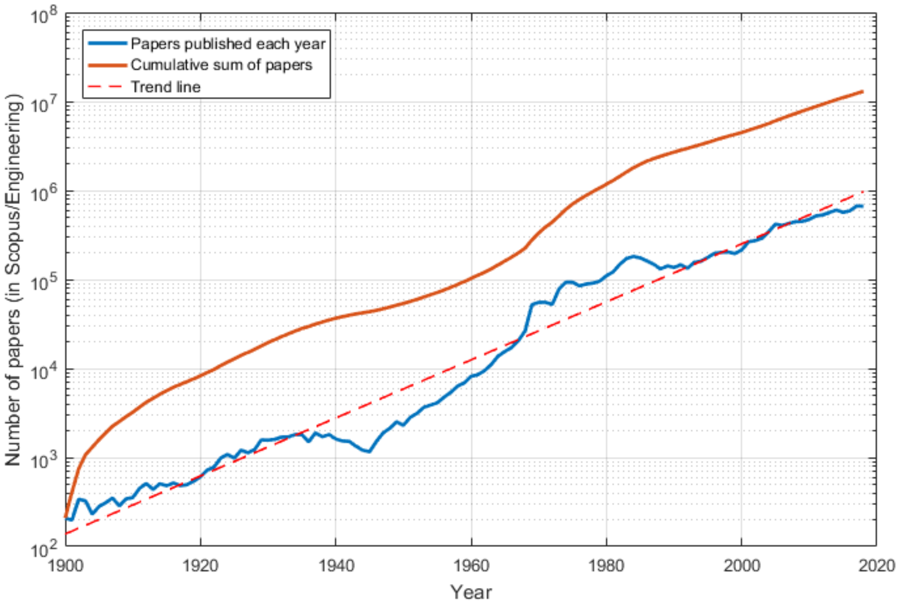


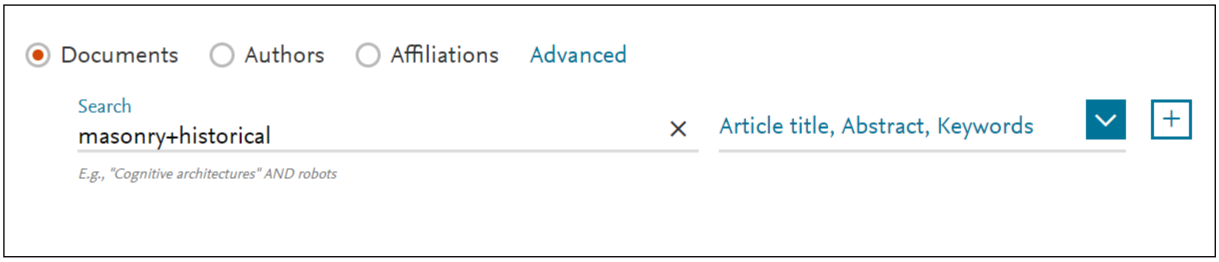
Figure 1: Basic search of the Scopus database with the search terms "masonry+historical".

For writing a literature review paper, a researcher nowadays needs to analyze a vast amount of research papers which is a highly demanding task. It is nearly impossible to read all the relevant papers and manually extract all the important information using traditional reading methods and the task keeps getting even harder as the production of scientific papers continues to grow exponentially. To deal with this problem, new automated techniques have evolved, called bibliometric analysis, bibliometrics, scientometrics, scientific mapping etc., where with the aid of computer algorithms, an automated analysis of a vast amount of research papers is possible. These techniques can significantly help an individual researcher on exploring the literature, writing literature reviews and can even automate, to some extent, these processes [6].

The present study is a continuation and improvement of a previous relevant work [7], but the application and focus is now on a slightly different subject. The objective is to use automated machine-learning techniques to analyze a big database of papers published since 1990 regarding the conservation of masonry buildings considered as historical heritage. A database of nearly three thousand papers obtained from Scopus database [1] were investigated. In a first stage, the papers are analyzed using basic statistics in order to describe the evolution of the research during the past thirty years. In the second stage, Genetic Algorithms (GA) were implemented to create bibliometric maps for the visualization of co-word analysis that was performed using the provided database. Bibliometric maps take into account associations among keywords, authors or others (e.g. references), through their distances on a two-dimensional map, revealing significant information about how the objects studied are inter-related, i.e. appearing simultaneously in research papers. The obtained diagrams constitute comprehensive maps of relevant characteristics among the investigated literature such as similarities between author keywords or author names. The maps are constructed using a rigorous methodology that involves the mapping of each item (for example, a keyword) to a two-dimensional (x, y) point. The distances between the items represent their dissimilarity for a specific characteristic. The numerical procedure for the construction of the bibliometric map involves an optimization task solved by means of GA [8]. The obtained result is a powerful tool for data analysis, which provides a deep insight of relevant characteristic of a large database in a very short time.

2 PAPERS DATASET AND DATA COLLECTION METHODOLOGY

The present study uses data that have been collected from the Scopus database [1]. Figure 2 shows a basic Scopus search, where we search for the terms “masonry+historical” in the *Article title*, *Abstract*, *Keywords*. The equivalent query string is “TITLE-ABS-KEY (masonry+historical)”. **TITLE-ABS-KEY** is the default search field in Scopus, where **TITLE** is the title of a manuscript, **ABS** is its abstract (a condensed summary of the full-text) and **KEY** are the Author keywords or Index keywords, to be explained later.



● Documents ○ Authors ○ Affiliations Advanced

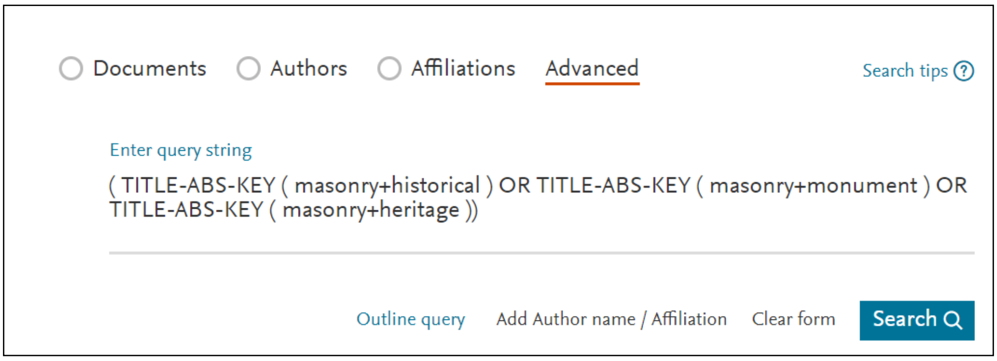
Search: masonry+historical ×

Article title, Abstract, Keywords

E.g., "Cognitive architectures" AND robots

Figure 2: Basic search of the Scopus database with the search terms “masonry+historical”.

Figure 3 shows an advanced Scopus search, where we search for the terms “masonry+historical” or “masonry+monument” or “masonry+heritage” in the Article title, Abstract, Keywords. The equivalent query string is “TITLE-ABS-KEY (masonry+historical) OR TITLE-ABS-KEY (masonry+monument) OR TITLE-ABS-KEY (masonry+heritage)”.



(TITLE-ABS-KEY (masonry+historical) OR TITLE-ABS-KEY (masonry+monument) OR TITLE-ABS-KEY (masonry+heritage))

Figure 3: Advanced search with the search terms
“masonry+historical” OR “masonry+monument” OR “masonry+heritage”.

In the study, we decided to limit the search to the years from 1990 (inclusive) to 2019 (inclusive). It has to be noted that the indexing of year 2019 is obviously not complete, which may also be the case for the year 2018, due to delays in indexing. Nevertheless, we decided to use both these years in our study to be able to include the latest trends in the field. The full query used is the following:

```
( TITLE-ABS-KEY ( masonry+historical )  OR  TITLE-ABS-KEY ( masonry+monument ) OR  TITLE-ABS-KEY ( ma-
sonry+heritage ))  AND  (LIMIT-TO ( PUBYEAR ,  2019 )  OR  LIMIT-TO ( PUBYEAR ,  2018 )  OR  LIMIT-TO
( PUBYEAR ,  2017 )  OR  LIMIT-TO ( PUBYEAR ,  2016 )  OR  LIMIT-TO ( PUBYEAR ,  2015 )  OR  LIMIT-TO
( PUBYEAR ,  2014 )  OR  LIMIT-TO ( PUBYEAR ,  2013 )  OR  LIMIT-TO ( PUBYEAR ,  2012 )  OR  LIMIT-TO
( PUBYEAR ,  2011 )  OR  LIMIT-TO ( PUBYEAR ,  2010 )  OR  LIMIT-TO ( PUBYEAR ,  2009 )  OR  LIMIT-TO
( PUBYEAR ,  2008 )  OR  LIMIT-TO ( PUBYEAR ,  2007 )  OR  LIMIT-TO ( PUBYEAR ,  2006 )  OR  LIMIT-TO
( PUBYEAR ,  2005 )  OR  LIMIT-TO ( PUBYEAR ,  2004 )  OR  LIMIT-TO ( PUBYEAR ,  2003 )  OR  LIMIT-TO
( PUBYEAR ,  2002 )  OR  LIMIT-TO ( PUBYEAR ,  2001 )  OR  LIMIT-TO ( PUBYEAR ,  2000 )  OR  LIMIT-TO
( PUBYEAR ,  1999 )  OR  LIMIT-TO ( PUBYEAR ,  1998 )  OR  LIMIT-TO ( PUBYEAR ,  1997 )  OR  LIMIT-TO
( PUBYEAR ,  1996 )  OR  LIMIT-TO ( PUBYEAR ,  1995 )  OR  LIMIT-TO ( PUBYEAR ,  1994 )  OR  LIMIT-TO
( PUBYEAR ,  1993 )  OR  LIMIT-TO ( PUBYEAR ,  1992 )  OR  LIMIT-TO ( PUBYEAR ,  1991 )  OR  LIMIT-TO
( PUBYEAR ,  1990 ) )
```

The query was made on **4 June 2019** and returned **3293 results** (papers) in total. Scopus provides a lot of information for each entry (paper), which includes but is not limited to the following: *Authors*, *Title*, *Year*, *Source title*, *Volume*, *Issue*, *Cited by*, *DOI*, *Authors with affiliations*, *Abstract*, *Author keywords*, *Index Keywords*, *Publisher*, *ISSN*, among others. The full information was extracted first in csv format and then it was converted to MS Excel xlsx compressed format.

2.1 Papers per year

Figure 4 shows the total number of papers published for each year (left vertical axis, blue color). Not all these papers contain keywords. The solid blue line shows all papers, while the dashed blue line shows only the papers with (author) keywords. Especially in the past or depending on the journal/conference where the paper is published, some papers have no keywords at all. If a paper has no keywords, then it is not included in the sum of the dashed line, but it is included in the one of the solid line. In our case, out of 3293 papers, 2437 of them have (author) keywords while the others (856 papers) have no keywords.

We see that there is a significant increase in the number of papers published from 1990 to 2019 in the field. In 1990, only 3 papers had been published in the area, while the corresponding number of papers for 2017, 2018 and 2019 is 359, 354 and 259, respectively. The decrease in 2019 (and possibly 2018) is because indexing for these years is not yet complete in Scopus.

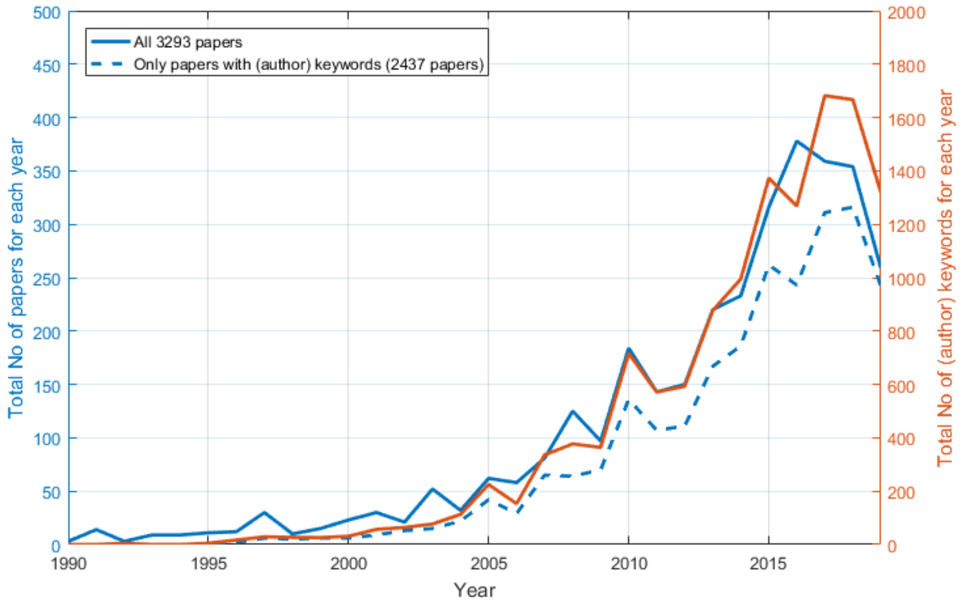


Figure 4: Total number of papers and total number of keywords, for each year (1990-2019).

3 KEYWORD ANALYSIS

3.1 Author keywords vs Index keywords

Scopus distinguishes two kinds of keywords: Author keywords and Index keywords. Author keywords are chosen by the author(s) as the keywords which, in their opinion, best reflect the contents of their document, while Index keywords are keywords chosen by content suppliers and are standardized based on publicly available vocabularies. Unlike Author keywords, the Index keywords take into account synonyms, various spellings, and plurals. Scopus has no influence over either Author or Index keywords because these are both determined by third parties.

AUTHKEY is the term used for author keywords, **INDEXTERMS** is used for only index keywords, while **KEY** is used for both author keywords and index keywords. In the present study, we used both types of keywords (with KEY) in our initial search in the database. So, the search result (3293 papers) is obtained based on both types of keywords. Yet, in the continuation of the study, **we use only the Author keywords** in our analysis, unless otherwise stated.

Figure 4 (above) depicts the total number of (author) keywords of papers for each year (right vertical axis, orange color). A significant increase is revealed regarding the number of keywords of published papers, following the same trend as the papers. In 1997, the total number of keywords of published papers was 29 (with 30 published papers of which 6 had author keywords), while the corresponding number for 2017, 2018 and 2019 is 1682, 1667 and 1319, respectively.

Figure 5 shows the average number of keywords per paper for each year. The blue line considers all papers (with or without author keywords), while the orange line takes into account only the papers with keywords. The orange line shows that the keywords/paper ratio remains almost constant with values ranging from 5 to 6 in most cases. This means that authors use 5 to 6 keywords (on average) in most cases, lately and also in the past. The blue line can be misleading as it implies that more keywords are used lately, but this is not the case. The increasing trend of the blue line is because of the relativity decreasing number of papers with no author keywords at all, over the years (in the past, relativity more papers had no keywords), which makes the average ratio go up.

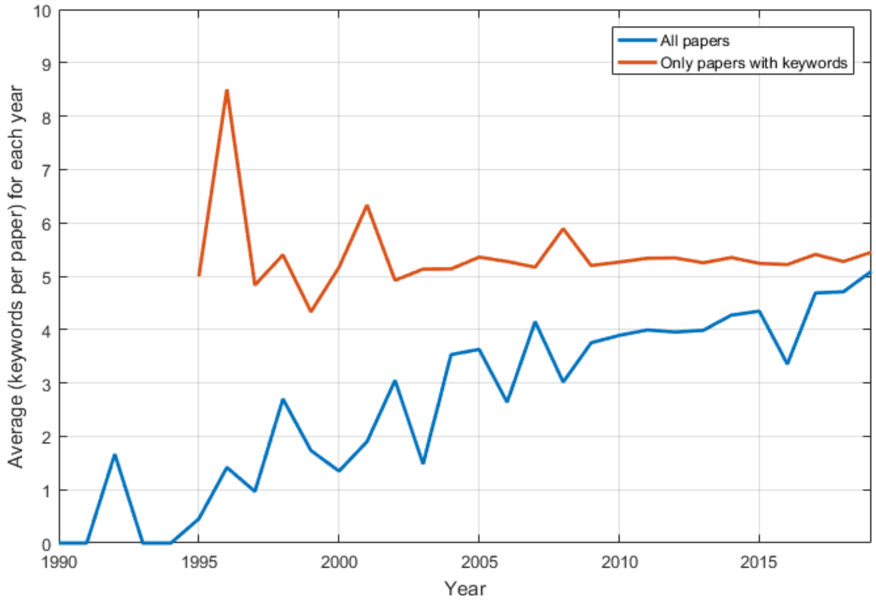


Figure 5: Average number of (keywords per paper), for each year (1990-2019).

3.2 Top keywords

Figure 6 presents the number of occurrences of the top-15 keywords in the total 3293 papers (2437 papers with keywords) in the period 1990-2019 (30 full years). The top-5 keywords are "masonry", "cultural heritage", "seismic vulnerability", "masonry structures" and "strengthening", as shown in the figure. "Masonry" outperforms all other keywords, which is expected and makes sense due to the nature of the query made, as the word "masonry" was present in all the three alternatives of the search ("masonry+historical" or "masonry+monument" or "masonry+heritage").

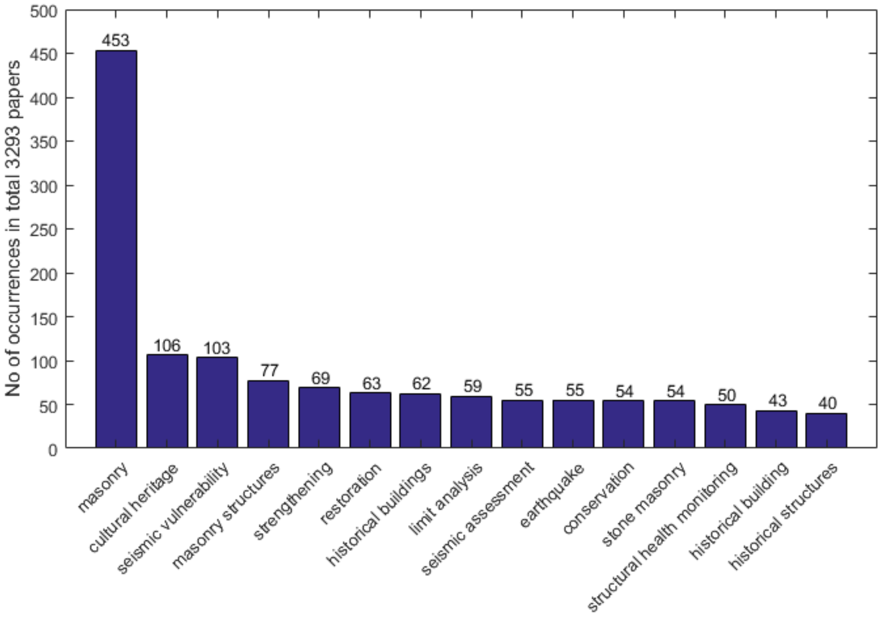


Figure 6: Number of occurrences of each of the top-15 keywords in the total 3293 papers.

3.3 Time series of the top keywords

Figure 7 depicts the time series of the occurrences of the top-10 keywords for all 3293 papers in the period 1990-2019, i.e. how many times each keyword appeared in each year. In Figure 7, it is shown that all keywords exhibit an increasing trend in their occurrences from past to present. This is mainly due to the general increase in the number of papers and keywords, as we approach from the past years to the latest years (see Figure 4).

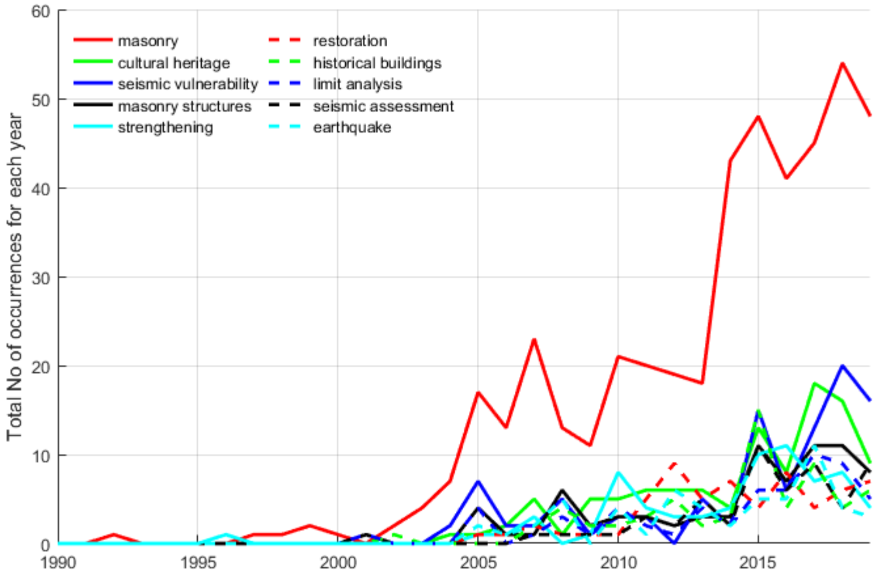


Figure 7: Total number of occurrences of each of the top-10 keywords, for each year.

Figure 8 presents the same time series in a normalized way, where the number of occurrences of each keyword has been divided by the total number of papers (with keywords) for each year. Papers with no keywords have not been taken into account in this figure. Thus, Figure 8 presents the time series of the *average* number of keyword occurrences *per paper*, for each of the top-10 keywords, for a more objective representation. Since before 1995 there were years with no papers with keywords, to avoid division by zero problems, in this figure we focus on years from 1995 until 2019.

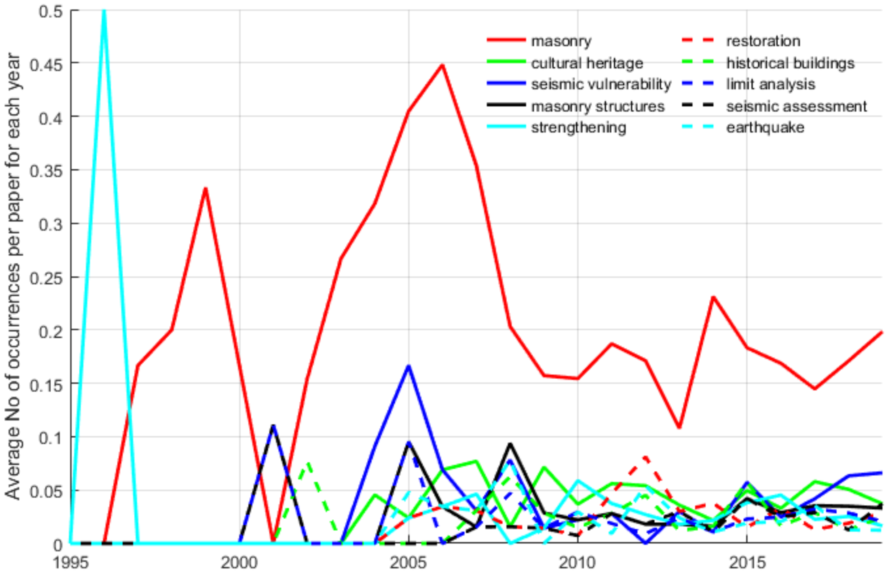


Figure 8: Average number of keyword occurrences per paper (with keywords), for each of the top-10 keywords, for each year.

Figure 9 depicts the matrix of Pearson Correlation coefficients for the normalized time series presented in Figure 8. Correlation between sets of data is a measure of how well they are related. In the particular case of Figure 9, if a cell has a value close to 1 (close to yellow color) then there is a strong positive relationship between the normalized time series of the corresponding keywords, i.e. the two time series show similar behavior (either increase together or decrease together with time). If a cell has a value close to -1 then again there is a strong relationship between the time series of the corresponding keywords, but in the opposite direction, i.e. the occurrence of one keyword increases with time while the other decreases. If a cell has a value close to zero then there is no correlation between the corresponding keywords.

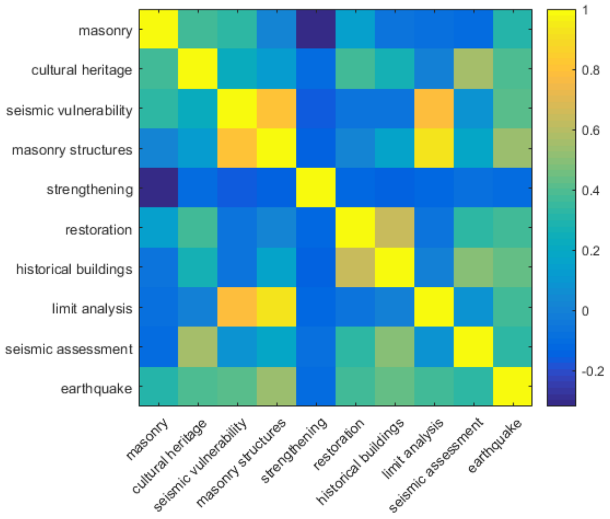


Figure 9: Color representation of the Pearson Correlation coefficients for the normalized time series of the top-10 keywords.

By examining Figure 9 it can be observed that the time series of some specific keywords have a good correlation with some others, while for many others there is no correlation. For example, the time series of the pairs “seismic vulnerability” and “limit analysis” show a strong correlation with *r*=0.78. The normalized time series for these two keywords are depicted in Figure 10 where it is clearly shown that there is a correlation between the two.

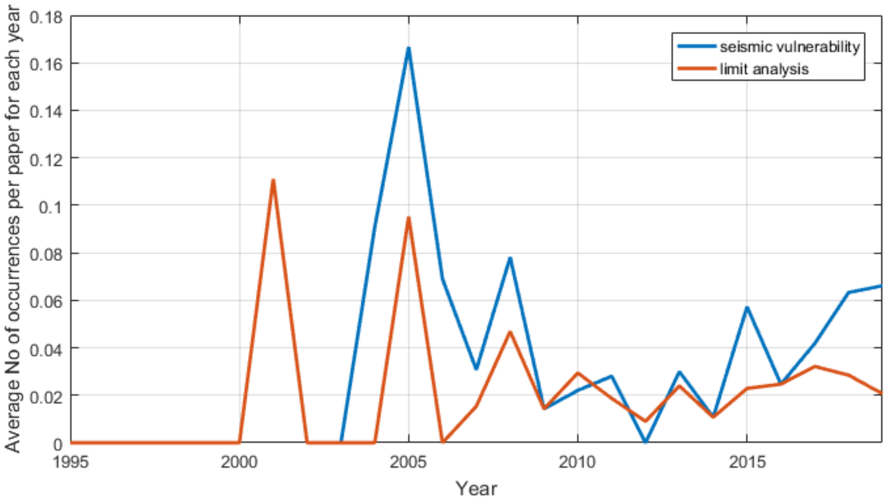


Figure 10: Average number of (keywords per paper), for each year (1990-2019).

3.4 Co-occurrence Matrix for the top keywords

It is interesting to observe the co-occurrence of keywords in papers. Some keywords tend to have simultaneous occurrences (be present in the same paper) while others tend not to co-exist. The co-occurrence matrix for the top-10 keywords is depicted in Figure 11 where the colors correspond to a scale from zero to 32, indicating the number of simultaneous occurrences of the keywords in papers of the database. The diagonal of the matrix has been set to zero for better presentation of the results.

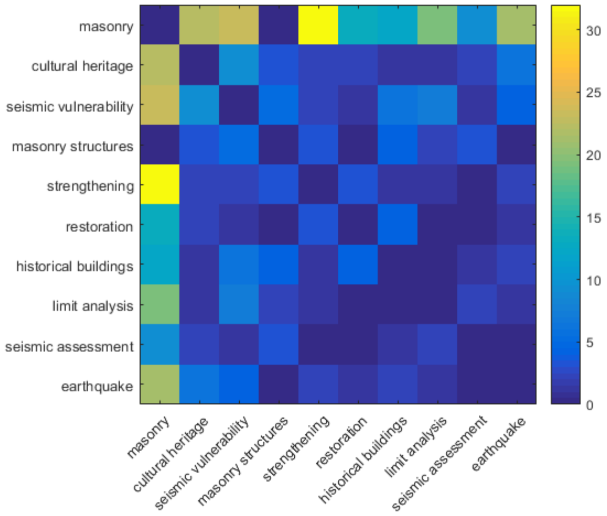


Figure 11: Colored representation of the co-occurrence matrix for the top-10 keywords.

The importance of the co-occurrences stems from their link to the conceptual association among the keywords. For example, the pair of keywords “masonry” and “strengthening” has a high number of co-occurrences (32), which is also the case for the pairs “masonry” and “cultural heritage” (22) and “masonry” and “seismic vulnerability” (23) which clearly indicates that the masonry related literature deals with the improvement of the structural characteristics of existing rather than new buildings. Similarly, the co-occurrence of the keywords “cultural heritage” and “seismic vulnerability” is 9, while the association of “seismic vulnerability” with “limit analysis” is 7. These results are significant from a statistics perspective, since they are based on a large number of papers (2437 papers with keywords) which cover a wide range of years (practically 1995-2019 since before 1995 there are many years with no papers with keywords).

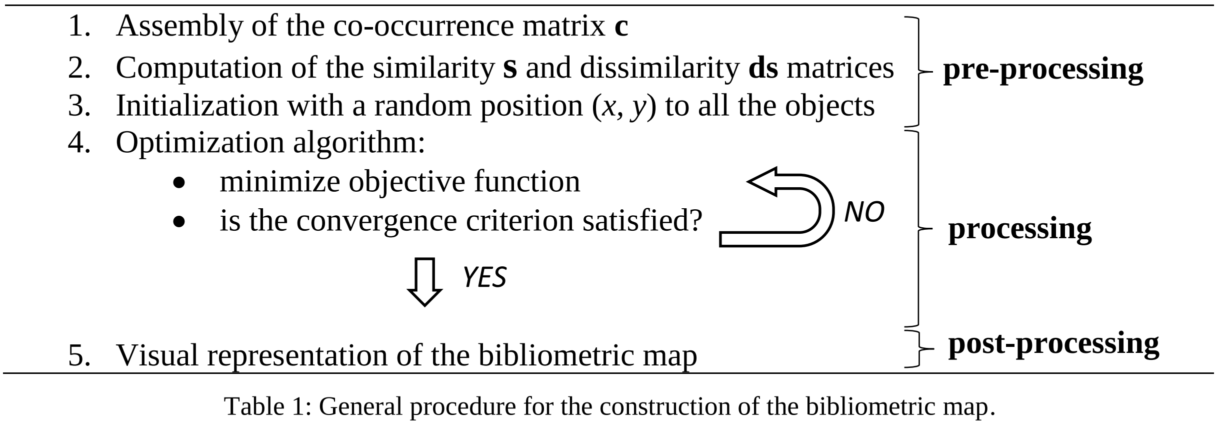
Although meaningful and interesting conclusions can be made just by looking at the co-occurrence matrix of Figure 11 (which depicts the co-occurrence of the top-10 keywords only), it is very difficult to interpret this matrix globally, especially if the total number of unique keywords (6222 items in our case) was taken into account which would make the co-occurrence matrix extremely large (6222×6222). For a high number of keywords, the associations between them can be further analyzed utilizing the bibliometric maps methodology which will be presented in the following section.

4 BIBLIOMETRIC MAPS

A bibliometric map is a visual representation of the solution of the multidimensional scaling problem [9]. It is based on the assembly and further processing of the co-occurrence matrix. The procedure is generic; thus, the term “object” will be used to denote either keywords, authors or references from the studied database. This work considers maps in two-dimensions and each object represents a point with coordinates (*x*, *y*). A pair of objects that have a high value of co-occurrence should be close to each other whereas objects with low co-occurrence must be distant from one another. For a set of *n* objects, there is a total of

$$t = \frac{n \cdot (n-1)}{2} \tag{1}$$

unique pair of combinations (or distances) between them. The exact solution of this problem usually does not exist due to the impossibility of matching *t* specific distances in only two-dimensions. Instead, optimization algorithms are applied to find approximated solutions. The general procedure implemented for the construction of the bibliometric map is presented in Table 1 followed by a detailed description of each step.



1. Assembly of the co-occurrence matrix **c**
2. Computation of the similarity **s** and dissimilarity **ds** matrices
3. Initialization with a random position (*x*, *y*) to all the objects
4. Optimization algorithm:
    - minimize objective function
    - is the convergence criterion satisfied?
5. Visual representation of the bibliometric map

Table 1: General procedure for the construction of the bibliometric map.

Similarity Matrix: In order to facilitate the construction of the maps, the co-occurrence matrix **c** is scaled to a range of [$s_{\min}$, $s_{\max}$] with a linear transformation. The scaled values of the co-occurrence matrix are denoted as the similarity matrix **s** and each term is computed as

$$s_{ij} = \frac{s_{\min}(c_{\max} - c_{ij}) + s_{\max}(c_{ij} - c_{\min})}{c_{\max} - c_{\min}} \quad (2)$$

where $c_{\max}$ and $c_{\min}$ are the maximum and minimum values of the co-occurrence matrix **c**. In this study, for the case of the top-15 keywords, it is $c_{\min}$=0 and $c_{\max}$=32. The parameters $s_{\min}$ and $s_{\max}$ have been set to $s_{\min}$=0.25 and $s_{\max}$=10 as will be described in detail later on.

Dissimilarity Matrix: The terms of the dissimilarity matrix **ds** are computed from the similarity matrix **s** as follows

$$ds_{ij} = \frac{1}{s_{ij}} \quad (3)$$

There are other methods that could be used to define the dissimilarity values from the similarity values, for example a linear transformation. Nevertheless, the above non-linear transformation has been chosen as an attempt to consider the influence of the low average number of co-occurrences among all objects into the solution. This comes from the fact that as more objects are studied, there are more low values (or even zeros) in the **c** matrix. A plot of the function of Eq. (3) is given in Figure 12.

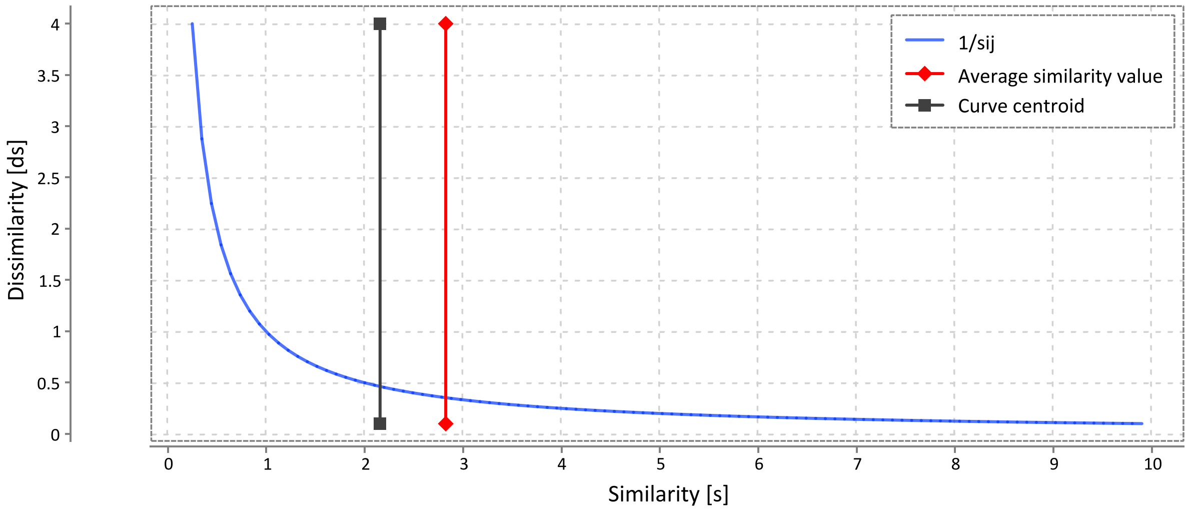


Figure 12: Non-linear mapping of similarity values into dissimilarity values for $s_{\min}$=0.25 and $s_{\max}$=10.

The average value of similarity and the centroid of the curve are represented with vertical lines. These two lines lie close to one another and it is an indicator that Eq. (3) suits the task. It provides low dissimilarity values between objects with a higher similarity than the average and high dissimilarity values to objects with lower similarity than the average.

The parameters $s_{\min}$ and $s_{\max}$ have a high influence on the shape of the curve. In our case, by using $s_{\min}$=0.25 and $s_{\max}$=10, the maximum and the minimum distances between any two points in the map are fixed to 1/10=0.1 and 1/0.25=4 units, respectively. Thus, avoiding zero or infinite values which can produce difficulties in the graphical representation or the numerical procedure. Furthermore, these values have shown better performance among other tested sets of parameters.

Optimization Problem: For the construction of the bibliometric map, the objective function to minimize is the mean absolute error. It is calculated by comparing, for each pair of objects, their real dissimilarity value with their measured distance in the two-dimensional map. In order to restrict the search space and to facilitate the visualization, a constraint is imposed that limits the coordinate values to a range of zero to two times the maximum dissimilarity value. The optimization problem is then formulated as follows

$$\text{minimize:} \quad f(\boldsymbol{x}) = \frac{1}{t}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{\left|ds_{ij}-d_{ij}\right|}{ds_{ij}}, \tag{4}$$

$$\text{with:} \quad \boldsymbol{x} = \{x_1, y_2, x_1, y_2, \ldots, x_n, y_n\}$$

$$d_{ij} = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2}$$

$$\text{subjected to:} \quad 0 \le x_i \le {}^{2}\!/_{s_{\min}} \quad , \quad 0 \le y_i \le {}^{2}\!/_{s_{\min}}$$

Considering that the matrix *ds* is symmetric, and that the information of the diagonal is omitted since the similarity of an object with itself is meaningless to the problem; therefore, to reduce the number of operations, the summation in Eq. (4) is only computed for the non-repeated values of *ds* and when $i \neq j$.

Optimization Algorithm: The Non-dominated Sorting Genetic Algorithm (NSGA-II) [8] was used to solve the optimization problem. The implementation was done by adapting the open source package MOEA Framework [10]. Genetic Algorithms are a key element of machine learning techniques and they use operators inspired in nature such as mutation, crossover and selection. By adjusting these operators properly, they can generate high quality solutions for all kind of optimization problems.

The vector of decision variables ***x*** is first populated with random values and then it is processed by the optimizer. A maximum number of 10,000 function evaluations has been chosen as the termination criterion. It has been noticed, for this particular problem, that a higher number does not improve the results.

4.1 Bibliometric map for the top-15 keywords

Utilizing the described approach, a bibliometric map is constructed to analyze the co-occurrence of the top-15 keywords (see Figure 6) in the provided database. The results are shown in Figure 12.

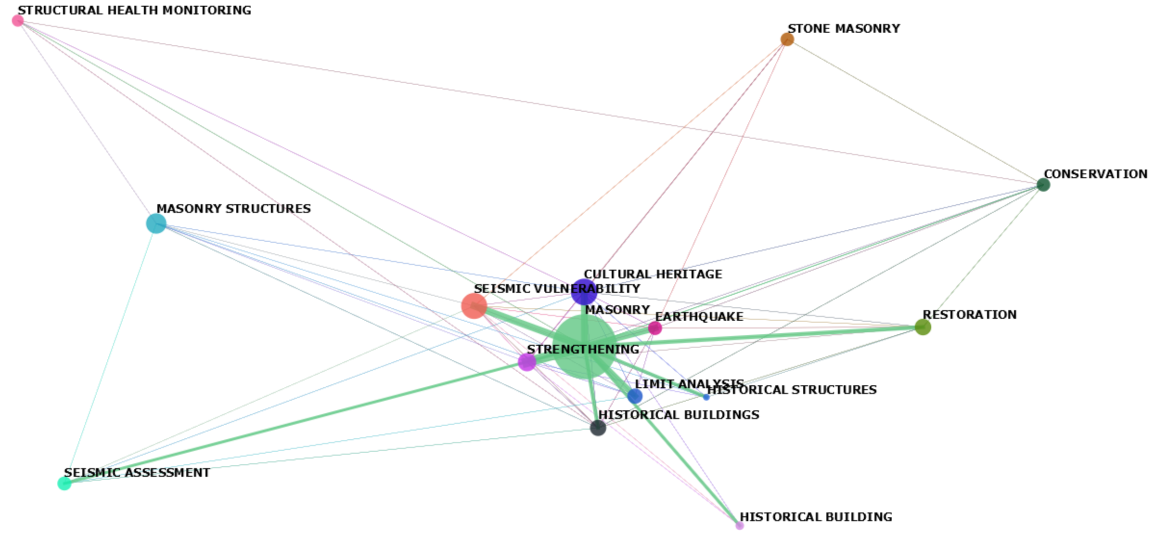


Figure 13: Bibliometric map for the top-15 keywords.

The characteristics of the map are summarized below:

- Each object (i.e. keyword in the case of Figure 13), is represented as a bubble with a position (x_i, y_i) and a specific color.
- The distance between any two objects is an indicator of their dissimilarity.
- The area of each bubble is proportional to its number of occurrences.
- The co-occurrences between each pair of objects is represented with a line between the objects. The thickness of each line can be proportional to the number of co-occurrences between the connected objects. In Figure 13 the latter characteristic (line thickness) is only implemented and shown for the central keyword “masonry” to avoid image clutter and for better clarity.

All the information is embedded on the map and a clear visual representation is achieved. As expected, the keyword “masonry”, which was used as the main query to obtain the database, appears in the middle surrounded by words that share a high similarity value with it (high number of co-occurrences).

According to the map, the keyword “masonry” is closely related to the keywords “strengthening”, “earthquake”, “cultural heritage” and “historical buildings”. Thus, a quick conclusion can be made that most of the literature regarding the conservation of masonry historical buildings, deals with their rehabilitation due to the damage received during earthquakes, especially since keywords related to earthquakes were not included in the original query.

4.2 Bibliometric map for the top-50 keywords

The analysis of even more keywords can reveal additional information and lead to the identification of new patterns, especially if a proper digital visualization is employed e.g. with zooming and panning capabilities.

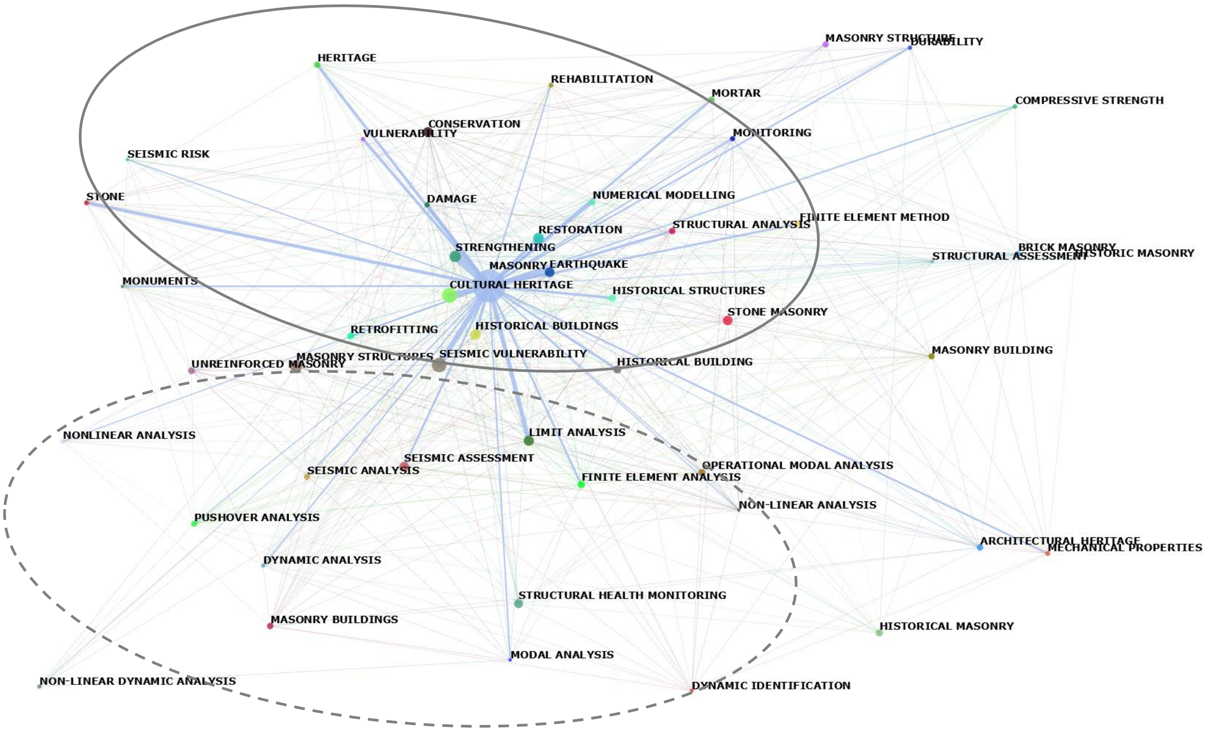


Figure 14: Bibliometric map for the top-50 keywords.

A new bibliometric map for the top-50 keywords is shown in Figure 14 using the same visual characteristics as described before. The position of all the keywords shows concordance with the dissimilarity matrix, therefore, the proposed optimization method also proves to be effective for large numbers of decision variables.

By inspecting the obtained map, some areas have been manually identified containing keywords close to each other (clusters) that can be related to specific areas of research. For example, the solid-lined cluster in Figure 14 contains words related to the restoration of masonry buildings, such as “heritage”, “vulnerability”, “stone masonry”, “historical building” and “rehabilitation”. On the other hand, the dashed-lined cluster has mostly words associated with structural dynamics, like “pushover analysis”, “finite element analysis”, “non-linear analysis”, “dynamic analysis”, among others. This observed characteristic is a further indicator of a successful implementation.

4.3 Bibliometric map for the top authors

Following the same procedure, a bibliometric map is constructed to analyze the top-50 authors and their co-occurrence (i.e. co-authorship) among papers of the provided database. The obtained map is shown in Figure 15. In this investigation, all papers (3293 papers) have been taken into account, even the one with no author keywords.

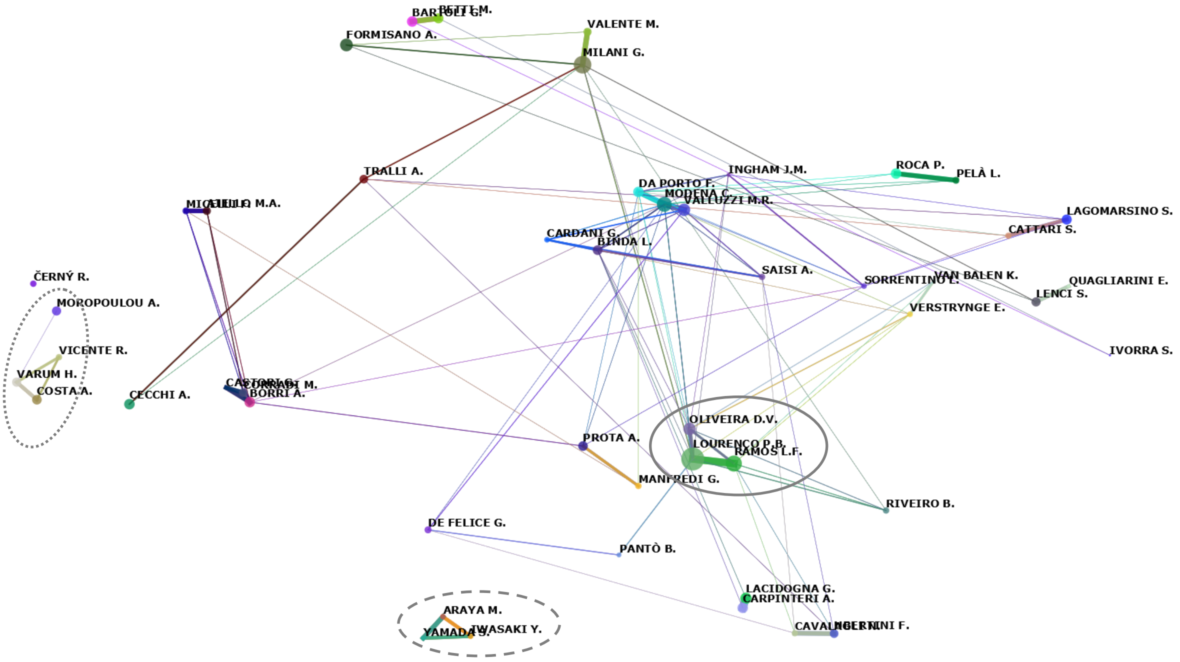


Figure 15: Bibliometric map for the top-50 authors.

By browsing the map, well-defined clusters of specific authors can be found. For example, in the middle cluster defined by a solid line in the figure, “Laurenco P.B.”, “Ramos L.E.” and “Oliveira D.V.” appear to be strongly related (“Laurenco P.B.” being the most significant contributor with a high number of co-authored papers) and share multiple connections to other authors/clusters. Similarly, in the bottom area (dashed-lined cluster) “Yamada S.”, “Araya M.”, and “Iwasaki Y.” also share strong connections but only with each other, thus, creating an isolated research group. A similar group is also identified on the left with the dotted line. Such connectivity suggests the existence of large and small research groups working in the field.

4.4 Computational time and efficiency

The java computer language was used for this implementation because of its vast set of tools (either native or external open-sourced) for mathematics, two-dimensional rendering and the creation of user-friendly interfaces. All numerical tests were carried out on a regular personal computer with an Intel i7-6700HQ @2.60GHz processor and 16 GB or RAM.

Due to the stochastic nature of the genetic algorithms and the use of random numbers, a different solution is found every time the algorithm is executed. Furthermore, the obtained objective value (or error) varies depending also on the number of objects studied. The error can never be expected to be minimized to zero for all practical applications, because of the two-dimensional limitation of the problem (2D map). In addition, multiple solutions exist for the same problem, i.e. corresponding to a different rotation of the map. Rotation is a transformation of the map for which the value of the objective function remains unchanged, but the positions of the objects will be different. Thus, it can be said that every execution of the algorithm will produce similar maps, but with a different random rotation.

As the number of objects increases, the pre-processing, processing and post-processing parts demand more computations; however, the most time-consuming task is the optimization procedure which consumes about 80% of the total execution time. According to Eq. (4), the required operations that must be performed by the optimizer grow exponentially. Figure 16 shows the needed computational time vs the total number of objects considered and confirms the increasing trend.

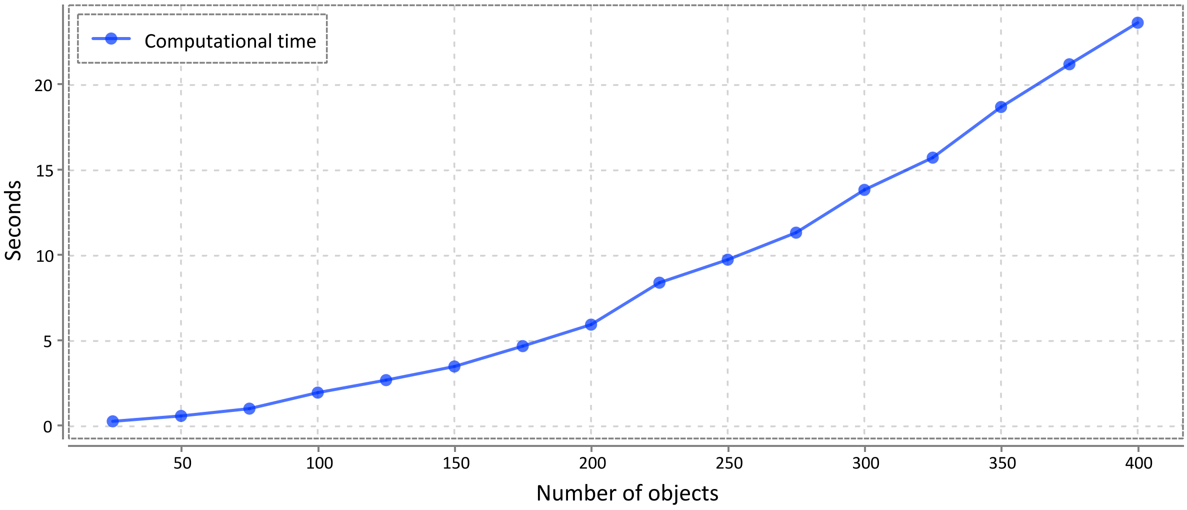


Figure 16: Computational time required to construct the bibliometric map.

5 CONCLUSIONS

A new approach to the multidimensional scaling problem using Genetic Algorithms was applied for the construction of bibliometric maps. With the presented numerical procedure, a sophisticated tool was developed for the analysis of large databases of research papers, with the aim of identifying relevant characteristics such as the associations between keywords, authors, references, etc. The database used contained 3293 papers downloaded from Scopus database [1]. The word “masonry” in combination with “historical”, “monument” and “heritage” was used as query. In the case of the author maps, all papers have been taken into consideration, while in the case of keyword maps, only the 2437 papers containing author keywords were taken into account. The data is processed, and the map is rendered in just a few seconds revealing interesting features. For example, by looking into the co-occurrence of keywords, it was found that seismicity and numerical modeling are topics closely related to the investigated query. This strongly suggest that the damage in historical masonry buildings comes from mostly earthquake excitations. Additionally, by analyzing the co-authorship, groups of researchers working together in the field or in isolation were identified.

The quality of scientific research can be improved by using machine learning techniques for bibliometrics. Therefore, the outlook of this work is to improve the presented methodology and the developed software, in order to create a powerful tool with an intuitive graphical user interface that allows a quick processing of any database and a smooth visualization of the results.

REFERENCES

[1] Elsevier. *About Scopus*. Available from: https://www.elsevier.com/solutions/scopus [Accessed on 24 May, 2019],

[2] Jinha, A.E., *Article 50 million: an estimate of the number of scholarly articles in existence.* Learned Publishing, 2010. **23**(3): p. 258-263.

[3] Bornmann, L. and R. Mutz, *Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.* Journal of the Association for Information Science and Technology, 2015. **66**(11): p. 2215-2222.

[4] Van Noorden, R. *Global scientific output doubles every nine years*. Available from: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html [Accessed on 15 Mar., 2017], 07 May 2014.

[5] Wikipedia: The Free Encyclopedia. *'Scopus'*. Available from: https://en.wikipedia.org/wiki/Scopus [Accessed on 24 May 2019,

[6] van Eck, N.J. and L. Waltman, *Software survey: VOSviewer, a computer program for bibliometric mapping.* Scientometrics, 2010. **84**(2): p. 523-538.

[7] Plevris, V., et al., *Literature review of masonry structures under earthquake excitation utilizing machine learning algorithms*, in *6th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering (COMPDYN 2017)*. 2017, Eccomas Proceedia: Rhodes, Greece. p. 2685-2694.

[8] Deb, K., et al., *A fast and elitist multiobjective genetic algorithm: NSGA-II.* IEEE Transactions on Evolutionary Computation, 2002. **6**(2): p. 182-197.

[9] Borg, I. and P.J.F. Groenen, *Modern Multidimensional Scaling*. 2 ed. Springer Series in Statistics. 2005, New York: Springer-Verlag.

[10] *MOEA Framework*. Available from: http://moeaframework.org/ [Accessed on 24 May, 2019],