Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v15n1p232

Football analytics: a bibliometric study about the last decade contributions
By Cefis

Published: 20 May 2022

Electronic Journal of Applied Statistical Analysis
Vol. 15, Issue 01, May 2022, 232-248
DOI: 10.1285/i20705948v15n1p232

Football analytics: a bibliometric study about the last decade contributions

Mattia Cefis *[a]

[a] *University of Brescia, Department of Economics and Management, Brescia (Italy)*

Published: 20 May 2022

Machine learning and digitization tools are exponentially increasing in these last years and their applications are reflected in different areas of our life: in particular, this article has the aim to focus on football (i.e. soccer for Americans), the most practised sport in the world. Due to needing of professional teams, analytics tools in football are becoming a crucial point, in order to help technical staff, scouting and clubs management in policy evaluation and to optimize strategic decisions. In this article we propose an original bibliometric analysis about football analytics in the decade 2010-2020, thanks the powerful R package *Bibliometrix* and the well-known bibliometric database SCOPUS. The main goal is to understand better what already exist in football analytics literature and what not, in order to suggest future researchers to find new topics or to refine existing tools. Furthermore, our intention is to show some results starting from the sources production distribution, then focus on the most productive research groups and their countries, discover the most dynamic authors and highlight topics trend thanks keywords, during these last ten years. Finally, three relevant articles that summaries the most important themes are presented.

keywords: Football, Soccer, Analytics, Bibliometric.

*Corresponding author: mattia.cefis@unibs.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

Nowadays, we can considering football clubs as real firms, while until some years ago we were in the so-called patronage era. Following this question, now the objective for football clubs is to optimize their own financial statements, in order to avoid any penalty from football authorities (for example, UEFA[1] for European clubs). In fact, as summary, the main earnings for each football club (Canova and Canepa, 2016), derive from:

- Pay TV
- Stadium tickets and official merchandising
- Sponsor
- Players' transfer market

With regard to this, it is logical to affirm that a successful team, with a good game-system and which often plays international competitions (for example, UEFA Champions League) causes a virtuous circle (Canova and Canepa, 2016): new sponsors and fans, more UEFA bonus and increasing in the players' market value.

So, we can say that all this virtuous circle is strongly influenced from sports results; for this reason, for a football club is extremely important to optimize them. In this last decade, for many sports and also for football is developing a digital revolution, where the crucial theme is: how to optimize players and team performance, in order to reach positive results on the pitch. Many teams and researchers are trying to answer this question.

1.1 Previous works and guideline of the paper

Until now, bibliometric reviews on sports have been focused on different topics but not directly on football analytics: for example there is a bibliometric analysis on sports science (Vigneshwaran and Kalidasan, 2018), some others focused on technology of the sport (Belfiore et al., 2019), until the more recently focalized on the role of social media in sports (López-Carril et al., 2020). So, following the introduction made in the previous paragraph and in order to provide a guide for football analysts and data scientists, our goal is to propose an original overview of the literature about football analytics, thanks to an innovative approach: the Bibliometrix R package (Aria and Cuccurullo, 2017). This interesting tool let us to automate the stages of data-analysis and data-visualization both, about literature, managing data directly from the famous bibliographic database SCOPUS[2]. As previous step we tried to take in consideration also another famous database (i.e. Web of Science), but merging different databases is one of the most challenging topics of bibliometrics literature: in fact SCOPUS and Web of Science have very different records and many metadata, such as authors' names, affiliations, and references, that are stored with not compatible formats. Furthermore, we

[1] www.uefa.com

[2] www.scopus.com

noticed that our query (Sec. 2) applied on SCOPUS produced 215 documents as output, while from Web of Science only 73, of which 67 already found in the SCOPUS database; for these reasons we decided to adopted just documents from SCOPUS.

After this introduction, now we specify the organisation of this paper: we will present query and data used for the bibliographic analysis (Sec. 2), followed from a presentation of the results (Sec. 3) and a discussion about the main articles in Sec. 4. Finally, a conclusion is given in Sec. 5.

2 Data extraction and preparation

As introduced in Sec. 1, data were extracted from SCOPUS. The goal was to collect all documents about football or soccer analytics, searching these words in the title, abstract and keywords of each article: before the year 2010 we observed that scientific production produced maximum one article per year, and as consequence the decision to focus just on the last ten years (decade 2010-2020, with more significant production); furthermore, we kept in consideration only documents in English language. This query (1) ran on July, the 26^{th} of 2021.

$$\begin{aligned}(Football\, or\, Soccer)\, and\, analytics\\ not\\ (Rugby\, or\, Cricket\, or\, Hockey\\ or\, American\, Football\\ or\, Australian\, Football)\end{aligned} \tag{1}$$

Initially, query (1) included just the first row, but after some inspections we noticed that in the output there were included some bias articles (for example, about American or Australian football, or other sports): for this reason, in order to automate the extraction, we attached below the first row of (1) the others five ones. By the complete (1) we could exclude noisy documents from our research, then the dataset was converted (thanks a special function provided by the *Bibliometrix* package) into a data-frame, with cases corresponding to articles and variables to field tags.

In the final dataset we obtained a total of 215 documents over the last decade. It is not a high number, but we must take in mind that soccer is one of the last sports where analytics achieved: in practise, as we will see in Sec. 3, football analytics revolution is on the cutting edge just from the last years. Before beginning the bibliometric analysis, we adjusted by hand some typos in the keywords and in the authors' names from the dataset, in order to avoid redundancy and misunderstanding in the results.

3 Results and Analysis

In this paragraph we will analyse results of this bibliometric analysis focusing on different aspects: in Sec. 3.1 we will see an overview about the scientific production in this last decade, while in Sec. 3.2 some statistics about the authors are showing; in Sec. 3.3 we will focus on the authors' keywords, while in Sec. 3.4 an in depth analysis about the most productive countries and universities is provided; eventually, some others graphs are shown in Sec. 3.5.

3.1 Overview results

As said in Sec. 1, this analysis was performed thanks the *Bibliometrix* package of R software; for first, here we show some general results, in order to understand how the bibliometric dataset is composed and the documents production trend over the last decade. In Tab. 1 we can see a preliminary classification of the documents: it's clear the prevalence of conference papers and articles.

Table 1: Documents classification.

Document types	Nb. of docs
Article	78
Book chapter	2
Conference paper	113
Conference review	11
Data paper	1
Editorial	3
Review	7
Total	215

In order to give an overview of the most relevant sources, considering all documents listed in Tab. 1, the plot in Fig. 1 is presented: the most relevant sources (i.e. with more than 15 documents) are the Journal of Lecture Notes in Computer Sciences, the CEUR workshop proceedings and the International Journal of Sports Science and Coaching.

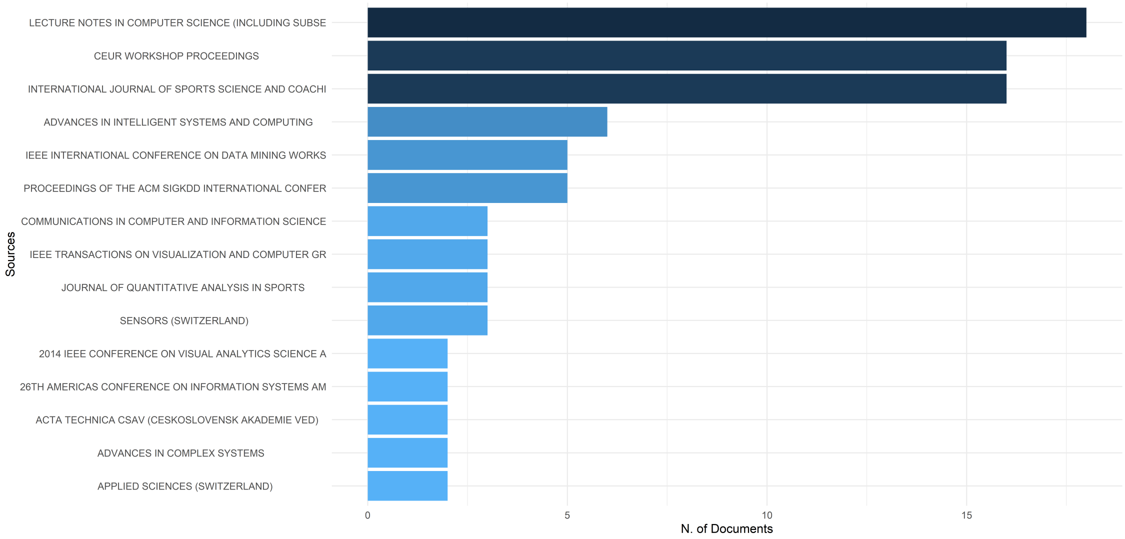

Figure 1: The most relevant sources in football or soccer analytics.

In Fig. 2 instead we can see the time-series of documents production over the last decade: this evolution shows us a significant growing in the football analytics production, with a peak in 2019 and a stabilization in 2020.

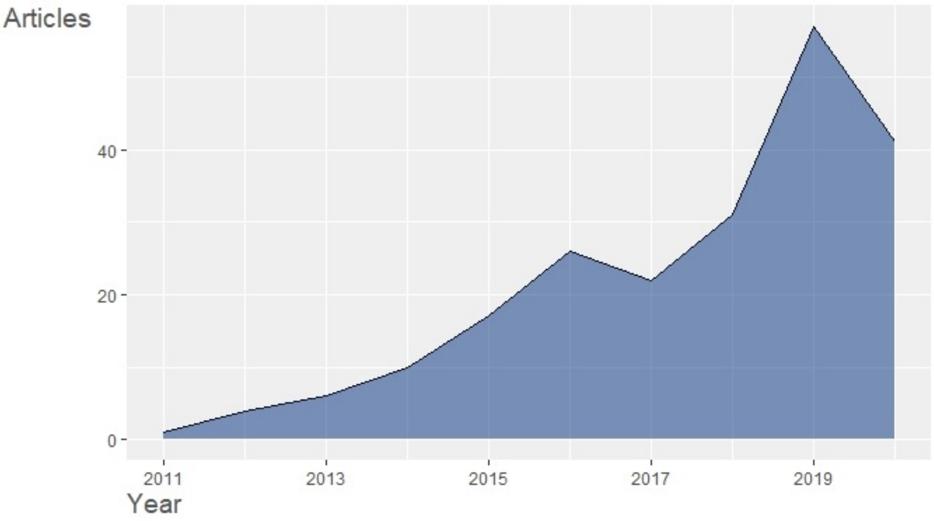

Figure 2: Annual scientific production.

All this let us to underline how soccer analytics is an emerging and attractive topic in the research world.

3.2 Authors analysis

In this paragraph we underline the most active authors; in Tab. 2 is suggested their ranking by the well-known dominance factor: it is a ratio indicating the fraction of multi-authored articles in which a scholar appears as the first author. Consequently, an index near to 1 indicates very high dominance (i.e. in this table are considered authors with dominance factor greater than 0.10).

Table 2: Authors' ranking by dominance factor.

Ranking	Name	Dominance factor
1	Stein M.	0.73
2	Bransen L.	0.50
3	Pappalardo L.	0.50
4	Fernandez J.	0.25
5	Lucey P.	0.20
6	Stensland H.	0.20
7	Cintia P.	0.17
8	Davis J.	0.14
9	Halvorsen P.	0.14
10	Janetzko H.	0.13
11	Van Haaren J.	0.11

Now, here below (Fig. 3) we propose an interesting plot that take in consideration not only the volume of the authors' production, but also the number of citations per year over the last decade: for this reason there is not perfect correspondence between Tab. 2 and Fig. 3 authors. In addition, take in consideration that the diameter of circles is proportional to the number of published articles, while their darkness is proportional to the total number of citations received per year.

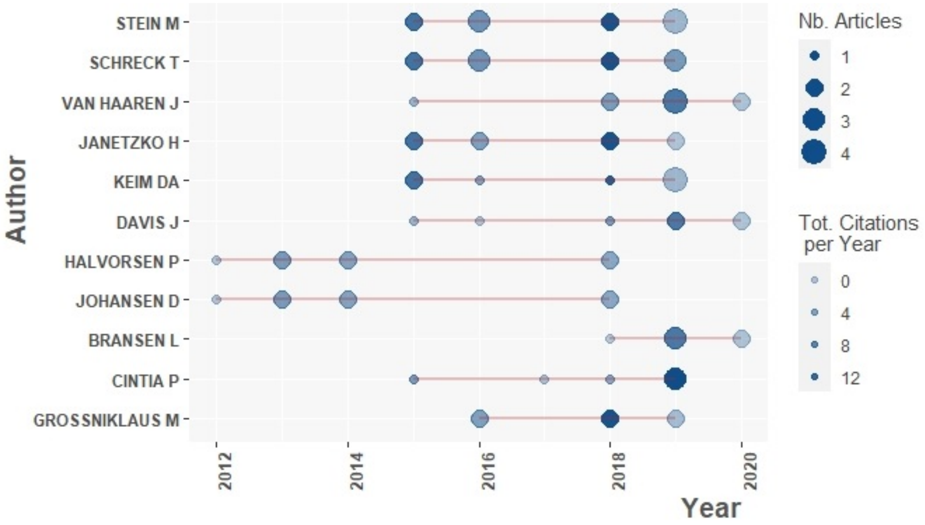

Figure 3: Top-Authors' production-citations over the years.

From Fig. 3 we can emphasize that activity of the most productive and cited authors is concentrated in the last five years, except for Halvorsen and Johansen. For example Cintia (University of Pisa -Italy) had an increasing of his production and obtained more than 10 citations in 2019, whereas Van Haaren (University of Leuven -Belgium-) contributed with more than 5 articles and received 15 citations in the last three years; remarkable also the contribution offered by Schreck (University of Graz -Austria-) and Stein (University of Konstanz -Germany-, with also the highest dominance factor, see Tab. 2) between 2015 and 2019 (more than 10 articles and 20 citations received for each one).

3.3 Keywords analysis

In this paragraph the aim is to investigate about research topics, show what are the most relevant keywords used from authors and their connection thanks to different plots, well documented in Cobo et al. (2011). As preliminary analysis, in Fig. 4 are shown the most used keywords from the authors, thanks a word cloud plot (i.e. the words size is proportional to their frequency). It is interesting to notice how, excepting the keywords used in the initial query (i.e. football, soccer and analytics, that we expected in this result), there are also "sports" (the most used one) and typical analytics tools such as "data mining", "learning systems", "visualization", "artificial intelligence" and "machine learning".

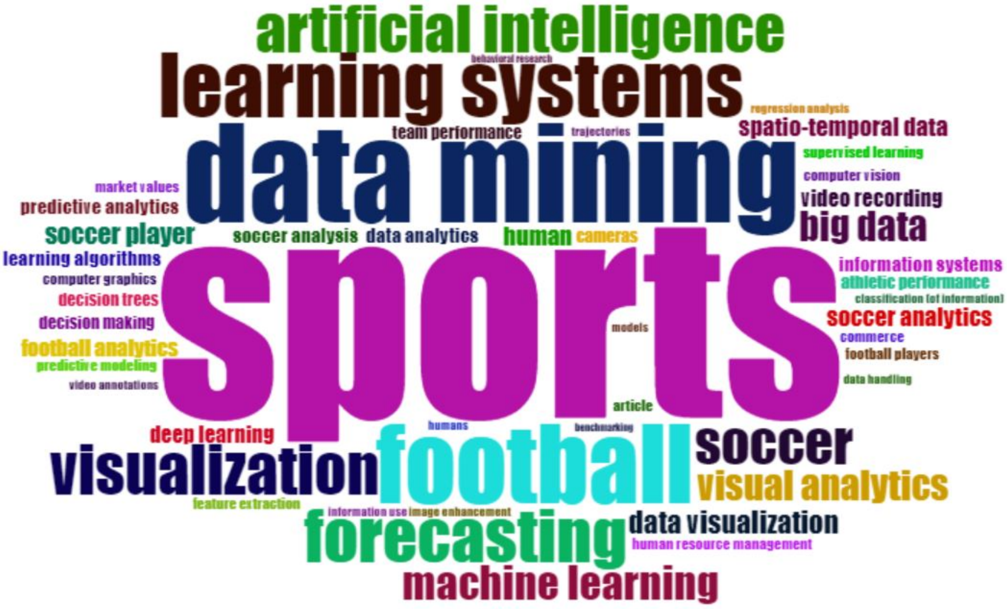

Figure 4: The most used keywords from authors.

For the next, *Bibliometrix* allows using the *conceptualStructure* function to perform multiple correspondence analysis to draw a conceptual structure of the field and K-means clustering to identify clusters of documents that express common concepts, all summarised by a network plot (Fig. 5); this graphic let us to explain co-occurrence, where keywords and rectangles size are proportional to the production, while thickness of ties to the strength of co-occurrence. Different colours represent clusters, created from a K-means clustering procedure, to identify groups of documents that express common concepts (Aria and Cuccurullo, 2017). In particular, co-word analysis aims to map the conceptual structure of a framework using the word co-occurrences (i.e. in this case the keywords) in a bibliographic collection.

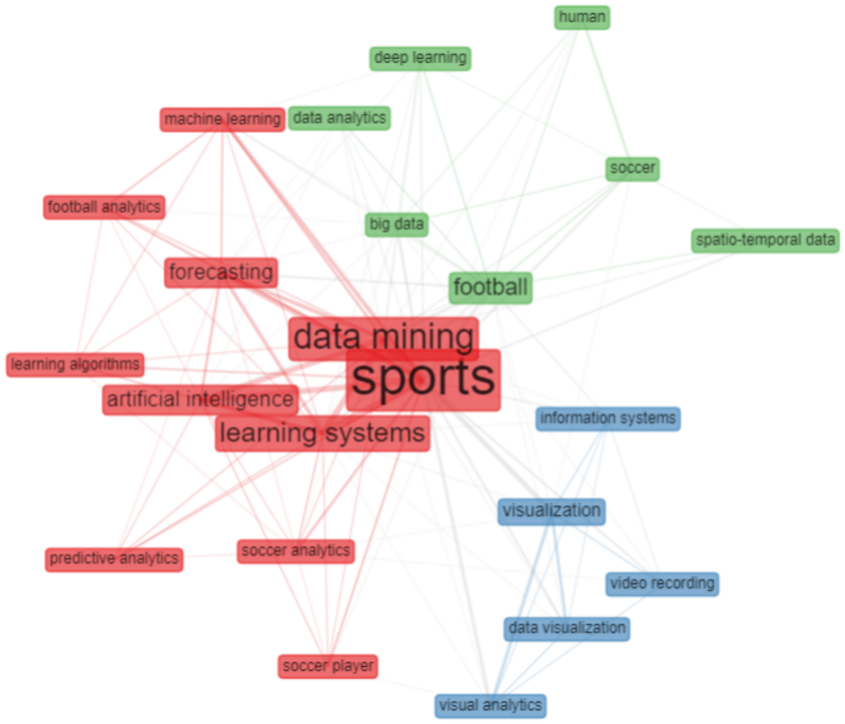

Figure 5: Keywords co-occurrence network.

From Fig. 5 we can highlight how the red cluster is the most representative (i.e. 11 keywords), with focus on technical tools (i.e., machine learning, data mining, forecasting, artificial intelligence, player analysis and prediction) while the blue one is focalized on visualization tools and the green one on more general topics such as big data and deep learning.

Now, in order to represent co-occurrences network in a simpler view (i.e. a 2- dimensions plot), we can see the thematic map (Fig. 6; for this plot we must take in consideration that the words used in the initial query (i.e. football and soccer) have been excluded, in order to have a clearer interpretation. As comment, this graphic lets us to understand:

- In the top-right quadrant (high density and centrality) we can see the motor themes.

- In the bottom-right quadrant (high density and low centrality) there are the basic themes.

- In the top-left quadrant (low density and high centrality) we find niche themes.

- In the bottom-left quadrant (low density and low centrality) there are emerging or discovering themes.

Take in mind that circle size is proportional to the cluster word (i.e. in this case keywords) occurrences. From Fig. 6 we see how technical tools are the motor, they are often

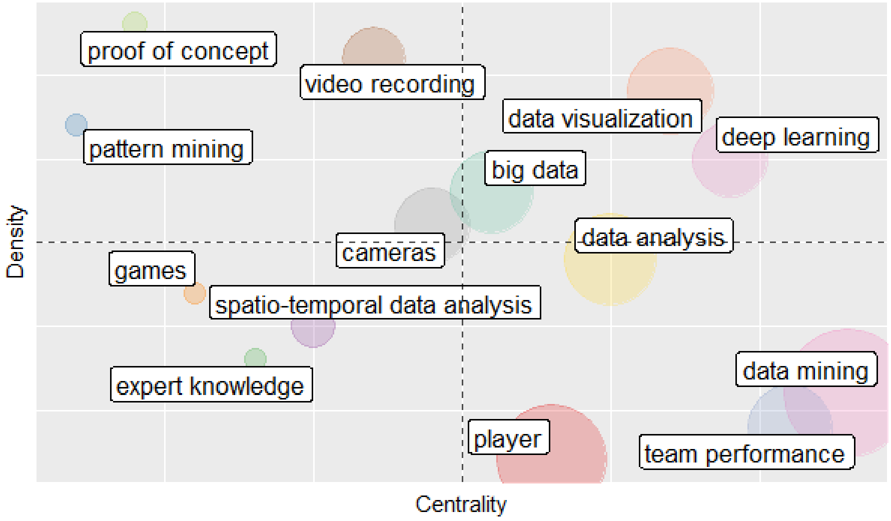

Figure 6: The thematic plot.

applied for basic themes (i.e., player, team performance and data mining), while niche themes are mainly video recording and cameras; finally, considerable emerging themes are spatio-temporal (also called as data tracking analysis), that is strictly related with video recording and cameras themes. Also expert knowledge is a crucial emerging theme, since it could be very useful in comparison with analytic results. As insight, we analyse in Fig. 7 the top-five keywords evolution over the last decade. It is interesting since we can see the topics trend applied into football analytics research.

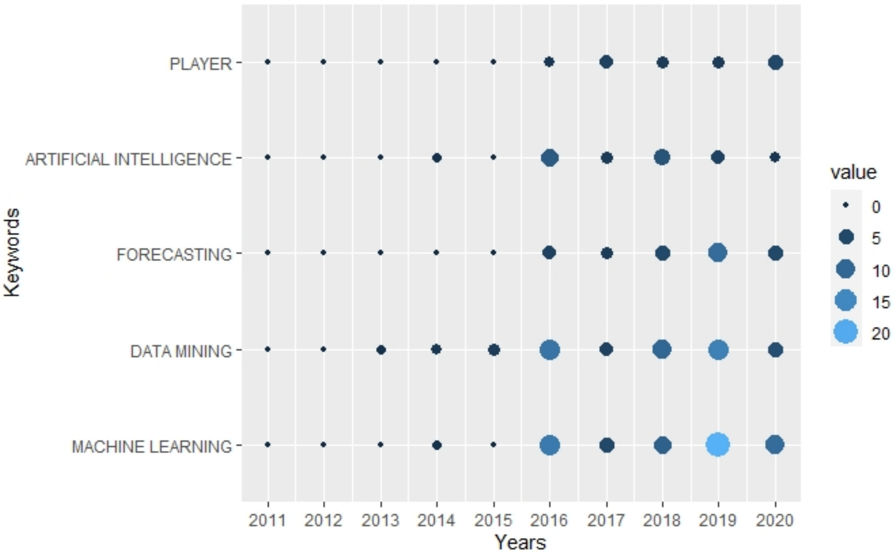

Figure 7: Keywords evolution over the ten years.

It's important to highlight that in Fig. 7 circle and brightness are proportional to the number of contributes. We emphasize the increasing of employment for these keywords, moreover until the year 2016, then a little decreasing and a new increasing in the last two years: the most employed keywords are machine learning and data mining.

3.4 Countries analysis

Now, in this section attention is relied on countries analysis, in order to discover what are the most productive ones and the network of universities collaboration. Notice that in Fig. 8 the Multiple Country Publications (MCP) indicates, for each country, the number of documents in which there is at least one co-author from a different country and so it measures the international collaboration intensity of a country; instead, the Single Country Publications (SCP) index measures the number of documents in which author and co-authors are from the same country. We can see how Germany, USA and Italy are the most productive countries, with an interesting difference: while for Germany and USA a part of their production derive from collaboration with authors from other countries, Italy do not contribute with anyone. Austria, Belgium and China shows a higher rate of collaboration with other countries (MCP) than their own SCP.

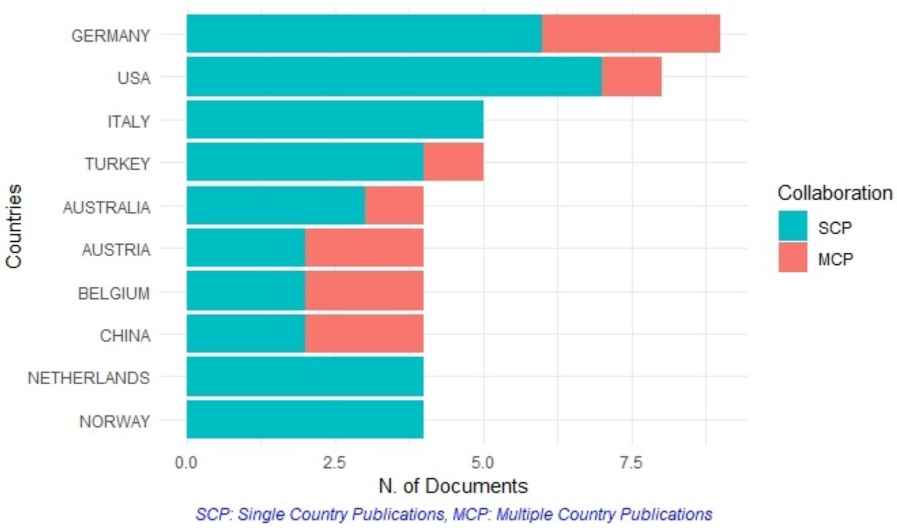

Figure 8: The most productive countries.

In order to have a clearer idea than before about countries collaboration and their rate of production we can see a summary plot in Fig. 9, the country collaboration map. Thanks this graphic, the darkness of each country is proportional to each own production (i.e. grey states have no production), while lines thickness among countries is proportional to their collaboration rate. This plot emphasize the relevant relationship between USA and Australia (i.e. the strongest one), and some others intercontinental relations respectively among centre Europe and Brazil, Spain and Japan. Since football analytics is an emerging theme, there are many countries with zero or very poor relations, for example Canada, Argentina, Italy, north Europe, middle East and India. It could be proficient to encourage a more global cooperation for the next years.

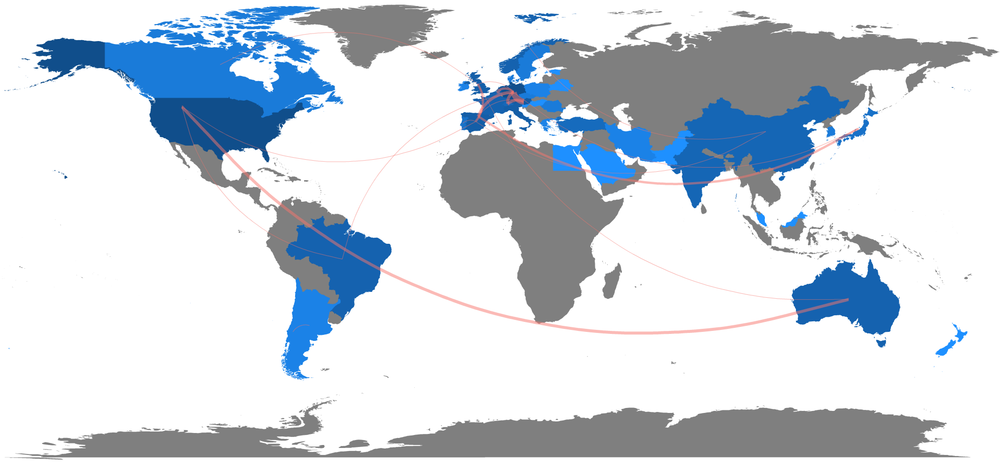

Figure 9: Country collaboration map.

In the following we present the research groups collaboration network: we must keep in mind the guidelines explained in the Sec. 3.3. Furthermore, we show just networks with at least two research groups involved in.

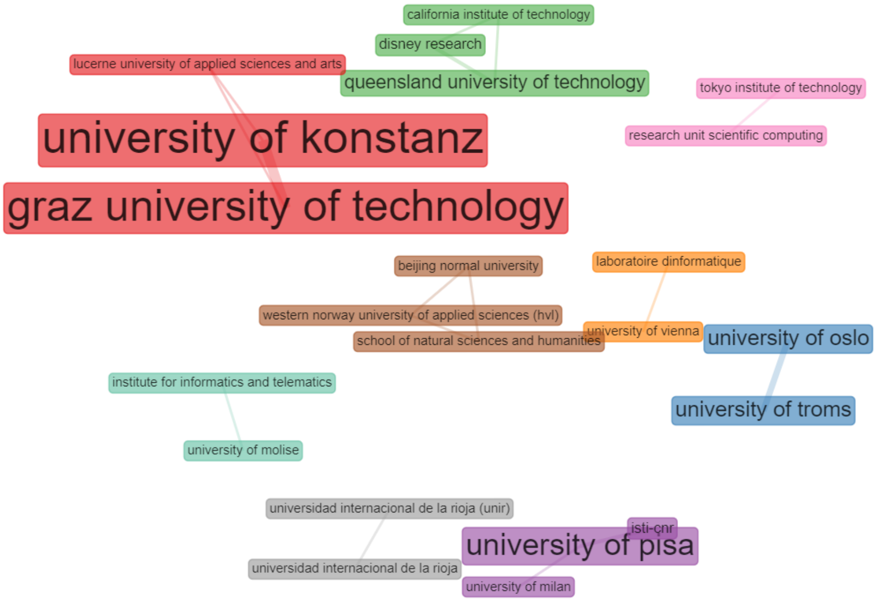
california institute of technology
disney research
lucerne university of applied sciences and arts
queensland university of technology
tokyo institute of technology
university of konstanz
research unit scientific computing
graz university of technology
beijing normal university
laboratoire dinformatique
western norway university of applied sciences (hvl)
university of vienna
university of oslo
school of natural sciences and humanities
institute for informatics and telematics
university of troms
university of molise
universidad internacional de la rioja (unir)
isti-cnr
university of pisa
universidad internacional de la rioja
university of milan

Figure 10: Research collaboration network.

We can see some clear clusters, where the red, green, brown and purple are the most representative ones (i.e. clusters with more than two research groups linked):

- Red cluster is a European group: it is composed from Dutch, Austrian and Switzerland universities.
- The green is an intercontinental cluster (i.e. research groups from USA and Australia).
- Brown cluster is another intercontinental group among China and Norway.
- Purple cluster is an example of single country group: here we find only Italian research institutes.

Each one of the remaining little clusters is composed mainly from strictly continental research groups or from single group.

3.5 In-depth analysis

In this final paragraph, we propose an in-depth analysis thanks two interesting graphs: in Fig. 11 we can see a three-field plot (Cobo et al., 2011; Aria and Cuccurullo, 2017), where there are linked authors, keywords and sources, taking in consideration only the articles; in Fig. 12 instead we can see thematic evolution between first and last five years of the decade, with focus just on conferences.

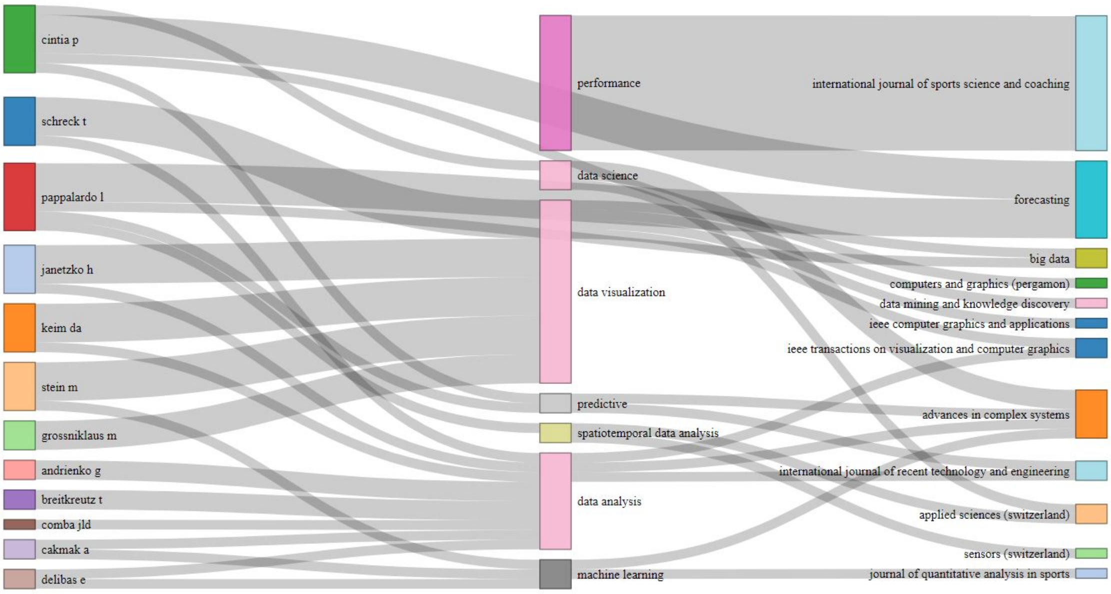

Figure 11: Three-fields articles plot: connection between authors, keywords and sources.

In Fig. 11 we can see top authors-keywords-sources linkage: note that the height of rectangles is proportional to the number of documents produced. Notice that the best source for football analytics article is the International Journal of Sport Science and Coaching while data visualization and performance are the most relevant keywords.

In the last plot (Fig. 12) we can see keywords thematic evolution, where height of rectangles is proportional to the number of documents produced.

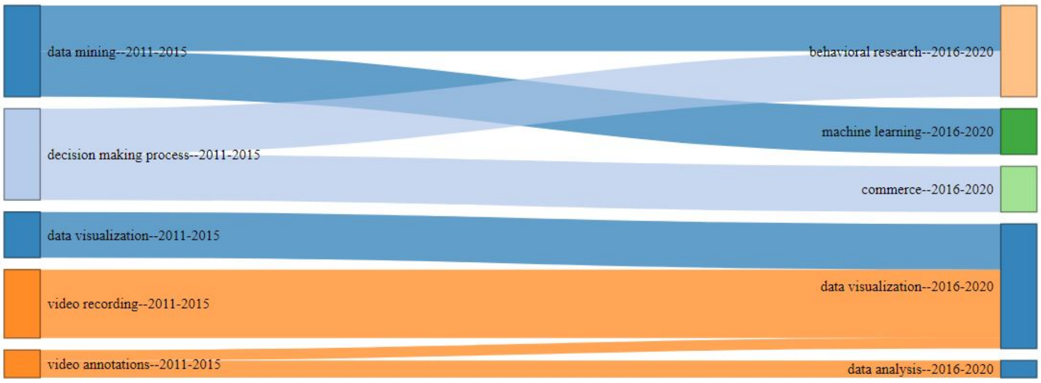

Figure 12: Thematic conference trend over the last ten years.

We can explain the graphic, for example, in this way: conferences focused on data mining in the first five years, in the last half of the decade (2016-2020) moved respectively to behavioural research and machine learning topics. In addition to this, it is interesting to underline the recently growing of data visualization topic conferences (higher rectangle in the last five years than before).

4 Discussion

As viewed until now, we can say that football analytics is an increasing topic in sports research; in particular, the most interconnection keywords used by authors bright us to sum up three crucial theme for football analytics:

- *Technical keywords and tools*, where the recurring ones are data mining, machine learning and artificial intelligence, tools nowadays applied in many branch of our life.
- *Data visualization*, because presentation of results is fundamental in every sector, and also football does not do exception; for instance, it is crucial to show results in the simplest and incisive way to the coaches and technical staff.
- *Performance* is the core of each analysis, in fact nowadays it is crucial for a football team to be able optimizing it.

Since Soccer analytics is an emerging topic, there aren't too much collaboration between research groups yet. For example Italy has productions just in its own country; the main groups are located within continent, except some sporadic case (for example USA and Australia, China and Norway).

Here below are presented, as insights, three relevant articles, produced from authors with the highest dominance factor (see Tab. 2):

- Stein et al. (2019): their work, result from a collaboration of different research groups (from Germany, Portugal, Austria and Switzerland), has its own focus on movement and visual analysis on soccer. They suggest a tool that covers the automatic detection of region-based faulty movement behaviour, as well as the automatic suggestion of possible improved alternative movements. They compare their work with experts knowledge, with an interesting result: an agree index of 83%. So, we can say that their approach could effectively supports analysts and coaches investigating matches. This contribution was published on the Journal of Sports Sciences.

- Pappalardo et al. (2019): this is an example of single-country work, in fact it is produced just from Italian research groups. The aim of this work is to create a sort of synthetic index in order to evaluate objectively football players performance, thanks some event-match data and machine learning/big data techniques. The final goal is to support teams in scouting, and so to evaluate players impartially. This article was published on the ACM Transactions on Intelligent Systems and Technology.

- Bransen and Van Haaren (2018): these Dutchmen authors, thanks artificial intelligence and learning systems, propose a novel approach to measure players' on-the-ball contributions from passes during games. Their method measures the expected impact of each pass on the scoreline. This document was published on the 5th Workshop on Machine Learning and Data Mining for Sports Analytics.

5 Conclusion

As summary, we have seen the important growing of football analytics topics over the last ten years, although it is a niche theme. We have shown how the main goal for researchers and football teams is to support policy-evaluation, thanks the more recent techniques of machine learning and artificial intelligence. Furthermore, another relevant topic is data visualization, in order to show results in the simplest and efficient way to the people without an analytics background. It has been illustrated how this topic has not involved an intercontinental collaboration yet, except some sporadic cases. The final goal of this paper is to guide researchers and practitioners in this new frontier of football research, highlighting the importance of this data-driven revolution.

The direction is traced, we have seen what already exist, but since this is a ”young” theme, there are also many emerging topics to improve and investigate. Eventually, it could be interesting to encourage an exchange between researchers and teams' experts, in order to create a bridge with the club needs: experts and statisticians collaboration could be the future for football.

Acknowledgement

I'm very glad to thank all the Department of Economics and Management of my university and the BDSports[3] group for their precious advice and support.

References

Aria, M. and Cuccurullo, C. (2017). bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4):959–975.

Belfiore, P., Ascione, A., and Di Palma, D. (2019). Technology and sport for health promotion: A bibliometric analysis. *Journal of Human Sport and Exercise*, 10(4):932–942.

Bransen, L. and Van Haaren, J. (2018). Measuring football players' on-the-ball contributions from passes during games. In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, pages 3–15. Springer.

Canova, L. and Canepa, C. (2016). La scienza dei goal: numeri e statistica applicati allo sport più bello del mondo. *La scienza dei goal*, pages 1–174.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., and Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of informetrics*, 5(1):146–166.

López-Carril, S., Escamilla-Fajardo, P., González-Serrano, M. H., Ratten, V., and González-García, R. J. (2020). The rise of social media in sport: a bibliometric analysis. *International Journal of Innovation and Technology Management*, 17(06):2050041.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., and Giannotti, F. (2019). Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–27.

Stein, M., Seebacher, D., Marcelino, R., Schreck, T., Grossniklaus, M., Keim, D. A., and Janetzko, H. (2019). Where to go: Computational and visual what-if analyses in soccer. *Journal of sports sciences*, 37(24):2774–2782.

Vigneshwaran, G. and Kalidasan, R. (2018). Study of publications output on sports science–a bibliometric analysis. *Ganesar College of Arts and Science*, pages 256–260.

[3] www.bodai.unibs.it/bdsports