Environmental Modelling and Software 107 (2018) 199–209

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Time series analysis with explanatory variables: A systematic literature review

Paula Medina Maçaira, Antônio Marcio Tavares Thomé*, Fernando Luiz Cyrino Oliveira, Ana Luiza Carvalho Ferrer

Pontifícia Universidade Católica do Rio de Janeiro, Industrial Engineering Department, Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, RJ, 22451-900, Brazil

ARTICLE INFO

Keywords:
Regression analysis
Artificial intelligence
Exogenous variables
Forecast scenarios

ABSTRACT

Time series analysis with explanatory variables encompasses methods to model and predict correlated data taking into account additional information, known as exogenous variables. A thorough search in literature returned a dearth of systematic literature reviews (SLR) on time series models with explanatory variables. The main objective is to fill this gap by applying a rigorous and reproducible SLR and a bibliometric analysis to study the evolution of this area over time. The study resulted in the identification of the main methods of time series that incorporate input variables per knowledge area and methodology. The largest number of papers belongs to environmental sciences, followed by economics and health. Regression model is the method with the highest number of applications, followed by Artificial Neural Networks and Support Vector Machines, which experienced rapid and recent growth. A research agenda in time series analysis with exogenous variables closes the paper.

1. Introduction

Time series modelling involves the analysis of a dynamic system characterised by inputs and outputs series, which relates to a function. Regardless of their ultimate purpose, the various techniques in this field have the mutual goal of reproducing the output series with reliability and accuracy from the estimation of the function and input series. Time series techniques can essentially be divided into two sets of methods: univariate and multivariate. In the case of univariate approaches, the output series is explained by a constant portion and/or trend, seasonality, and in many cases, by the series lagged in time. Multivariate methods, on the other hand, use the influence of other variables on the behaviour of the output series to obtain better results in the representation of the transfer function.

One of the main areas of application of such methods is in the environmental science. Espey et al. (1997) estimated an econometric model to evaluate the price elasticity of residential demand of water using rate structure, location, season and others exploratory variables. Nunnari et al. (2004) compared several statistical techniques for modelling SO_2 concentration using the information of wind direction, wind speed, solar radiation, temperature and relative humidity. Andriyas and McKee (2013) used biophysical conditions in farmers' fields and the irrigation delivery system during the growing season to anticipate irrigation water orders. Lima et al. (2014) developed a forecasting model for the water inflow incorporating the effect of climate variables like precipitation and El Niño.

There are studies showing the advantages and disadvantages of using both univariate and multivariate approaches. The work of Athanasopoulos et al. (2011) performed a competition between univariate and multivariate methods for predicting the international demand of tourists. Sfetsos and Coonick (2000) compared the approaches on predicting solar radiation, and Porporato and Ridolfi (2001) forecasted river flows. There are equally reviews of methods and techniques of both univariate and multivariate time series applied to specific areas. Milionis and Davies (1994), for example, reviewed regression methods and stochastic models applied to the analysis of air pollution. Ljung (1999) describes the theory, methodology, practice of ARMAX models, and nonlinear black box models, among other. Durbin and Koopman (2012) published a clear and comprehensive introduction to the state space approach to time series analysis together with a historical background, and Haykin (1999) presents an extensive state-of-the-art review of neural networks. Young (2011) offers an introduction to recursive estimation and demonstrates its various forms and its use as an aid in the modelling of stochastic, dynamic systems in time series. However, there is a paucity of systematic literature reviews of methods of time series analysis using exogenous variables in the modelling structure,

* Corresponding author.
E-mail addresses: paulamacaira@aluno.puc-rio.br (P.M. Maçaira), mt@puc-rio.br (A.M. Tavares Thomé), cyrino@puc-rio.br (F.L. Cyrino Oliveira), analcferrer@gmail.com (A.L. Carvalho Ferrer).

https://doi.org/10.1016/j.envsoft.2018.06.004
Received 16 May 2017; Received in revised form 30 April 2018; Accepted 1 June 2018
Available online 22 June 2018
1364-8152/

regardless of the application area or methodology.

Given the relevance of time series analysis with exogenous variables, the main objective is to fill the gap identified in the existing literature on models that use explanatory variables, providing a "science map" through a systematic literature review (SLR) and bibliometric analysis. According to Small (1999, p. 799) "a map of science is a spatial representation of how disciplines, fields, specialities, and individual documents or authors are related to one another" and "can provide insight into a contemporaneous state of knowledge." Thus, this study seeks to guide researchers and practitioners from various knowledge areas to identify what are the main areas of application of time series analysis with exogenous variables, who are the most prolific authors by knowledge area, what are the most influential papers, and what are the main methods used.

This paper comprises this introduction, followed, in section 2, by a description of the methods applied to the literature review and bibliometric analysis. Section 3 presents the results from citation, co-citation, and co-word analyses; and section 4 encompasses discussions and conclusions with implications for practice and future research.

2. Methodology

This section presents the methods and basic statistics extracted from the SLR, as well as the bibliometric methods applied to the longitudinal analysis of thematic areas.

2.1. Systematic review and basic statistics

Thomé et al. (2016a) developed a step by step approach to conduct an SLR in operations management consisting of eight steps: (i) planning and formulating the problem, (ii) searching the literature, (iii) data gathering, (iv) quality evaluation, (v) data analysis and synthesis, (vi) interpretation, (vii) presenting the results, and (viii) updating the review.

In Step 1, planning and formulating the problem, the research team comprised of the co-authors of this paper assembled and defined the scope of the review, the conceptualisation of the main research topic and the following research questions (RQ):

RQ 1. Who are the most prolific authors in time series analysis with explanatory variables?

RQ 2. Which are the most influential works in time series analysis with explanatory variables?

RQ 3. Which are the main themes in time series analysis with explanatory variables and how did they evolve?

RQ 4. Which are the main methods applied in time series analysis with explanatory variables and how did they evolve?

Following the second SLR step, a literature search was carried out in seven stages comprising the selection of databases, definition of keywords, review of abstracts, criteria for inclusion/exclusion of papers, full-text review, and backward and forward search in selected papers/references. The Scopus database was chosen because it is one of the largest abstract and citation databases of research literature containing over 53 million records and almost 22,000 titles from 5000 publishers (HLWIKI, 2017). The choice of Scopus is justified on the grounds of the large coverage of research domains intended by the present SLR (see for example Mongeon and Paul-Hus, 2016). There was no time restriction for the search. Table 1 shows the number of papers included in each phase of the keyword search.

Keywords should be broad to do not artificially restrict the number of studies and specific enough to bring only the studies related to the topic (Cooper, 2010). The first keywords were "time series" and "exogenous* variable", generating 244 papers. The search was later expanded with synonyms of "exogenous" in time series analysis, leading to 2020 papers. This was intended to comply with the two criteria for searching studies in SLR suggested by Petticrew and Roberts (2006, p. 81), sensitivity to retrieve everything of relevance, and specificity to leave behind the irrelevant. The papers' exclusion criteria are the language of the paper and document type, resulting in 1930 papers written in English. The limitation of document type to article, review, and articles in press narrowed the selection to a final set of 1547 documents included in the bibliometric analysis. The full bibliographic reference is available upon request to the lead author.

Table 1
Number of papers by keyword search.

Keyword search	No. of papers
"time series" AND"exogenous* variable"	244
"time series" AND ("exogenous* variable" OR "explanat* variable")	830
"time series" AND ("exogenous* variable" OR "explanat* variable" OR "input variables")	1323
"time series" AND ("exogenous* variable" OR "explanat* variable" OR "input variables" OR "predict* variable")	1575
"time series" AND ("exogenous* variable" OR "explanat* variable" OR "input variables" OR "predict* variable" OR "explicative variable")	1579
"time series" AND ("exogenous* variable" OR "explanat* variable" OR "input variables" OR "predict* variable" OR "explicative variable" OR "dependent variable")	2020

Table 2 presents the top 10 areas with the greatest concentration of papers from 28 knowledge areas identified. As expected, there is a large array of subject areas, ranging from computer science to medicine. Another consistent result is the concentration of 20% of the papers in the environmental area (Environmental Science & Earth and Planetary Sciences) given the complexity of natural phenomena series and the need to use models that incorporate outside information.

Environmental Science, Mathematics, and Medicine are responsible for approximately 31% of the total number of papers. Together with Social Sciences and Economics, Econometrics and Finance, the first five areas correspond to approximately 50% of the total.

Fig. 1 illustrates the number of papers published by year. The first paper appeared in 1967 (Scott Jr. and Heady, 1967), about the demand for new investment in farm buildings. In the next three years, there were no publications. It is worth noting that until 1995 the total number of published articles per year was consistently lower than 20. Publications peaked at 27 in 1996. Five years later, in 2001, the number of publications reached the mark of over 40 published articles per year. There has been a fast-growing trend since then, reaching a record of 150 publications in 2015, and 124 publications in 2016.

When analysing the distribution of papers by journals, a large variety of 151 different sources emerges, as expected in such a multidisciplinary field. The top 20 journals in the number of citations are in Table 3. Together, they account for 73% of the total number of citations. The highest ranked journal in the number of citations relates to the economy, where time series models are extensively used. The Journal of Econometrics is also the journal with the highest number of

Table 2
Top ten subject areas.

Scopus categories	No. of papers	Percentage of contribution
Environmental Science	314	12%
Mathematics	266	10%
Medicine	250	9%
Social Sciences	247	9%
Economics, Econometrics and Finance	242	9%
Earth and Planetary Sciences	203	8%
Agricultural and Biological Sciences	202	8%
Engineering	196	7%
Computer Science	190	7%
Business, Management and Accounting	127	5%

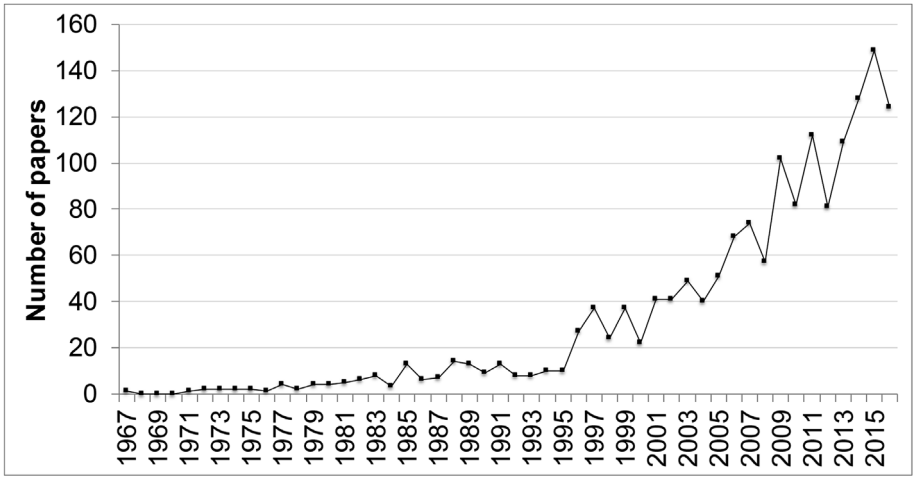

Fig. 1. Number of papers published per year.

Table 3
Numbers of papers and citations per source.

Source	No. of papers	No. of citations
Journal of Econometrics	24	1923
American Journal of Political Science	4	1186
Journal of the American Statistical Association	23	1066
Water Resources Research	20	812
IEEE Transactions on Power Systems	9	701
Journal of Climate	8	625
Neurocomputing	16	548
International Journal of Forecasting	19	521
Remote Sensing of Environment	11	489
Statistics in Medicine	6	455
Annual Review of Political Science	2	424
Journal of Hydrology	22	376
International Journal of Climatology	12	374
Ecological Modelling	6	351
Expert Systems with Applications	11	342
American Political Science Review	2	309
European Journal of Political Research	5	282
International Organization	3	247
Atmospheric Environment	11	231
Fisheries Research	11	217
Water Research	3	195
Information Sciences	3	189
ICES Journal of Marine Science	7	187
Academic Emergency Medicine	3	186
Ecological Economics	2	184
Accident Analysis and Prevention	9	156
Water Resources Management	6	155
Tourism Management	5	153
Environmental Modelling and Software	8	152
Neural Computing and Applications	6	152

publications in the theme. It is equally worth mentioning the large presence of journals related to the environment (Environmental Modelling and Software, Water Resources Research, Journal of Climate, Remote Sensing of Environment, Journal of Hydrology, International Journal of Climatology, Ecological Modelling, Atmospheric Environment, Water Research, Ecological Economics, and Water Resources Management). Together, they represent approximately 30% of the total number of citations and 39% of papers, respectively. However, Spearman's rank correlation between the number of publications and the total number of citations equals 0.57, an average value showing that some of the most prolific journals are not necessarily the most influential ones.

The third and fourth step of the SLR is data gathering and quality evaluation. A computer template was used for data gathering and the calculation of simple frequencies and means. The restriction of selected papers to peer-reviewed journals provided an initial gauge of the quality of the papers retrieved for analysis. The fifth SLR step is data analysis and synthesis. An inductive approach to qualitative content analysis was adopted (Seuring and Gold, 2012; Thomé et al., 2016a) and combined with quantitative bibliometric analysis of co-citations and co-occurrence of keywords, described in subsection 2.2. Subsequently, the sixth step, interpretation, refers to a qualitative research synthesis coupled with bibliometric indicators from the co-citation and co-word analyses. This study consists of the seventh step of the SLR approach, presentation of results. The step eight, updating of the review, lies past the purview of this study.

2.2. Bibliometric analysis

Bibliometric measures and indicators can be employed to carry out performance analyses and generate scientific maps. These procedures quantify and measure the performance and impact of scientific research (Cobo et al., 2011). Three specific analyses are carried out: citation, co-citation, and co-word analyses.

Citation analysis is the first and simplest analysis performed in the database. This analysis involves the study of the most influential authors, measuring the number of papers and citations by knowledge areas, institution and country. Co-occurrence is the study of mutual appearances of pairs of units over a number of bibliographic records. There are different types of bibliometric analysis of co-occurrence: co-citation, co-author, and co-word analyses being the most common ones (Persson et al., 2009). For instance, if document B and C cite document A, they both appear as co-citing A. This method can be used to describe the backbone of a research area, mapping the network of co-citation among the most influential authors. It allows the identification of the main research streams.

2.2.1. Software and data availability

There is a large number of software available for bibliometric analysis in literature reviews, as evidenced by Cobo et al. (2012). For the citation, co-citation, and co-word analyses, software BibExcel (Persson et al., 2009) (http://homepage.univie.ac.at/juan.gorraiz/bibexcel/), Pajek (De Nooy et al., 2005) (http://mrvar.fdv.uni-lj.si/pajek), and SciMAT (Cobo et al., 2012) (http://sci2s.ugr.es/scimat/) were used. All software is available free-ware for non-profit academic use. BibExcel was used to prepare the matrix of co-citation and generate the clusters, Pajek to prepare the network analysis, and SciMAT to analyse dynamic maps of longitudinal co-word analysis of themes using Callon's thematic bipartite graphs or strategic diagrams (Cobo et al., 2012), depicting the density and centrality of co-occurrences of author's keywords. The full dataset containing 1547 documents included in the bibliometric

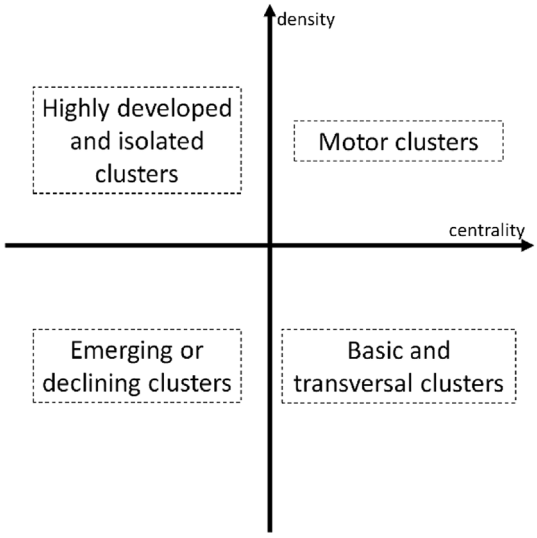

Fig. 2. Callon's strategic diagram.
Source: Adapted from Cobo et al. (2012).

analysis is made available at Mendeley repository (Thomé et al., 2017).

2.2.2. Callon's strategic diagram: bibliometric indexes

Fig. 2 shows Callon's strategic diagram (Callon et al., 1991).

Following Callon's strategic diagram, in Fig. 2, the motor clusters (top right) represent the core themes of the research area, with high centrality and high density. The basic and transversal clusters (bottom right) are central themes that are also key to the research field but are not well developed as they combine high centrality with low density. Emerging or declining themes (bottom left) present low centrality and low density; meaning, they do not relate well to other themes and are not well represented in the research area. Highly developed and isolated clusters (top left) are usually well-researched areas with high density, commonly denoting classic themes in the research field.

The similarity index measures co-occurrence. It considers the number of documents in which two keywords occur and the number of documents in which each one occurs. Clusters are formed in SciMAT with the application of the "simple centre algorithm" (Cobo et al., 2011; Thomé et al., 2016b). In the "simple centre algorithm", the centrality and density are calculated as a linear function of the similarity index. The similarity index is calculated as $e_{ij} = {C_{ij}^2}/{C_i C_j}$, with C_{ij} being the number of documents in which two keywords i and j co-occur, C_i and C_j being the number of documents in which each one occurs. In one hand, the centrality is a measure of the interaction among networks of keywords calculated as $C = 10 \sum e_{Kh}$, where k is a keyword belonging to the theme and h is a keyword belonging to the other theme. On the other hand, the density measures the internal strength of the network or theme calculated as $d = 100\left(\frac{\sum e_{ij}}{w}\right)$, where i and j are keywords belonging to the theme and w is the total number of keywords in the theme.

3. Results from citation, co-citation, and co-word analyses

3.1. Citation analysis

In an attempt to answer the first research question, "Who are the most prolific authors in time series analysis with explanatory variable?", the citation analysis scrutinises the number of citations and number of papers by author. Table 4 depicts the ranking of authors by number of citations, number of papers, institution, country, and subject area. In Table 4, the number of papers per author and the number of citations per author is the number of articles and the sum of all citations from a given author appearing in the database, respectively.

The most prolific author (Beck, N.) is also the author with the highest number of citations; however, the number of documents and number of citations might not correlate. In fact, Spearman's rank correlation among the number of papers and citations is not significant($p = 0.13$). Another look at Table 4 reveals that the main areas for these authors are Social Science and Economics and they are located mainly in the US and UK.

Beck's most cited work found in the database (Beck et al., 1998) is co-authored with Katz and Tucker. It presents a solution to analyse time-series cross-section data using logit analyses and applies the proposed methodology to estimate the relationship between economic interdependence, democracy, and peace. It is worth mentioning that Beck, Katz, and Tucker are frequent co-authors in studies with time series cross-section data. Meese and Rogoff (1983) have the second most cited paper in the database; their paper compares the out-of-sample forecasting accuracy of various structural and time series exchange rate models. The work of Anderson and Hsiao (1982) comes next in number of citations; it presents a statistical analysis of time series regression models for longitudinal data with and without lagged dependent variables. Wilby, Wigley and Conway co-authored the work that has more citations in the database for each author (Wilby et al., 1998); it calibrates a range of different statistical downscaling models using both observed and general circulation models to generate and compare daily precipitation time series. As expected, most frequently cited authors published earlier, since these authors had more time to obtain citations than authors that are more recent did.

3.2. Co-citation analysis

The co-citation analysis addresses the second research question: "Which are the most influential works in time series analysis with explanatory variables?" The co-citation presents the most influential works in time series analysis with exogenous variables, that is, the backbone of the studied area.

The network in Fig. 3 illustrates the co-citation relationships among the top 30 documents with the largest number of citations, that is, the most influential works in time series analysis with explanatory variables. Documents appear in Fig. 3 chronologically from top to bottom. The sizes of the circumference of the circles are proportional to the number of citations, and the width of the lines connecting the network is proportional to the number of co-occurrences.

The following describes the major contribution of each node of the co-citation tree. Granger (1969) defined the difference between causality and feedback by using the simple example of bivariate models. He introduced the use of causal lag and causal strength to analyse the direction of causality between two related variables. Nash and Sutcliffe (1970) discussed the river flows from rainfall, evaporation, and other factors by observing the R^2 coefficient from a linear regression.

The popular work of Box and Jenkins (1970) introduced the autoregressive integrated moving average (ARIMA) time series models to characterise and predict time series observations at equally spaced periods. This classic model was revised by the same authors in 1976. There are three different methods derived from the ARIMA model: autoregressive (AR), moving average (MA), and AR moving average (ARMA) models. AR models are applied to a time series that can be represented by its own past values, while MA models are used to represent a time series where the past errors (disturbances) determine its future values. ARMA models are appropriate when a series is a function of unobserved errors as well as its own past behaviour. The difference between ARIMA and ARMA is that the first one is applied to series that are not stationary over time, that is, with some trend, and thus need to be differentiated (the "I" term) to become stationary.

The Akaike information criterion (AIC) was originally developed under the name "an information criterion" by Akaike (1973) and received its full denomination of AIC in 1974, along with a formal

Table 4
Number of citations and papers published by authors, institution, country, and subject area.

Authors	No. of citations	No. of papers	Institution	Country	Subject area
Beck N.	1668	5	New York University	USA	Social Sciences
Katz J.	1300	3	California Institute of Technology	USA	Social Sciences
Meese R.	1267	1	UC Berkeley Haas School of Business	USA	Economics
Rogoff K.	1267	1	Harvard University	USA	Economics
Tucker R.	1064	1	Vanderbilt University	USA	Social Sciences
Hsiao C.	669	2	University of Southern California	USA	Economics
Anderson T.	656	1	Stanford University	USA	Mathematics
Wigley T.	583	3	National Center for Atmospheric Research	USA	Earth and Planetary Sciences
Wilby R.	566	2	Loughborough University	UK	Environmental Science
Conway D.	428	2	London School of Economics and Political Science	UK	Environmental Science

definition of concepts. The AIC assists in model selection to estimate the relative quality of a model based on information loss and parsimony.

Schwarz (1978) tried to solve the problem of selecting one from a number of models of different dimensions, and Ljung and Box (1978) presented a modification to the lack of fit test proposed by Box and Pierce (1970).

The Dickey-Fuller test derives representations for the limiting distributions of the AR coefficient estimator and tests the null hypothesis of whether a unit root is present (Dickey and Fuller, 1979). Later, Philips and Perron (1988) developed a unit root test called Phillips–Perron test that makes a nonparametric correction to the *t*-test statistic present in the Dickey-Fuller test.

Heteroscedasticity in the disturbances of a properly specified linear model leads to inefficient parameter estimates, which results in faulty inferences when testing statistical hypotheses. White (1980) developed the so-called White test to verify if there is heteroscedasticity in the disturbances of a linear model.

Still dealing with heteroscedasticity, Engle (1982) introduced a new class of stochastic process called AR conditional heteroscedasticity (ARCH) that does not assume constant variance for the error term. This type of process assumes that the error variance follows an AR model. In 1986, Bollerslev (1986) extended the idea of Engle (1982) by introducing a more general class of process called generalised AR conditional heteroscedasticity (GARCH), which allows a much more flexible lag structure and admits an ARMA model to the error structure.

Entering in the artificial intelligence (AI) area, the first main reference appears in 1985 with the study of Takagi and Sugeno (1985), which suggested a mathematical tool to describe a system that can represent highly nonlinear relations fuzzily. The main contribution of the study is dealing with fuzzy system representation as a linear system. Equally, in the area of AI, Rumelhart et al. (1986) presented a procedure called error propagation, whereby the gradient can be determined by individual units of the network based only on locally available information.

Working with the definition of co-integration, Engle and Granger (1987) suggested a test for estimating co-integration relations using regression and combining the problems of unit root tests and test with parameters unidentified under the null. In Johansen's study (1988), instead of working with regression estimates to test co-integration, the author derived a maximum likelihood estimator that takes into account the error structure of the underlying process, which is not possible with the regression estimates. The method proposed earlier was modified by adding linear restrictions on the co-integration vectors and AR weights in the method of Johansen and Juselius (1990). Johansen (1991) presented maximum likelihood estimators and likelihood ratio tests for co-integrations in Gaussian vector AR models, which allow for the introduction of the constant term and seasonal dummies in the model.

In 1989, Harvey (1989) published a book that provided a unified and comprehensive theory of structural time series models and included a detailed treatment of the Kalman filter. Various applications illustrated the properties of the models and methodological techniques.

Returning to references that deal with AI methods, Hornik et al. (1989) took the first step in a rigorous general investigation on the capabilities and properties of multilayer feedforward networks by establishing that this particular type is a class of universal approximation.

The book written by McCullagh and Nelder (1989) provided a

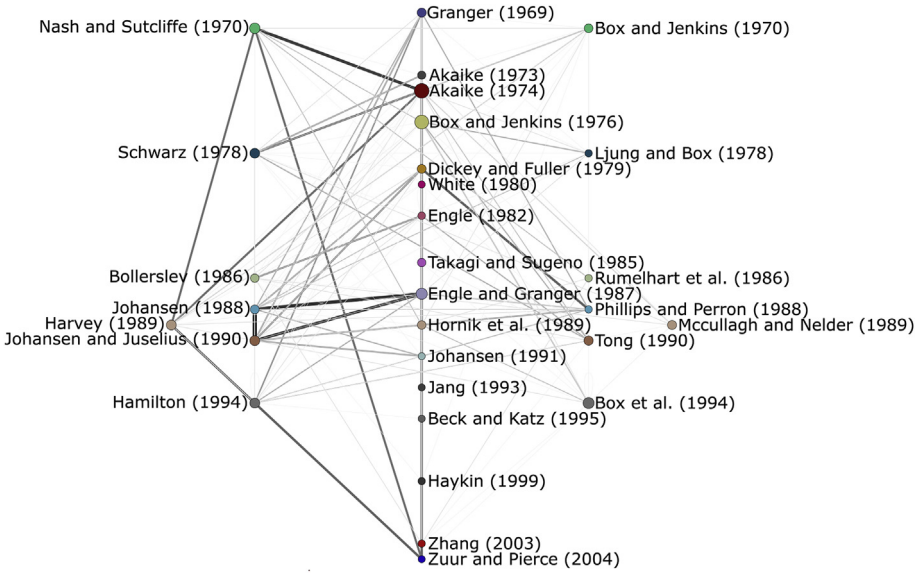

Fig. 3. Co-citation network of time series analysis with exogenous variables.

unified treatment of methods for the analysis of diverse types of data. The work focused on examining the way a response variable depends on the combination of explanatory variables.

In his book, *Non-Linear Time Series: A Dynamical System Approach*, Tong (1990) introduced the nonlinear time series theory and provided state-of-the-art research at that time. His book bridged the gap between linear and chaotic time series analysis, becoming one of the main references in nonlinear time series.

The fuzzy modelling and identification, first explored by Takagi and Sugeno (1985), is the main inspiration for the architecture called adaptive-network-based fuzzy inference system (ANFIS) developed by Jang (1993). By using a hybrid learning procedure, the proposed ANFIS can construct an input-output mapping based on both health studies knowledge and stipulated input-output data pairs.

In 1994, two books that covered the tools for modelling and analysing time series and introduced the latest developments that have occurred in the field over the past decade were published. Box et al. (1994) published a new edition of the classical *Time Series Analysis: Forecasting and Control* from Box and Jenkins (1970), while Hamilton (1994) explored the first edition of a book that synthesises the advances in economic and financial time series.

The work of Beck and Katz (1995) is the first reference that appears in the co-citation network that deals with the transversal study, a technique widely used in medical research and social sciences. Their work examined some issues in the estimation of time series cross-sectional models and proposed a new method.

A book presenting a detailed analysis of artificial neural networks (ANNs) did not appear until 1999 when Haykin (1999) presents an extensive state-of-the-art ANN main methods and background.

With the growth of computer power processing and consequently the use of AI methods in time series, hybrid models, such as the one from Zhang (2003), which combine the classical time series models, ARIMA, AI, and ANNs to provide better error measures in time series forecasting appeared. This methodology, in particular, takes advantage of the unique strength of ARIMA and ANN models in linear and nonlinear modelling.

The most recent reference in the co-citation network is from Zuur and Pierce (2004), which used a technique called dynamic factor analysis, presented in Zuur et al. (2003), to estimate common trends in time series of squid catch per unit of effort instead of applying ARIMA models. The method also allows the determination of the relationship between independent and exploratory variables such as sea surface temperature and the North Atlantic Oscillation index influence in the time series of squid catch.

With the co-citation analysis, it is possible to conclude that the works that primarily influence the researches in the studied area are also the most important references in time series methods and provided the main statistical tests. The co-citation tree from Fig. 3 offers a chronological guide to the study of time series methods with exogenous variables.

3.3. Co-word analysis

The co-word analysis intends to answer the last two research questions: “Which are the main themes in time series analysis with explanatory variables and how did they evolve?” and “Which are the main methods applied in time series analysis with explanatory variables and how did they evolve?” In order to allow a better understanding of the evolution of themes and methods and to portray research that is more recent in the field, documents were subdivided into four successive periods: 1967–1998, 1999–2007, 2008–2012, and 2013–2016, with 246, 423, 434, and 444 documents, respectively. The choice of the duration of the periods was made from the equalisation of the numbers of articles from the most recent to the oldest, providing the final three periods with an average of around 400 articles and the last with about 250. The main themes in the research area by period are in Table 5, with the number of core and secondary documents retrieved for analysis.

Table 5
Number of papers and citations by period, thematic cluster and type of document.

Period	Thematic cluster	Number of documents			
		Core	Secondary	Total	Core with up to 80% citations
1967–1998	Forecasting	16	33	49	6
	Health studies	10	17	27	3
	Developed countries	2	5	7	1
	Kalman filter	2	5	7	1
	Subtotal	**30**	**60**	**90**	**11**
1999–2007	Health studies	57	61	118	27
	Mathematical models	47	83	130	18
	Air pollutant	15	55	70	8
	Subtotal	**119**	**199**	**318**	**53**
2008–2012	Health studies	42	54	96	18
	Forecasting	63	64	127	24
	Algorithms	7	32	39	4
	Statistical models	6	38	44	3
	Subtotal	**118**	**188**	**306**	**49**
2013–2016	Health studies	32	46	78	15
	Forecasting	55	87	142	17
	Decision trees	4	6	10	2
	Costs	5	32	37	1
	Algorithms	4	36	40	2
	Subtotal	**100**	**207**	**307**	**37**
Total		**367**	**654**	**1021**	**150**

Core documents present at least two co-occurrences of keywords, while secondary documents present a single co-occurrence of keywords in the thematic cluster. Full-text reading of the core documents corresponding to up to 80% of total citations within the cluster provided the basis for the analysis of each thematic cluster. For example, in the 1967–1998 period, the six out of 16 core documents regrouping up to 80% of total citations for the forecasting cluster (first line of Table 5) were content analysed for the identification of themes and methods for the forecasting cluster in this period. Fig. 4 shows the number of keywords per period and presents their evolution by showing the number of outgoing and incoming keywords and the number and percentage of keywords that remain from one period to the next.

As expected, the number of keywords grows throughout the periods, paralleling the increase in the number of documents over the years. The number of keywords increases from 130 in the first period (1967–1998) to 377 in the fourth period (2013–2016), a 290% growth. Of the 130 words that appear in the first period, 92 (71%) remain for the second period, and 194 are added, totalling 286 words. For the third period, 218 (76%) remain and 171 new words are included, accounting for 389 in total. For the fourth period, 241 (62%) keywords from the third

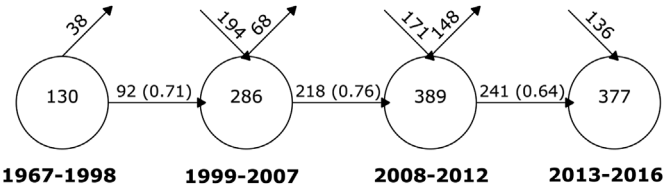

Fig. 4. Evolution of keywords by periods.

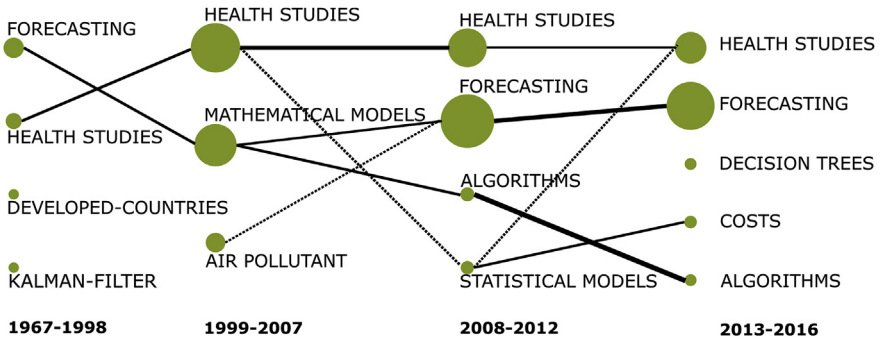

Fig. 5. Evolution of thematic clusters.

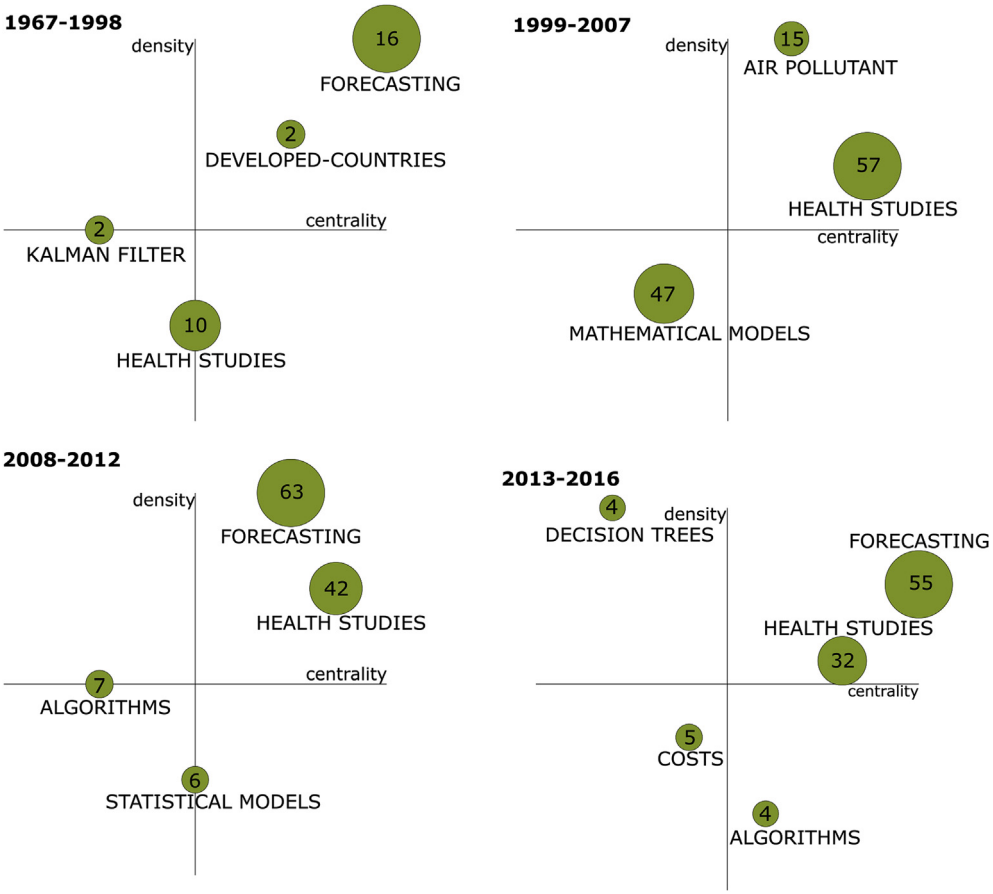

Fig. 6. Strategic diagrams.

period remain, and 136 new words appear, resulting in 377 keywords. This can be indicative that the areas that use explanatory variables in time series are diversifying and increasing over time and hence, this is not yet a consolidated field.

The most frequently used keyword in all the periods is human, followed by forecast, regression-analysis, neural-networks and algorithms. The words mathematical-models, demography and developing countries appear with high frequency in the first period. Air-pollution, ozone, temperature, climate-change and environmental-monitoring are frequently used words from the second to the last period. In the last period, the word risk-assessment and economic-growth are used more prominently than in the previous periods.

Fig. 5 shows the evolution of themes per period, based on the strength of the association among themes from one period to the next. The inclusion index measures the strength of the association, represented by the thickness of the lines connecting each cluster. The inclusion index varies from zero to one, where zero means that the thematic areas are not connected, and is measured as $\frac{\#U \cap V}{\min(\#U, \#V)}$, where U and V are disjoint sets (themes in different periods). The continuous lines represent a name association between thematic clusters, i.e., either two clusters have the same name in consecutive periods or one thematic cluster contains the other one, and the dotted lines represent associations on aspects other than the name. The sphere sizes correspond to the number of core documents present in the cluster.

From Table 5 and Fig. 5, in 1967–1998, forecasting represents 53% of the total number of core documents in the period, while in 1999–2007 health studies and mathematical models are the main themes responsible for 48% and 39% of core documents, respectively. In the following two periods, 2008–2012 and 2013–2016, the main theme is again forecasting, accounting for 53% and 55% of the core documents, respectively. The second main theme is health studiesis in both periods, responsible for 36% and 32% of core documents,

respectively.

The forecasting cluster in 1967–1998 continues as mathematical models in 1999–2007 and as forecasting for the remaining two periods. The health studies area remains throughout all the periods but merges with statistical models in the third period, which split into costs and health studies in the following period. In the first period, two isolated clusters are present: developed countries and Kalman filter. An isolated theme, decision trees, appears only in the last period. The algorithms cluster in 2008–2012 originates from mathematical models and remains with the same name in the next period.

The forecasting theme gathers studies that focus on the evaluation of prediction methods, health studies concentrates on works related to health behaviour, developed countries gathers works applied to countries such as Sweden and Australia, and Kalman filter focuses on works that use the filter. Papers that applied AI techniques form the mathematical models and algorithms clusters. Air pollutant cluster concentrates on works that deal with the influence of air quality in the development of diseases. Papers that applied decision trees methodology form the decision trees theme and costs cluster presents works related to prices and costs.

Fig. 6 synthesizes the evolution of strategic thematic diagrams for the four periods, where the sizes of the spheres are proportional to the number of core documents in each cluster and period.

According to the positioning on the Callon's diagram of Fig. 6, motor clusters evolved from forecasting and developed countries in 1967–1998 to health studies and air-pollutant in 1999–2007, to health studies and forecasting in 2008–2012 and 2013–2016. Since 2008, the core thematic clusters appear with a clear concentration in the health area and forecasting methods and techniques.

3.4. Main methods in time series analysis

In order to identify the main methods applied to time series with explanatory variables, a thorough reading of 150 articles corresponding to core documents responding to 80% of total citations was carried out. Only the methods that incorporate explanatory variables were reviewed, that is, if an article used exponential smoothing, multiple linear regression, and dynamic regression, only the last two methods were attributed to that article.

Through the analysis, it was possible to identify 30 different types of methods present in the core documents. The method applied more often was regression models, followed by ANNs, Box and Jenkins' ARIMA with the incorporation of explanatory variable (ARIMAX), support vector machines (SVMs), and structural models, which together are present in 70% of the papers. Fig. 7 displays the temporal evolution of the top five methods. Despite the predominance of regression models and ANNs, the use of SVMs is increasing over time.

The top five methods are summarised and discussed next.

3.4.1. Regression models

According to the co-word analysis, in the period 1967–1998, regression models were the second most applied method, led by Goldstein et al. (1994), who proposed a model that can incorporate explanatory variables to time series data, where the measurements are made close together in time, resulting in a possible correlation in the residuals. Goldstein et al. (1994) were classified in the health studies cluster. From 1999 to the present day, regression-type models lead the number of applications. In 1999–2007, the most cited work that uses regression models is Marcellino et al. (2006), who compared forecasts from linear univariate and bivariate models to forecast monthly macroeconomic time series. In 2008–2012, the study of Sistrom et al. (2009), the most cited work among regression models in this period, was classified as health studies. Sistrom et al. (2009) used piecewise linear regression to determine the effects of computerised order entry with integrated decision support on the growth of outpatient procedures. In 2013–2016 the paper from Ishak et al. (2013) led the work in regression models and SVMs, focused on the development of a hybrid model that uses two regression models, ANNs and SVMs, to better estimate wind speed using hydro-meteorological parameters.

3.4.2. Artificial neural networks (ANN)

In 1967–1998 ANNs are led by the work of Jorquera et al. (1998), classified in the forecasting cluster, which compared methods of linear time series, ANNs, and fuzzy models to forecast daily maximum ozone levels. In the following periods, González et al. (2005) proposed an input-output hidden Markov model to analyse and forecast electricity spot prices. Coman et al. (2008) evaluated a dynamic model versus a static one, using multilayer perceptron (MLP) in the context of predicting hourly ozone concentrations. Fernandez et al. (2009) used fuzzy neuro networks to improve wastewater flow-rate forecasting. Guresen et al. (2011) evaluated the effectiveness of ANN models in stock-market predictions. Nourani et al. (2013) applied a feed-forward neural network to model a rainfall-runoff process on a daily and multi-step ahead-of-time scale.

3.4.3. ARIMAX

In 1967–1998, the most prevalent method in number of papers is ARIMAX, where the work of Yang et al. (1996) is the most cited paper. In their work, classified in the forecasting cluster, a new evolutionary programming approach to identify the ARMAX model for one day to one week ahead hourly load demand forecasts is proposed. The study of Low et al. (2006), is the most cited work that uses ARIMAX

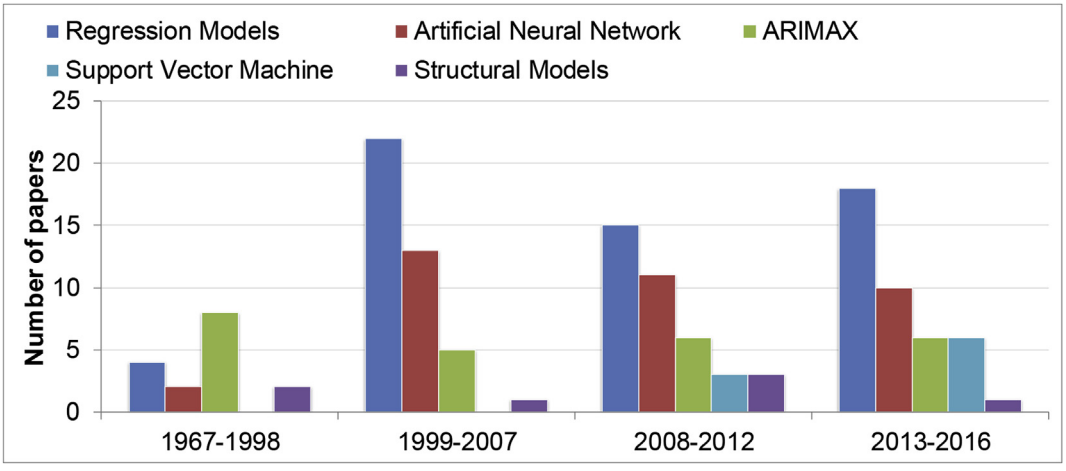

Fig. 7. Number of papers per methods by period.

methodology in the second period, using stroke admissions as the response variable and day of the week, holidays, September 11, and other counts and levels as explanatory variables. In the third period, Pisoni et al. (2009) used a nonlinear AR model with exogenous variables to forecast peak air pollution levels. The study by Marcilio et al. (2013), is the leading work in ARIMAX methods in the last period, used generalised linear models, generalised estimating equations, and seasonal AR integrated moving average with and without the effect of mean daily temperature as a predictive variable to forecast daily emergency department visits.

3.4.4. Support vector machine (SVM)

According to the co-word analysis the two most prominent works that apply SVR are Lu et al. (2009), that applied independent component analysis and Support Vector Regression to forecast financial time series, and Ishak et al. (2013) that use SVR to accurately estimate wind speed.

3.4.5. Structural models

In structural models, the most prominent work in the period 1967–1998 is from Velicer et al. (1996), who developed a theoretical model that attempts to define more appropriate multivariate sets of dependent variables in the study of health behavior change, allocated in the health studies cluster. The second period is led by Hernán et al. (2002), who used marginal structural models to estimate the effect of zidovudine therapy on mean CD4 count among HIV-infected men. In the third period, Athanasopoulos and Hyndman (2008) used a regression framework to estimate important economic relationships for domestic tourism demand in Australia and innovation state space models to forecast the same. The main work that applied structural models in the fourth period is from Bergel-Hayat et al. (2013), which highlighted the link between weather conditions and the risk of a road accident.

Table 6 summarises the strengths and weaknesses of each method presented in Fig. 7, as well as the most prominent papers for the same periods, extracted from Table 5 and Fig. 5.

4. Conclusion

This study applied a rigorous and reproducible SLR protocol resulting in the selection of 1547 papers related to time series analysis with explanatory variables. Time series analysis with exogenous variables is not a recent methodology, with the first publication dating back to 1969. However, there is an exponential growth in the number of publications after 1996. The increasing number of research subject areas that apply the models, ranging from computer science to medicine, shows the prominence of time series analysis with explanatory variables in a much-diversified array of disciplines. The largest number of papers belongs to environmental sciences, as expected given the complexity of modelling and forecasting of such data as saying El Niño phenomenon, sunspot, water runoff, inflows and stream flows, concentration of CO_2, air quality.

Four research questions guided the analysis. The citation analysis was applied to answer the first question “Who are the most prolific authors in time series analysis with explanatory variables?” Citation analysis shows that the most prolific author in the research area is Beck, N., based on the number of papers. He is equally the author with the larger number of citations. However, there is not a direct correlation between the number of papers and citations. The analysis equally evidenced the most cited research in time series with exogenous variables from social sciences and economic research originated in the US and UK-based institutions.

The co-citation analysis answered the second question “Which are the most influential works in time series analysis with explanatory variables?” Results present the most influential works in the area of interest in chronological order and show that they are the references for the most prominent time series methods and the main statistical tests. Co-citation analysis shows the backbone of seminal work in time series analysis with explanatory variables and can be a chronological guide to any researcher that wants to deepen into the subject.

The last two questions, “Which are the main themes in time series analysis with explanatory variables and how did they evolve?” and “Which are the main methods applied in time series analysis with explanatory variables and how did they evolve?” were answered by applying co-word analysis and dividing the database into four different periods. The co-word analysis shows the main contemporary research streams and their evolution. Forecasting and health studies themes are the prevalent themes over the four periods. Regarding the main methods, Box and Jenkins' ARIMA with the incorporation of explanatory variables led in the first period (1967–1998), followed by regression models. Currently (2013–2016), regression models retain the

Table 6
Strengths and weaknesses of time series methods.

Method	Strengths	Weaknesses	References
Regression models	Does not require high computational power. The relationship between the exogenous and response variables is open and most of the time, interpretable.	Is sensitive to outliers. It presupposes that the model errors are independent and follow a normal distribution with zero mean and constant variance.	Goldstein et al. (1994) Marcellino et al. (2006) Sistrom et al. (2009) Ishak et al. (2013)
ANN	High data processing power due to its massively distributed structure and its ability to learn and therefore to generalise, producing suitable outputs for inputs that were not present during the training.	Requires a high computational power and a large amount of data in the training process. Is considered a black-box model, since the relationship between exogenous and response variables is impossible to be interpret.	Jorquera et al. (1998) González et al. (2005) Guresen et al. (2011) Nourani et al. (2013)
ARIMAX	Based on a methodology already established in time series analysis. Easy to find several formal tests to verify the model suitability.	Assumes that the relationship between variables is linear. Presupposes that the model errors are independent and follow a normal distribution with zero mean and constant variance.	Yang et al. (1996) Low et al. (2006) Pisoni et al. (2009) Marcilio et al. (2013)
SVM	Efficient in solving nonlinear problems. Robustness towards a small number of data points.	Requires a high computational power and a large amount of data in the training process. Lack of transparence in the results.	Lu et al. (2009) Ishak et al. (2013)
Structural models	Handles missing data in an efficient way. Handles easily structural breaks, shifts and time-varying parameters.	The determination of the parameters' initial values can be a challenge. Assumes that the relationship between variables is linear. It presupposes that the model errors are independent and follow a normal distribution with zero mean and constant variance.	Velicer et al. (1996) Hernán et al. (2002) Athanasopoulos and Hyndman (2008) Bergel-Hayat et al. (2013)

Table 7
Key answers to the Research Questions (RQ).

RQ	Analyses	Main answers
1	Citation	The author with the highest number of papers is Beck, N., but the one with highest *h*-index is Wigley, T., showing that the authors with highest number of papers in the field are not necessarily the most cited ones.
2	Co-citation	The most influential works are the references for the most prominent time series methods (Box and Jenkins, 1970; Haykin, 1999; Takagi and Sugeno, 1985 etc.) and the main statistical tests (Akaike, 1974; Ljung and Box, 1978; Philips and Perron, 1988 etc.).
3	Co-word	Four periods (1967–1998, 1999–2007, 2008–2012 and 2013–2016) are enough to analyse the themes' evolution. The main themes in all the periods are forecasting and health studies.
4	Co-word	The main methods, in number of applications, are Regression models, Artificial Neural Networks, ARIMAX, Support Vector Machine and Structural models. In the first period, ARIMAX leads in the number of papers but loses place to the Regression model. Support Vector Machine grows after the third period.

first place as the method of choice and ANNs hold the second place in the number of applications, followed by SVMs. The growth of AI methods can be attributed to the increase in the processing capacity of computers and their attractive features, being semiparametric learning machines, permitting universal approximation of arbitrary linear and nonlinear functions from examples without a priori assumptions on the model structure and often outperforming conventional statistical approach methods (Crone et al., 2006).

Table 7 summarises the answers found for each one of the Research Questions (RQ) drawn in subsection 2.1.

The incorporation of exogenous variables in the forecasting models are more prevalent in research aiming at improving forecast accuracy, identifying the most important co-variables that affects the response (e.g., causality, relationships among variables), verifying the elasticity between the output variable and their inputs, and generate synthetic scenarios for the dependent variable taking into account different scenarios for the independent variables. Considerable effort is equally devoted to the choice of the set of exogenous variables themselves, as well as to the methods to forecast and simulate different scenarios.

This review has some limitations due to the broad scope covering different research areas, to the choice of the search parameters (keywords and exclusion criteria), and of the citations databases (Scopus alone), therefore offering a limited sample of relevant articles. These limitations lead the path to new research avenues. First, as the review is not focused on a specific research area, further in-depth analysis and mapping of the time series methods with exogenous variables in the more prevalent research areas of environmental sciences, economics and health seem to be opportune. Second, the main methods outlined here could be further described with guidelines as to when and in which cases such methods should or should not be applied. Third, the analysis of the evolution of themes and time series methods by discipline and periods could be pursued further. It could lead to a much more detailed analysis of differentials in the research front of time series with explanatory variables in different disciplines. Fourth, the combination of other databases, beyond Scopus, can lead to a more exhaustive review, despite the large intersection between documents figuring in different citation databases. Fifth, the inclusion of non-peer-reviewed material such as thesis and dissertations can enrich the analysis and lead to the inclusion of new methods and applications yet to be published in peer-reviewed journals, circumventing the risk of publication bias associated with peer-reviewed literature.

Acknowledgement

The authors acknowledge the support of the following Brazilian agencies: National Council for Scientific and Technological Development, Grants# 304931/2016-0, 404682/2016-2, and 443595/2014-3, and the Research Support Foundation of the State of Rio de Janeiro (FAPERJ), Grants # E-26/203.252/2017, and 202.806/2015.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2018.06.004.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Second Int. Symp. Inf. Theory, pp. 267–281.
Akaike, H., 1974. A new look at the statistical model identification. Autom. Control. IEEE Trans. 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.
Anderson, T.W., Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. J. Econom. 18, 47–82. https://doi.org/10.1016/0304-4076(82)90095-1.
Andriyas, S., McKee, M., 2013. Recursive partitioning techniques for modeling irrigation behavior. Environ. Model. Software 47, 207–217. https://doi.org/10.1016/j.envsoft.2013.05.011.
Athanasopoulos, G., Hyndman, R.J., 2008. Modelling and forecasting australian domestic tourism. Tourism Manag. 29, 19–31.
Athanasopoulos, G., Hyndman, R.J., Song, H., Wu, D.C., 2011. The tourism forecasting competition. Int. J. Forecast. 27, 822–844. https://doi.org/10.1016/j.ijforecast.2010.04.009.
Beck, N., Katz, J., Tucker, R., 1998. In: In: Beck, Nathaniel, Katz, Jonathan N., Tucker, Richard (Eds.), Taking Time Seriously: Time-series-cross-section Analysis with a Binary Dependent Variable, vol. 42. Published by : Midwest Political Science Association Stable, pp. 1260–1288. http://www.jstor.org/stable/2.
Beck, N., Katz, J.N., 1995. What to do (and not to do) with time-series cross-section data. Am. Polit. Sci. Assoc. 89, 347–634.
Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: weather effects. Accid. Anal. Prev. 60, 456–465.
Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31, 307–327. https://doi.org/10.1016/0304-4076(86)90063-1.
Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis, Forecasting and Control. Holden Day, San Franciso.
Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. Prentice Hall.
Box, G.E.P., Pierce, D.A., 1970. Distribution of Residual Autocorrelations in Autoregressive-integrated Moving Average Time Series Models, vol. 65. pp. 1509–1526.
Callon, M., Courtial, J.P., Laville, F., 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. Scientometrics 22 (1), 155–205.
Cobo, M., López-Herrera, A., Herrera-Viedma, E., Herrera, F., 2011. An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the Fuzzy Sets Theory field. J. Informetr 5.
Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2012. SciMAT: a new science mapping analysis software tool. J. Am. Soc. Inf. Sci. Technol. 63, 1609–1630.
Coman, A., Ionescu, A., Candau, Y., 2008. Hourly ozone prediction for a 24-h horizon using neural networks. Environ. Model. Software 23, 1407–1421. https://doi.org/10.1016/j.envsoft.2008.04.004.
Cooper, H.M., 2010. Research Synthesis and Meta-analysis: a Step-by-step Approach, fourth ed. Sage, Thousand Oaks. Thousand Oaks.
Crone, S.F., Guajardo, J., Weber, R., 2006. A study on the ability of support vector regression and neural networks to forecast basic time series patterns. IFIP Int. Fed. Inf. Process 217, 149–158.
De Nooy, W., Mrvar, A., Bagatelj, V., 2005. Exploratory Network Analysis with Pajek. Cambridge University Press.
Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimates for autoregressive time series with a unit root. J. Am. Stat. Assoc. 366, 427–431.
Durbin, J., Koopman, S.J., 2012. Time series analysis by state space methods. Oxf. Stat. Sci. 38.
Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50, 987–1008.
Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica 2, 251–276.
Espey, M., Espey, J., Shaw, W.D., 1997. Price elasticity of residential demand for water: a meta-analysis. Water Resour. Res. 33, 1369–1374. https://doi.org/10.1029/97wr00571.
Fernandez, F.J., Seco, A., Ferrer, J., Rodrigo, M.A., 2009. Use of neurofuzzy networks to improve wastewater flow-rate forecasting. Environ. Model. Software 24, 686–693.
Goldstein, H., Healy, M.J.R., Rasbash, J., 1994. Multilevel time series models with applications to repeated measures data. Stat. Med. 13, 1643–1655.

González, A.M., San Roque, A.M., García-González, J., 2005. Modeling and forecasting electricity prices with input/output hidden Markov models. IEEE Trans. Power Syst. 20, 13–24.
Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.
Guresen, E., Kayakutlu, G., Daim, T.U., 2011. Using artificial neural network models in stock market index prediction. Expert Syst. Appl. 38, 10389–10397.
Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press.
Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.
Haykin, S., 1999. Neural Networks: a Comprehensive Foundation. Prentice Hall, Englewood CliKs.
Hernán, M.A., Brumback, B.A., Robins, J.M., 2002. Estimating the causal effect of zidovudine on Cd4 count with a marginal structural model for repeated measures. Stat. Med. 21, 1689–1709.
HLWIKI, 2017. HLWIKI Canada.
Hornik, K., Stinchcomb, M., White, H., 1989. Multilayer feed-forward networks are universal approximators. Neural Network. 2, 359–366.
Ishak, A.M., Remesan, R., Srivastava, P.K., Islam, T., Han, D., 2013. Error correction modelling of wind speed through hydro-meteorological parameters and mesoscale model: a hybrid approach. Water Resour. Manag. 27, 1–23.
Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man. Cybern 3, 665–685.
Johansen, S., 1988. Statistical analysis of cointegration vectors. J. Econ. Dynam. Contr. 12, 231–254. https://doi.org/10.1016/0165-1889(88)90041-3.
Johansen, S., Juselius, K., 1990. Maximum likelihood estimation and inference on cointegration with applications to the demand for money. xford Bull. Econ. Stat 52, 169–210.
Johansen, S.J., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregression models. Econometrica 59, 1551–1580.
Jorquera, H., Pérez, R., Cipriano, A., Espejo, A., Letelier, M.V., Acuña, G., 1998. Forecasting ozone daily maximum levels at santiago, Chile. Atmos. Environ. 32, 3415–3424.
Lima, L.M.M., Popova, E., Damien, P., 2014. Modeling and forecasting of Brazilian reservoir inflows via dynamic linear models. Int. J. Forecast. 30, 464–476. https://doi.org/10.1016/j.ijforecast.2013.12.009.
Ljung, G.M., Box, G.E., 1978. On a measure of lack of fit in time series models. Biometrika 65. https://doi.org/10.1093/biomet/65.2.297.
Ljung, L., 1999. System Identification: Theory for the User. Prentice Hall PTR.
Low, R.B., Qureshi, A.I., Dunn, V., Stuhlmiller, D.F.E., Dickey, D.A., 2006. The relation of stroke admissions to recent weather, airborne allergens, air pollution, seasons, upper respiratory infections, and asthma incidence, september 11, 2001, and day of the week. Stroke 37, 951–957.
Lu, C.-J., Lee, T.-S., Chiu, C.-C., 2009. Financial time series forecasting using independent component analysis and support vector regression. Decis. Support Syst. 47, 115–125.
Marcellino, M., Stock, J.H., Watson, M.W., 2006. A comparison of direct and iterated multistep Ar methods for forecasting macroeconomic time series. J. Econom. 135, 499–526.
Marcilio, I., Hajat, S., Gouveia, N., 2013. Forecasting daily emergency department visits using calendar variables and ambient temperature readings. Acad. Emerg. Med. 20, 769–777.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall/CRC.
Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies. Do they fit out of sample? J. Int. Econ. 14. https://doi.org/10.1016/0022-1996(83)90017-X.
Milionis, A.E., Davies, T.D., 1994. Regression and stochastic models for air pollution-I. Review, comments and suggestions. Atmos. Environ. 28, 2801–2810. https://doi.org/10.1016/1352-2310(94)90083-3.
Mongeon, P., Paul-Hus, A., 2016. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106, 213–228.
Nash, J.E., Sutcliffe, J.V., 1970. river flow forecasting through conceptual models Part I-a discussion of principles*. J. Hydrol 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Nourani, V., Baghanam, A.H., Adamowski, J., Gebremichael, M., 2013. Using self-organizing maps and wavelet transforms for space-time pre-processing of satellite precipitation and runoff data in neural network based rainfall-runoff modeling. J. Hydrol 476, 228–243.
Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R., Chatterton, T., 2004. Modelling SO2 concentration at a point with statistical approaches. Environ. Model. Software 19, 887–905. https://doi.org/10.1016/j.envsoft.2003.10.003.
Persson, O., Danell, R., Wiborg Schneider, J., Astrom, R., Danell, R., Larsen, B., Wiborg Schneider, J., 2009. How to use Bibexcel for various types of bibliometric analysis. In: Celebrating Scholarly Communication Studies: a Festschrift for Olle Persson at His 60th Birthday.
Petticrew, R., Roberts, H., 2006. Systematic Reviews in the Social Sciences. A Practical Guide. Blackwell, Malden, MA.
Philips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Pisoni, E., Farina, M., Carnevale, C., Piroddi, L., 2009. Forecasting peak air pollution levels using narx models. Eng. Appl. Artif. Intell. 22, 593–602.
Porporato, A., Ridolfi, L., 2001. Multivariate nonlinear prediction of river flows. J. Hydrol 248, 109–122. https://doi.org/10.1016/S0022-1694(01)00395-X.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, pp. 318–362.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464. https://doi.org/10.1214/aos/1176344136.
Scott Jr., J.T., Heady, E.O., 1967. Regional demand for farm buildings in the United States. Am. J. Agric. Econ. 49.
Seuring, S., Gold, S., 2012. Conducting content-analysis based literature reviews in supply chain management. Supply Chain Manag. An Int. J. 17, 544–555.
Sfetsos, a., Coonick, A.H., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol. Energy 68, 169–178. https://doi.org/10.1016/S0038-092X(99)00064-X.
Sistrom, C.L., Dang, P.A., Weilburg, J.B., Dreyer, K.J., Rosenthal, D.I., Thrall, J.H., 2009. Effect of computerized order entry with integrated decision support on the growth of outpatient procedure volumes: seven-year time series analysis. Radiology 251, 147–155.
Small, H., 1999. Visualizing science by citation mapping. J. Am. Soc. Inf. Sci. 50, 799–813.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modeling and control. IEEE Trans. Syst. Man, Cybern 15, 116–132.
Thomé, A.M.T., Scavarda, L.F., Scavarda, A., 2016a. Conducting systematic literature review in operations management. Prod. Plann. Contr. 27, 408–420.
Thomé, A.M.T., Scavarda, L.F., Scavarda, A., Thomé, F.S.S., 2016b. Similarities and contrasts of complexity, uncertainty, risks, and resilience in supply chains and temporary multi-organization projects. Int. J. Proj. Manag. 34, 1328–1346.
Thomé, A.M.T., Maçaira, P., Cyrino Oliveira, F.L., Ferrer, A.L.C., 2017. Citation Data in Time Series, Mendeley Data, V1. http://dx.doi.org/10.17632/4svhhzrw8d.1.
Tong, H., 1990. Non-linear Time Series: a Dynamical System Approach. Oxford University Press, New York.
Velicer, W.F., Rossi, J.S., Prochaska, J.O., Diclemente, C.C., 1996. A criterion measurement model for health behavior change. Addict. Behav. 21, 555–584.
White, H., 1980. A heteroskedasticity-consistent covariance-matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.
Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J., Wilks, D.S., 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water Resour. Res. 34, 2995–3008. https://doi.org/10.1029/98wr02577.
Yang, H.-T., Huang, C.-M., Huang, C.-L., 1996. Identification of armax model for short term load forecasting: an evolutionary programming approach. IEEE Trans. Power Syst. 11, 403–408.
Young, P.C., 2011. Recursive Estimation and Time–series Analysis: an Introduction for the Student and Practitioner. Springer–Verlag, Berlin.
Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.
Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R., Beukema, J.J., 2003. Estimating common trends in multivariate time series using dynamic factor analysis. Environments 14, 665–685.
Zuur, A.F., Pierce, G.J., 2004. Common trends in northeast Atlantic squid time series. J. Sea Res. 52, 57–72.