ROLE AND IMPACT OF THE SCIENTIST IN THE SOCCER ECONOMIC WORLD

Fabio Romero Manrique, J. Sebastián Hernández Aguilera & Andrés D. Pérez

{fromerom, jhhernandezag, anperezpe}@unal.edu.co

Universidad Nacional de Colombia

ABSTRACT

The multidisciplinarity allows engineers, mathematicians and physicists, among others to work together in pursuit of a common goal from different approaches. This article presents the scientist's role and impact in the soccer economic world. In addition, this paper presented a review looking to find relationships or correlations between data variables by using exploratory data analysis, statistical indicators calculation, spatial relationships, correlations, word clouds and term-frequency methods. As a result, thoughts were obtained for the scientists' impact in the soccer economic world. Likewise, a view on the behaviour and development of scientific production for the last 9 years related to the soccer economic world.

Key words: Analytics, Soccer, Economic and Analysis.

1. INTRODUCTION

It is very important to identify how science is not a study field which only feeds and develops itself, but is also a multidisciplinary study. This multidisciplinarity allows engineers, mathematicians and physicists, among others to work together in pursuit of a common goal from different approaches as said by Morillo et al. [1] their work.

Sports is a very important field, it has an impact in several aspects, such as economy, culture and entertainment. One of the most important of these is the economic world, because even if that’s not so evident, there are many investments in this field. One of the most famous sports is soccer, having a presence around the whole world playing a very important role from countries such as England, Spain, France, USA and Germany.

But, the real question here is: How does the soccer field move the economic world? Maybe it is not clear the amount of money invested on soccer teams, but behind a game there are plenty of factors being involved in the same game, factors like the sponsors of the teams and the amount of stuff being sold by the different sponsors, the selling tickets earnings and the bets made around a game.

Keeping this in mind one of the points of interest could be how to determine the way to improve the investments and earnings and what are the variables playing a role so this can be possible; is here where science starts to play as a determinant integrant of the analysis. Is science the one in charge of analysing these variables and determining the pros and cons of the problem, making the exercise a measurable problem giving objective quantitative and qualitative values.

A measurement of this impact is how involved and developed is the scientific production on the field and how has its behaviour been in the last years. This is why it is considered the research level of this topic.

Among these researches you can find the one developed by Rein et al. [2] named "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science" where you will find an overall point of view of how science interferes in sports. Brooks et al.[3] on their work named "Using machine learning to draw inferences from pass location data in soccer" show an example of how to use scientific knowledge and skills to determine ways to improve the playing development of a team and its impact in the economic world.

"Constraint programming and machine learning for interactive soccer analysis" an article written by Duque et al. [4], shows several tools to be used in order to analyze the game development and the way to understand it from different perspectives.

However, for our knowledge, although there are different works related to sport analytics and its influence and/or relationship with the economic sector, there is no work that reviews the impact of scientific production in this field on the economic world. This information can help to define a baseline and to provide a clear and organized perspective of the actual state-of-the-art status for soccer data analysis.

The remainder of the article is organized as follows: Section 2 explains in detail the proposed methodology. Section 3 presents the results. Finally, section 4 reports the discussion, main conclusions and future works.

2. METHODOLOGY

Figure 1. shows the overall methodology for this project. This consists of four main stages, which are: data acquisition, data processing, data analytics and data visualization.

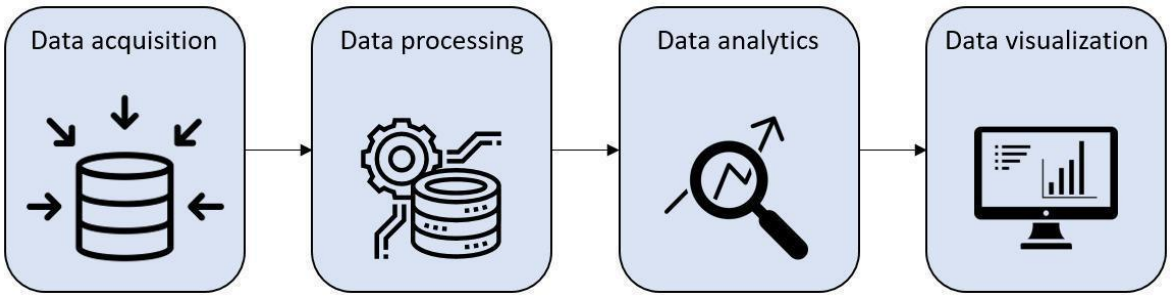

Figure 1. Overall workflow for the project methodology. Contains four stages from left to right which are: Data acquisition, processing, analytics and visualization.

Data acquisition

This stage focuses on manual unstratified and systematic collecting state-of-the-art works metadata from different public and private databases using Google Scholar that tackle soccer analytics on different domains such as: medical, biomedical, physic, statistic and financial, among others, for the last 8 years. These works with associated information (search query, publication year, cites, type, title, abstract, publication platform or event and authors) are stored in a database for further processing.

Data processing

For this stage, once the data has been collected, using tools such as software or programming languages (Excel and Python), the data is being processed and organized into an interpretable and structured table, adjusting formats, generating processed text fields with standardized title and abstract in lowercase, without special characters.

Data analytics

With structured information, different algorithms and analysis tools such as: exploratory data analysis, statistical indicators calculation, spatial relationships, correlations, word clouds and term-frequency were implemented.

Data visualization

Once the findings are clear, using different visualization tools such as, Python and Office Suite, will be presented coherent and explainable graphs and figures that summarize the obtained results. This document is proposed to be organized as follows: Section 2 explains in detail the proposed method. Section 3 shows the experimental evaluation of the method including the dataset, the experimental setup, the results and discussion. Finally, section 4 reports the main conclusions and future works.

Note: this dataset as well as source code and other files, can be found in this Github repository: https://github.com/adpzz/unal-sfi-proy

3. RESULTS

This section presents the results obtained after the development of the project.

3.1. Annual production

For the annual production, both the percentage distribution of scientific production within the dataset (see Figure 2) and the disaggregated breakdown at type level for the annual production were reviewed (see Figure 3).

As presented in Figure 2., it's evident that the highest concentration of scientific production (approximately 51% of the data) corresponds to the periods of 2014 and 2015. On the other hand, the

remaining 49% is distributed with similar proportions in the different periods except for the years 2019 -2021.

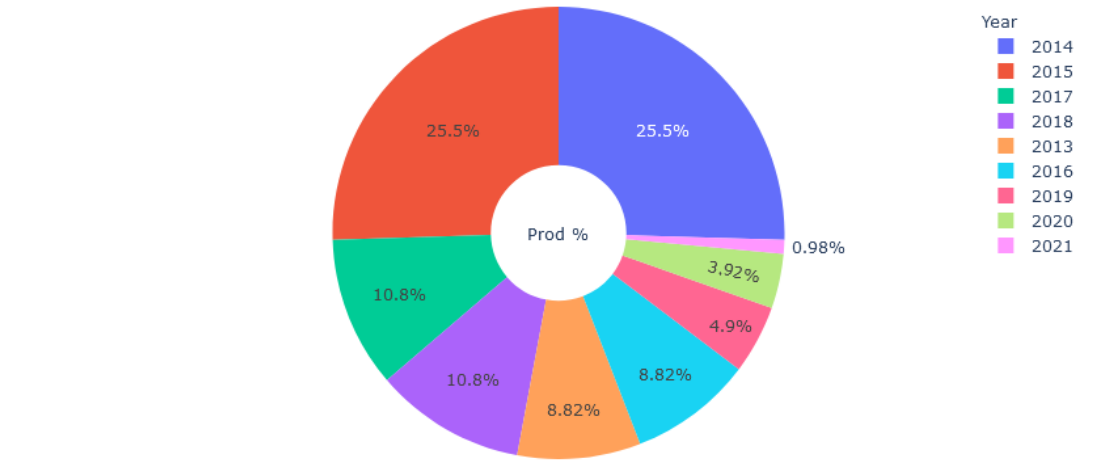

Figure 2. 2013 - 2021 Annual production review at percentage level on the acquired dataset

For Figure 3., a disaggregation at type level is presented, showing that the scientific production of both books and theses stopped after 2016. On the other hand, it's also notable a higher prevalence or representation of production of scientific articles, which has ceased or considerably mitigated with respect to previous periods for the last 3 periods (2019, 2020 and 2021).

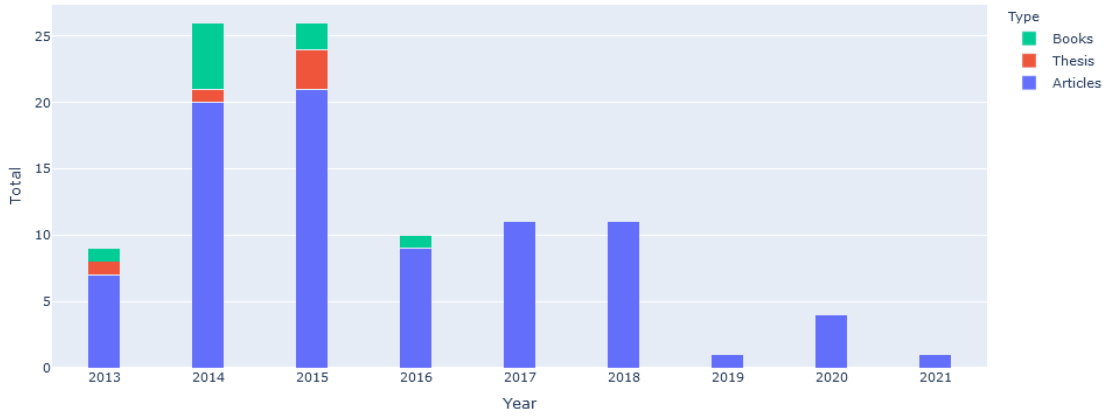

Figure 3. 2013 - 2021 Annual production review desegregated by type of production (books, theses and articles)

3.2. Annual citations

For the annual citations, was reviewed both the annual distribution of citations and some citation statistics at the article level within the year group (see Figure 4.). Figure 5, on the other hand, presents a spatial review seeking to validate some relationship between title and abstract lengths versus the number of citations for each year.

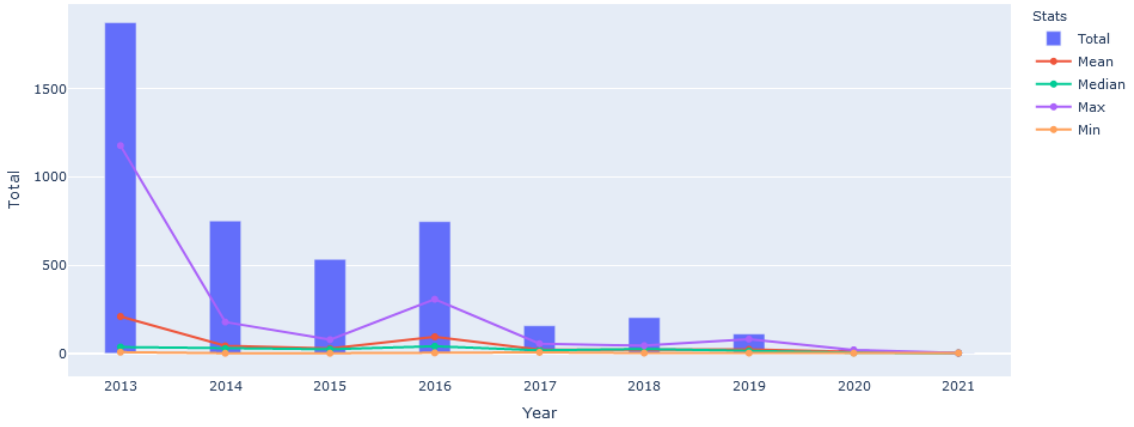

Figure 4. 2013 - 2021 Citation review for annual output along with indicators. Total - Total citations for that year. Max - Maximum number of citations for that year. Min - Minimum number of citations for that year. Mean - Average number of citations for that year. Median - Median number of citations for that year

As shown in Figure 5, there is a distribution at the title and abstract length level centered on short texts, evidencing a preference of researchers for this characteristic. Similarly, it is not very noticeable that short or long texts combined in any way can influence the number of citations an article receives.

For the text analysis, the information was pre-processed, seeking to standardise the text to lowercase, without any special characters or punctuation.

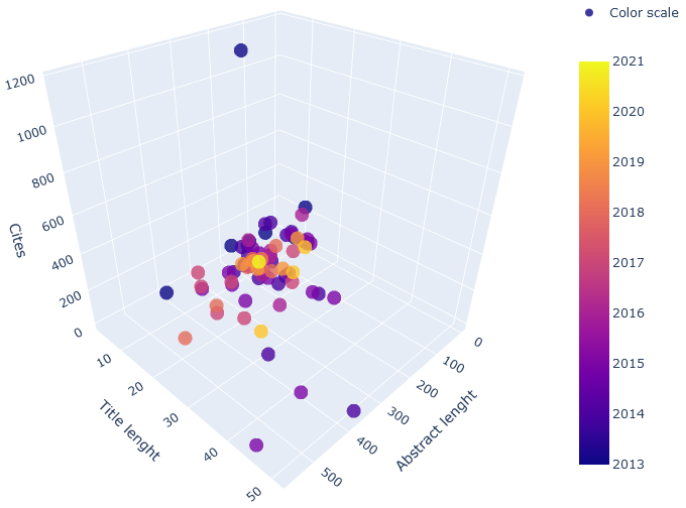

Figure 5. 2013 - 2021 Review of the relationship and distribution between total citations, text length and abstract length for each of the years

3.3. Word cloud and term-frequency analysis

A word cloud analysis was performed for both article titles and abstracts, seeking to obtain both a visual representation of the most dominant terms and a review at term frequency level.

3.3.1 Titles

For the title word cloud, the text is preprocessed by removing stopwords, which are meaningless words such as articles, pronouns, prepositions, etc. As can be seen in Figure 6, according to the visual representation some of the dominant terms would be "soccer", "analytics", "analysis", "data", "performance" and "player", among others. Then, when compared to the term frequency review in Figure 7, it is evident that these terms do indeed have frequencies above 10 within the dataset.

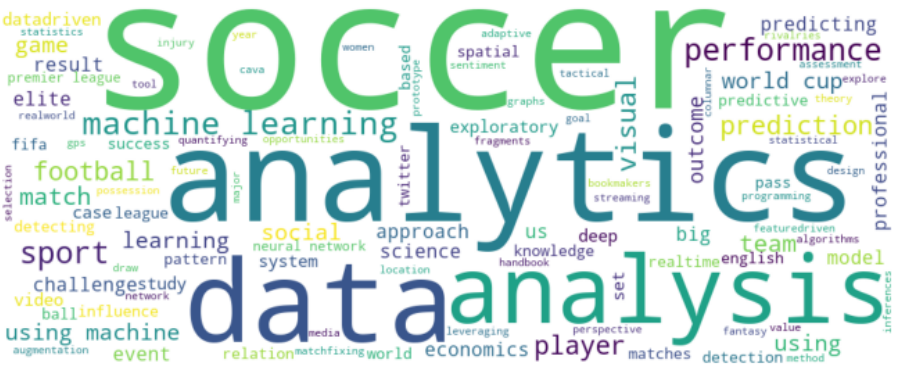

Figure 6. Word cloud for the title terms

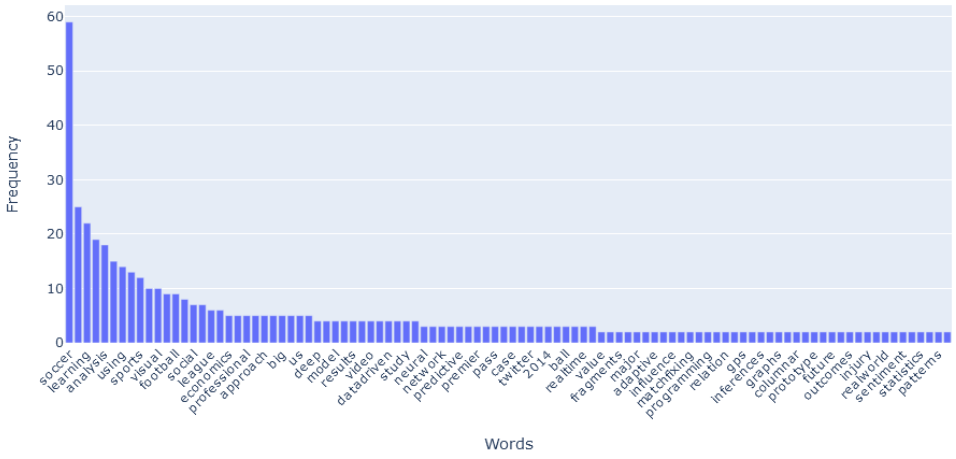

Figure 7. Review of the top 100 most frequent terms in the title word cloud

3.3.2 Abstract

As with the titles, a corpus was constructed for the abstracts to produce a word cloud. The text is preprocessed by removing stopwords, which are meaningless words such as articles, pronouns, prepositions, etc. As can be seen in Figure 8, according to the visual representation some of the dominant terms would be "player", "soccer", "data", "team", "performance" and "analysis", among others. Then, when compared to the term frequency review in Figure 9, it is evident that these terms do indeed have frequencies above 10 within the dataset.

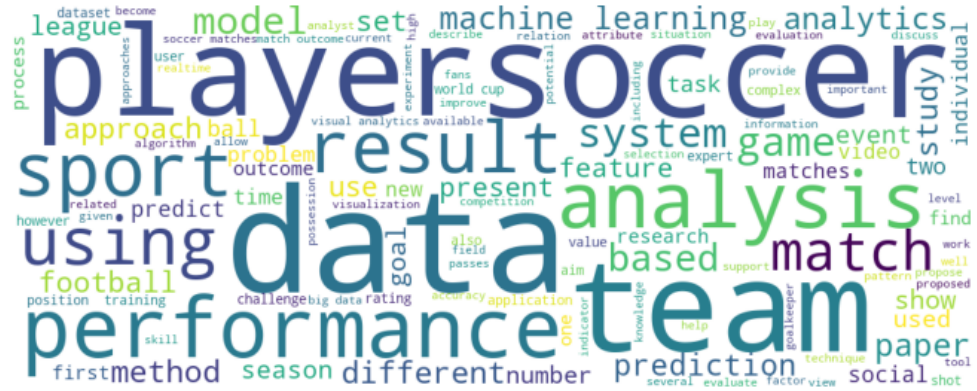

Figure 8. Word cloud for the abstract terms

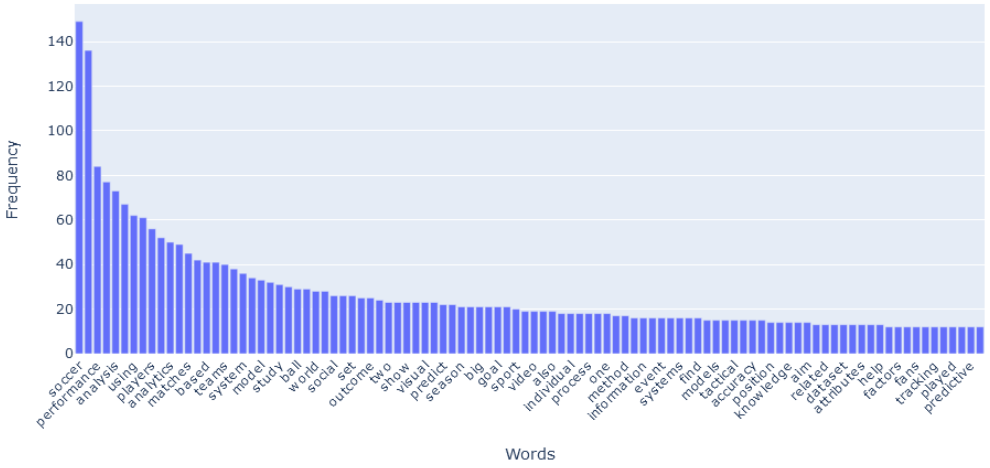

Figure 9. Review of the top 100 most frequent terms in the abstract word cloud

As noticed, terms in title are related or being congruent with the abstract terms.

3.5. Annually authors performance

Finally, the authors were analyzed, finding the number of authors per year (see Figure 10), as well as which of these authors generated the most scientific output (see Figure 11).

As shown in Figure 10, the periods with the highest number of authors correspond to 2014 and 2015, coinciding with the periods of greatest scientific production (see Figure 3).

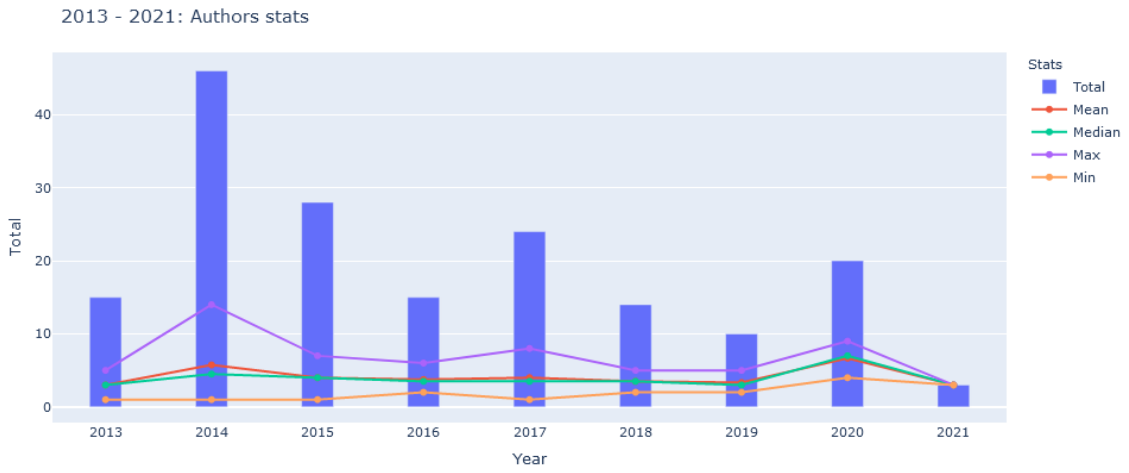

Figure 10. Revision of the number of authors for the annual production together with the indicators. Total - Total number of authors for that year. Max - Maximum number of authors for that year. Min - Minimum number of authors for that year. Mean - Average number of authors for that year. Median - Average number of authors for that year

On the other hand, it is evident that the majority of authors have only 1 publication related to this field, on the other hand, around 25% of the authors show interest and permanence giving continuity to their research.

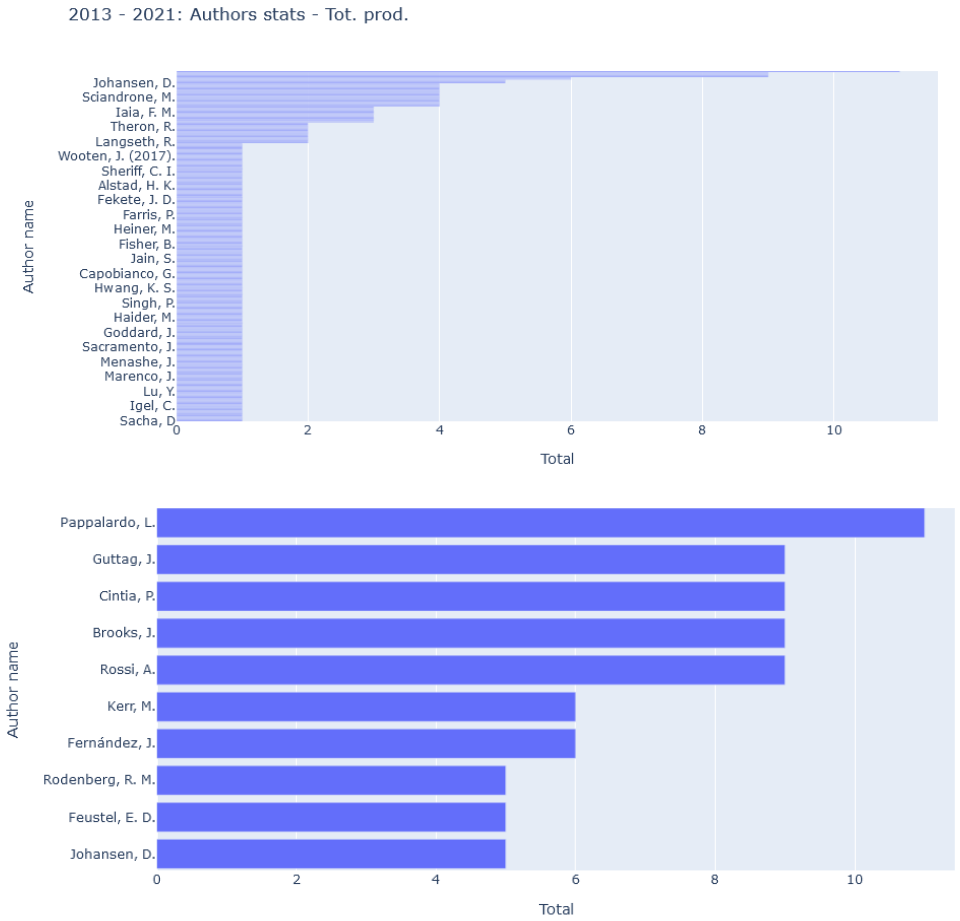

Figure 11. Review of the number of products of each author in the database. Upper- Overview. Lower- View of authors with 5 or more publications

4. CONCLUSIONS AND FUTURE WORK

The current database shows a significant decrease in scientific production over the last 3 years, which could be related to the current world situation.

The years of greatest academic production were 2014 and 2015, which shows a particular global interest in this sector with considerable capital injection.

There is evidence of a decrease in the number of citations of the articles in the database possibly related to the low scientific production and the production's aging.

It is evident that the length of the title or abstract does not have a considerable impact on the number of citations of an article. However, authors mostly prefer abstracts and short titles.

Based on the word clouds, it is evident that the titles and their abstracts are largely congruent. On the other hand, the prevalence of terms such as analytics, data, performance and analysis, among others, shows the high impact of scientists in this field.

There is a direct correlation between the number of authors and the number of publications. However, it could be inferred that quantity does not represent quality and/or interest, when considering the number of authors vs. the number of citations obtained.

It would be interesting to extend the dataset both to a larger number of periods and to a search with additional terms. This in order to increase the representativeness and to generate a longitudinal study. On the other hand, expand the number of algorithms, methods and analysis techniques with the aim of finding more and possibly diverse relationships with features not evidenced and/or considered in this study.

ACKNOWLEDGMENT

We are grateful for the support of Professor Rafael German Hurtado Heredia, as well as to the other professors and guests who participated somehow in the development of the Fundamentals of Research course.

BIBLIOGRAPHY

[1]. Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and technology. vol. 54, no 13, pp. 1237-1249.

[2]. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus, vol 5, no. 1, pp. 1-13.

[3]. Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol *9, no.* 5, pp. 338-349.

[4]. Duque, R., Díaz, J. F., & Arbelaez, A. (2016, May). Constraint programming and machine learning for interactive soccer analysis. In *International Conference on Learning and Intelligent Optimization*, Springer, Cham, pp. 240-246.