The effect of web of science subject categories on clustering: the case of data-driven methods in business and economic sciences

Berndt Jesenko¹ · Christian Schlögl¹

Received: 15 October 2020 / Accepted: 20 May 2021 / Published online: 23 June 2021
© The Author(s) 2021, corrected publication 2021

Abstract

The primary goal of this article is to identify the research fronts on the application of data-driven methods in business and economics. For this purpose, the research literature of the business and economic sciences Subject Categories from the Web of Science is mapped using BibExcel and VOSviewer. Since the assignment to subject categories is done at the journal level and since a journal is often assigned to several subject categories in Web of Science, two mappings are performed: one without considering multiple assignments (broad view) and one considering only those (articles from) journals that have been assigned exclusively to the business and economic sciences subject categories and no others (narrow view). A further aim of this article is therefore to identify differences in the two mappings. Surprisingly, engineering sciences play a major role in the broad mapping, in addition to the economic sciences. In the narrow mapping, however, only the following clusters with a clear business-management focus emerge: (i) Data-driven methods in management in general and data-driven supply chain management in particular, (ii) Data-driven operations research analyses with different business administration/management focuses, (iii) Data-driven methods and processes in economics and finance, and (iv) Data-driven methods in Information Systems. One limitation of the narrow mapping is that many relevant documents are not covered since the journals in which they appear are assigned to multiple subject categories in WoS. The paper comes to the conclusion that the multiple assignments of subject categories in Web of Science may lead to massive changes in the results. Adjacent subject areas—in this specific case the application of data-driven methods in engineering and more mathematically oriented contributions in economics (econometrics) are considered in the broad mapping (not excluding subject categories from neighbouring disciplines) and are even over-represented compared to the core areas of business and economics. If a mapping should only consider the core aspects of particular research fields, it is shown in this use case that the exclusion of Web of Science-subject categories that do not belong to the core areas due to multiple assignments (narrow view), may be a valuable alternative. Finally, it depends on the reader to decide which mapping is more beneficial to them.

Keywords Data-driven · Business and economic sciences · Management · Science mapping · Research fronts · WoS subject categories

Introduction

Data-driven technologies, i.e. technologies that use large amounts of data to generate added value, play a prominent role nowadays. The main reason for the rapid expansion of data-driven technologies is the high degree to which data has been created, shared and used in recent years. A further important step was achieved when the first "mass digitization" projects (e.g. Google's Mass Digitization Project at Oxford), i.e. the conversion of entire traditional libraries into machine-readable files, were accomplished (Coyle, 2006). As soon as signals are converted into a digital format, they can be used to create structured data sets. This step is known as 'datafication' and provides the capability to perform macro-level analysis of relevant trends and patterns, that are impossible when using an analogue format of the data being investigated (Mayer-Schönberger & Cukier, 2013). In Google's Mass Digitization Project, accordingly, the conversion of the vast amount of text strings into sequences of related words and the resulting capability of examining the frequency of occurrence of these sequences over a long-term period was the process of datafication (De Mauro et al. 2016).

Data-driven methods can be used in many disciplines. The aim of this paper is to give a first overview on the use of data-driven technologies in business and economic sciences. Furthermore, the effect of the (multiple) assignment of journals to subject categories, which is a common practice in Web of Science, is analyzed. There are several publications that have investigated the implications of the system of multiple assignments of publications and journals to subject categories in WoS. Leydesdorff and Bornmann (2016) show, using a case study of the WoS-category "information science & library science" and those WoS-categories that are assigned to journals in the field of "science and technology studies", that the classification system of the WoS-categories is not sufficiently clear to bear the bibliometric normalization in the evaluation practice. Boyack and Klavans (2011) even note that far too few journals are truly disciplinary, therefore classification is not purposeful, and consequently question the validity of any classification system for journals and disciplines when used for research evaluation. However, journal classification systems have been used as a basis for comparison and categorization of journals for a long time and although there were numerous other classification systems developed (e.g. Glänzel & Schubert, 2003; Klavans & Boyack, 2010; Perianes-Rodriguez & Ruiz-Castillo, 2017), the Thomson Reuters' Subject Categories are probably the most widely used such classification system (Boyack & Klavans, 2011). The approach used in this study is a new, rather radical attempt that aims to improve the visibility of research fronts within a particular scientific discipline, in this case business administration, by excluding journals with multiple attributions in its analysis.

It must be stated that this approach is very likely to eliminate important publications from the results, i.e. it also entails a loss of information, which is the main limitation. However, if the aim is to highlight the core aspects of a specific topic as clearly as possible within a particular scientific discipline, the chosen method can be considered as an alternative to conventional approaches, as the following use case will demonstrate.

Data, methodology and tools

This article focuses on the data sources covered by the Web of Science Core Collection which is composed by the Science Citation Index Expanded (SCI-EXPANDED), the Social Sciences Citation Index (SSCI), the Arts & Humanities Citation Index (A&HCI) and the Emerging Sources Citation Index (ESCI). Although Science Citation Index Expanded (SCI-EXPANDED) and Social Sciences Citation Index (SSCI) make up the majority of the publications found, a few relevant publications are also found in the other mentioned indices. In order to consider only publications (articles, reviews and conference proceedings) with a background in the business and economic sciences, the search was limited to the following Web of Science subject categories:

- Business
- Business, Finance
- Economics
- Management
- Operations Research & Management Science
- Public Administration
- Computer Science, Information Systems.

The search was conducted on works published in all languages before 2020, obtaining 2528 records (date of retrieval 28/01/2020).¹ For these documents the major research fronts should be determined.

Research fronts are clusters of documents on a specific topic that can usually be identified by an analysis of bibliographic couplings (Huang and Chang 2014). They are characterized by a growing research activity on a particular topic (Schiebel, 2012). Hence, research fronts are of particular interest for identifying and illustrating current research topics and emerging trends (Persson, 1994) and therefore of main interest in this article. If research fronts are based on bibliographic couplings, they are time-invariable, i.e. new publications do not retrospectively change the result of an analysis of bibliographic couplings at a certain point in time (Morris et al. 2003). Furthermore, research fronts are not defined by individuals, but result from the sum of contributions of certain groups of authors (De Moya Anegón et al. 1998).

Data analysis was conducted with BibExcel and VOSviewer. The tool BibExcel was developed at the University of Umea in Sweden and is free for non-commercial use. It is capable of reading bibliographic data from common literature databases. In addition, bibliographic networks from co-citations, bibliographic couplings and co-authorships can be created. In Bibexcel it is possible to perform most types of bibliometric analysis and it allows easy interaction with other software. For example, it is possible to create Pajek files, which are subsequently used in VOSviewer when creating the network files. The strength of the program lies in its flexibility in both data management and analysis (Persson et al. 2009) and offering functions that can be used to improve data quality (e.g. "Convert Upper

¹ The corresponding search string was: TS=(data-driven) AND WC=(Business OR Business, Finance OR Economics OR Management OR Operations Research Management Science OR Public Administration OR Computer Science, Information Systems) Refined by: DOCUMENT TYPES: (ARTICLE OR PROCEEDINGS PAPER OR REVIEW) AND [excluding] PUBLICATION YEARS: (2020) Indexes=SCI-EXPANDED, SSCI, A&HCI, ESCI Timespan=All years.

Lower Case" or "Keep only author's first initial"). BibExcel is also capable of extracting bibliographic couplings from reference lists that are exported as text files (Pilkington, 2018). However, BibExcel has no visualization function. Instead, the data can be exported and visualized with other tools such as VOSviewer, UCINET, SPSS or Pajek (Cobo et al. 2011). BibExcel therefore serves exclusively to prepare the exported data for visualization.

The reason for choosing VOSviewer to display the bibliometric data in this article is the significant display quality, the choice of the display density of the links (Belli and Balta 2019) and the possibility to set a minimum total link strength (TLS) of an article to ensure that only substantial publications are included in the network. Another reason for choosing VOSviewer was that it uses a unified approach to mapping and clustering that avoids inconsistencies between the results produced by mapping and clustering techniques (for a detailed description see van Eck & Waltman, 2010; Waltman et al. 2010). In order to consider only publications which are inter-related to a certain degree, the minimal TLS was set to 20. This means, that only those papers with at least 20 bibliographical couplings to the other investigated publications were considered. As a consequence, out of the 2528 publications only 1234 remained in the following analysis.

Results

Mapping 1: multiple assignment of web of science subject categories

Figure 1 shows the results of the mapping for those publications which were assigned to one of the seven before-mentioned subject categories. Since journals (and therefore also their included publications) are usually assigned to more than one subject category in Web of Science, it is most likely that further subject categories are involved (this is meant by multiple assignment of Web of Science subject categories) (Figs. 2, 3).

As can be seen in Table 1, technical subject categories can even outweigh several subject categories from business and economic sciences in numbers of assigned publications. In particular, *Engineering, Electrical & Electronic; Communication; Computer Science, Artificial Intelligence; Engineering, Industrial; and Computer Science, Software Engineering* are thematically distant from business economics. Their high proportion suggests some technological orientation when it comes to the application of data-driven applications and methods in business and economic sciences. The question arises whether this is also reflected in the resulting mapping.

Figure 1 shows the result of the mapping procedure. A circle or node represents a publication on the topic under investigation. The more similar two publications are in terms of their bibliographic coupling, the closer they are to each other on this map. The size of the nodes depends on the level of the TLS. The higher the TLS (the more bibliographic couplings an article shares with the other articles in the cluster), the larger the node is. The nodes are colored according to the assigned cluster.

As can be seen, there are six more or less specific clusters as regards contents (see Table 2).

Data-driven methods in Engineering Sciences (red cluster)

The red cluster is strongly dominated by the engineering sciences. From the 454 publications of this cluster, 359 were assigned to the Computer Science research area and 203

Fig. 1 Data-driven methods in business and economic sciences (multiple assignment of WoS subject categories)

to Engineering. The most frequent business management research area is Operations Research & Management Science with 125 assignments. Accordingly, the business management reference in this cluster is comparatively low.

This impression is reinforced when we look at the journals that frequently appear in this cluster: 69 articles were published in IEEE Access, an Open Access journal published by the Institute of Electrical and Electronics Engineers, and 49 articles in Expert Systems with Applications.

An analysis of the TOP 10 publications by total link strength (TLS) gives an indication of which contents are primarily covered in this cluster. Process monitoring and control play a central role. For instance, the article with the highest TLS (*“Statistical learning methods applied to process monitoring: an overview and perspective”*) by Weese et al. (2017) provides a review of the state of the art of data-driven multivariate statistical process monitoring. Some of the monitoring and control methodologies, which are dominated by data mining techniques and are promising for monitoring large and diverse data sets, are highlighted.

Quality prediction in manufacturing processes is another example of how the analysis of massive amounts of data can influence modern manufacturing environments. For instance, the article *“Imbalanced classification of manufacturing quality conditions using*

Fig. 2 Data-driven methods in business and economic sciences (multiple assignment of WoS subject categories)—red and yellow clusters only

cost-sensitive decision tree ensembles “ by Kim et al. (2018) deals with the problem of classifying production process conditions into normal and defective products. Such a quality analysis data set is usually not balanced because the error rate is low in practice. To solve this unbalanced classification problem, a cost-sensitive decision tree ensemble algorithm is used to increase the small number of defective cases and to assign higher costs to the misclassification of defective products than to normal products.

The detection and diagnosis of faults in industrial chemical processes is the subject of another TOP 10 article from this cluster (Ragab et al. 2017). Logical Analysis of Data (LAD) is used for this purpose. This classification technique from machine learning uncovers hidden knowledge in industrial data sets by revealing interpretable patterns linked to underlying physical phenomena. The patterns are then combined to build a decision model that can be used to diagnose errors during process operation and explain the potential causes of these errors.

Data-driven fault prediction and diagnosis is also the main research focus in several other publications from this cluster (e. g. Riaz et al. 2018; Diallo et al. 2018; Zhang et al. 2018; Xu et al. 2017a, b).

Data-driven predictive maintenance of machine equipment (yellow)

The local proximity and the partly overlaps between the yellow and the red cluster in the network suggest that there is also a proximity in terms of content.

Table 1 Top-10 subject categories of retrieved documents

Subject category	Number (#)	% Out of 1234
Computer science, information systems	543	44.0
Management	468	37.9
Engineering, electrical and electronic	301	24.4
Communication	227	18.4
Economics	204	16.5
Computer science, artificial intelligence	140	11.4
Engineering, industrial	130	10.5
Business	100	8.1
Mathematics	100	8.1
Computer science, software engineering	87	7.1

Table 2 Data-driven methods in business and economic sciences (main clusters)

Cluster	#
Data-driven methods in engineering sciences (red)	454
Data-driven methods in economics and econometrics (green)	226
Data-driven methods in management in general and data-driven supply chain management in particular (blue)	212
Data-driven predictive maintenance of machine equipment (yellow)	204
Data-driven decision making and optimization in operations research and management science (purple)	104
Data-driven control and automation processes (turquoise)	20

As in the red cluster, the two research areas *Computer Science* (158) and *Engineering* (137) are particularly frequently represented in the yellow cluster. The business economics related research areas *Operations Research & Management Science* (57) and *Business & Economics* (8) are even more underrepresented here than in the red cluster.

The situation is similar with regards to the publishing journals. *IEEE Access* is by far the most frequently used journal with 63 publications. The other journals (*Reliability Engineering & System Safety*, *Expert Systems with Applications*, *IEEE Transactions on Multimedia and Multimedia Tools and Applications*) are also application-oriented and have a strong reference to engineering sciences.

In terms of content, data-driven maintenance of machines and machine equipment plays a particularly strong role. Since current research on predictive maintenance is still in a primarily experimental stage, most of the work is carried out using several open data sets. A review article with the title “*Data-driven methods for predictive maintenance of industrial equipment: A survey*” is the publication with the highest TLS in the yellow cluster. It presents a comprehensive survey on the applications and attempts to provide a better understanding of the existing works recently published (Zhang et al. 2019).

The majority of the publications contained in this cluster is dedicated to the data-driven estimation of the remaining life time of machines. “*Remaining useful life estimation—A review on the statistical data driven approaches*” by Si et al. (2011) is the article with the second highest TLS in the cluster providing an overview of modelling developments

for estimating the remaining useful life. The overview focuses on data-driven statistical approaches that rely only on available previous observational data and statistical models. A more recent article by Li et al. (2018) entitled “*Remaining useful life estimation in prognostics using deep convolution neural networks*” proposes a newer data-driven approach for predicting the remaining useful life using deep convolution neural networks. Several other publications in this cluster (e. g.; Al-Dahidi et al. 2019; Zhang et al. 2015; Ellefsen et al. 2019; Hu et al. 2012) have a similar orientation.

Even the data-driven prediction of the remaining useful life of certain goods is the subject of some publications. For example, those of bearings in the articles entitled “*Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction*” by Li et al. (2019), “*Prediction of bearing remaining useful life with deep convolution neural network*” by Ren et al. (2018) or “*Real-time bearing remaining useful life estimation based on the frozen convolutional and activated memory neural network*” by Chen et al. (2019a, b, c).

Data-driven control and automation processes (turquoise)

The smallest cluster is also strongly technically oriented. Computer Science with 16 entries and Automation & Control Systems with 9 are the most frequently assigned Research Areas. This thematic direction can also be derived from the publishing journals. The two top journals are IEEE Access (7 articles) and International Journal of System Science (5 articles). In 16 of the 20 publications included in the cluster, the term “control” is used in the title. Data-driven control and automation technology therefore plays the central role in this cluster.

The top article of this cluster (“*From model-based control to data-driven control: Survey, classification and perspective*”) by Hou and Wang (2013) can be considered an overview of problems and challenges inherent in the theory of data-driven model-based control. Existing methods and applications are presented with corresponding classifications and findings.

The more recent publications take a very application-oriented and specific approach in this context. For example, the model-free adaptive control as a data-driven control method and its concrete application to a launcher vehicle is the subject of one article (Duan et al. 2019). Another paper entitled “*Deterministic continuous-time Virtual Reference Feedback Tuning (VRFT) with application to PID design*” by Formentin et al. (2019) presents a data-driven control design method that is not based on a model of a plant. In this context a PID autotuner is developed and its effectiveness is tested on an experimental brake-by-wire system and improved performance compared to the design based on a benchmark model is shown.

Data-driven methods in management in general and data-driven supply chain management in particular (blue)

The blue cluster is characterized by its focus on management areas (general management, but also somewhat more technical areas such as data-driven process management and supply chain management in the sense of optimizing and managing supply chains using data-driven approaches). This is also reflected in the frequencies of the assigned research areas. While *Business & Economics* is the most frequent research area with 104 publications,

Computer Science (66 articles), *Engineering* (66 articles) and *Operations Research & Management Science* 63 (articles) follow, nearly equally located, with some distance.

If we look at the distribution of the publishing journals, the business management relevance becomes even more pronounced. Almost all TOP journals, with the exception of *IEEE Access* and *Technological Forecasting and Social Change*, are purely business management journals. The background of the *International Journal of Production Research*, the *International Journal of Production Economics* and the *Journal of Manufacturing Systems* indicates topics from process and production management.

A look at the TOP publications of this cluster reveals a similar picture. “A multidisciplinary perspective of big data in management research “ by Sheng et al. (2017) is a review article on data-driven management and has the highest TLS in the cluster. It tracks scholarly work’s perspectives on big data in the management domain over the past decade. The article suggests that research should use both structured and unstructured big data to promote the understanding of the value of big data as a basis for managerial organizational decisions and to improve the competitiveness of enterprises. However, in order to generate the full value from the data, it is important that companies formulate and implement a data-driven strategy in advance. Further publications with a similar approach on (data-driven) general management are, for instance, from Carillo (2017), Akhtar et al. (2019), and Urbinati et al. (2019).

The article entitled “Data-driven supply chain capabilities and performance: A resource-based view” by Yu et al. (2018) explores how data-driven supply chains affect the overall performance of enterprises. The authors come to the conclusion that a data-driven supply chain has a significantly positive effect on the financial performance. Another paper deals with the dimensions and sub-dimensions of a “Supply Chain Analytics Capability”. The authors model their overall impact on supply chain agility and business performance (Wamba & Akter, 2019). There are several more publications which also investigate different contexts with data-driven supply chains (e. g. Yu, Jabobs, Chaves and Feng 2019, Chavez et al. 2017).

Data-driven procedures in economics and econometrics (green)

The green cluster has a strong link to economics and econometrics which distinguishes it significantly from the other clusters. If we take a look at the Research Areas, *Business & Economics* is by far the most frequently one (192 out of 226 articles). The fact that *Mathematical Methods in Social Sciences* and *Mathematics* are the second and third most frequent research areas in this cluster suggests a strong mathematical and in particular econometric orientation. A closer look at the journals confirms this perception. Most papers were published in the *Journal of Econometrics*, followed by *Econometric Theory*, *Journal of Business & Economic Statistics* and *Empirical Economics*.

In the article with the highest TLS of this cluster “A review on variable selection in regression analysis” by Desboulets (2018) several variable selection procedures are examined to provide an overview of the existing literature for practitioners. The article explains which methods should be used for which model purposes and which are the most important differences among them. Besides theoretical and statistical methods, data-driven methods provide an alternative for automatic model selection.

In another paper of this cluster by Qian and Su (2016) the problem of determining the number of structural changes in multiple linear regression models using group-fused lasso is investigated. A method is presented which can correctly determine the unknown number

of breaks with a probability of tending to one. Furthermore, a data-driven method for determining the tuning parameter is also proposed and validated.

A data-driven cross-validation approach is the focus of another publication entitled “*Nonparametric estimation of conditional quantile functions in the presence of irrelevant covariates*” by Chen et al. (2019). The aim is to simultaneously determine the optimal smoothing parameters and to remove irrelevant covariates. For this purpose, the conditional quantile regression function is estimated in advance using the kernel method based on the check function.

Also further publications (e.g., “*Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide*” by Chen and Qiu (2016), “*Shrinkage estimation of regression models with multiple structural Changes*” by Qian and Su (2016) or “*Nonparametric/semiparametric estimation and testing of econometric models with data dependent smoothing parameters*” by Li and Li (2010)) go in the direction of data-driven nonparametric/semiparametric estimation and testing of econometric models.

Data-driven decision making and optimization in Operations Research and Management Science (purple)

The distribution of Research Areas in this cluster indicates a clear focus on Operations Research & Management. Out of the 104 publications, 84 articles were assigned to this subject category.

Data driven robust optimization in general and the data driven newsvendor problem as a special aspect of it are the main topics covered in this cluster. 6 of the 11 top-ranked publications by TLS deal with data-driven robust optimization, 3 with a data-driven approach to the newsvendor model.²

In the publication “*Data-driven robust optimization*” by Bertsimas et al. (2014), a novel approach to the use of data is proposed in order to design uncertainty sets for robust optimization using statistical hypothesis testing. The authors come to the conclusion that data-driven sets significantly outperform traditional robust optimization techniques. A number of other publications also deal with the superiority of data-driven robust optimisation compared to traditional optimization in different contexts (e. g. Esfahani & Kuhn, 2018; Kang et al. 2018; Fernandes et al. 2016; Chen et al. 2018).

Traditionally, the newsvendor problem is solved by assuming a distribution. In reality, however, the decision-maker does not know the true distribution of the demand. The highest ranked paper by TLS in this cluster entitled “*A data-driven newsvendor problem: From data to decision*” by Huber et al. (2019) presents data-driven solution methods based on machine learning and quantile regression that do not require the assumption of a specific demand distribution. The authors conclude that machine learning approaches are clearly superior to traditional methods, provided the data set is large enough. Also Ban and Rudin (2019) come to the conclusion that single-step machine-learning algorithms are superior for solving the newsvendor problem. Several other publications from this cluster also cover this topic (e.g. Levi et al. 2015; Sachs & Minner, 2014; Saghaian & Tomlin, 2016).

² The newsvendor problem is a model used in applied economics to determine optimal inventory levels.

Mapping 2: consideration of subject categories exclusively from the business and economic sciences

In the second mapping, only journals assigned solely to subject categories from business and economic sciences were considered. In order to avoid articles also from other subject categories due the multiple assignment strategy of Web of Science, they were explicitly excluded.³ This resulted in 694 publications in total, which are exclusively assigned to one of the business and economics subject categories.⁴ Since a minimum TLS of 20 was also specified in this mapping for reasons of comparability, there remained only 166 articles in the mapping.⁵ They distribute over the WoS-subject categories of business and economic sciences as follows (see Tables 3, 4, Fig. 3):

As can be seen, the most apparent difference to the first broader mapping is that two clusters with a stronger technological focus (data-driven predictive maintenance of machine equipment (yellow) and data-driven control and automation processes (turquoise) are no longer included. The remaining clusters have changed, in some cases significantly, in terms of content. A much stronger relation to business and economic sciences is evident.

With 69 included publications, the blue cluster is the largest in this mapping. In terms of content, not much has changed compared to the previous mapping, but the focus is more distinct. The use of data-driven methods in business administration and management in general (7 out of the top-10-publications), and data-driven supply chains (3 out

³ TS=(data-driven) AND WC=(Business OR Business, Finance OR Economics OR Management OR Operations Research Management Science OR Public Administration OR Computer Science, Information Systems) Refined by: DOCUMENT TYPES: (ARTICLE OR PROCEEDINGS PAPER OR REVIEW) AND [excluding] PUBLICATION YEARS: (2020) AND [excluding] WEB OF SCIENCE CATEGORIES: (TRANSPORTATION OR COMMUNICATION OR ENVIRONMENTAL STUDIES OR HOSPITALITY LEISURE SPORT TOURISM OR ENGINEERING ELECTRICAL ELECTRONIC OR TRANSPORTATION SCIENCE TECHNOLOGY OR SOCIAL WORK OR TELECOMMUNICATIONS OR GEOGRAPHY OR COMPUTER SCIENCE CYBERNETICS OR MATHEMATICAL COMPUTATIONAL BIOLOGY OR DEVELOPMENT STUDIES OR ETHICS OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR ENGINEERING MULTIDISCIPLINARY OR OPTICS OR ENGINEERING INDUSTRIAL OR MATHEMATICS APPLIED OR SOCIAL SCIENCES INTERDISCIPLINARY OR COMPUTER SCIENCE THEORY METHODS OR CHEMISTRY MULTIDISCIPLINARY OR INDUSTRIAL RELATIONS LABOR OR COMPUTER SCIENCE SOFTWARE ENGINEERING OR ENGINEERING CIVIL OR INTERNATIONAL RELATIONS OR SOCIAL SCIENCES MATHEMATICAL METHODS OR REGIONAL URBAN PLANNING OR RADIOLOGY NUCLEAR MEDICINE MEDICAL IMAGING OR CHEMISTRY MEDICINAL OR SOCIOLOGY OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR GEOGRAPHY PHYSICAL OR BEHAVIORAL SCIENCES OR ENGINEERING MANUFACTURING OR ENERGY FUELS OR EDUCATION SCIENTIFIC DISCIPLINES OR MATHEMATICS INTERDISCIPLINARY APPLICATIONS OR ENVIRONMENTAL SCIENCES OR FOOD SCIENCE TECHNOLOGY OR INFORMATION SCIENCE LIBRARY SCIENCE OR ENGINEERING ENVIRONMENTAL OR FORESTRY OR MEDICAL INFORMATICS OR PSYCHOLOGY APPLIED OR MULTIDISCIPLINARY SCIENCES OR HEALTH POLICY SERVICES OR NURSING OR COMPUTER SCIENCE HARDWARE ARCHITECTURE OR POLITICAL SCIENCE OR NUTRITION DIETETICS OR STATISTICS PROBABILITY OR AGRICULTURAL ECONOMICS POLICY OR PHARMACOLOGY PHARMACY OR HEALTH CARE SCIENCES SERVICES OR URBAN STUDIES OR PSYCHOLOGY EXPERIMENTAL OR AUTOMATION CONTROL SYSTEMS) Indexes = SCI-EXPANDED, SSCI, A&HCI, ESCI Timespan = All years.

⁴ In our analysis, we also considered Computer Science, Information Systems as one of the business and economic sciences subject categories.

⁵ We also performed another mapping with a TLS value of 10 to check the validity of this mapping. Due to the lower TLS value 292 publications were considered. The second mapping resulted in the same main clusters with the difference that the demarcation between them was less clear.

Table 3 Subject categories of documents (exclusively from business and economic sciences)

Subject category	Number (#)	% Out of 166
Management	99	59.6
Business	46	27.7
Economics	28	16.9
Computer science, information systems	21	12.7
Business, finance	20	12.1

Table 4 Clusters after exclusion subject categories other than from business and economic sciences

Cluster	#
Data-driven methods in management in general and data-driven supply chain management in particular (blue)	68
Data-driven OR analysis with different business administration/management focuses (purple)	45
Data-driven methods and processes in economics and finance (green)	30
Data-driven methods in information systems (red)	23

Fig. 3 Data-driven methods in business and economics sciences (after exclusion of WoS-subject categories other than from business and economic sciences)

of the top-10-publications) in particular, form the core of the blue cluster. The analysis of this cluster also shows that top publications from journals that have (also) been assigned to other non-business subject areas (*Engineering*) and, as a consequence, are not included in this mapping though can have a high significance for business management research. This is the case for the two top publications by Sheng et al. (2017) and Yu et al. (2018) (TLS 421 and 381, respectively) from the previous clustering that are also assigned to the subject areas *Engineering* and *Manufacturing*.

13 publications that appeared in the top-20, ranked by TLS, in the previous mapping (without exclusion of the non-business subject categories) are also included in the top-20 publications of this mapping. The overlaps in content are therefore considerable, which is why the individual publications are not discussed in more detail anymore.

The purple cluster is now the second largest with 42 publications. Its analysis demonstrates the benefits of excluding multiply assigned subject areas which are outside the core interest. 19 of the top-20 publications are assigned to the Web of Science-Categories *Management* and *Operations Research & Management Science*. Compared to the previous clustering, many publications were eliminated that had been assigned to the categories *Mathematics*, *Applied* and *Mathematics* with a much stronger mathematical focus. Furthermore, in the previous analysis only half of the top 10 publications had been assigned to the management sector. The content focus and, namely how data-driven operations research analyses are used in business and management, is therefore more clearly recognizable and easier to identify. Now a much stronger focus on business administration/inventory management is visible, the originally strong mathematical/statistical emphasis is fading into the background. Chen et al. (2019a, b, c), for example, propose a data-driven algorithm that learns continuously from demand data during application and simultaneously supports managers in supply and pricing decisions. The algorithm should result in the company actively experimenting with pricing and inventory to collect demand information with minimal loss of profit. In another publication entitled "*Quantile forecasting and data-driven inventory management under nonstationary demand*" by Cao and Shen (2019) a data-based approach for the determination of stock levels (in the framework of the newsvendor problem and its multi-periodic extension) is presented. It is shown that the data-driven approach outperforms existing statistical methods and provides good solutions.

In the green cluster, there has also been a significant reduction in the number of publications included (from originally 226 to 31). Because the WoS-categories *Social Sciences*, *Mathematical Methods* and *Statistics & Probability* have been removed, the initially very strong econometric component is almost completely eliminated. This is also underlined by the fact that only 4 of the original Top 20 publications according to TLS are included in the new green cluster. After viewing the top publications, the content can now be described as the use of data-driven processes in economics and finance. Pricing and applied economics are central topics in this cluster. For example, one publication by Byrne et al. (2019) investigates the movement of global yield curve dynamics using a data-driven hierarchical factor model. This data-driven approach is capable of determining the driving forces of yield curve dynamics and producing plausible estimates of term premiums. In another article, the effects of the driving factors for natural gas prices in the period from 1999 to 2017 are examined (Ji et al. 2018). A data-driven approach is used to reveal the simultaneous relationships between natural gas, oil and other factors. In contrast to most previous studies, the data-driven approach established a long-term equilibrium relationship between oil and natural gas yields. The publication entitled "*Network causality structures among Bitcoin and other financial assets: A directed acyclic graph approach*" also uses a data-driven method to reveal the simultaneous and lagged relationships between Bitcoin and other asset classes. The results of the analysis suggest that the Bitcoin market is rather isolated and that no specific asset plays a dominant role in influencing it (Ji et al. 2018a).

The new red cluster consists mainly of those publications that belong to the WoS category *Computer Science*, *Information Systems*. They are not assigned to any other Computer Science category but in some cases also to a business and economics subject category. The result is a very heterogeneous cluster, which consists on the one hand of publications with a computer science background, and on the other hand also contains articles with a stronger business and management background. 5 of the TOP-10 publications were published in the journal "Information Sciences", which has a strong focus on informatics. The business publications cover topics ranging from innovative data-driven forecasting models (Luna & Ballini, 2011; Yu et al. 2019a, b) and market segment analysis (Dolnicar and Leisch 2014;

Boztug & Reutterer, 2008) to text mining (Lim & Maglio, 2018). It is also noteworthy that the publications are on average considerably older than those in other clusters. 13 of the 24 included articles were published before 2017.

Conclusion

As was shown in our contribution, the multiple assignment of (articles in) journals to subject categories, which is common practice in Web of Science, has a strong impact on science mapping. In the case of data-driven applications and methods in business and economic sciences, six clusters could be found, which can be roughly classified to three business and economics and three technical clusters. More than half of the publications were assigned to more technical disciplines, although the original search was limited to Web of Science subject categories from business and economic sciences. This is due to the fact that journals were also assigned to more technical subject categories, in the case of our study also from engineering sciences.

Therefore, we excluded subject categories outside business and economic sciences explicitly before performing the second mapping. As a result, clusters with a stronger technical focus (*data-driven predictive maintenance of machine equipment* and *data-driven control and automation processes*) did not appear in the mapping anymore. Only four clusters with a main focus on business and economic sciences remained: (i) *data-driven methods in management in general and data-driven supply chain management in particular*, (ii) *data-driven operations research analysis with different business administration/management focuses*, (iii) *data-driven methods and processes in economics and finance* (iv) *data-driven methods in information systems*. These four clusters were partly much smaller than those in the broader mapping. However, in our case they showed a stronger relationship to business and economic sciences since mathematical aspects and core topics in computer science were not covered by them anymore.

It may be concluded that the multiple assignment to subject categories can have a great impact on the result of the mapping. In particular, neighboring topics (clusters) can be identified in the broader mapping which would not have been anticipated otherwise. If one prefers to identify only the core aspects of the research fields under investigation, our case shows that limiting the analysis to the core subject categories of a specific scientific discipline by excluding all categories that do not uniquely relate to it can have a positive impact on the visibility and clarity of the clusters. However, it must be pointed out that these potential achievements may come at the cost that highly relevant publications in terms of content might not be considered because they appear in journals that are also assigned to other Web of Science categories not associated with the scientific discipline under investigation. In our opinion both mappings are beneficial. It mainly depends on the focus of the readers which mapping they prefer. Readers with a broader focus will prefer the first mapping. However, if they have a narrower view and are primarily interested in the core aspects of a certain topic, they will find the second mapping more relevant.

Funding Open access funding provided by University of Graz.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article

are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akhtar, P., Frynas, G., Mellahi, M., & Ullah, S. (2019). Big data-Savvy teams' skills, big data-driven actions and business performance. *British Journal of Management*, 30(2), 252–271.
- Al-Dahidi, Ss., Ayadi, O., Alrbai, M., & Adeeb, J. (2019). Ensemble approach of optimized artificial neural networks for solar photovoltaic power prediction. *IEEE Access*, 7, 81741–81758.
- Ban, G. Y., & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1), 90–108.
- Belli, S., & Baltà, J. (2019). Stocktaking scientific publication on bi-regional collaboration between Europe 28 and Latin America and the Caribbean. *Scientometrics*, 121(3), 1447–1480.
- Boyack, K. W., & Klavans, R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines. In *13th International Conference of the International Society for Scientometrics and Informetrics*, pp. 123–133.
- Boztug, Y., & Reutterer, T. (2008). A combined approach for segment-specific market basket analysis. *European Journal of Operational Research*, 178(1), 294–312.
- Bertsimas, D., Gupta, V., & Pascholdis, I. C. (2014). Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153, 595–633.
- Byrne, J. P., Cao, S., & Korobilis, D. (2019). Decomposing global yield curve co-movement. *Journal of Banking and Finance*, 106, 500–513.
- Cao, Y., & Shen, Z.-J.M. (2019). Quantile forecasting and data-driven inventory management under nonstationary demand. *Operations Research Letters*, 47(6), 465–472.
- Carillo, K. D. A. (2017). Let's stop trying to be “sexy”—Preparing managers for the (big) data-driven business era. *Business Process Management Journal*, 23(3), 598–622.
- Chavez, R., Yu, W., Jacobs, M. A., & Feng, M. (2017). Data-driven supply chains, manufacturing capability and customer satisfaction. *Production Planning and Control*, 28(11–12), 906–918.
- Chen, B., Chao, X., & Ahn, H.-S. (2019). How to build competencies for a data-driven business: Keys for success and seeds for failure. *Operations Research*, 67(4), 1035–1052.
- Chen, X., Li, D., Li, Q., & Li, Z. (2019). Nonparametric estimation of conditional quantile functions in the presence of irrelevant covariates. *Journal of Econometrics*, 212(2), 433–450.
- Chen, X., & Qiu, Y. J. (2016). Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide. *Annual Review of Economics*, 8, 259–290.
- Chen, Z., Peng, S., & Liu, J. (2018). Data-driven Robust chance constrained problems: a mixture model approach. *Journal of Optimization Theory and Applications*, 179, 1065–1085.
- Chen, Z., Tu, X., Hu, Y., & Li, F. (2019). Real-time bearing remaining useful life estimation based on the frozen convolutional and activated memory neural network. *IEEE Access*, 7, 96583–96593.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the Association for Information Science and Technology*, 62(7), 1382–1402.
- Coyle, K. (2006). Mass digitization of books. *Journal of Academic Librarianship*, 32(6), 641–645.
- De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65(3), 122–135.
- De Moya Anegón, F., Contreras, E., & Corrochano, M. (1998). Research fronts in library and information science in Spain (1985–1994). *Scientometrics*, 42(2), 229–246.
- Desboulets, L. (2018). A review on variable selection in regression analysis. *Econometrics*, 6(4), 45.
- Dolnicar, S., & Leisch, F. (2014). Using graphical statistics to better understand market segmentation solutions. *International Journal of Market Research*, 56(2), 97–120.
- Duan, L., Hou, Z.-S., You, X., Jin, S.-T., & Lu, K.-F. (2019). Data-driven model-free adaptive attitude control approach for launch vehicle with virtual reference feedback parameters tuning method. *IEEE Access*, 7, 54106–54116.
- Ellefsen, A. L., Ushakov, S., Æsøy, V., & Zhang, H. (2019). Validation of data-driven labeling approaches using a novel deep network structure for remaining useful life predictions. *IEEE Access*, 7, 71563–71575.

- Esfahani, P. M., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171, 115–166.
- Fernandes, B., Street, A., Valladao, D., & Fernandes, C. (2016). An adaptive robust portfolio optimization model with loss constraints based on data-driven polyhedral uncertainty sets. *European Journal of Operational Research*, 255(3), 961–970.
- Formentin, S., Campi, M.-C., Carè, A., & Savaresi, M. (2019). Deterministic continuous-time Virtual Reference Feedback Tuning (VRFT) with application to PID design. *Systems and Control Letters*, 127, 25–34.
- Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357–367.
- Hou, Z.-S., & Wang, Z. (2013). From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235, 3–35.
- Hu, C., Youn, B.-D., Wang, P.-F., & Yoon, J.-T. (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. *Reliability Engineering and System Safety*, 103, 120–135.
- Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, 278(3), 904–915.
- Huang, M.-H., & Chang, C.-P. (2014). A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation. *Scientometrics*, 102(3), 2041–2057.
- Ji, Q., Bouri, E., Gupta, R., & Roubaud, D. (2018a). Network causality structures among Bitcoin and other financial assets: A directed acyclic graph approach. *The Quarterly Review of Economics and Finance*, 70, 203–213.
- Ji, Q., Zhang, H. Y., & Geng, J. B. (2018b). What drives natural gas prices in the United States?—A directed acyclic graph approach. *Energy Economics*, 69, 29–88.
- Kang, Z.-L., Li, X., Li, Z., & Zhu, S. (2018). Data-driven robust mean-CVaR portfolio selection under distribution ambiguity. *Quantitative Finance*, 19(1), 105–121.
- Kim, A., Oh, K., Jung, J.-Y., & Kim, B. (2018). Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles. *International Journal of Computer Integrated Manufacturing*, 31(8), 701–717.
- Klavans, R., & Boyack, K. W. (2010). Toward an objective, reliable and accurate method for measuring research leadership. *Scientometrics*, 82(3), 539–553.
- Levi, R., Perakis, G., & Uichanco, J. (2015). The data-driven newsvendor problem: New bounds and insights. *Operations Research*, 63(6), 1294–1306.
- Leydesdorff, L., & Bornmann, L. (2016). The operationalization of “fields” as WoS subject categories (WCs) in evaluative bibliometrics: The cases of “library and information science” and “science and technology studies.” *Journal of the Association for Information Science and Technology*, 67(3), 707–714.
- Li, D., & Li, Q. (2010). Nonparametric/semiparametric estimation and testing of econometric models with data dependent smoothing parameters. *Journal of Econometrics*, 157(1), 179–190.
- Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering and System Safety*, 172, 1–11.
- Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability Engineering and System Safety*, 182, 208–218.
- Lim, C., & Maglio, P. P. (2018). Data-driven understanding of smart service systems through text mining. *Service Sciences*, 10(2), 154–180.
- Luna, I., & Ballini, R. (2011). Top-down strategies based on adaptive fuzzy rule-based systems for daily time series forecasting. *International Journal of Forecasting*, 2(3), 708–724.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.
- Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54(5), 413–422.
- Persson, O. (1994). The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science*, 45(1), 31–38.
- Qian, J., & Su, L. (2016). Shrinkage estimation of regression models with multiple structural changes. *Economic Theory*, 32(6), 1376–1433.
- Ragab, A., El Koujok, M., Poulin, B., Amazouz, M., & Yacout, S. (2017). Fault diagnosis in industrial chemical processes using interpretable patterns based on logical analysis of data. *Expert Systems with Applications*, 95, 368–383.
- Ren, L., Sun, Y., Wang, H., & Zhang, L. (2018). Prediction of bearing remaining useful life with deep convolution neural network. *IEEE Access*, 6, 13041–13049.

- Riaz, S., Arshad, A., & Jiao, L. (2018). Rough noise-filtered easy ensemble for software fault prediction. *IEEE Access*, 6, 46886–46899.
- Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2017). A comparison of the Web of Science and publication-level classification systems of science. *Journal of Informetrics*, 11(1), 32–45.
- Persson, O., Danell, R. & Schneider, J. W. (2009) How to use Bibexcel for various types of bibliometric analysis. In Åström, F., Danell, R., Larsen, B. and Schneider, J.W. (eds), *Celebrating Scholarly Communication Studies*. Leuven, Belgium: ISSI, pp. 9–24.
- Pilkington, A. (2018). Bibexcel-Quick Start Guide to Bibliometrics and Citation Analysis. In: Daim, T. & Pilkington, A. (Ed.) *Innovation Discovery: Network Analysis Of Research And Invention Activity For Technology Management*. London: World Scientific Publishing.
- Sachs, A.-L., & Minner, S. (2014). The data-driven newsvendor with censored demand observations. *International Journal of Production Economics*, 149, 28–36.
- Saghafian, S., & Tomlin, B. (2016). The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, 64(1), 167–185.
- Schiebel, E. (2012). Visualization of research fronts and knowledge bases by three-dimensional areal densities of bibliographically coupled publications and co-citations. *Scientometrics*, 91(2), 557–566.
- Sheng, J., Amankwah-Amoah, J., & Wang, X. (2017). A multidisciplinary perspective of big data in management research. *International Journal of Production Economics*, 191, 97–112.
- Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation—A review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1), 1–14.
- Urbinati, A., Bogers, M., Chiesa, V., & Frattini, F. (2019). Creating and capturing value from big data: A multiple-case study analysis of provider companies. *Technovation*, 84–85, 21–36.
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.
- Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629–635.
- Wamba, F. S., & Akter, S. (2019). Understanding supply chain analytics capabilities and agility for datarich environments. *International Journal of Operations and Production Management*, 39(6–8), 887–912.
- Weese, M., Martinez, W., Megahed, F. M., & Jones-Farmer, L. A. (2017). Statistical learning methods applied to process monitoring: an overview and perspective. *Journal of Quality Technology*, 48(1), 4–24.
- Xu, X., Sun, Y., Wan, J., Liu, X., & Song, Z. (2017a). Industrial big data for fault diagnosis: Taxonomy, review, and applications. *IEEE Access*, 5, 17368–17380.
- Xu, Z., Hu, C., Yang, F., Kuo, S.-H., Goh, C.-K., Gupta, A., & Nadarajan, S. (2017b). Data-driven inter-turn short circuit fault detection in induction machines. *IEEE Access*, 5, 25055–25068.
- Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019a). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, 35(1), 213–223.
- Yu, W., Jacobs, M. A., Chavez, R., & Feng, M. (2019b). Data-Driven Supply chain orientation and financial performance: The moderating effect of innovation-focused complementary assets. *British Journal of Management*, 30(2), 299–314.
- Zhang, D., Qian, L., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, 6, 21020–21031.
- Zhang, W., Yang, D., & Wang, H. (2019). Data-Driven Methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, 13(3), 2213–2227.
- Zhang, Z., Si, X., Hu, C., & Kong, X. (2015). Degradation modeling-based remaining useful life estimation: A review on approaches for systems with heterogeneity. *Proceedings of the Institution of Mechanical Engineers, Part o: Journal of Risk and Reliability*, 229(4), 343–355.

Authors and Affiliations

Berndt Jesenko¹ · Christian Schlögl¹

✉ Berndt Jesenko

berndt.jesenko@campus02.at; berndt.jesenko@edu.uni-graz.at

Christian Schlögl

christian.schloegl@uni-graz.at

¹ Institute of Operations and Information Systems, University of Graz, Universitätsstraße 15/F3, 8010 Graz, Austria