Big Data in Sports: A Bibliometric and Topic Study

Ana Šuštaršič, Mateja Videmšek, Damir Karpljuk
Faculty of Sport, University of Ljubljana, Slovenia
Ivan Miloloža
Faculty of Dental Medicine and Health, Josip Juraj Strossmayer University of Osijek, Croatia
Maja Meško
University of Maribor, Faculty of Organizational Sciences, Slovenia

Abstract

Background: The development of the sports industry was impacted by the era of Big Data due to the rapid growth of information technology. Unfortunately, that has become an increasingly challenging Issue. **Objectives:** The purpose of the research was to analyze the scientific production of Big Data in sports and sports-related activities in two databases, Web of Science and Scopus. **Methods/Approach:** Bibliometric analysis and topic mining were done on 51 articles selected after four exclusion criteria (written in English, journal articles, the final stage of publication, and a detailed review of all full texts). The software tool used was Statistica Data Miner. **Results:** We found that the first articles appeared in Scopus in 2013 and WoS in 2014. USA and China are countries which produced the most articles. The most common research areas in WoS and Scopus are Public environmental and occupational health, Medicine, Environmental science ecology, and Engineering. **Conclusions:** We conducted that further research and literature review will be required as this is a broad and new topic.

Keywords: Big Data; sport; bibliometric study; topic study; health care management; services; decision making
JEL classification: C8
Paper type: Original article

Received: 09 May 2021
Accepted: 19 Nov 2021

Citation: Šustarić, A., Videmšek, M., Karpljuk, D., Miloloža, I., Meško, M. (2022), "Big Data in Sports: A Bibliometric and Topic Study", Vol. 13 No. 1, pp. 19-34.
DOI: 10.2478/bsrj-2022-0002

Introduction

The phenomenon of Big Data has emerged and has been investigated extensively so far (Zhao et al., 2015). Due to the speed growing data volume, various devices, i.e., the Internet of Things and Social Networks, are used. With a growing population, about 2 billion people are connected to the Internet, and 5 billion use various mobile devices (Worldometer, 2021). These billions of people, through multiple devices, will produce an enormous amount of data (Khan et al., 2014). The estimation in 2013 was that in 2020 50 billion devices will be connected to the Internet. Therefore, data production would be 44 times greater than in 2009 (Sagiroglu et al., 2013). As information is transferred and shared at the speed of light on optical fibre and wireless networks, the volume of data and marker growth speed increase (Khan et al., 2014, Che et al., 2013). This data is known as the term Big Data. Big Data is "a collection of data that is huge in volume yet growing exponentially with time. It is data with such a large size and complexity that no traditional data management tools can store it or process it efficiently" (Favaretto et al., 2020). The pervasive nature of digital technologies and the broad range of data-reliant applications have also made this expression widespread across disciplines, including sociology, medicine, sports science, biology, engineering, economics, management, and information science (De Mauro et al., 2016).

In response to the ever-increasing network data challenges, there is a globally noticeable development trend of convergence with big data frameworks, network analytical modelling, link or path prediction, and recommendation systems (Li et al., 2019). Big Data is characterized and defined differently by different authors. Some authors (Dijcks, 2012; Schroeck et al., 2012, Chen et al., 2014) define Big Data with four characteristics (volume, velocity, variety and value). Another author (Suthaharan, 2013) argues that “data can be defined using three data characteristics (cardinality, continuity and complexity)”. They are also defined with three characteristics, i.e. volume, velocity and variety (Laney, 2021). Furthermore, Big Data can be described in the five V's, which stand for value, volume, velocity, variety, and validity) (Oguntimilehin et al., 2014) or even more precisely by ten characteristics: venue, variety, volume, variability, value, validity, veracity, velocity, vocabulary, and vagueness (Borne, 2021).

These characteristics provide a research horizon for researchers and practitioners to effectively manage Big Data. Three main characteristics of Big Data will be presented below. Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, sensors, images, videos, etc. (Rajeshwari Sreenivasan, 2017). Variety is one of the important characteristics of Big Data. It refers to structured, unstructured, and semistructured data gathered from multiple sources (De Mauro et al., 2016, 2015). Nowadays, data comes in various forms, for example, emails, PDFs, images, videos, audio, SM spots, transactions, and long data (Owais et al., 2016). The third characteristic is velocity. This is the speed at which the data is generated, stored, analyzed, and processed (Al Nuaimi et al., 2015). Velocity essentially refers to the speed at which data is being created in real-time. It comprises the rate of change in a broader prospect, linking incoming data sets at varying speeds and activity bursts. It is challenging to analyze Big Data that is generated at a very high velocity (Morbetz, 2021).

Big Data is “the most valuable and powerful fuel that can run the massive IT industries of the 21st Century. Big Data is the most widespread technology used in almost every business sector (Chen et al., 2014)”. It has gained interest and application in various fields such as government, healthcare and medicine, retail sector, agriculture, research, online and social media, telecommunication and banking

(Singh et al., 2015). It can also be applied in a different scientific discipline (Khan et al., 2014).
Lifestyle is the typical way of life of an individual, group, or culture-dependent on many different factors. It is divided into a healthy and unhealthy lifestyle. Sports and physical activity have an important role in a healthy lifestyle. In the last two decades, the importance of physical health has grown, and people nowadays give close and thoughtful attention to the healthy development of their bodies. Many studies show a positive correlation between sports and life quality (Snedden et al., 2019, Wu et al., 2017, Marker et al., 2017).

The period of Big Data has also impacted sport development, and it still has influence. However, it has influenced professional sports and various fields connected with the sport and sports industry (Liu et al., 2020). Advanced Big Data technique has brought changes in sports. In sports, Big Data has generated new opportunities and challenges with the aim of spreading sports data (Patel et al., 2020). Big Data services closely related to it, including exercise performance, training statics, data connected with body characteristics, health data, analysis, etc., can effectively help athletes in daily training, planning, and developing sports skills and motor capabilities. Consequently, Big Data is becoming indispensable and helps win the competition (Pappalado et al., 2019).

Big Data is common in professional sports but is not the only one. The connection between Big Data and different sports fields can be found in many areas that are not directly related to sports (Cheng et al., 2021). The mentioned areas of research on the use of sport and Big Data are, for example, engineering (Kim et al., 2020), medicine (Hayano et al., 2019, Al-Mallah et al., 2014) and public health (Park et al., 2020), social and computer science (Hou et al., 2017), business and management (Khazaeli et al., 2016), environmental science (Phan et al., 2020) and some others.

In short, given the topic's interest, it would be informative to research trends in Big Data utilization in sports to identify trends and topics that emerge in the literature. The study about Big Data and sport was driven by the five research questions that guided us through the work: (RQ1) What has been production status over time?; (RQ2) Which countries produce the most papers?; (RQ3) What are the most common areas of research?; (RQ4) Which are the most frequent words and phrases in papers?; and (RQ5) What is the analysis according to the "Topic mining" method?

Methodology

A literature review on using Big Data in sports and sports-related activities was done, were articles published on or before February 25th were analyzed. Two databases were used: Scopus and Web of Science (WoS). The following steps were used for gathering literature. First, research platforms were used to identify relevant papers. Second, we reviewed and tried different keywords based on the read literature. Third, a logical combination of keywords was used to find relevant documents. This was a combination of keywords: ((“Big Data”) AND (“sport activity” OR “physical activity” OR “professional sport”)). The Boolean operators "AND" and "OR" were used.

Our study used only articles published in peer-reviewed journals. This was our first exclusion criterion of choosing, and the review was limited to open-access documents and documents written in English. The first search before exclusion criteria resulted in 344 results, of which 166 were published in WoS and 178 were published in Scopus. Based on our exclusion criteria, the search resulted in 88 articles, of which 64 were published in WoS and 24 were published in Scopus. One article was published in both databases.

We performed a detailed review of all abstracts and full texts to select the articles that are focused on our research topic and not just mention it sporadically. Based on this criterion, 51 of 88 articles were selected, and a bibliometric analysis was conducted on 51 articles. The next step was text mining which was conducted using Provalis Wordstat. Finally, we analyzed the texts according to the clusters identified by topic mining.

Results

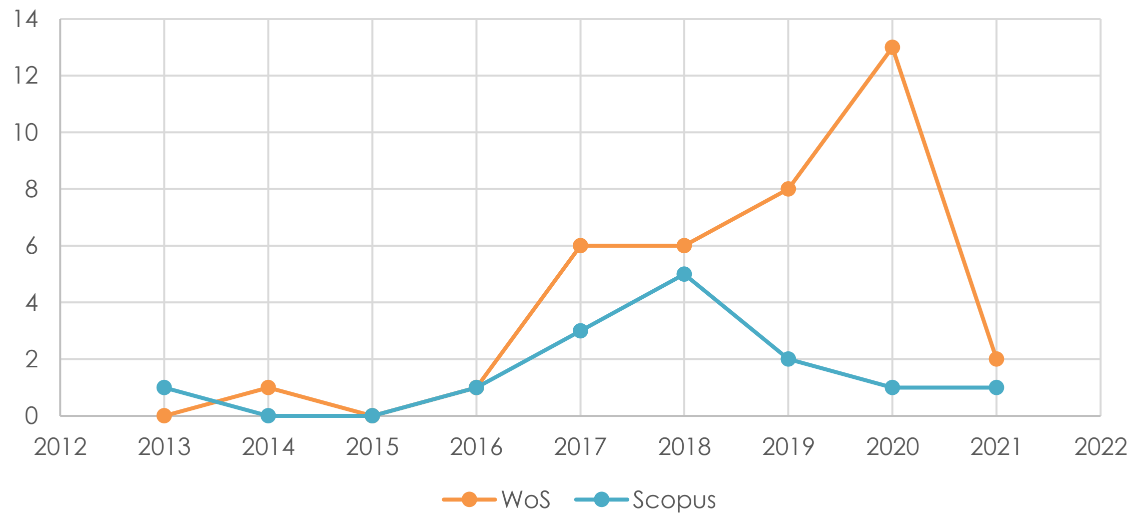
This literature review presented insight into Big Data's incidence in sports and targeted 51 articles published on or before February 25th, 2020, published in two databases, namely Scopus and Web of Science.

Bibliometric analysis

Considering the variable "year production", a similar evolution was observed in both databases. The first contributions to Big Data in sports were detected in WoS 2014 and Scopus one year earlier – in 2013. The number of contributions has been increasing since the year 2016. A change was recorded in 2015 because there were no published articles in both database. In WoS were the higher rate of published articles in 2020 (13 articles; 35,1 %), 2019 (8 articles; 21,6 %) and 2018 and 2017 (both 6 articles; 16,2 %). In Scopus were the higher rate of published articles in 2018 (5 articles; 35,7 %), 2017 (3 articles; 21,4 %) and 2019 (2 articles, 14,3 %). The results showed the maximum production peaks in each database; the maximum was reached in Scopus in 2018 and WoS two years later, in 2020.

Figure 1 shows the number of publications on "Big Data and sport" from 2013 to 2021.

Figure 1
Scientific yearly production in WoS and Scopus (# of publications)



Source: Author's work

Considering the variable "country" that produced the most literature, the first place was different in WoS and Scopus. While the USA was first in WoS (43,2 %), it was second in Scopus (35,7 %). China produced the most articles in Scopus (42,8 %). There were three countries with the same percentage (14,3 %) in the database Scopus in third place, i.e. Australia, the Netherlands and India. China which took first place in Scopus in WoS was in second place with England. Their percentage was 13,5 %. South Korea, with 10,8 %, took third place. In WoS, Australia produced 8,1 % of literature and the Netherlands 5,4 % of literature. Both countries were in third place in Scopus. India, which shared third place with Australia and the Netherlands, had not produced any articles in the WoS database.

Table 1 shows countries with the highest production in WoS and Scopus.

Table 1
Countries with the highest production in WoS and Scopus

Country	WoS		Scopus		Both databases	
	N	%	N	%	N	%
USA	16	41,2	5	35,7	21	41,2
China	5	13,5	6	42,8	11	21,5
England	5	13,5	-	-	5	9,8
South Korea	4	10,8	-	-	4	7,8
Australia	3	8,1	2	14,3	5	9,8
Germany	3	8,1	-	-	3	5,8
Finland	2	5,4	-	-	2	3,9
France	2	5,4	-	-	2	3,9
the Netherlands	2	5,4	2	14,3	4	7,8
Spain	2	5,4	-	-	2	3,9
India	-	-	2	14,3	2	3,9
Canada	1	2,7	1	7,1	2	3,9

Source: Author's work

For the variable "organizations", the two databases' institutions with more references differed. Besides, those that appeared in WoS were not found in Scopus. In Scopus, we analyzed 14 articles, and all articles had a different institution. In WoS, Johns Hopkins University and the University of London had three references (8,1 %), and twelve articles had two references (5,4 %).

Articles used in the analysis were published in different journals in both databases. In Scopus, only one journal published more than one contribution, Boletin Tecnico Technical Bulletin (21,4 %), and other journals published one contribution. In the database WoS, three journals have published more than one contribution, i.e. four articles International Journal of Environmental Research and Public Health (10,8 %) and two articles (5,4 %) in BMC Public Health and Sensors.

The variable "research areas" was similar in both databases because they were mostly connected with medicine, public health, computer and environmental science and engineering. The four most common research areas in WoS were 27,0 % for public environmental and occupational health, 16,2 % for environmental science ecology, 10,8 % for computer science, engineering and science technology other topics and 8,1 % for other topics were also three areas: chemistry, medical informatics and psychology. Other research areas in WoS have had less than eight per cent. The three most common research areas in Scopus were two areas with 19,2%: engineering and medicine, three areas with 11,5 %: computer science, materials science and social sciences and three areas with 7,7 %: biochemistry, genetics and molecular biology, health professions and business, management and accounting. Other research areas in Scopus have had less than seven per cent.

Table 2 presents the most common research areas in both databases in percentages.

Table 2
The most common research areas in WoS and Scopus

Researching areas	WoS	Scopus
	%	%
Public environmental and occupational health	27,0	-
Medicine	-	19,2
Environmental science ecology	16,2	-
Computer science	10,8	11,5
Engineering	10,8	19,2
Science, technology, other topics	10,8	-
Chemistry	8,1	-
Medical informatics	8,1	-
Psychology	8,1	-
Material science	-	11,5
Social science	-	11,5
Biochemistry, genetics and molecular biology	-	7,7
Health professions	-	7,7
Business, management and accounting	-	7,7

Source: Author's work

Topic mining

To discover the most frequent topics found in abstracts, we used the phrase and word extraction process and cluster analysis functions provided by the software, namely WordStat Provalis. The analysis was done with a method by Column Frequency-Inverse Document Frequency (TF-IDF), which shows the importance of each phrase within the collection of papers. A phrase with higher TF-IDF values is highly important (the last column in Table 3).

Figure 2 shows the word cloud of the most frequent words found in 51 abstracts.

Figure 2
Wordcloud of the most frequent word

Source: Author's work

The most commonly found word was: data (frequency 137), analysis (frequency 54), health and study (both with frequency 51), based (frequency 51) and sports (frequency 40). Five words with the highest TF-IDF were sports (32,2), cycling (30,3), weather (24,6), exercise (22,0) and obesity (21,6). Table 3 presents the following most frequently used phrases, where the frequency of occurrence was higher or equal to 4.

Table 3
Most frequent phrases in papers (frequency 4+)

Phrase	F	NO. C.	% C.	L	TF • IDF	Phrase	F	NO. C.	% C.	L	TF • IDF
Heart rate	11	4	7,84%	2	12,2	Human movement	5	1	1,96%	2	8,5
Machine learning	9	7	13,73%	2	7,8	Kcal min	5	1	1,96%	2	8,5
Senility death ratio	8	1	1,96%	3	13,7	Regression models	5	4	7,84%	2	5,5
Wearable devices	8	5	9,80%	2	8,1	Smartphone app	5	2	3,92%	2	7,0
Bicycle usage	6	2	3,92%	2	8,4	Social media	5	2	3,92%	2	7,0
Chronic disease	6	2	3,92%	2	8,4	Tai chi	5	1	1,96%	2	8,5
Colorectal cancer	6	1	1,96%	2	10,2	Time series	5	4	7,84%	2	5,5
Decision making	6	4	7,84%	2	6,6	Weather conditions	5	1	1,96%	2	8,5
Deep learning	6	4	7,84%	2	6,6	Behaviour change	4	1	1,96%	2	6,8
Google street view	6	3	5,88%	3	7,4	Cancer survivors	4	1	1,96%	2	6,8
Green parks	6	1	1,96%	2	10,2	College students based	4	2	3,92%	3	5,6
Sports activities	6	5	9,80%	2	6,1	Cultural space	4	1	1,96%	2	6,8
Case studies	5	1	1,96%	2	8,5	Geographic information	4	4	7,84%	2	4,4
Energy expenditure	5	3	5,88%	2	6,2	Green spaces	4	2	3,92%	2	5,6
Fit project	5	1	1,96%	2	8,5	Neural networks	4	2	3,92%	2	5,6
Health outcomes	5	2	3,92%	2	7,0	Older adults	4	2	3,92%	2	5,6

Note: Legend: F – frequency, No. C. – number of cases, % C. – percentages of cases, L – length
Source: Author's work

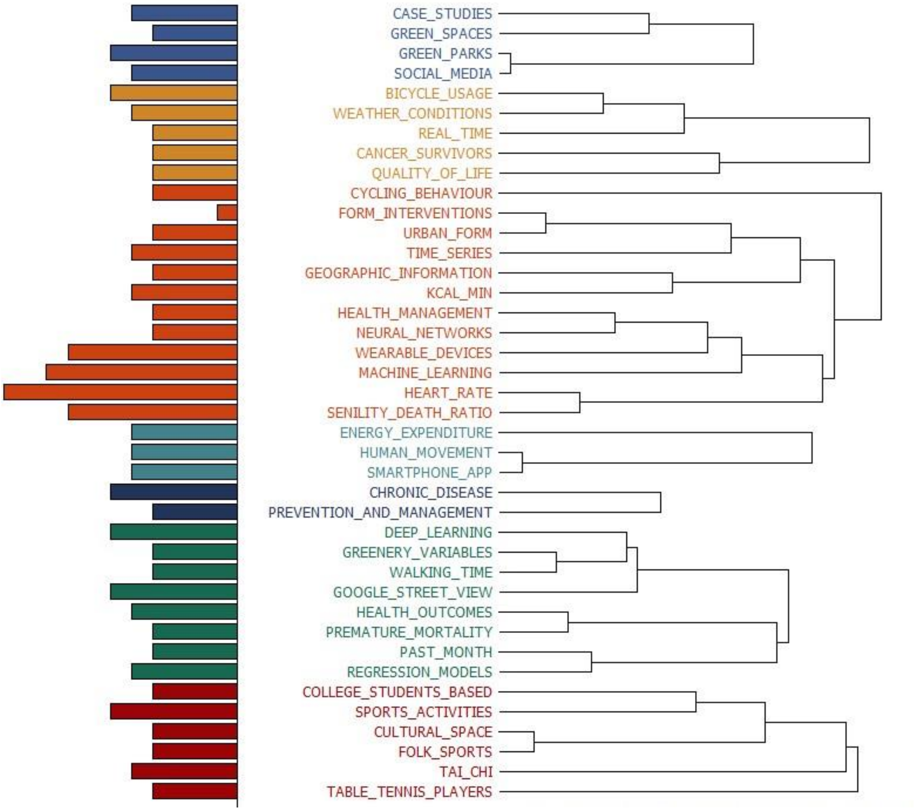
The most used phrases: heart rate (frequency 11), machine learning (frequency 9), senility death ratio (both with frequency 8) and bicycle usage, chronic disease, colorectal cancer, decision making, deep learning, Google street view, green parks and sports activities (all with frequency 6). Four phrases with the highest TF-IDF were senility death ratio (13,7), heart rate (12,2) and colorectal cancer and Google street view (both 10,2).

Figures 3 and 4 present the cluster analysis results that identified seven topics concerning sports and sports activities related to Big Data:

- *Cluster 1* includes abstracts with the co-occurring phrases: case studies, green spaces, green parks, and social media.
- *Cluster 2* includes abstracts with the co-occurring phrases: bicycle usage, weather conditions, real-time, cancer survivors, and quality of life.
- *Cluster 3* includes abstracts with the co-occurring phrases: cycling behaviour, form interventions, urban form, time-series, geographic information, kcal/min, health management, neural networks, wearable devices, machine learning, heart rate, and senility death ratio.
- *Cluster 4* includes abstracts with the co-occurring phrases: energy expenditure, human movement, and smartphone app.
- *Cluster 5* includes abstracts with the co-occurring phrases: chronic disease and prevention and management.

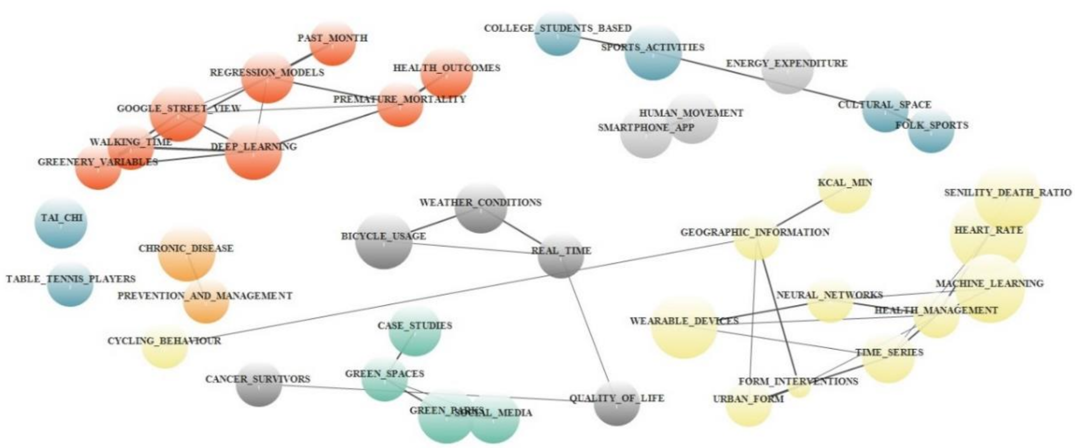
- *Cluster 6* includes abstracts with the co-occurring phrases: deep learning, greenery variables, walking time, Google street view, health outcomes, premature mortality, past month, and regression models.
- *Cluster 7* includes abstracts with the co-occurring phrases: college student-based, sports activities, cultural space, folk sports, tai chi, and table tennis players.

Figure 3
Cluster analysis of phrases



Source: Author's work

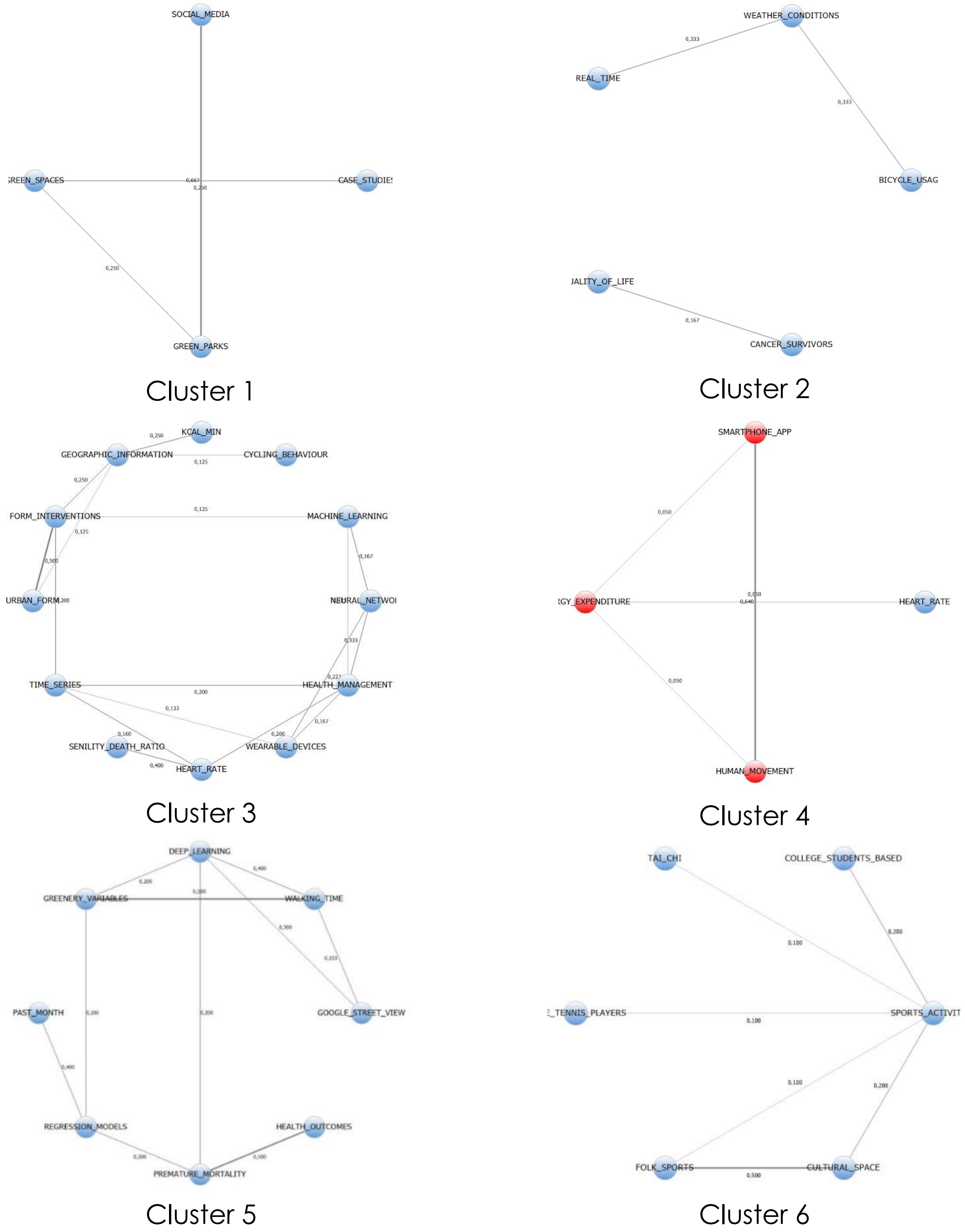
Figure 4
Mapping of clusters



Source: Author's work

We did the additional analysis with the link analysis feature to visualize the connections between keywords. In our case, we used phrases instead of keywords, which offers a high level of interactivity, and allows us to explore relationships and detect underlying patterns and structures of co-occurrences. A line between elements shows a relationship between elements, and the thickness of this line represents the strength of this relationship. Figure 5 presents six network graphs indicating most often co-occurred phrases within each cluster.

Figure 5
Proximity plots of phrases in selected clusters



Source: Author's work

Table 4 presents an overview of the most salient topics from 51 abstracts using state-of-the-art automatic topic extraction techniques. Column sign with Coh. (coherence) shows the percentage of coherence explained, which means the smaller the segment, the lower percentage. Column F (frequency) displays the total frequency of all items listed in the keyword column, and column sign with C displays the number of cases containing at least one of the items listed in the column with keywords. The topic with the highest total frequency and the highest number of cases is the topic "devices data" (percentage of cases is 92,16 %).

Table 4
Topic mining

NO	TOPIC	KEYWORDS	Coh.	F	C	% C
1	MANAGEMENT PREVENTION	management; prevention; change; lifestyle; disease; support; knowledge; program; health;	0,423	69	31	60,78%
2	SMARTPHONE APP HUMAN MOVEMENT	human; smartphone; analytics; people; healthcare; human movement; smartphone app;	0,322	32	17	33,33%
4	HEALTHCARE ASSOCIATIONS	proposed; collected; services; results; variables; paper; built; healthcare; show; associations; system;	0,411	73	44	86,27%
5	INTERVENTIONS CHALLENGES	provide; challenges; individual; understanding; time; volume; interventions; information; healthy; large;	0,427	82	43	84,31%
6	SENSORS DESIGN	sensors; design; people; life; monitoring; technologies; user; support;	0,395	50	32	62,75%
7	DEVICES DATA	devices; data; activity; daily; patterns; collected; frequency; wearable; analysis;	0,407	144	47	92,16%
8	URBAN TRAVEL	bus; mode; urban; travel; mile; city; metro; gsv;	0,423	65	15	29,41%
9	RISK FACTORS	factors; risk; networks; volume; cross; healthy; prediction; population; life; large; study; results; obesity; risk factors;	0,422	98	43	84,31%
10	HOME CONFINEMENT	home; confinement; activity; monitoring; home confinement;	0,381	38	19	37,25%
11	WOMEN MEN	women; men; ratio; death; age; higher; low;	0,530	61	16	31,37%
12	DEVELOPMENT TRAINING	development; training; system; performance; technology; user; quality; lifestyle; proposed; psychological; sports; sports activities;	0,425	73	34	66,67%
13	ACCELEROMETER ACTIVITIES	accelerometer; living; compared; feature; participants; measures; activities;	0,416	50	29	56,86%
14	PARK WEATHER	park; findings; users; behaviour; weather; show; features; distance; study; feature; results; related;	0,426	85	39	76,47%
15	ENVIRONMENTAL RUNNING	environmental; mile; distance; performance; running;	0,338	38	18	35,29%
16	GREEN BEHAVIOR	greenery; traditional; green; methods; activities; daily; show; variables; life; low; time; behaviour;	0,379	83	43	84,31%
17	STUDENTS SPORTS	students; sports; physical; fitness; technology; paper; analysis; activities; based; change; Big Data analysis;	0,428	124	45	88,24%
18	CITY SOCIAL	built; social; methods; analytics; survey; traditional; studies; challenges; understanding; collected; city; knowledge;	0,388	74	41	80,39%
19	REGRESSION MODELS	walking; GPS; survey; models; neighbourhood; greenery; index; regression models;	0,372	53	21	41,18%

20	MODEL PREDICTION	model; increasing; user; variables; show; provide; paper; distance; models; frequency; prediction; change; training; networks;	0,435	88	45	88,24%
21	CANCER EXERCISE	cancer; exercise; patients; individual; factors; total; years; social;	0,385	75	33	64,71%
22	RESEARCH OBESITY	research; influence; obesity; health; understanding; knowledge;	0,394	67	38	74,51%
23	HEART RATE	rate; heart; death; metro; healthcare; analysis; low; performance; heart rate;	0,402	63	34	66,67%
24	GOOGLE STREET VIEW	street; view; level; environment; mortality; built; associations; obesity; green; Google street view;	0,457	77	29	56,86%
25	MODELS DEVELOPMENT	features; including; models; disease; learning; features; techniques; studies; low; development;	0,410	62	39	76,47%
26	ROAD CYCLING	road; cycling; weather; city; cities;	0,401	50	12	23,53%
27	WALKABILITY CITIES	walkability; cities; framework; index; related; city;	0,379	36	23	45,10%
28	POPULATION QUALITY	behaviour; increasing; metro; low; large; population; associations; environmental; quality; patterns;	0,397	60	36	70,59%
29	OUTCOMES CARDIOVASCULAR	outcomes; population; age; years; related; knowledge; clinical; cardiovascular; influence; design; studies; mortality;	0,433	71	37	72,55%
30	WEARABLE DEVICES	wearable; devices; healthcare; large; techniques; risk; monitoring; services; technology; wearable devices;	0,409	58	32	62,75%

Note: Legend: Coh. – coherence, F – frequency, C – cases, % C – % cases
Source: Author's work

Bibliometric analysis of the literature shows us that Big Data in sports occurs differently. The term "sport". We found that the term "sport" occurs in a very wide range, not just as the "sport" that we all imagine (e.g. running, group workout, tennis …). As we have already written and shown in Table 4, the topic "devices data" has the highest percentage of cases. The group of topics with a high percentage of cases (% C) is related to medicine, health and public health, namely topics "healthcare associations" (86,27 %), "interventions challenges" (84,31 %), "risk factors" (84,31 %), "research obesity" (74,51 %), "outcomes cardiovascular" (72,55 %), "heart rate" (66,67 %) and "cancer exercise" (64,71 %). Research related to this topic has analyzed different sizes and types of data. Some authors (Kharabian Masouleh et al., 2018) have determined the effects of common cardiovascular risk factors on vulnerable grey matter networks in a large and well-characterized population-based cohort that involved 616 healthy elderly of the LIFE-Adult-Study. Researchers (Hayano et al., 2019) have examined the regional difference in senility death ratio with the regional differences in heart rate variability and physical activity in 108,865 men and 136,536 women. Other researchers (Nguyen Quynh et al., 2017) examined the relationship between these neighbourhood characteristics and obesity and diabetes diagnoses (Type 1 and Type 2). They collected 422,094 tweets and leveraged administrative and clinical records on 1.86 million individuals. Phan and co-researchers (Phan et al., 2020) wanted to discover the associations between select neighbourhood-built environment indicators (e.g. crosswalks) and health outcomes (e.g. obesity, diabetes), and premature mortality, at the state level. Using deep learning techniques, they utilized 31,247,167 images collected from Google Street View to create indicators for neighbourhood-built environment characteristics. These four studies show the extent and size of Big Data and the different spectrum of use. However, a smaller initial number of participants or measuring devices does not necessarily mean a smaller number of final data. Several different data can be obtained from one measured participant or measuring device or monitored for a longer period (Wang et al., 2021, Sung-Un et al., 2020, Raywood et al., 2020)

Discussion

Big Data has become of utmost importance in competitive and professional sports, and its processing leads to sports analytics, which is part of data science (Kaur et al., 2020). The effective use of Big Data opens substantial doors because it can influence the results and, consequently, athletes' careers. Athletes can gain the best from individual and personalized training sessions, which is made possible by a large database. Thus, the coach can focus on the athlete's weaker side and precise goals (Morgulev et al., 2019, Liu, 2019).

Table 4 shows that the topics "development training", "model prediction", and "models' development" have a relatively high percentage of cases, between 66,67 % and 88,24 %. All three topics are linked to professional sports and play a key role. The topic of “model prediction” stands out among these three topics, as it is not only related to professional sports but many research areas; for example, rehabilitation and medicine (Kokkotis et al., 2020, Emig et al., 2020), recreation and urban mobility – prediction of physical activity (Goel et al., 2018, Saez et al., 2016).

As described in the “Results" section, the databases first appeared in 2013 in Scopus and in 2014 in WoS. In 2013, the results of a consensus meeting on non-communicable chronic disease prevention were sponsored by the International Olympic Committee in April 2013 in Lausanne. They wanted to encourage new creative approaches that leverage and integrate evidence through the support of Big Data, technology, and design thinking in different areas (Sports and Exercise Medicine, lifestyle, modern technology, prevention etc.) (Matheson et al., 2013). In WoS in 2014 was published research about The FIT Project (the Henry Ford Exercise Testing Project), which “is unique in its combined use of directly measured clinical exercise data retrospective collection of medical history and medication treatment data at the time of the stress test” with the help of administrative and medicine databases for a prognosis for physical activity and health factors (Al-Mallah et al., 2014). In both articles, we can observe that it is a topic related to health and lifestyle concerning prediction.

Conclusion

The purpose of the study was to conduct bibliometric analysis and data mining using Statistica Data Miner analysis on the use of Big Data in sports and sports-related activities. We used two databases, WoS and Scopus, from which we selected 51 articles. The last part of the methodology was analysis according to the clusters identified by the topic mining. Our study has shown interesting connections between Big Data and sport (research areas, applicability, different sample size, etc.). However, the study has certain limitations. The main limitations are i) only two databases were used; ii) the search was limited to articles written in English; iii) only three sports-related keywords were used. We believe that the number of articles will increase, as the topic of “Big Data in sport” is a relatively new topic that will be further developed.

References

1. Al Nuaimi, E., Al Neyadi, H., Mohamed, N., Al-Jaroodi, J. (2015), "Applications of big data to smart cities", Journal of Internet Services and Applications, Vol. 6 No. 1, pp. 25.
2. Al-Mallah, M. H., Keteyian, S. J., Brawner, C. A., Whelton, S., Blaha, M. J. Rationale (2014), "Design of the Henry Ford Exerclse Testing Project (The FIT Project)", Clinical Cardiology, Vol. 37 No. 8, pp. 456-461.
3. Borne, K. (2021), "Top 10 List – The V's of Big Data. Data Science Central – a Community for big data practitioners", available at https://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data (8 March 2021)
4. Che, D., Safran, M., Peng, Z. (2013), "From big data to big data mining: challenges, issue, and opportunities", Database Systems for Advanced Applications, Vol. 19 No. 2, pp. 1-15.
5. Chen, M., Mao, S., Liu, Y. (2014), "Big data: A survey", Mobile Networks and Application, Vol. 19 No. 2, pp. 171-209.
6. Cheng, Y., Song, Y. (2021) "Sports big data analysis based on cloud platform and its impact on sports economy", Mathematical Problems in Engineering, Vol. 21 No. 2, pp. 1-12.
7. De Mauro, A., Greco, M., Grimaldi, M. (2015), "What is big data? A consensual definition and a review of key research topics", AIP Conference Proceedings, Vol. 1644 No. 1, pp. 97-104.
8. De Mauro, A., Greco, M., Grimaldi, M. (2016), "A formal definition of big data based on its essential features", Library Review, Vol. 63 No. 1, pp. 122-135.
9. Dijcks, J. (2012), "Oracle: Big data for the enterprise", Redwood Shores, USA, pp. 1–16.
10. Emig, T., Peltonen, J. (2020), "Human running performance from real-world big data", Nature communications, Vol. 11 No. 1, pp. 4936-4945.
11. Favaretto, M., De Clercq, E., Schneble, C. O., Elger, B. S., Fischer, F. (2020), "What is your definition of Big Data? Researcher's understanding of the phenomenon of the decade", PLoS One, Vol. 15 No. 2, pp. 1-20.
12. Goel, R., Garcia, L. M. T., Goodman, A., Johnson, R., Aldred, R., Murugesan, M., Brage, S., Bhalla, K., Srinivasan, M. (2018), "Estimating city-level travel patterns using street imagery: A case study of using Google Street View in Britain", PLoS One, Vol. 13 No. 5, pp. 1-22.
13. Hayano, J., Kisohara, M., Yoshida, Y., Sakano, H., Yuda, E. (2019), "Association of heart rate variability with regional difference in senility death ratio: ALLSTAR big data analysis", SAGE Open Medicine, Vol. 7 No. 2, pp. 1-7.
14. Hou, X., Jiang, J. (2017), "Analysis on the mental quality and performance of table tennis players based on the cloud computing big data", Technical Bulletin, Vol. 55 No. 1, pp. 348-354.
15. Kaur, G., Jagdev, G. (2020), "Analyzing and Exploring the Impact of Big Data Analytics in Sports Science", in Indo - Taiwan 2nd International Conference on Computing, Analytics and Networks, Indo-Taiwan ICAN, pp. 218-224.
16. Khan, N., Yaqoob, I., Targio Hashem, I. A., Inayat, Z., Mahmud Ali, W. K., Alam, M., Shiraz, M., Gani, A. (2014), "Big Data: Survey, Technologies, Opportunities, and challenges", The Scientific World Journal, Vol. 2014 No. 2, pp. 1-18.
17. Kharabian Masouleh, S., Beyer, F., Lampe, L., Loeffler, M., Luck, T., Riedel-Heller, S., Schroeter, M. L., Stumvoll, M., Villringer, A., Witte, A. V. (2018), "Gray matter structural networks are associated with cardiovascular risk factors in healthy older adults", Journal of Cerebral Blood Flow & Metabolism, Vol. 38 No. 2, pp. 360-372.
18. Khazaeli, M., El Kari, C. (2016), "The effects of technology and big data in sports industry", in Proceedings of the 2016 Industrial and Systems Engineering Research Conference, ISERC 2016-2020, pp. 2404-2409.
19. Kim, S. W., Lee, K., Sohn, J. S., Cha, S. W. (2020), "Product development using online customer reviews: A case study of the South Korean subcompact sport utility vehicles market", Applied Sciences, Vol. 10 No. 19, pp. 1-12.
20. Kokkotis, C., Moustakidis, S., Giakas, G., Tsaopoulos, D. (2020), "Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients", Applied Sciences, Vol. 10 No. 19, pp. 1-23.
21. Laney, D. (2012), "3D Data Management: Controlling Data Volume, Velocity and Variety", available at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (15 March 2021)

22. Li, J., Deng, K., Huanh, X., Xu, J. (2019), "Analysis and application of location-aware big complex network data", Complexity, Vol. 2019 No. 1, pp. 1-2.
23. Liu, G., Luo, Y., Schulte, O., Kharrat, T. (2020), "Deep soccer analytics: learning an action-value function for evaluating soccer players", Data Mining and Knowledge Discovery, Vol. 34 No. 2, pp. 1531-1559.
24. Liu, H. (2019), "Opportunities, challenges and Countermeasures for the development of China's sports industry in the era of big data", Journal of Physic: Conference Series, Vol. 1237 No. 2, pp. 1-14.
25. Marker, A. M., Steele, R. G., Noser, A. E. (2017), "Physical activity and health-related quality of life in children and adolescents: A systematic review and meta-analysis", Health Psychology, Vol. 37 No. 1, pp. 893-903.
26. Matheson, G. O., Klügl, M., Engebretsen, L., Bendiksen, F., Blair, S. N., Börjesson, M., Budgett, R., Derman, W., Erdener, U., Ioannidis, J. P. A., Khan, K. M., Martinez, R., van Mechelen, W., Mountjoy, M., Sallis, R. E., Schwellnus, M., Shultz, R., Soligard, T., Steffen, K., Sundberg, C. J., Weiler, R., Ljungqvist, A. (2013), "Prevention and Management of Noncommunicable Disease", Clinical Journal of Sport Medicine, Vol. 23 No. 2, pp. 419-429.
27. Mobertz, L. (2019), "The Four V" s of Big Data", available at https://www.bigdataframework.org/ four-vs-of-big-data/ (24 March 2021)
28. Morgulev, E., Azar, O. H., Lidor, R. (2018), "Sports analytics and the big-data era", International Journal of Data Science and Analitics, Vol. 5 No. 4, pp. 156-159.
29. Nguyen Quynh C., Brunisholz, K. D., Yu, W., McCullough, M., Hanson, H., Litchman, M. L., Li, F., Wan, Y., VanDerslice, J. A., Wen, M., Smith, K. R. (2017), "Twitter-derived neighborhood characteristics associated with obesity and diabetes˝, Scientific Report, Vol. 7 No. 1, pp. 1-10.
30. Oguntimilehin A., Ademola E. O. (2014), "A Review of Big Data Management, Benefits and Challenges", A Review of Big Data Management, Benefits and Challenges, Vol. 5 No. 6, pp. 433-438.
31. Owais, S. S., Hussein, N. S. (2016), "Extract Five Categories CPIVW from the 9V "s Characteristics of the Big Data", International Journal of Advanced Computer Science and Applications, Vol. 7 No. 3, pp. 254-258.
32. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E. (2019), "A public data set of spatio-temporal match events in soccer competitions", Scientific Data, Vol. 6 No. 1, pp. 1-15.
33. Park, S. U., Ahn, H., Dong, K., So, W. (2020), "Big Data Analysis of Sports and Physical Activities among Korean Adolescents", International Journal of Environmental Research and Public Health, Vol. 17 No. 15, pp. 5577-5589.
34. Patel, D., Shah, D., Shah, M. (2020), "The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports", Annals of Data Science, Vol. 7 No. 1, pp. 1-16.
35. Phan, L., Yu, W., Keralis, J. M., Mukhija, K., Dwivedi, P., Brunisholz, K. D., Javanmardi, M., Tasdizen, T., Nguyen, Q. C. (2020), "Google Street View Derived Built Environment Indicators and Associations with State-Level Obesity, Physical Activity, and Chronic Disease Mortality in the United States", International Journal of Environmental Research and Public Health, Vol. 17 No. 10, pp. 3659-3669.
36. Rajeshwari Sreenivasan, R. (2017), "Characteristics of Big Data – A Delphi study", Newfoundland, Faculty of Business Administration Memorial University of Newfoundland, pp. 13-34.
37. Raywood, E., Douglas, H., Kapoor, K., Filipow, N., Murray, N., O'Connor, R., Stott, L., Saul, G., Kuzhagaliyev, T., Davies, G., Liakhovich, O., Van Schaik, T., Furtuna, B., Booth, J., Shannon, H., Bryon, M., Main, E. (2020), "Protocol for Project Fizzyo, an analytic longitudinal observational cohort study of physiotherapy for children and young people with cystic fibrosis, with interrupted time-series design", BMJ Open, Vol. 10 No. 10, pp. 1-10.
38. Saez, Y., Baldominos, A., Isasi, P. (2016), "A Comparison Study of Classifier Algorithms for Cross-Person Physical Activity Recognition", Sensors, Vol. 17 No. 1, pp. 66-92.
39. Sagiroglu, S., Sinanc, D. (2013), "Big data: a review", in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13), San Diego, California, USA, pp. 42-47.
40. Schroeck, M., Schockley, R., Smart, J., Morales D. R. (2012), "Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data", executive report,

IBM Institute for Business Value and Said Business School at the University of Oxford, Somers, USA.
41. Singh, J., Singla, V. (2015), "Big Data: Tools and Technologies in Big Data", International Journal of Computer Application, Vol. 112 No. 15, pp. 6-10.
42. Snedden, T. R., Scerpella, J., Kliethermes, S. A., Norman, R. S., Blyholder, L., Sanfilippo, J., McGuine, T. A., Heiderscheit, B. (2019), "Sport and Physical Activity Level Impacts Health-Related Quality of Life Among Collegiate Students", American Journal of Health Promotion, Vol. 33 No. 5, pp. 675-682.
43. Sung-Un, P., Hyunkyun, A., Dong-Kyu, D., Wi-Young, S. (2020), "Big Data Analysis of Sports and Physical Activities among Korean Adolescents", International Journal of Environmental Research and Public Health, Vol. 17 No. 15, pp. 5577-5588.
44. Suthaharan, S. (2013), "Big Data Classification: Problems and challenges in network intrusion prediction with machine learning", ACM Sigmetrics, Vol. 41 No. 4, pp. 70-73.
45. Wang, S., Scheider, S., Sporrel, K., Deutekom, M., Timmer, J., Krose, B. (2021), "What are good sitiations for running? A machine learning study using mobile and geographical data", Frontiers in Public Health, Vol. 8, pp. 1-15.
46. Worldometer (2021), "Current world population", available at: https://www.worldometers.info/world-population/ (6 March 2021)
47. Wu, X. Y., Han, L. H., Zhang, J. H., Luo, S., Hu, J. W., Sun, K. (2017), "The influence of physical activity, sedentary behavior on health-related quality of life among the general population of children and adolescents: A systematic review", PLoS One, Vol. 12 No. 11, pp. 1-29.
48. Zhao, D., Wei, L., Wang, Z., Du., Y. (2015), "Modeling and analysis in marine big data: advandes and challenges", Mathematical Problems in Engineering, Vol. 15 No. 2, pp. 1-15.

About the authors

Ana Šuštaršič is an assistant professor at the Faculty of Sports, University of Ljubljana and a doctoral student. Her doctoral dissertation will explore the connection between infertility and obesity, and lifestyle. Her research interests include the areas of exercise for the elderly, exercise during and after pregnancy, exercise for overweight and obese people, recreational exercise suitable for different groups of people, and exercise for preschool children and people with special needs. She conducts various pieces of training to obtain the title of recreation instructor. She authored or co-authored some scientific papers. The author can be contacted at **ana.sustarsic@fsp.uni-lj.si**

Mateja Videmšek is a full professor at the Faculty of Sport, University of Ljubljana. She received PhD in kinesiology. Her research interests include the areas of sports exercise for preschool children, sports exercise for children and adolescents with special needs, social contexts of kinesiology, exercise during and after pregnancy, exercise for overweight and obese people and andragogy. She has also participated in various projects and authored or co-authored various scientific papers published in professional and academic journals. The author can be contacted at **mateja.videmsek@fsp.uni-lj.si**

Damir Karpljuk is a full professor and a dean at the Faculty of Sport, University of Ljubljana. He received PhD in kinesiology. His research interests include the bio-psycho-social aspects of the elderly (chronic diseases, acute movement injuries), sports training for children and adolescents with special needs, and social contexts of kinesiology and martial arts. He authored or co-authored various scientific papers published in professional and academic journals. The author can be contacted at **damir.karpljuk@fsp.uni-lj.si**

Ivan Miloloža is an associate professor in the Faculty of Dental Medicine and Health, Josip Juraj Strossmayer University of Osijek. He is vice dean for inter-institutional collaboration and development. He is the author and co-author of many scientific articles, professional papers and lectures at international conferences. The author can be contacted at **ivan.miloloza@fdmz.hr**

Maja Meško, PhD, is a full professor of management at the Faculty of Management, University of Primorska, and full professor of human resource management at the Faculty of organizational sciences, University of Maribor. She received PhD in kinesiology. Her research interests include management, psychology in management, occupational health and management. She has also participated in various projects. She authored or co-authored various scientific papers published in professional and academic journals. The author can be contacted at **maja.mesko@gmail.com**