Data-driven scientific research based on public statistics: a bibliometric perspective

Jorge-Eusebio Velasco-López; Ramón-Alberto Carrasco; Manuel J. Cobo; Gema Fernández-Avilés

Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87085

Recommended citation:

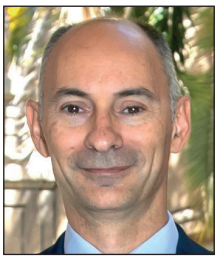
Velasco-López, Jorge-Eusebio; **Carrasco, Ramón-Alberto**; **Cobo, Manuel J.**; **Fernández-Avilés, Gema** (2023). "Data-driven scientific research based on public statistics: a bibliometric perspective". *Profesional de la información*, v. 32, n. 3, e320314.

https://doi.org/10.3145/epi.2023.may.14

Manuscript received on 4th August 2022
Accepted on 8th March 2023

José-Eusebio Velasco-López ✉
https://orcid.org/0000-0003-2392-5690

Universidad Complutense de Madrid
Facultad de Estudios Estadísticos
Avda. Puerta de Hierro, s/n
28040 Madrid, Spain
jorge.velasco.lopez@ine.es



Ramón-Alberto Carrasco
https://orcid.org/0000-0001-7365-349X

Universidad Complutense de Madrid
Facultad de Estudios Estadísticos
Avda. Puerta de Hierro, s/n
28040 Madrid, Spain
ramoncar@ucm.es

Manuel J. Cobo
https://orcid.org/0000-0001-6575-803X

Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (*DaSCI*)

Universidad de Granada
Periodista Daniel Saucedo Aranda, s/n
18071 Granada, Spain
mjcobo@ugr.es

Gema Fernández-Avilés
https://orcid.org/0000-0001-5934-1916

Universidad de Castilla-La Mancha
Facultad de Ciencias Jurídicas y Sociales de Toledo
San Pedro Mártir, s/n
45071 Toledo, Spain
gema.faviles@uclm.es

Abstract

Official statistics provide information on different areas of citizens' lives and are widely used in scientific research as a source of data due to their open data nature and quality assurance. In this context, a bibliometric analysis is carried out using all *Scopus* publications from 1960 to 2020 that use official statistics as data sources. Thus, 10,777 publications are analyzed using the *SciMAT* bibliometric analysis software, providing a complete conceptual analysis of the main research topics in the literature through the quantification of the main bibliometric performance indicators, identifying the most important authors, organizations, countries, sources, and intellectual structures corresponding to the main fields of research and bringing classification by subject area as an innovation to the methodology.

Keywords

Official statistics; Co-word analysis; Strategic diagram; Science mapping analysis; Bibliometric analysis; *SciMAT*.

Funding

This paper has been partially supported by:

- *Spanish State Research Agency* through project PID2019-105381GA-I00/AEI/10.13039/501100011033 (*iScience*).
- *University of Castilla-La Mancha, Faculty of Law and Social Science* at Toledo, Code: 00440710.

1. Introduction

Today's society is undergoing an accelerated process of digital transformation, or datafication. big data and open data are the most important pillars of this transformation. While big data is defined by volume, open data is defined by its public accessibility for research, new ventures, patterns and trends analysis, data-driven decisions (**Moreno**; **Carrasco**; **Herrera-Viedma**, 2019), and complex problem-solving. A key source of open data is official statistics, which are statistical operations published as a public good by government agencies or other public bodies, such as international organizations (*United Nations*, 2021).

Scientific researchers are among the potential users that may benefit from the data provided by official statistics. They are usually familiar with the statistical methods used and may want to delve into statistical facts and observations for analytical purposes, studying the cause-and-effect interrelationships among different phenomena. In this context, we focus on examining the areas in which there has been scientific research published that employs official statistics as data sources (in whole or in part).

Advancing the knowledge of any scientific discipline requires an examination of the theoretical and empirical contributions of academics and professionals over time. As such, a periodic analysis of the accumulated knowledge is necessary to understand the current state of a discipline (**Rodríguez-López** *et al.*, 2020). After conducting a literature review, we have found no studies similar to the one proposed here. The present article thus seeks to fill this gap and provide a better understanding of the state of the question through an exhaustive bibliometric study involving an analysis of the co-occurrence of keywords and the mapping of all published articles related to official statistics and indexed in the *Scopus* bibliographic database (**Mukherjee** *et al.*, 2022).

The expected result is to discover which themes can be defined as "central" for research on the basis of official statistics in the period examined, which themes play a minor, or "non-central" role, and what changes can be observed during the different periods for all the themes. A performance analysis will be carried out.

In the rest of the article, section 2 provides a description of the methodology applied for the bibliometric analysis, the dataset, and the main tools used. Section 3 presents the bibliometric analysis and section 4 presents a discussion of the relationship detected between official statistics, open data, and scientific research. Finally, section 5 summarizes the most relevant results of the performance and content analysis conducted and outlines possible lines of future research.

2. Methodology and dataset

Below, we detail the methodology employed in our study, with a description of the general context, the specific methodology, the computer tool, and the dataset used.

2.1. Science mapping analysis

Science mapping or bibliometric mapping is a spatial representation of how disciplines, fields, specialties, documents, or authors are related to one another (**Small**, 1999). It has been widely used to reveal the hidden relationships among key elements (documents, authors, institutions, topics, etc.) in different research fields (**Cobo** *et al.*, 2011a; **Rodríguez-Ledesma** *et al.*, 2015; **Rodríguez-López** *et al.*, 2020; **Galán**; **Carrasco**; **LaTorre**, 2022)

Science mapping analysis can be performed with several software tools (**Cobo** *et al.*, 2011b). *SciMAT* was presented by **Cobo** *et al.* (2012) as a powerful tool that integrates most of the advantages of the available science mapping software tools and was designed according to the science mapping analysis approach presented in **Cobo** *et al.* (2011c). It combines both performance analysis tools and science mapping tools to analyze a research field and detect and visualize its conceptual subdomains (particular topics/themes or general thematic areas) and its thematic evolution.

Therefore, in this contribution, *SciMAT* was employed to develop a co-word science mapping analysis (**Callon** *et al.*, 1983; **Batagelj**; **Cerinšek**, 2013; **Börner**; **Chen**; **Boyack**, 2003). In line with **Cobo** *et al.* (2011c), the analysis is performed following a four-stage methodology:

1) Detection of the research themes. For each of the periods analyzed, research themes are detected by applying a clustering algorithm (**Coulter**; **Monarch**; **Konda**, 1998). over a normalized co-word network (**Callon** *et al.*, 1983). The similarity between the keywords is assessed using the equivalence index (**Callon**; **Courtial**; **Laville**, 1991).

2) Visualization of the research themes and thematic networks. The detected themes are visualized through different visualization instruments (Figure 1): the strategic diagram (**He**, 1999) and the thematic network (**Cobo** *et al.*, 2011C). Each theme can be characterized by two measures (**Callon**; **Courtial**; **Laville**, 1991): centrality, which measures the degree of interaction of a network with other networks; and density, which measures the internal strength of the network. Given both measures, a research field can be visualized as a set of research themes, mapped in a two-dimensional strategic diagram, Figure 1(a), and classified into four groups:

- Motor themes (quadrant Q1): They present strong values of centrality and density. These are well-developed themes that are important or central to structuring a research field.
- Highly developed and isolated subjects (quadrant Q2): These are strongly related, highly specialized, and peripheral, but they do not have the appropriate background or importance for the field.

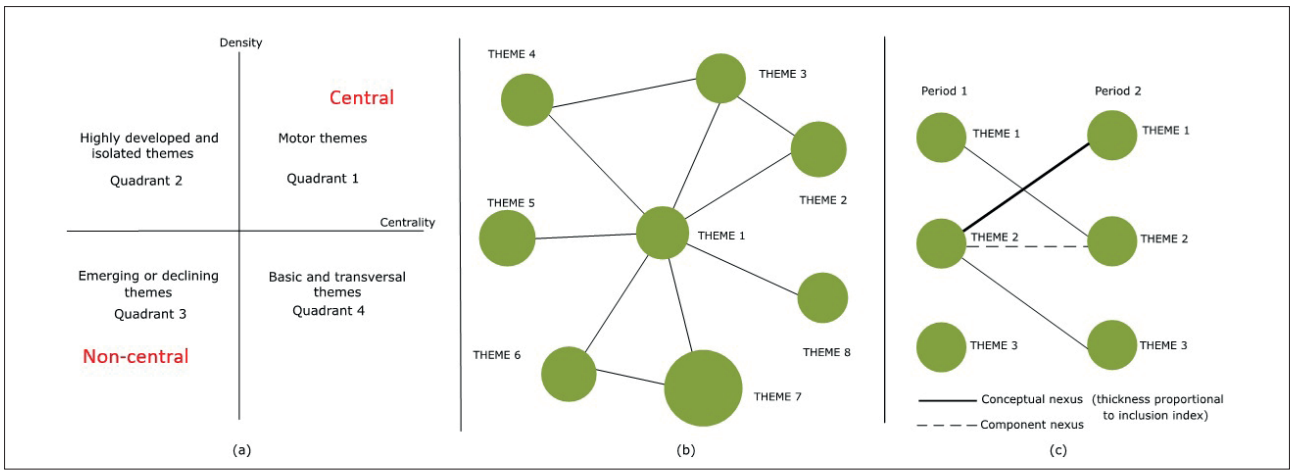


Figure 1. (a) Strategic diagram, (b) thematic network, and (c) thematic evolution

- Emerging or declining themes (niche) (quadrant Q3): They present both low density and low centrality. They are underdeveloped and marginal to the field of research.
- Basic and transversal themes (quadrant Q4): These themes are important for the research field, but internally they are not extensively developed.

In this study, we refer to the themes on the right-hand side (motor and basic) as central and those on the left-hand side (niche and peripheral) as non-central.

3) Discovery of thematic areas. In this stage, the evolution of the research themes over a set of periods is first detected and then analyzed to identify the main general areas of evolution in the research field, their origins, and their interrelationships. Their evolution over the whole period is then measured as the overlapping of clusters from two consecutive periods. For this purpose, the inclusion index (**Sternitzke**; **Bergmann**, 2009) is used to detect conceptual nexuses between research themes in different periods, thus allowing the identification of the thematic areas in a research field. The thematic evolution map, Figure 1(c), depicts the current development of research that uses official statistics, applying a longitudinal perspective covering the entire period examined. In the visualization, the solid lines mean that the linked cluster shares the main item, while a dotted line means that the themes share elements that are not the main item. The thickness of the lines is proportional to the inclusion index value, and the size of the circles is proportional to the number of published documents associated with each cluster. Section 3.3 will show the thematic fluctuations.

4) Performance analysis. In this stage, the relative contribution of the research themes to the whole research field is measured (quantitatively and qualitatively) and used to establish the most prominent, productive, and highest-impact subfields. Some of the bibliometric indicators to use are the number of published documents, the number of citations, and different types of h-index (**Alonso** *et al.*, 2009; **Hirsch**, 2005; **Martínez** *et al.*, 2004). For each theme, the performance measures are computed based on the documents associated with it.

The methodology described above is extended by introducing a new step in which each theme detected in the analysis phase is assigned to a certain main area of knowledge, through a classification procedure based on a cluster network analysis (thematic network, Figure 1b). After determining this area of knowledge, we then add this new information to our maps from the previous stage. A theme can cut across different fields of knowledge, although there is always one field that predominates, to which it is therefore assigned. Given that the visualization techniques applied already make use of the positioning and the size of the object to graphically represent the properties of the objects (circles that represent a theme), in this study we assign colors to these objects according to the main field of knowledge to which they belong. The colors associated with each knowledge area are defined in Table 1.

Table 1. Subject areas with their associated colors

Medicine	**Psychology**	**Sociology**	**Economy**
Microbiology	**Arts**	**Environment**	**Business and management**

2.2. Dataset

Scopus is one of the world's largest abstract and citation databases of peer-reviewed research literature, with over 25,000 titles from more than 7,000 international publishers and extensively used in bibliometric analyses. To identify possible changes in the themes identified in research based on official statistics, a global search by title, abstract, and keywords on articles in the above-mentioned *Scopus* database is carried out. The results are then analyzed using a co-word analysis in the following steps. Note that no filter by starting year is applied, since the purpose of the study is to analyze the entire *Scopus* database, which starts in 1960.

It is empirically observed that entering the term GOVERNMENT DATA distorts the search, yielding many results about E-GOVERNMENT, which is not directly related to official statistical information. Likewise, the term PUBLIC DATA does not adequately capture the use of official statistical data sources. Therefore, the search eventually implemented is as follows:

> A key provider of open data is official statistics, which are statistical operations published by government agencies or other public bodies such as international organizations

TITLE-ABS-KEY ("OFFICIAL STATISTICS") OR TITLE-ABS-KEY ("OFFICIAL DATA") OR TITLE-ABS-KEY ("GOVERNMENT STATISTICS") OR TITLE-ABS-KEY ("PUBLIC STATISTICS") OR TITLE-ABS-KEY ("NATIONAL STATISTICS").

A total of 10,995 results are obtained. By conducting a manual review, we discard 218 of them that relate to E-GOVERNMENT and other non-related topics, considering the remaining 10,777 of interest and suitable for the study.

This selection is exported in Research Information Systems (RIS) format, and then imported into *SciMAT* (**Cobo** *et al.*, 2012). It includes all the citation information, the abstract along with the keywords, as well as the references, all of which are used to build the knowledge base for a subsequent science mapping analysis. Approximately 10% of the documents are found to be lacking any keywords (and therefore will not be analyzed). In addition, some keywords without meaning in this context, such as stop words or words with a very broad and general meaning, are removed to improve the analysis. Considering that our focus is on understanding the development of the themes, we exclude methodological terms, codes, years, geographic information, and dates, although some information could also be learned from those terms. The stop words are selected following the procedure proposed by **Jacob** and **Brust** (2019). The next step is time slicing, whereby the entire time frame of the sample (1960-2020) is divided into the following four periods:

- 1960-2000 (1,885 documents);
- 2001-2010 (2,527);
- 2011-2015 (2,486);
- 2016-2020 (3,878).

In this study, we configure the keyword frequency threshold and the co-occurrence frequency threshold for each period: (9, 2), (13, 2), (11, 2), and (17, 2), respectively.

3. Bibliometric analysis

Next, the bibliometric analysis is carried out using the methodology and dataset presented in section 2.

3.2. Performance statistical analysis

From the documents retrieved for this study, a total of 33,078 keywords have been identified, with the most frequent ones presented in Table 2.

Table 2. Keywords with corpus frequency higher than 200

Keywords	Frequency	Keywords	Frequency
Mortality	1,219	Morbidity	286
Epidemiology	677	Health	278
Sex-Difference	530	Geography	277
Cause-of-Death	502	Suicide	245
Population	499	Pregnancy	244
Economics	466	Health-Care-Policy	230
Demography	389	Sex-Factors	229
Public-Health	382	Economic-Analysis	228
Socioeconomics	353	Cancer	225
Health-Survey	347	Hospitalization	222
Age-Distribution	333	Time-Factors	215
Migration	330	Psychology	210
Socioeconomic-Factors	329	Gender	210
Health-Care-System	307	Survival-Rate	207
Age-Factors	291	Employment	206

Documents with more than 500 citations are shown in Table 3, along with their respective authors, citations, and year of publication. The table is ordered by subject area and the Field-Weighted Citation Impact (FWCI) metric (**Purkayastha** *et al.*, 2019), which is the ratio of the total citations received by the denominator output and the total citations expected based on the average of the subject field.

Table 3. Documents with more than 500 citations, sorted by subject and FWCI

Subject	FWCI	Document	Citations
Business and Management	22.75	(**Gustavsson** *et al.*, 2011)	970
	18.22	(**Olesen** *et al.*, 2012).	828
Economy	47.55	(**Henderson**; **Storeygard**; **Weil**, 2012)	553
Environment	27.89	(**Giri** *et al.*, 2011)	1,271
	9.20	(**Liu** *et al.*, 2015)	627
	8.23	(**Chen**, 2007)	564
Medicine	129.00	(**Allison** *et al.*, 1999)	1,463
	58.41	(**Hippisley-Cox** *et al.*, 2008)	861
	43.95	(**Llewelyn** *et al.*, 2004)	907
	29.88	(**Peto** *et al.*, 2000)	1,006
	17.19	(**Unal**; **Critchley**; **Capewell**, 2004)	539
	10.27	(*Multicentre Aneurysm Screening Study Group*, 2002)	909
	7.79	(**Bernstein** *et al.*, 2000)	575
Sociology	49.41	(**Gillborn**, 2008)	520
	44.32	(**Van-Dijk**; **Hacker**, 2003)	635
	25.01	(**Grantham-McGregor** *et al.*, 2007)	1,599
	13.96	(**Abraido-Lanza** *et al.*, 1999)	692
	6.87	(**Savage** *et al.*, 2013)	546

The most prolific countries in terms of document output can be seen in Table 4, together with the affiliation organizations accounting for the highest number of documents in the sample (all of them public institutions in the United Kingdom).

Table 4. Countries with more than 300 documents published and affiliation organizations with more than 75 documents

Country	Documents	Affiliation	Documents
United Kingdom	2,292	*University College London*	163
United States	1,579	*Office for National Statistics London*	163
Germany	636	*University of Oxford*	128
Spain	630	*University of Manchester*	123
Russian Federation	625	*King's College London*	112
Italy	587	*London School of Hygiene-Tropical Medicine*	106
China	420	*Imperial College London*	103
Sweden	339	*University of Cambridge*	90
Australia	333	*University of Oxford-Medical Science Division*	81
Brazil	328	*University of Bristol*	81
France	319	*University of Southampton*	78

3.2. Thematic analysis

To analyze for each period the most prominent themes in research based on official statistics, a strategic diagram is provided for each period. Performance analysis is used to improve the results.

3.2.1. First period (1960-2000)

According to the strategic diagram presented in Figure 2a, during this period the research activity was focused on 22 themes, with the following 12 central themes (motor themes plus basic themes): *Economic-Factors*, *Developed-Countries*, *Developing-Countries*, *Employment*, *Contraceptive-Usage*, *Geography*, *Social-Policy*, *Pregnancy*, *Cause-of-Death*, *Forensic-Medicine*, *Life-Table*, *Health-Survey.*

Note that for all strategic diagrams in this section, the color of the circles corresponds to the classification of the themes obtained from the cluster network analysis, the size of the circles reflects the number of documents found for each theme, and the quadrant where they are located indicates the centrality and density of the theme, as outlined in Figure 1. The performance measures of the themes are shown separately for each of the four periods in Tables 5, 6, 7, and 8, sorted by the number of documents. Each table shows the number of documents assigned to the themes with more than 100 documents, the citations, and the h-index of the documents associated with the theme. Also, a summary of the central (basic and motor) and non-central themes and their subject area can be found.

(a) First period (1960-2000)

(b) Second period (2000-2010)

(c) Third period (2010-2015)

(d) Fourth period (2015-2020)

Figure 2. Strategic diagram of the four periods

According to the performance measures, the following themes stand out with an h-index greater than 28 and a number of citations greater than 3,000: *Developing-Countries*, *Geography*, *Economic-Factors*, *Cause-Of-Death*, *Forensic-Medicine*, *Developed-Countries*, *Prevention*, *Health-Survey*, *Employment*, and *HIV-Infections.* Most of these fall within the central themes.

Note that themes related to Medicine and Sociology predominate, both in motor and basic themes as well as in non-central ones (niche and peripheral)

Non-central themes (i.e., peripheral or niche) within the research are shown in the left section of Figure 2a. The themes *Congenital-Malformation*, *Cancer-Incidence*, *History-Of-Medicine*, and *Substance-Related-Disorders* (all health-related) in the upper left quadrant can be classified as isolated and developed themes. The themes of *Prevention*, *HIV-Infections*, *Hospitalization*, *Morbidity*, *Economic-Analysis*, and *Primary-Health-Care* in the lower left quadrant are related to emerging or declining themes.

3.2.2. Second period (2001-2010)

In this period, the research delved into 23 themes (see Figure 2b). In this case, according to the strategic diagram, 11 central themes can be identified (motor themes plus basic themes*): Mortality*, *Socioeconomics*, *Health-Care-Policy*, *Cardiovascular-Risk*, *Mental-Disorders*, *Health-Behavior*, *Health-Care-Quality*, *Treatment-Outcome*, *Health-Statistics*, *Cost-Of-Illness*, *Ethnicity*, and *Cancer*. However, of the motor themes, *Mortality*, *Socioeconomics*, *Health-Care-Policy* and *Cardiovascular-Risk* are the most influential as they are well developed and central to research based on official statistics during this period.

> We investigate which themes can be defined as "central" and "peripheral" of the research based on official statistics in the period examined and, if a change in trend is observed in these themes, which may be the reasons for such changes

Table 5. Performance measures of the themes (central and non-central) of the first period (1960-2000), sorted by the number of documents. Note: *Non-central themes

Name	Docs.	h index	Citations	Subject area	Summary
Developing-Countries	419	42	9,881	Sociology	Topics related to mortality and health, the status of women, and education
Geography	396	41	6,333	Medicine	Epidemiology related to gender, ethnicity, age, and geographic factors
Economic-Factors	304	33	5,015	Sociology	Economic factors that influence the socioeconomic status and working conditions
Cause-Of-Death	300	46	10,131	Medicine	Different types of diseases (coronary, respiratory, cancer...) and their associated cause of death
Forensic-Medicine	270	36	5,370	Medicine	Ethological and psychological aspects that can lead to suicide, etc.
Developed-Countries	260	28	2,923	Sociology	Sociodemographic characteristics in developed countries (type of family, migration, marriage)
Prevention	202	31	3,646	Medicine*	Prevention of work accidents and related aspects such as diagnosis and injuries
Health-Survey	157	32	4,819	Medicine	Uses of health statistics in demographic and medical studies
Employment	145	29	3,097	Sociology	Sociological and psychological aspects of employment and unemployment
HIV-Infections	143	30	3,365	Medicine*	Determinants of HIV infections according to cultural, educational, and racial characteristics
Pregnancy	142	26	2,334	Medicine	Aspects related to infant mortality and characteristics of the pregnant mother and newborn
Contraceptive-Usage	142	17	1,861	Medicine	Use of contraceptive measures and sexual behavior and related aspects such as abortions, fertility, and access to healthcare
Morbidity	135	26	3,687	Sociology*	Hospitalized illnesses according to lifestyles, race, and cost of stay
Hospitalization	127	22	1,896	Medicine*	Medical issues related to the length of stay in hospitals
Social-Policy	110	17	1,918	Business and Management	Social, population, health, and public policies

The *Mortality* theme presents the highest performance indicators (h-index of 87), with the highest number of documents and citations. *Socioeconomics*, *Health-Care-Policy*, *Ethnicity*, *Health-Statistics*, *Residence-Characteristics*, *Treatment-Outcome*, *Cardiovascular-Risk* are also important themes, all with more than 9,000 citations, h-index values greater than 45, and a significant number of documents. All these themes are framed within the central themes.

According to the number of documents, it is observed that themes related to Medicine predominate, as in the first period. Now, however, themes related to Business and Management (mostly public policies) emerge as the second most important group. Sociology moves to third place, followed by themes related to Psychology, Environment, and Economy. Medicine and Sociology are distributed among all the quadrants, while Business and Management does not appear in the second one.

Non-central themes within the research are shown in the left section of Figure 2b. The *Antidepressant-Agent*, *Occupational-Accident*, *Traffic-Accident*, *Environmental-Monitoring*, and *International-Trade* themes in the upper left quadrant can be classified as isolated and developed themes. We observe that only the first two are related to health, but the first theme is related to the Environment and two more related to the Economy appear. The themes *Residence-Characteristics*, *Pregnancy*, *Groups-By-Age*, *Length-Of-Stay*, *HIV-Infections*, and *Public-Policy* in the lower left quadrant relate to emerging or declining themes.

3.2.3. Third period (2011-2015)

In this period, the research delved into 23 themes (Figure 2c). In this case, according to the strategic diagram, 12 central themes can be identified (motor themes plus basic themes): *Mortality*, *Cancer-Registry*, *Diabetes-Mellitus*, *Health-Care-Disparity*, *Population*, *Hospital*, *Health-Care-Cost*, *Myocardial-Infarction*, *Suicide*, *Educational-Status*, *Organization-And-Management*, and *Urban-Population*. However, of the motor themes, *Cancer-Registry*, *Diabetes-Mellitus*, and *Health-Care-Disparity* are the most influential as they are well developed and central to the research based on official statistics during this period.

The *Mortality* theme registers the highest performance indicator (h-index of 56) with the highest number of documents and citations. *Population*, *Health-Care-Disparity*, *Health-Care-Cost*, *Suicide*, *Educational-Status*, and *Hospital* are also important themes, with more than 6,000 citations, an h-index greater than 38, and a notable number of documents. All these themes are framed within the central themes.

The themes identified are classified into large thematic areas, where those related to Medicine and Sociology predominate in terms of the number of documents, as in the first period. Themes related to Business and Management (mostly

Table 6. Performance measures of the themes (central and non-central) of the second period (2000-2010), sorted by the number of documents. Note: *Non-central themes

Name	Docs.	h index	Citations	Subject area	Summary
Mortality	667	87	30,032	Medicine	Topics related to causes of death
Socioeconomics	500	68	19,156	Sociology	Topics involving socio-cultural and economic aspects of the population
Health-Care-Policy	311	49	8,823	Business and Management	Topics related to health policies
Ethnicity	274	48	9,442	Sociology	Social and demographic aspects related to ethnicity
Health-Statistics	251	50	8,977	Medicine	Topics related to health quality
Residence-Characteristics	242	47	9,859	Sociology*	Socioeconomic characteristics of dwellings according to whether they are rural or urban
Treatment-Outcome	233	54	12,102	Business and Management	Results of medical treatment
Health-Care-Quality	230	48	7,199	Business and Management	Health quality in terms of accessibility and other factors
Mental-Disorders	218	45	7,813	Psychology*	Social, legal, and economic aspects of mental disorders
Occupational-Accident	209	39	5,723	Sociology*	Accidents by type of employment
Pregnancy	195	40	6,833	Medicine*	Medical and social aspects related to motherhood
Cost-Of-Illness	183	42	5,977	Business and Management	Cost of diseases by type
Health-Behavior	182	45	8,083	Psychology	Health behavior according to socioeconomic and mental health factors
Cancer	175	47	8,181	Medicine	Cancer-related aspects (risk, mortality, incidence)
Environmental- Monitoring	166	41	5,721	Environmental*	Environmental supervision concerning contamination and protection
Cardiovascular-Risk	165	47	9,826	Medicine	Cardiovascular risk factors and their consequences
Antidepressant- Agent	163	45	6,898	Medicine*	Consequences and risks of using antidepressant medications
Traffic-Accident	156	36	4,406	Medicine*	Topics related to traffic accidents
International-Trade	121	24	2,261	Economy*	Topics related to international trade
Public-Policy	133	27	2,358	Business and Management*	Public policies based on economic analysis

public policies) are still important, but not as much as in the previous period. Economic and Environmental themes are less important but continue to appear, while it is notable that those related to Psychology disappear altogether. Themes are distributed throughout all the quadrants, except the Environment, which remains a peripheral and developed theme, as in the previous period. Likewise, as was the case in the previous period, Management themes do not appear in the second quadrant.

Non-central themes within the research are shown in the left section of Figure 2c. The *Greenhouse-Gases*, *Animals*, *Satellite-Imagery*, and *Internet* themes in the upper left quadrant can be classified as isolated and developed themes. We note that the themes related to the Environment are located here, as well as one related to the use of the internet. The themes *Residence-Characteristics*, *Pregnancy*, *Groups-By-Age*, *Length-Of-Stay*, *HIV-Infections*, and *Public-Policy* in the lower left quadrant relate to emerging or declining themes.

3.2.4. Fourth period (2016-2020)

In this period, the research delved into 22 themes (see Figure 2d). In this case, according to the strategic diagram, 12 central themes can be identified (motor themes plus basic themes): *Mortality*, *Hypertension*, *Hospital-Admission*, *Primary-Medical-Care*, *Disease-Severity*, *Socioeconomics*, *Health-Care-Cost*, *Virology*, *Environmental-Monitoring*, *Land-Use*, *Hospital*, and *Legislation-And-Jurisprudence.* Of the motor themes, *Mortality* and *Hypertension*, are the most influential because they are well-developed and central to research based on official statistics during this period.

> "A new step is introduced in the traditional mapping process, consisting of each theme detected in the analysis phase of the process being assigned to a certain main area of knowledge"

The *Mortality* theme registers the highest performance indicator (h-index of 38), while *Health-Care-Cost*, *Economic-Analysis*, *Socioeconomics*, *Land-Use*, *Hospital-Admission*, and *Virology* are also important themes, with

Table 7. Performance measures of the themes (central and non-central) of the third period (2010-2015), sorted by the number of documents.
Note: *Non-central themes

Name	Docs.	h index	Citations	Subject area	Summary
Mortality	492	56	13,283	Medicine	Morbidity, epidemiology, and causes of death by sex and age
Population	306	43	7,842	Sociology	Studies on cultural and social issues affecting the population
Health-Care-Disparity	265	43	7,396	Sociology	Socioeconomic issues that affect access to healthcare facilities
Health-Care-Cost	264	41	8,501	Business and Management	Cost of healthcare related to socioeconomic factors
Suicide	239	39	6,008	Psychology	Social, medical, and psychological aspects of suicide
Educational-Status	239	38	6,553	Sociology	Pregnancy and access to healthcare according to educational status
Decision-Making	222	34	4,073	Business and* Management	Political aspects of the Sustainable Development Goals (SDG)
Hospital	210	41	6,296	Medicine	Access to healthcare according to sociodemographic aspects
Employment	206	30	3,387	Sociology*	Economic aspects of employment
Animals	172	34	5,459	Microbiology*	Environmental protection and animal epidemics
Organization-And- Management	162	28	3,363	Business and Management	The political organization of public health
Diabetes-Mellitus	139	31	3,515	Medicine	Various diseases related to diabetes-mellitus together with risk factors
Urban-Population	129	31	3,493	Sociology	Social aspects related to the urban population
Greenhouse-Gases	123	34	4,298	Environmental*	Climate aspects of greenhouse gases
Satellite-Imagery	119	35	4,743	Environmental*	Use of satellite images for monitoring of agriculture and ecosystems
Myocardial- Infarction	116	29	3,432	Medicine	Topics related to myocardial infarction
Cancer-Registry	104	31	3,090	Medicine	Treatment, survival, and other aspects related to cancer registries
Infant-Mortality	103	25	1,877	Medicine*	Infant mortality related to vaccinations and cardiovascular risk

more than 3,000 citations, an h-index greater than 26, and a notable number of documents. Most of these themes fall within the central themes.

For this last period, the themes obtained are classified into large thematic areas, showing that the areas of Medicine and Sociology are still important in terms of number of documents, as in all previous periods, but now the most important field is related to Business and Management (mostly public policy). In addition, the Environment gains prominence while the Economy loses it. Finally, new areas appear, notably Microbiology. Sociology and Medicine mainly lie in the first and third quadrants, whereas Environment is now distributed throughout all quadrants. As in the previous period, Business and Management themes do not appear in the second quadrant.

Non-central themes within the research are shown in the left section of Figure 2d. The *Crops* and *Energy-Utilization* themes in the upper left quadrant can be classified as isolated and developed themes. We note that this is the quadrant that contains themes related to the Environment and agriculture. The *Ethnicity*, *Economic-Analysis*, *Education*, *Population-Distribution*, *Big-Data*, *Disease-Surveillance*, *Mental-Health*, and *Benchmarking* themes in the lower left quadrant are related to emerging or declining themes.

3.3. Conceptual evolution map

Using *SciMAT*, an analysis of the themes detected in each period is carried out, considering their keywords and their evolution over time.

The evolution map (Figure 3) depicts the current development of research using official statistics, applying a longitudinal perspective covering the entire period examined. Few thematic fluctuations can be found here, although there is more variation in themes that are not basic or motor. In Figure 3, only thick dashed lines (connecting groups in adjacent periods that share individual keywords) have been left, while we have removed the themes that are not basic or motor and that did not show any connections with adjacent periods, with themes assigned to their corresponding main thematic area through the color of the circle.

Table 8. Performance measures of the themes (central and non-central) of the fourth period (2016-2020), sorted by the number of documents. Note: *Non-central themes.

Name	Docs.	h index	Citations	Subject area	Summary
Mortality	663	38	6,776	Medicine	Causes of death
Health-Care-Cost	477	30	4,217	Business and Management	Cost of healthcare
Economic-Analysis	436	27	3,578	Economy*	The financial crisis and its consequences
Socioeconomics	453	27	3,863	Sociology	Socioeconomic aspects (poverty, health) according to population profiles
Education	395	26	2,949	Sociology*	Influence of education and other factors such as age and sex on well-being and employment
Legislation-And- Jurisprudence	343	25	2,599	Business and Management	Public policies related to trade, health, and crime
Environmental-Monitoring	334	28	2,923	Environmental	Environmental issues
Land-Use	331	28	3,134	Environmental	Land use and Environmental aspects
Hospital-Admission	320	31	3,886	Business and Management	Management aspects of public health and the treatments used
Virology	318	26	3,114	Microbiology	COVID-19 related issues
Ethnicity	311	20	2,124	Sociology*	Relationship between ethnicity and health and social aspects such as social class and migration
Big-Data	293	17	1,457	Sociology*	Alternative data sources to obtain statistical indicators
Primary-Medical-Care	259	25	3,117	Business and Management	Aspects of primary health care system by types of disease
Crops	231	23	2,144	Environmental*	Aspects related to agriculture and means of analysis
Disease-Severity	226	25	2,969	Medicine	The severity of various diseases
Energy-Utilization	216	24	1,867	Environmental*	Energy efficiency
Hypertension	184	23	1,844	Medicine	Hypertension as a cause of cardiovascular events, diabetes, and other diseases
Hospital	159	18	1,563	Business and Management	Aspects related to accidents and illnesses and hospital stays
Population-Distribution	149	15	982	Environmental*	Lifestyle influence on climate change
Disease-Surveillance	103	17	1,081	Medicine*	Follow-up of some diseases

Considering the above, we can establish some of the major trends in the themes according to the related subject areas.

Concerning the field of knowledge to which themes belong, we make the following observations:

- For the field of Medicine, an especially relevant timeline is *Cause-of-Death* (period 1)/ *Mortality* (period 2)/ *Mortality* (period 3)/ *Mortality* (period 4), dealing with aspects such as causes of death, epidemiology, cancer mortality, hospitalization, and treatment outcomes.
- For the field of Sociology, we observe the timeline *Economic-Factors* (period 1)/ *Socioeconomics* (period 2)/ *Health-Care-Disparity* (period 3)/ *Socioeconomics* (period 4). It covers aspects related to the social sciences, especially sociology, and features interconnections with other fields such as Economy and Medicine.
- For the field of Business and Management, we identify the timeline *Social-Policy* (period 1)/*Health-Care-Policy*, *Cost-Of-Illness* (period 2)/*Health-Care-Cost* (period 3)/ *Health-Care-Cost* (period 4), covering aspects related to healthcare management.
- There are also timelines related to the Environment starting in period 2: *Environmental-Monitoring* (period 2) / *Greenhouse-Gases* (period 3)/ *Energy- Utilization* (period 4). It covers aspects related to environmental protection.
- The relationships between different periods are very solid, although there are also themes that have a different assigned field of knowledge (subject area) in different periods. This is because, as has been pointed out before, themes do not necessarily belong exclusively to a single field of knowledge; in fact, they usually relate to several, although they remain assigned to the main one. The themes with the highest rate of such cross-links are *Mortality*, *Socioeconomics*, and *Health-Care-Cost*, which again represent some of the main central themes within the research based on official statistics over time. Other themes develop quite independently with only a few cross-links throughout the entire period examined, such as *Animals* and *Crops*.

Regarding the themes:

- The cross-links between the themes in the upper half of Figure 3 (central research themes) are stronger than those in the lower half (non-central themes). Furthermore, we can identify which specific central and non-central themes

occur more frequently and which other themes, unusual or not at all related, hardly appear throughout the period examined.

- Some themes appear throughout almost all the periods examined, such as *Mortality*, *Socioeconomics*, or *Health-Care-Cost*, with these generally being central themes. Other themes have developed very recently and are notably central, such as *Virology*, *Environmental-Monitoring*, *Hypertension*, and *Primary-Medical-Care*. They are all themes related to Medicine and the Environment. Specifically, *Environmental-Monitoring* and *International-Trade*, which had previously established hardly any cross-links in the network, emerged at the bottom of the diagram in the second period. In the third, the same happened with *Satellite-Imagery*, *Internet*, and *Animals.*
- It is also possible to identify topics that lose researchers' interest over time, such as *HIV-Infections*, *Developed-Countries* and *Developing-Countries*, *Contraceptive-Usage*, or *Traffic-Accidents.*
- Some themes gain more centrality and density over time, such as the *Cancer* cluster. This is the case until the fourth period, where it disappears and is integrated into that of *Mortality*, albeit with less importance.

Figure 3. Map of the evolution of all the clusters over the four periods

4. Discussion

As indicated by **Giovannini**, **Martins** and **Gamba** (2009), decision-making at any level is increasingly based on the culture of evidence. They also highlight the changes generated by technological developments, big data, and the datafication of society (**Harford**, 2014), which make statistical information more accessible to any potential user. Moreover, globalization means that the need for statistical information is expanding and is not restricted to governmental or political frameworks. Based on such shifts, these authors conclude that we are witnessing a change in the role of National Statistical Offices, from being information providers to facilitators of knowledge, which is why they are considered a public good. Given this public good character, official statistics may affect the decisions of all members of our society, including scientific researchers and the private sector. Since decisions are based on the available evidence (**García-Villar**, 2012), they may be distorted or affected if these statistics are poorly prepared, or people lack confidence in them. Since statistical output revolves around users' perception of trust, which is largely derived from the quality of statistics, official statisticians invest a lot of effort in ensuring it. The quality management system of the National Statistics Offices is aimed at establishing Quality Assurance Frameworks, inspired by the preeminent schemes at the international level, such as that of the *United Nations* or the one defined by *Eurostat* from the Code of Good Practices (**Radermacher**, 2014) of European Statistics.

Data typically needed for an investigation are collected in different ways. In addition to the specific ad-hoc generated or available data for a specific issue researchers may use official statistical sources. It is worth highlighting the value of open data for scientific work (**Lnenicka**; **Luterek**; **Nikiforova**, 2022), which can come from any of the abovementioned sources. In this sense, the data provided by official statistics are essentially open, pursuant to the legislation on data protection. The provision of open data is currently being promoted at an international level, with legislation such as *Directive (EU) 2019/1024* (on open data and reuse of public sector information), the *Data Governance Act*, the *Data Act* (linked to the *European Digital Strategy*), the creation of *European Data Spaces* and other legislative initiatives to promote high-value public sector datasets (HVDS), which include official statistics. Initiatives such as the *ODIN* project, by *Open Data Watch*, which seeks to assess the coverage and openness of official statistics, confirm the rise in reusable and free-use statistical data, but also indicate that there are challenges ahead in both availability and quality.

> Show relationship between the demand and supply of official statistics as a source of information that can be used as a basis for promoting from Public Statistics those areas of work that respond to the needs of academic research

Regarding their relevance, public statistics are constantly being updated, with national forums such as the *Higher Council of Statistics* and its *National Statistical Plans for Spain* (*España*, 2020), where these changes are legally approved. At an international level, forums of this kind can also be found in the field of official statistics, such as the *Conference of European Statisticians* (*CES*) of the *United Nations Economic Commission for Europe* (*Unece*). They analyze the production of statistics in pioneering areas where measurement is crucial but notoriously difficult, such as climate change, international migration, well-being, human capital, gender identity, social exclusion, and advances in the use of new data sources from the big data environment due to the appearance of new phenomena in society, such as social networks. Indeed, in our study, we can identify a series of themes studied using official statistics, which match these aforementioned recently emerging areas. Some of these themes that appear are *Big-Data*, which enters quadrant 4, linked to *Statistical Commission Decision* (*United Nations*, 2014), which promotes the practical use of big data sources; *Environmental-Monitoring* (*Unece*, 2020), *Sustainable-Development-Goals* (*Unece*, 2021), *Crops* (*Unece*, 2019), *Gender* (**Vikat**; **Jones**, 2014), *Land-Use* (*Unece*, 2017), among others. In addition, we note the capacity of official statistics to provide information on themes that arise unexpectedly; for example, *Virology*, closely related to terms such as *Covid-19*, *Coronavirus*, etc., stands out in the last period for its high density and centrality. Therefore, this study makes clear that societal demand for the availability of new topics leads to thematic shifts and updates over time.

This work may have various implications:

- For those responsible for public statistics, it confirms that their efforts to regularly update contents through statistical plans and to meet the unexpected needs that arise, such as the need to measure the effects of the covid pandemic, are reflected in the use of this information by social agents, and specifically by the scientific community. On the other hand, this study makes it possible to quantify the use that has been made of public statistical sources for research purposes in each area of knowledge. Where appropriate, this can enable statisticians to adjust or promote the statistical resources provided for this research. It has also been pointed out that National Statistical Offices are called on to play a leading role when it comes to coordinating open data publication policies that can serve citizens and the scientific community. To achieve this, it is essential to ensure that open data and interoperability are key criteria in the creation, implementation, and execution of national statistical platforms from now on. For this reason, the *United Nations Statistical Commission* is currently discussing how to update the fundamental principles of official statistics (**Georgiou**, 2017) and its statistical manuals to incorporate the principles of open data.
- For the scientific community, the efforts made by national statistics offices to guarantee quality data and make relevant open data available to society, mean public statistics can be valued as a source of open and reliable data for researchers, enabling them to test their theoretical models. In addition, this study allows researchers to find out whether this source of information and its thematic structure and evolution has been used previously in their research topic.
- It is also important to highlight the advantages that official statistics offer to data-driven public and private organizations. The conceptual architecture required for such enterprises has several layers (**Moreno**; **Carrasco**; **Herrera-Viedma**, 2019; **Galán**; **Carrasco**; **LaTorre**, 2022): data storage, insight generation, action, and outcomes. Due to the important role played by data (the first layer of this architecture), this study, which analyzes the use of official (and therefore high quality) data, can be useful when incorporating new sources into this architecture for decision-making.

5. Conclusion and future work

Applying a bibliometric perspective, this paper analyzes the use of official statistics as data sources, either in whole or in part, for scientific research. To that end, we examine 10,777 related articles from the period 1960-2020, extracted from the *Scopus* database. Due to the heterogeneous nature of the fields of knowledge of the themes found, and to simplify their categorization and interpretation, we propose an original classification of the themes, based on an analysis of their thematic network. The authors of this paper combine experience of official statistics with scientific expertise in bibliometrics, which has been essential for reaching and interpreting the conclusions of this study.

Regarding the thematic analysis, the following results can be highlighted:

- The main themes, those with the highest performance measures, have been described in detail for each period. They are closely related to Medicine, Business and Management and Sociology. Other high performance themes from different areas have also been found in recent periods, related to Psychology, Environment and Microbiology.
- We identify a progressive appearance of basic and motor themes from fields other than Medicine (*Mortality*…) and Sociology (*Socioeconomics*…), especially Business and Management (*Health-Care-Cost*…), which have become especially relevant in the last period. Themes related to Psychology also appear in the second period (*Health-Behavior*, *Suicide*, etc.), while the fourth period sees the appearance of *Virology* in the field of Microbiology, and other themes related to the Environment (*Land-Use*, *Environmental-Monitoring*).

- On the other hand, a few themes persist as non-central, with the ones related to the Economy being especially relevant. In addition to being scarce, they do not become central in any of the periods. However, from the second period on, Environmental themes that have been gaining centrality over time appear as non-central.

> "Capacity of official statistics to provide information on topics that arise spontaneously, as happened in the last period with the theme Virology, closely related to terms such as Covid-19"

From the longitudinal study and the analysis of the visualizations, the following findings can be noted:

- Several timelines have been identified that relate themes over time, generally from the same field, but in some cases intertwining fields of knowledge. The themes related to Medicine and Sociology persist in all periods as motor and central themes, especially the themes of *Mortality* and *Socioeconomics*, respectively. Concerning the field of Business and Management, the *Health-Care-Cost* theme stands out.
- We identify themes that lose researchers' interest over time (such as *HIV-Infections*, *Developed-Countries*, *Developing-Countries*, *Contraceptive-Usage*, or *Traffic-Accidents*) as well as those that gain greater centrality and density over time, such as the *Cancer* theme. Therefore, we can affirm that slight changes in trends have been observed in research themes based on official statistics.
- It can be said that there is a greater degree of thematic change within the non-central research area. Conversely, during the period analyzed, we find greater stability in the thematic structure within the central area of research.

We have also used the classification by fields of knowledge (subject areas) to determine the number of documents produced in each. In the first period, the studies on Medicine stand out, and in the second those on Sociology, focusing both on central themes as well as peripheral or niche ones. In the second period, in addition to the themes in the field of Medicine and Sociology, those in Business and Management and to a lesser extent Psychology are prominent. The same pattern occurs in the third period, where Environmental themes also begin to appear. In the last period, topics related to Medicine cease to be the most prolific and are overtaken by Business and Management themes in terms of the number of documents produced, with themes related to the Environment also becoming much more relevant. It is also noteworthy that Business and Management issues, despite their importance during all periods, are never in the second quadrant of the strategic diagram (isolated and developed), which is not the case for the other important areas.

Furthermore, we have summarized the most relevant papers, authors, countries, and institutions. Some of the most relevant results are the following:

- the most prolific countries are the UK and the US, followed by Germany, Spain, and the Russian Federation;
- the organizations with the highest number of affiliations are University College of London, Office for National Statistics and the University of Oxford;
- the studies with the highest number of citations are **Grantham-McGregor** *et al.* (2007) and **Allison** *et al.* (1999) from the field of Medicine, and those with the highest FWCI index are **Hippisley-Cox** *et al.* (2008) also from the field of Medicine and **Gillborn** (2008) from the field of Sociology.

The limitations of this research need to be considered. Word co-occurrence analysis enables a broad examination of a topic to gain initial insights into a research area. Therefore, the results of our analysis only scratch the surface of the underlying structure of the research based on official statistics. Our central objective has been to provide an overall impression of the thematic structure of the research based on official statistics. Another critical aspect is the choice of keywords as the basis of the analysis as there are no general rules for assigning keywords when they are selected by the authors. Also, we noticed that approximately 10% of the documents do not contain keywords. In addition, some keywords that have no meaning in this context, such as stop words or words with a very broad meaning, were removed. Despite all this, it can be seen that the keywords examined have a notable signaling effect, with the common appearance of keywords enabling the identification of thematic networks and an underlying thematic structure.

Future studies could focus on finding more detailed and reliable findings by examining smaller scientific subdivisions, but our analysis takes on the challenge of examining all periods and scientific disciplines together. It would also be worth focusing on methodology, due to recent changes in official statistics (**Buskirk**; **Kirchner**, 2020; **Groves**, 2011) that are leading to a "Data Science Sampling Age", marked by triangulation, integration, innovation, and refinement. Similarly, another future research task would be to repeat the bibliometric study sometime later to assess the evolution of the findings of this study.

6. References

Abraído-Lanza, Ana F.; **Dohrenwend, Bruce P.**; **Ng-Mak, Daisy S.**; **Turner, J. Blake** (1999). "The Latino mortality paradox: a test of the 'salmon bias' and healthy migrant hypotheses". *American journal of public health*, v. 89, n. 10, pp. 1543-1548. *https://doi.org/10.2105/AJPH.89.10.1543*

Allison, David B.; **Fontaine, Kevin R.**; **Manson, JoAnn E.**; **Stevens, June**; **VanItallie, Theodore B.** (1999). "Annual deaths attributable to obesity in the United States". *Jama* v. 282, n. 16, pp. 1530-1538.
https://doi.org/10.1001/jama.282.16.1530

Alonso, Sergio; **Cabrerizo, Francisco-Javier**; **Herrera-Viedma, Enrique**; **Herrera, Francisco** (2009). "h-Index: A review focused in its variants, computation and standardization for different scientific fields". *Journal of informetrics*, v. 3, n. 4, pp. 273-289.
https://doi.org/10.1016/j.joi.2009.04.001

Batagelj, Vladimir; **Cerinšek, Monika** (2013). "On bibliographic networks". *Scientometrics*, v. 96, n. 3, pp. 845-864.
https://doi.org/10.1007/s11192-012-0940-1

Bernstein, Ira M.; **Horbar, Jeffrey D.**; **Badger, Gary J.**; **Ohlsson, Arne**; **Golan, Agneta** (2000). "Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction". *American journal of obstetrics and gynecology*, v. 182, n. 1, pp. 198-206.
https://doi.org/10.1016/S0002-9378(00)70513-8

Börner, Katy; **Chen, Chaomei**; **Boyack, Kevin W.** (2003). "Visualizing knowledge domains". *Annual review of information science and technology*, v. 37, n. 1, pp. 179-255.
https://doi.org/10.1002/aris.1440370106

Buskirk, Trent D.; **Kirchner, Antje** (2020). "Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis". In: Hill, Craig A.; Biemer, Paul P.; Buskirk, Trent D.; Japec, Lilli; Kirchner, Antje; Kolenikov, Stas; Lyberg, Lars E. *Big data meets survey science: A collection of innovative methods*, pp. 9-62. ISBN: *978 111 897 635 7*
https://doi.org/10.1002/9781118976357.ch1

Callon, Michel; **Courtial, Jean-Pierre**; **Laville, Francoise** (1991). "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry". *Scientometrics*, v. 22, pp. 155-205.
https://doi.org/10.1007/BF02019280

Callon, Michel; **Courtial, Jean-Pierre**; **Turner, William A.**; **Bauin, Serge** (1983). "From translations to problematic networks: An introduction to co-word analysis". *Social science information*, v. 22, n. 2, pp. 191-235.
https://doi.org/10.1177/053901883022002003

Chen, Jie (2007). "Rapid urbanization in China: A real challenge to soil protection and food security". *Catena*, v. 69, n. 1.
https://doi.org/10.1016/j.catena.2006.04.019

Cobo, Manuel J.; **López-Herrera, Antonio G.**; **Herrera, Francisco**; **Herrera-Viedma, Enrique** (2011a). "A note on the ITS topic evolution in the period 2000-2009 at T-ITS". *IEEE transactions on intelligent transportation systems*, v. 13, n. 1, pp. 413-420.
https://doi.org/10.1109/TITS.2011.2167968

Cobo, Manuel J.; **López-Herrera, Antonio G.**; **Herrera-Viedma, Enrique**; **Herrera, Francisco** (2011b). "Science mapping software tools: Review, analysis, and cooperative study among tools". *Journal of the American Society for Information Science and Technology*, v. 62, n. 7, pp. 1382-1402.
https://doi.org/10.1002/asi.21525

Cobo, Manuel J.; **López-Herrera, Antonio G.**; **Herrera-Viedma, Enrique**; **Herrera, Francisco** (2011c). "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field". *Journal of informetrics*, v. 5, n. 1, pp. 146-166.
https://doi.org/10.1016/j.joi.2010.10.002

Cobo, Manuel J.; **López-Herrera, Antonio G.**; **Herrera-Viedma, Enrique**; **Herrera, Francisco** (2012). "*SciMAT*: A new science mapping analysis software tool". *Journal of the American Society for Information Science and Technology*, v. 63, n. 8, pp. 1609-1630.
https://doi.org/10.1002/asi.22688

Coulter, Neal; **Monarch, Ira**; **Konda, Suresh** (1998). "Software engineering as seen through its research literature: A study in co-word analysis". *Journal of the American Society for Information Science*, v. 49, n. 13, pp. 1206-1223.
https://doi.org/10.1002/(SICI)1097-4571(1998)49:13<1206::AID-ASI7>3.0.CO;2-F

España (2020). "Real decreto 1110/2020, de 15 de diciembre, por el que se aprueba el Plan Estadístico Nacional 2021-2024". *BOE*, n. 340, 30 diciembre .
https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-17283

EU (2019). "Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida)". *Diario oficial de la UE*, 26 junio.
https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32019L1024&from=ES

EU (2022). "Reglamento (UE) 2022/868 del Parlamento Europeo y del Consejo de 30 de mayo de 2022 relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE) 2018/1724 (Reglamento de Gobernanza de Datos)". *Diario oficial de la Unión Europea*, n. 152, 3 junio.
https://www.boe.es/buscar/doc.php?id=DOUE-L-2022-80835

Galán, José-Javier; **Carrasco, Ramón-Alberto**; **LaTorre, Antonio** (2022). "Military applications of machine learning: A bibliometric perspective". *Mathematics*, v. 10, n. 9, 1397.
https://doi.org/10.3390/math10091397

García-Villar, Jaume (2012). "La estadística oficial como bien público: retos del presente". *Revista asturiana de economía*, n. 46, pp. 61-85.
https://dialnet.unirioja.es/servlet/articulo?codigo=4204191

Georgiou, Andreas V. (2017). "Towards a global system of monitoring the implementation of UN fundamental principles in national official statistics". *Statistical journal of the IAOS*, v. 33. n. 2, pp. 387-397.
https://doi.org/10.3233/SJI-160335

Gillborn, David (2008). *Racism and education: Coincidence or conspiracy?*. Routledge. ISBN: 978 0 203928424
https://doi.org/10.4324/9780203928424

Giovannini, Enrico; **Martins, J. Oliveira**; **Gamba, Michela** (2009). "Statistics, knowledge and governance". *Statistika*, v. 6, pp. 471-490.

Giri, Chandra; **Ochieng, Edward**; **Tieszen, Larry L.**; **Zhu, Zq**; **Singh, Ashbindu**; **Loveland, Tomas**; **Masek, Jeff**; **Duke, Norman** (2011). "Status and distribution of mangrove forests of the world using earth observation satellite data". *Global ecology and biogeography*, v. 20, n. 1, pp. 154-159.
https://doi.org/10.1111/j.1466-8238.2010.00584.x

Grantham-McGregor, Sally; **Cheung, Yin-Bun**; **Cueto, Santiago, Glewwe, Paul**; **Richter, Linda**; **Strupp, Barbara** (2007). "Developmental potential in the first 5 years for children in developing countries". *The lancet*, v. 369, n. 9555, pp. 60-70.
https://doi.org/10.1016/S0140-6736(07)60032-4

Groves, Robert M. (2011). "Three eras of survey research". *Public opinion quarterly*, v. 75, n. 5, pp. 861-871.
https://doi.org/10.1093/poq/nfr057

Gustavsson, Anders; **Svensson, Mikael**; **Jacobi, Frank**; **Allgulander, Christer**; **Alonso, Jordi**; **Beghi, Ettore**; **Dodel, Richard**; **Ekman, Mattias**; **Faravelli, Carlo**; **Fratiglioni, Laura**; **Gannon, Brenda**; **Jones, David H.**; **Jennum, Poul**; **Jordanova, Albena**; **Jönsson, Linus**; **Karampampa, Korinna**; **Knapp, Martin**; **Kolbelt, Gisela**; **Kurth, Tobias**; **Lieb, Roselind**; **Linde, Mattias**; **Ljungcrantz, Christina**; **Maercker, Andreas**; **Melin, Beatrice**; **Moscarilli, Massimo**; **Musayev, Amir**; **Norwood, Fiona**; **Preisig, Martin**; **Pugliatti, Maura**; **Rehm, Juergen**; **Salvador-Carulla, Luis**; **Schlehofer, Brigitte**; **Simon, Roland**; **Steinhausen, Hans-Christoph**; **Stovner, Lars J.**; **Vallat, Jean-Michel**; **Van-den-Bergh, Peter**; **Van-Os, Jim**; **Vos, Pieter**; **Xu, Weili**; **Wittchen, Hans-Ulrich**; **Jönsson, Bengt**; **Olesen, Jes** (2011). "Cost of disorders of the brain in Europe 2010". *European neuropsychopharmacology*, v. 21, n. 10, pp. 718-779.

Harford, Tim (2014). "Big data: A big mistake?". *Significance*, v. 14, n. 5, pp. 14-19.
https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2014.00778.x

He, Qin (1999). "Knowledge discovery through co-word analysis". *Library trends*, v. 48, n. 1, p. 133-159.
https://www.ideals.illinois.edu/items/8226

Henderson, J. Vernon; **Storeygard, Adam**; **Weil, David N.** (2012). "Measuring economic growth from outer space". *American economic review*, v. 102, n. 2, pp. 994-1028.
https://doi.org/10.1257/aer.102.2.994

Hippisley-Cox, Julia; **Coupland, Carol**; **Vinogradova, Yana**; **Robson, John**; **Minhas, Rubin**; **Sheikh, Aziz**; **Brindle, Peter** (2008). "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2". *British medical journal*, v. 336, pp. 1475-1482.
https://doi.org/10.1136/bmj.39609.449676.25

Hirsch, Jorge E. (2005). "An index to quantify an individual's scientific research output". *Proceedings of the National Academy of Sciences*, v. 102, n. 46, pp. 16569-16572.
https://doi.org/10.1073/pnas.0507655102

Jacob, Ulrike; **Brust, Oliver A.** (2019). "Confronting the anomaly: directions in (German) economic research after the crisis". *Science in context*, v. 32, n. 4, pp. 449-471.
https://doi.org/10.1017/S026988972000006X

Liu, Zhu; **Guan, Dabo**; **Wei, Wei**; **Davis, Steven J**; **Ciais, Philippe**; **Bai, Jin**; **Peng, Shushi**; **Zhang, Qiang**; **Hubacek, Klaus**; **Marland, Gregg**; **Andres, Robert J.**; **Crawford-Brown, Douglas**; **Lin, Jintai**; **Zhao, Hongyan**; **Hong, Chaopeng**; **Boden, Thomas A.**; **Feng, Kuishuang**; **Peters, Glen P.**; **Xi, Fengming**; **Liu, Junguo**; **Li, Yuan**; **Zhao, Yu**; **Zeng, Ning**; **He, Kebin** (2015). "Reduced carbon emission estimates from fossil fuel combustion and cement production in China". *Nature*, v. 524, n. 7565, pp. 335-338.
https://doi.org/10.1038/nature14677

Llewelyn, Charlotte A.; **Hewitt, Patricia E.**; **Knight, Richard S. G.**; **Amar, K.**; **Cousens, S.**; **Mackenzie, J.**; **Will, Robert G.** (2004). "Possible transmission of variant Creutzfeldt-Jakob disease by blood transfusion". *The lancet*, v. 363. n. 9407, pp. 417-421.
https://doi.org/10.1016/S0140-6736(04)15486-X

Lnenicka, Martin; **Luterek, Mariusz**; **Nikiforova, Anastasija** (2022). "Benchmarking open data efforts through indices and rankings: Assessing development and contexts of use". *Telematics and informatics*, v. 66, 101745.
https://doi.org/10.1016/j.tele.2021.101745

Martínez, Miguel-Ángel; **Herrera, Manuel**; **López-Gijón, Javier**; **Herrera-Viedma, Enrique** (2004). "H-Classics: Characterizing the concept of citation classics through". *Scientometrics*, v. 98, pp. 1971-1983.
https://doi.org/10.1007/s11192-013-1155-9

Moreno, Caio; **Carrasco, Ramón-Alberto**; **Herrera-Viedma, Enrique** (2019). "Data and artificial intelligence strategy: A conceptual enterprise big data arquitecture to enable market-oriented organizations". *IJIMAI*, v. 5, n. 6, pp. 7-14.
https://doi.org/10.9781/ijimai.2019.06.003

Mukherjee, Debmalya; **Lim, Weng-Marc**; **Kumar, Satish**; **Donthu, Naveen** (2022). "Guidelines for advancing theory and practice through bibliometric research". *Journal of business research*, v. 148, pp. 101-115.
https://doi.org/10.1016/j.jbusres.2022.04.042

Multicentre Aneurysm Screening Study Group (2002). "The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial". *The lancet*, v. 360, n. 9345, pp. 1531-1539.
https://doi.org/10.1016/S0140-6736(02)11522-4

Olesen, Jes; **Gustavsson, Anders**; **Svensson, Mikael**; **Wittchen, H.-U.**; **Jönsson, Bengt** (2012). "The economic cost of brain disorders in Europe". *European journal of neurology*, v. 19, n. 1, pp. 155-162.
https://doi.org/10.1111/j.1468-1331.2011.03590.x

Peto, Richard; **Darby, Sarah**; **Deo, Harz**; **Silcocks, Paul**; **Whitley, Elise**; **Doll, Richard** (2000). "Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies". *British medical journal*, v. 321, n. 7257, pp. 323-329.
https://doi.org/10.1136/bmj.321.7257.323

Purkayastha, Amrita; **Palmaro, Eleonora**; **Falk-Krzesinski, Holly J.**; **Baas, Jeroen** (2019). "Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR)". *Journal of informetrics*, v. 13, n. 2, pp. 635-642.
https://doi.org/10.1016/j.joi.2019.03.012

Radermacher, Walter J. (2014). "The European statistics code of practice as a pillar to strengthen public trust and enhance quality in official statistics". *Statistical and Social Inquiry Society of Ireland*, v. 43, pp. 27-33.
http://www.tara.tcd.ie/handle/2262/72773

Rodríguez-Ledesma, Antonio; **Cobo, Manuel J.**; **López-Pujalte, Cristina**; **Herrera-Viedma, Enrique** (2015). "An overview of animal science research 1945-2011 through science mapping analysis". *Journal of animal breeding and genetics*, v. 132, n. 6, pp. 475-497.
https://doi.org/10.1111/jbg.12124

Rodríguez-López, María-Eugenia; **Alcántara-Pilar, Juan-Miguel**; **Del-Barrio-García, Salvador**; **Muñoz-Leiva, Francisco** (2020). "A review of restaurant research in the last two decades: A bibliometric analysis". *International journal of hospitality management*, v. 87, 102387.
https://doi.org/10.1016/j.ijhm.2019.102387

Savage, Mike; **Devine, Fiona**; **Cunningham, Niall**; **Taylor, Mark**; **Li, Yaojun**; **Hjellbrekke, Johs**; **Le-Roux, Brigitte**; **Friedman, Sam**; **Miles, Andrew** (2013). "A new model of social class? Findings from the BBC's great British class survey experiment". *Sociology*, v. 47, n. 2, pp. 219-250.
https://doi.org/10.1177/0038038513481128

Small, Henry (1999). "Visualizing science by citation mapping". *Journal of the American Society for Information Science*, v. 50, n. 9, p. 799-813.
https://doi.org/10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G

Sternitzke, Christian; **Bergmann, Isumo** (2009). "Similarity measures for document mapping: a comparative study on the level of an individual scientist". *Scientometrics*, v. 78, p. 113-130.
https://doi.org/10.1007/s11192-007-1961-z

Unal, Belgin; **Critchley, Julia-Alison**; **Capewell, Simon** (2004). "Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000". *Circulation*, v. 109, n. 9, pp. 1101-1107.
https://doi.org/10.1161/01.CIR.0000118498.35499.B2

United Nations (2014). "Statistical Commission Decision 45/110 on big data and modernization of statistical systems". *Statistics Division*, Decision code: 45/110, chapter: I, section: B.
https://unstats.un.org/unsd/statcom/decisions-ref/?code=45/110

Unece (2017). *Guidance on land-use planning, the siting of hazardous activities and related safety aspects*. Geneva, Switzerland: United Nations Economic Commission for Europe.
https://unece.org/DAM/env/eia/Publications/2017/1735403E_Final_ENG_web.pdf

Unece (2019). *Use of satellite image and Earth observation data in official statistics*. Geneva, Switzerland: United Nations Economic Commission for Europe.
https://unece.org/DAM/stats/documents/ece/ces/2019/ECE_CES_2019_16-1906490E.pdf

Unece (2020). *Guidelines on the shared environmental information system reporting mechanism*. Geneva, Switzerland: United Nations Economic Commission for Europe.
https://unece.org/sites/default/files/2020-12/2014795E_WEB.pdf

Unece (2021). *Measuring and monitoring progress towards the Sustainable Development Goals*. Geneva, Switzerland: United Nations Economic Commission for Europe.
https://unece.org/sites/default/files/2021-04/2012761_E_web.pdf

United Nations (2021). *The handbook on management and organization of National Statistical Systems*.
https://unstats.un.org/capacity-development/handbook/html/topic.htm#t=Handbook%2FCover%2FCover.htm

Van-Dijk, Jan; **Hacker, Kenneth** (2003). "The digital divide as a complex and dynamic phenomenon". *The information society*, v. 19, n. 4, pp. 315-326.
https://doi.org/10.1080/01972240309487

Vikat, Andres; **Jones, Christopher** (2014). *Indicators of gender equality*. Unece: Geneva, Switzerland.
https://unece.org/sites/default/files/2022-02/ECE_CES_37_WEB.pdf