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

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%