Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2017) 2685-2694

ECCOMAS
Proceedia

COMPDYN 2017
6th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Rhodes Island, Greece, 15–17 June 2017

LITERATURE REVIEW OF MASONRY STRUCTURES UNDER EARTHQUAKE EXCITATION UTILIZING MACHINE LEARNING ALGORITHMS

Vagelis Plevris[1], Nikolaos Bakas[2], Gro Markeset[1] and John Bellos[2]

[1] Oslo and Akershus University College of Applied Sciences
Department of Civil Engineering and Energy Technology
Pilestredet 35, Oslo 0166, Norway
e-mail: {vagelis.plevris, gro.markeset}@hioa.no

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

Keywords: multidimensional scaling, bibliometric mapping, co-word analysis, citation network analysis, optimization, knowledge management.

Abstract. *This work aims to analyze and reveal critical features of the papers published since 1990 on the topic of masonry structures under earthquake loading. In particular, detailed information for nearly three thousand papers (exactly 2909) was extracted from the Scopus database [1], and investigated in two stages. Initially, the papers were analyzed in terms of simple statistics and keyword time series – as either raw or normalized data – in order to describe the evolution of the relevant research during the past twenty-seven years (1990-2016, inclusive). In a second phase, bibliometric maps of the papers were developed, regarding their similarities with respect to a variety of the papers' characteristics such as: author keywords and author names. The resulting diagrams constitute comprehensive maps of the relevant literature, with respect to the associations among the particular characteristics. The bibliometric maps were constructed based on a rigorous methodology, which converts each item (for example, keyword) to a two-dimensional (x, y) point on the bibliometric map. These distances between items reflect the dissimilarities between them, for a particular characteristic. The numerical procedure involved in the construction of the map is a constrained optimization problem which was formulated and solved with an efficient methodology.*

Peer-review under responsibility of the organizing committee of COMPDYN 2017.
doi: 10.7712/120117.5598.18688

1 INTRODUCTION

The number of research papers published in scientific journals during the last years has shown an exponential growth. Bornmann and Mutz [2] investigated the rate at which science has grown since the mid-1600s, identifying three essential growth phases in the development of science, which each led to growth rates tripling in comparison with the previous phase: from less than 1% up to the middle of the 18th century, to 2 to 3% up to the period between the two world wars, and 8 to 9% to 2010 [2]. Lately, the evolution of global scientific output, is equivalent to a doubling every nine years on average, as shown in a recent study by Van Noorden [3]. Thus, for writing a literature review paper, a researcher nowadays needs to analyze a vast amount of research papers which is a highly demanding task to do solely by reading the papers and manually extracting any important information. The task keeps getting even harder as the production of scientific papers continues to grow exponentially. Accordingly, new automated techniques have evolved, called bibliometric analysis, bibliometrics, scientometrics, scientific mapping etc., where with the aid of computer algorithms, an analysis of a vast amount of research papers is possible. Nowadays, these techniques can significantly help the individual researchers on exploring the literature, writing literature reviews or can even automate, to some extent, these processes [4].

The main purpose of such analyses is to construct bibliometric maps of the scientific field studied. Bibliometric maps, take into account associations among keywords, authors as well as references, through their distances on a two-dimensional map, revealing significant information about how the papers studied are inter-related, i.e. appearing simultaneously in research papers. Thus, the conclusions regarding the scientific field studied, are based on an extended database of papers, and through a rigorous computational procedure, the outcomes are documented precisely. On the other hand, such tools need to be used with caution and care and not to be forced to produce results or identify patterns where they simply do not exist. In a recent study published in Nature Methods [5], it is described how clustering analysis may be misleading, identifying non-existent patterns in data when used the wrong way.

In the present work, bibliometric maps were constructed based on a rigorous methodology, which converts each item to a two-dimensional (x, y) point on the bibliometric map. The distances between items reflect the dissimilarities between them. The numerical procedure is a constrained optimization problem where the objective function of the problem is directly related to the multidimensional scaling error. Thus, the researcher, is able to evaluate the performance of the clustering procedure, for the particular problem studied.

2 PAPERS DATASET AND THE METHODOLOGY TO OBTAIN IT

All the data used in the study have been taken from Scopus [1]. Scopus is a bibliographic database containing abstracts and citations for academic journal articles, covering nearly 22,000 titles from over 5,000 publishers, of which 20,000 are peer-reviewed journals in the scientific, technical, medical, and social sciences [1]. The total number of entries in the database today is more than 69 million, with 1.4 billion cited references dating back to 1970. Scopus is owned by Elsevier and is available online by subscription.

Figure 1: Searching the Scopus database with the query “Masonry+Earthquake”.

First, we searched the database using the query "Masonry+Earthquake", and the option "Article title, Abstract, Keywords" as shown in Figure 1. The query was made on 9 March 2017 and returned 3152 results (papers) in total. These results included also old papers which needed to be removed: 182 papers with year earlier than 1990 and another 11 old papers with no year information (193 old papers in total). They also included 50 new papers from year 2017 which needed also to be removed, since 2017 is not a full year yet and we cannot have complete information for it (all 2017 papers appearing in Scopus) until the first months of 2018. As a result, the final database included 3152-182-11-50=2909 papers in total, covering a period of 27 years, from 1990 (inclusive) to 2016 (inclusive). Scopus provides a lot of information which includes but is not limited to the following for each entry (paper): *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. The final xlsx file (2909 entries) had a size of 6.1 MB.

2.1 Papers per year

Figure 2 shows the total number of papers for each year (left vertical axis, blue color). We see that there is a significant increase in the number of papers published from 1990 to 2016 in the fields of "masonry" and "earthquake". In 1990, only 19 papers had been published in the area, while the corresponding number of papers for 2016 was 290 (and also 307 for 2015).

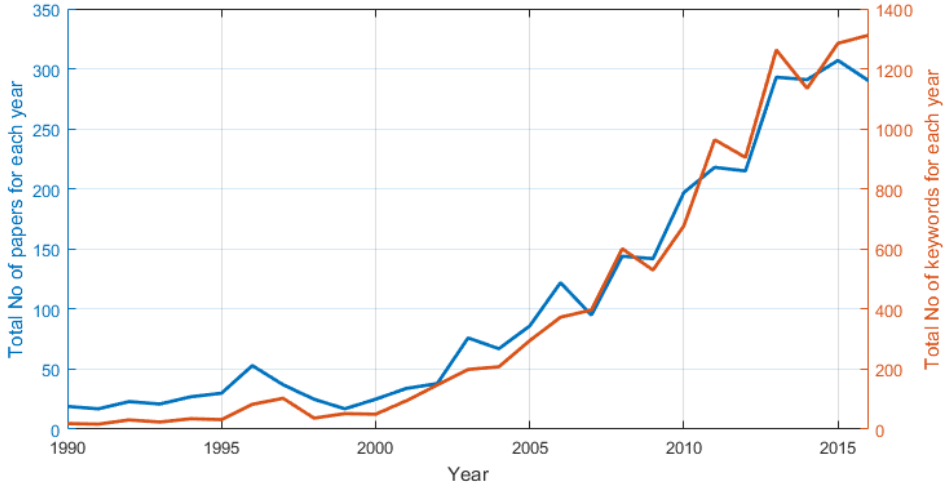

Figure 2: Total number of papers and total number of keywords, for each year (1990-2016).

3 KEYWORD ANALYSIS

Scopus provides rich information on the author keywords of each paper. Figure 2 (above) depicts the total number of 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 103, while the corresponding number for 2016 is 1312.

Another significant observation has to do with the average number of keywords per paper for each year, which is graphically depicted in Figure 3. We see that there is a significant increase in the number of keywords per paper, from a value of 1 for 1990 to a value of 4.52 for 2016. This means that researchers tend to use more keywords to describe their work, compared to the past. The average number of keywords used during the last 5 full years (2012 to 2016, inclusive) is 4.23.

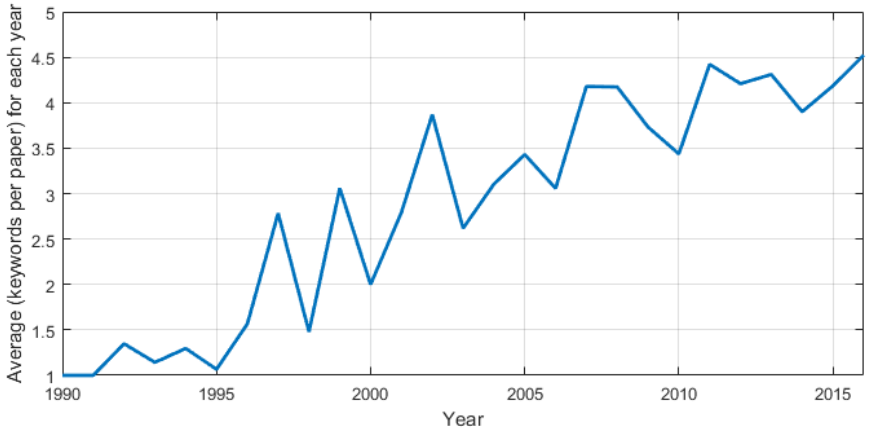

Figure 3: Average number of (keywords per paper), for each year (1990-2016).

3.1 Top-15 keywords

Figure 4 presents the number of occurrences of the top-15 keywords in the total 2909 papers in the period 1990-2016 (27 full years). The top-5 keywords are "masonry", "earthquake", masonry structure", "seismic performance" and "seismic vulnerability", as shown in the figure.

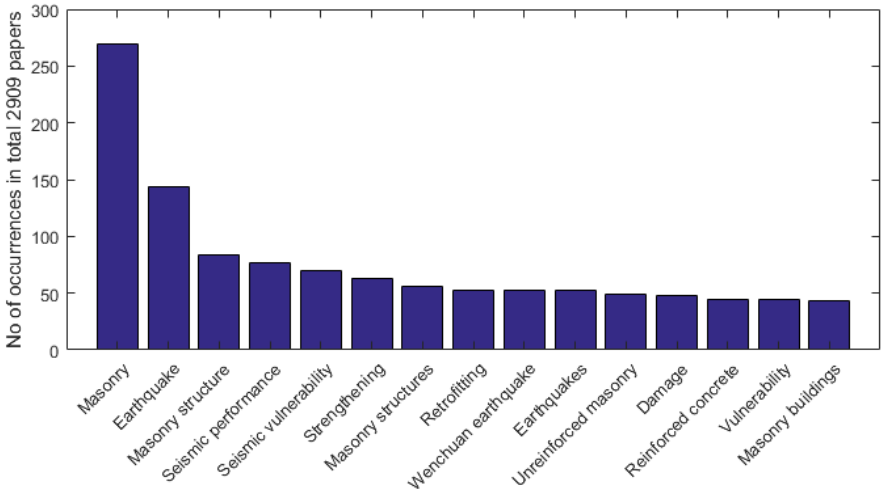

Figure 4: Number of occurrences of each of the top-15 keywords in the total 2909 papers.

3.2 Time series of the top-15 keywords

Figure 5 presents the time series of the occurrences of the top-15 keywords for all 2909 papers in the period 1990-2016, i.e. how many times each keyword appeared for each year. In Figure 5, 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 2).

Thus Figure 5 needs to be "corrected" (normalized) taking into account the number of papers (or keywords) for each year. Figure 6 presents the same time series, where the number of occurrences of each keyword has been divided by the total number of papers for each year. Thus, Figure 6 presents the time series of the average number of keyword occurrences *per paper*, for each of the top-15 keywords, aiming to a more objective interpretation.

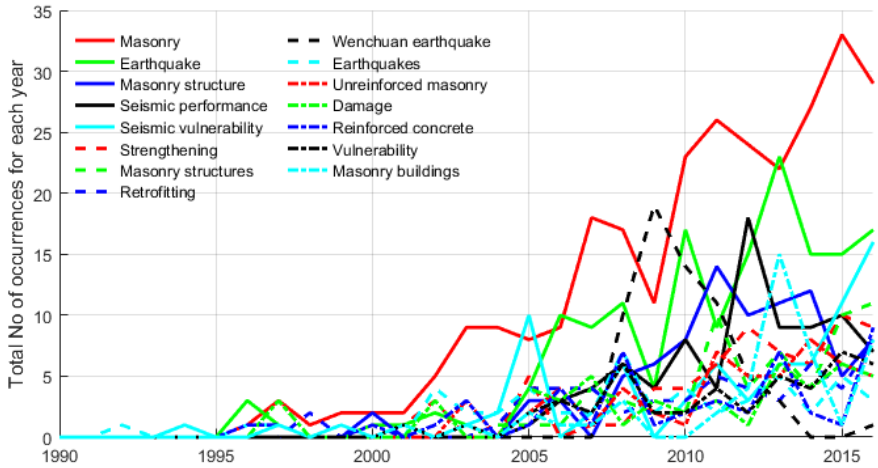

Figure 5: Total number of occurrences of each of the top-15 keywords, for each year.

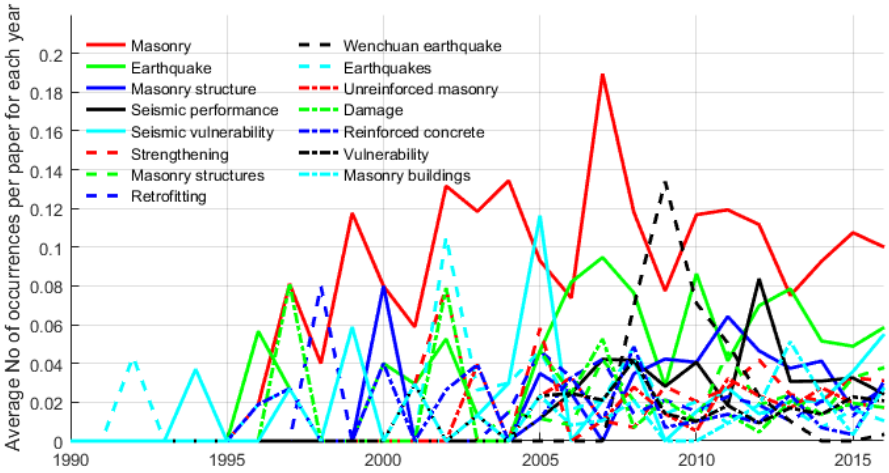

Figure 6: Average number of keyword occurrences per paper, for each of the top-15 keywords, for each year.

Figure 7 depicts the matrix of Pearson Correlation coefficients for the normalized time series presented in Figure 6. Correlation between sets of data is a measure of how well they are related. In the particular case of Figure 7, if a cell has a value close to 1 (close to yellow color) then there is a strong positive relationship between the 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.

By examining Figure 7 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 “earthquake” and “seismic performance” show a strong correlation with r=0.73, which is also the case for “unreinforced masonry” and “vulnerability” with r=0.65.

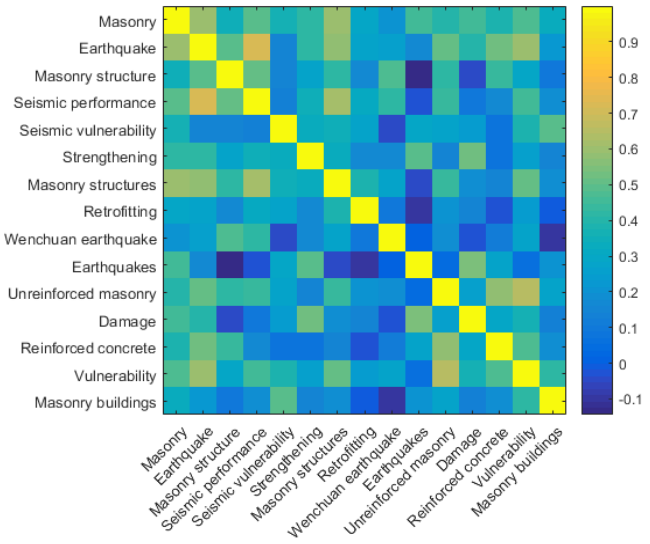

Figure 7: Color representation of the Pearson Correlation coefficients for the time series of the top-15 keywords.

3.3 Co-occurrence Matrix for the top-10 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 8 where the colors correspond to a scale from zero to 45, 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.

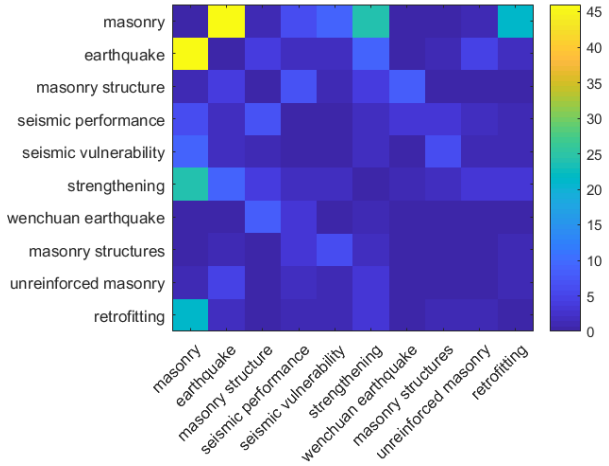

Figure 8: Colored representation of the co-occurrence matrix for the top-10 keywords.

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 (24), which is also the case for the pair “masonry” and “retrofitting” (21), 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 “earthquake” and “strengthening” is 9, while the association of “unreinforced masonry” with “earthquake” is 5, signifying that the seismic performance of masonry structures without reinforcement has comprehensively concerned the researchers.

These results are significant from a statistics perspective, since they are based on a large number of papers (2909) which cover a wide range of years (1990-2016).

Although meaningful and interesting conclusions can be made just by looking at the co-occurrence matrix of Figure 8 (which depicts the co-occurrence of 10 keywords only), it is very difficult to interpret this matrix globally, especially if the total number of unique keywords (4910 items) has been taken into account which would make the matrix extremely large (4910×4910). For such 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 [6, 7], which is based on the assembly and further processing of the co-occurrence matrix. In the present work, the steps of the numerical procedure implemented for the construction of the bibliometric map are presented in Table 1.

1. c_{ij}: The (i,j) entry of the co-occurrence matrix **c**
2. s_{ij}: The (i,j) entry of the similarity matrix **s**
3. ds_{ij}: The (i,j) entry of the dissimilarity matrix **ds**, $ds_{ij} = 1/s_{ij}$
4. d_{ij}: Distance on the map $d_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|$
5. e_{ij}: The (i,j) entry of the error matrix, $e_{ij} = \left| ds_{ij} - d_{ij} \right|$

   Objective function (to minimize): $f = \sum_i \sum_j e_{ij}^2$

   Optimization Algorithm
6. Are the optimality criteria satisfied?

   NO

   YES
7. End => Draw the bibliometric map

Table 1. Steps of the bibliometric map generation algorithm.

The procedure is generic; thus, the term "object" will be used, denoting either keywords, authors or references, etc. The similarity matrix **s** is generated from the co-occurrence matrix **c** with proper normalization and has values in the region [0, 1]. All the matrices **c**, **s**, **ds**, **d** and **e** are symmetric. The vector **x** describes the design variables of the optimization problem. It denotes a point in the 2D space with components (x_1, x_2) or simply (x, y). Ideally, all the errors e_{ij} should be minimized, but this cannot be the case in reality as we try to present complex multi-dimensional relationships in the 2D space.

The algorithm starts with the initial calculation of the co-occurrence table of the studied objects (keywords, authors or references). Accordingly, the similarity and dissimilarity matrices are obtained. Afterwards, the optimization algorithm initializes randomly the positions $\mathbf{x}_i = (x_i, y_i)$ of each object, and computes the distances between the elements on the bibliometric map. The objective function is the sum of squares of the absolute difference between the pairwise distances, and the corresponding dissimilarities. The optimal values of the positions $\mathbf{x}_i$ are utilized to visualize the results on the bibliometric map. For the solution of the optimization problem, the Levenberg-Marquardt (LM) algorithm, also known as the damped least-squares method, has been used, working specifically with loss functions which take the form of a sum of squared errors. The algorithm works efficiently without computing the exact Hessian matrix. Instead, it utilizes the gradient vector and the Jacobian matrix.

4.1 Bibliometric map for the top keywords

By means of the proposed approach, a graphical representation of the links between the keywords is accomplished, through the bibliometric map which is shown in Figure 9.

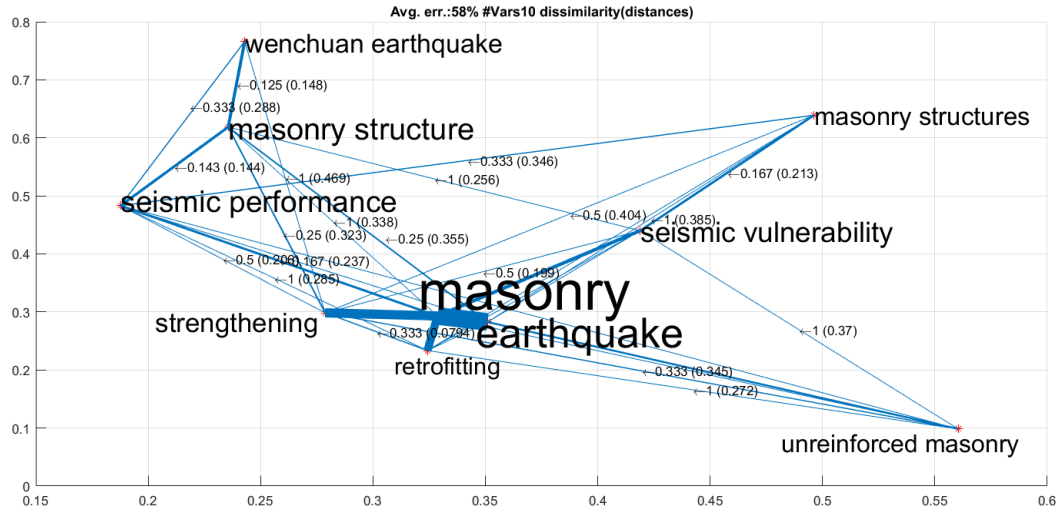

Figure 9: Bibliometric map for the top-10 keywords.

The bibliometric map has the following general attributes:

- Each object (in this particular case: *keyword*) is represented as a point on the 2D map, with its (x, y) coordinates.
- Each object's font size is proportional to the number of its occurrences.
- The objects with co-occurrences are connected with a line. The line thickness represents the link strength, which is proportional to the similarity (co-occurrence) between the objects.
- The distances between the objects are indicators of their dissimilarity. The exact value of the dissimilarity is written in the middle of each link with an indicative arrow (->), while the corresponding distance on the graph is written in parenthesis.

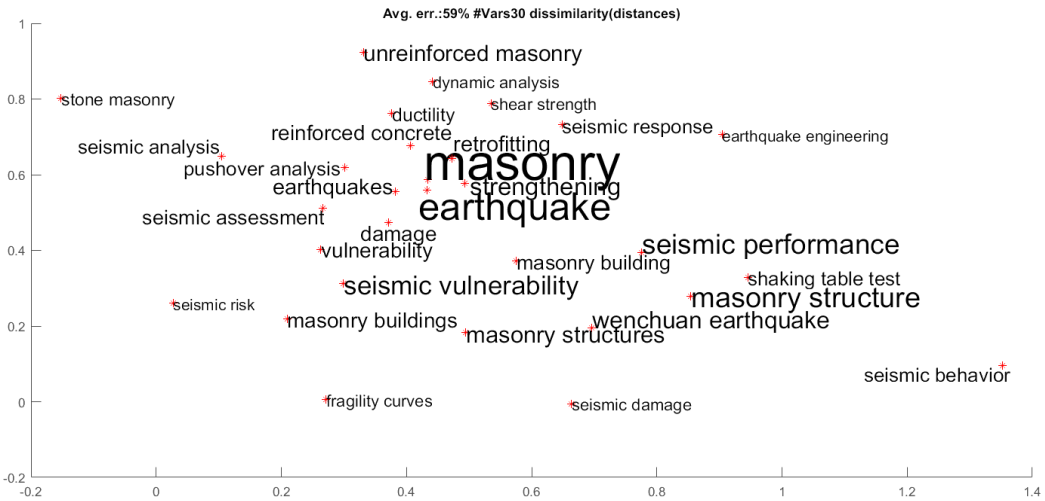

Figure 10: Bibliometric map for the top-30 keywords (no lines are included for better readability).

Hence, the entire information is embedded on the bibliometric map in an integrated manner. Figure 9 exhibits a clear representation of the top-10 keywords' map, with all the relative information (distances, dissimilarities, etc.) as mentioned above. In particular, it was found that "earthquake" and "masonry" keywords, exist at the center of the map, due to the particular formulation of the database query, suggesting that the procedure worked properly.

If we take more keywords into account, in particular 30 instead of 10, we end up with the map depicted in Figure 10. According to this map, the keywords most related with "masonry", found to be the: "strengthening", "retrofitting", "damage", "vulnerability" and "rehabilitation", among others as shown in the figure. This further strengthens the hypothesis that the majority of the papers regarding masonry structures, deal with existing structures and their rehabilitation, rather than new ones.

The analysis of even more keywords can reveal new information and identify other patterns, especially when proper visualization tools (zoom, hide etc.) are employed.

4.2 Bibliometric map for the top authors

If some authors tend to cooperate (write usually papers together) then there is a strong connection between them and their names appear together in papers. If we take the authors into account, instead of the keywords, and apply the same methodology, we end up with some interesting results. The bibliometric map for the top-30 authors, exhibits a sparse image, as shown in Figure 11. A sparse image means that the corresponding papers are written basically by individual researchers, rather than research groups. However, a specific cluster of authors was found, with authors J.M. Ingman, P.B. Lourenco, G. Magenes, S. Lagomarsino and others being the key players, as demonstrated in Figure 11. By using the proposed methodology, a researcher can easily and rapidly obtain valuable information regarding the clusters or cooperating groups among the scientific community, which is very significant for the literature review part of any research work as well as for review papers or surveys [8, 9].

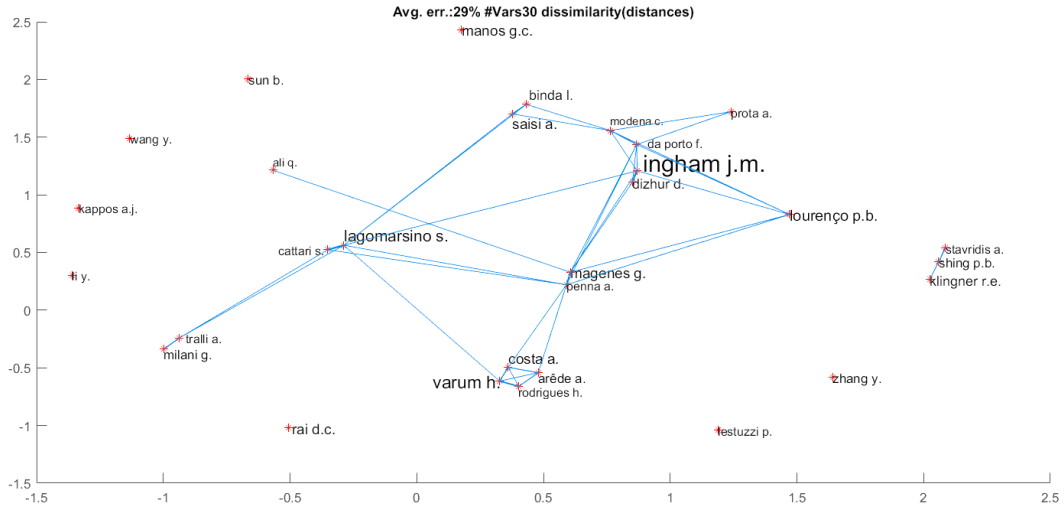

Figure 11: Bibliometric map for the top-30 authors.

The same can also be done with the references of each paper. A bibliometric map can be made for the references, revealing the relationship between papers that have been referenced by the papers of the database. Co-citations is a good indicator of similarities among papers [10, 11] which supplementary clarifies the bounds among disciplines as well as their interactions. This type of bibliometric map is outside the scope of the present paper and may be examined in a future work.

5 CONCLUSIONS

This work demonstrates a new approach to multidimensional scaling, particularly applied for the construction of bibliometric maps. Through a rigorous numerical procedure, an in-depth analysis of thousands of research papers is possible. The aim is to reveal the significant research topics in a subject area, as well as to identify associations between thematic areas, authors, references, institutions, etc. The overall management of this vast amount of information regarding scientific knowledge, is a demanding task, as the growth rate of the scientific output is exponential. Utilizing the prosed approach, the literature review part of any research work, can be thoroughly analyzed and documented, avoiding the focus on topics of minor interest and detecting interdisciplinary associations.

The specific application was made on the topic of masonry structures under seismic excitations, but the described procedure is generic and can be easily applied to other scientific areas as well. Two bibliometric maps for the top-10 and top-30 keywords were investigated. The results showed that the majority of papers regarding masonry structures, deal with existing structures and their rehabilitation, rather than new ones. The investigation of the bibliometric maps for the top-30 authors identified a cluster of authors who tend to cooperate with each other, and also other isolated authors who prefer to work on their own.

REFERENCES

[1] Elsevier. *Scopus Content Overview*. Available from: https://www.elsevier.com/solutions/scopus/content [Accessed on 24 Feb.],

[2] 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.

[3] 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.], 07 May 2014.

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

[5] Altman, N. and M. Krzywinski, *Points of Significance: Clustering.* Nat Meth, 2017. **14**(6): p. 545-546.

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

[7] Ahlgren, P., B. Jarneving, and R. Rousseau, *Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient.* Journal of the American Society for Information Science and Technology, 2003. **54**(6): p. 550-560.

[8] Boyack, K.W., R. Klavans, and K. Börner, *Mapping the backbone of science.* Scientometrics, 2005. **64**(3): p. 351-374.

[9] Leydesdorff, L. and I. Rafols, *A global map of science based on the ISI subject categories.* Journal of the American Society for Information Science and Technology, 2009. **60**(2): p. 348-362.

[10] Small, H. and E. Sweeney, *Clustering thescience citation index® using co-citations.* Scientometrics, 1985. **7**(3): p. 391-409.

[11] White, H.D., *Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists.* Journal of the American Society for Information Science and Technology, 2003. **54**(5): p. 423-434.