A Systematic Literature Review on Decomposition Approaches to Estimate Time Series Components

RICARDO ARAÚJO RIOS[1]
RODRIGO FERNANDES DE MELLO[1]

Institute of Mathematics and Computer Science
University of São Paulo
Avenida Trabalhador Sãocarlense, 400
São Carlos, SP, Brazil
P.O. Box 668
[1]{rios,mello}@icmc.usp.br

Abstract. The study and modeling of systems have called the attention of several researchers, who are interested in estimating rules to describe data behavior. However, before proceeding with this estimation, it is necessary to understand the intrinsic features embedded in data. When such features are not correctly analyzed, the model accuracy tends to decrease. A well-known way to perform this analysis is by the study of time series behavior according to their stochastic and deterministic components. Nevertheless, the time series decomposition into these components is not a simple task. In order to address this issue, we conducted a rigorous and well-structured search for scientific papers in different repositories. By analyzing the recovered papers, we drew relevant conclusions such as: which methods are commonly used to decompose time series; the frequency of published papers per year; and the gaps of each method. Moreover, we have also classified the most suitable studies to estimate the determinism and stochasticity present in time series. After conducting this study, we concluded the development of methods to decompose time series into stochastic and deterministic components is still an open problem.

Keywords: Time Series Analysis, Signal Decomposition, Systematic Literature Review

(Received January 28th, 2013 / Accepted March 3rd, 2013)

1 Introduction

The term system is commonly used in several domains such as Economy, Computer Science, Mathematics, Astronomy and Biology. In general, these areas study systems in order to understand their operations, the behavior of their components, the relationship with other systems, and their produced outputs [41, 43]. System behavior is consequence of the interaction of multiple interconnected entities, referred to as subsystems or components [41, 43]. Interconnections define the different degrees of dependency and relationship among components.

The analysis of the outputs produced by a system, usually named experimental data, has been attracting the attention of many researchers, who are interested in developing methods capable of modeling and understanding systems behavior [6]. By representing this behavior, one can take more accurate decisions (e.g., management decisions in companies), simulate future situations (e.g., drugs interacting with an organism), predict operations (e.g., stock market), estimate system states (e.g., weather forecasting) and their influences over other systems (e.g., climate effects on agriculture), detect the occurrence of faults and failures (e.g., problems in production lines), and so on.

Modeling methods usually assume that experimental data were obtained from independent and identically distributed (i.i.d.) processes [36]. In that sense, many

researchers analyze data assuming they fit to probability distributions like Poisson, Exponential, Erlang, and Normal. However, an important factor to be considered when studying real-world systems is that data usually present temporal dependencies, i.e., the instantaneous behavior of a system depends on past observations, making unfeasible the application of such methods [36].

The presence of temporal dependencies has motivated researchers to organize systems outputs as sequences of observations, which are referred to as time series. Consequently, by analyzing such time series, one can model and understand the system behavior, transitions, relationships to other systems, as well as estimate and predict future observations. The area responsible for performing this study is called Time Series Analysis [5, 36], which attempts to obtain a rule or function to represent observations. In order to obtain such rule, it is first necessary to study the features of time series and its implicit components which support accurate modeling. Among those features are stationarity, linearity and, finally, stochasticity and determinism [20].

A time series is said to be stationary when its observations are in a particular state of statistical equilibrium, i.e., they evolve over time around a constant average [5]. On the other side, linear time series are those whose observations are composed of linear combinations of past occurrences and noises. Such linear combinations are present on the model, map, or process that generated the series. In turn, non-linear time series are composed of non-linear combinations of past occurrences and noises. Finally, stochastic time series are composed of random observations and relations, which follow probability density functions and may change over time. In contrast, deterministic time series strictly dependent on past observations [5].

Although it is very important to consider all the previous features when modeling a time series, a well-known way to analyze time series is through the study of the determinism and stochasticity present in its observations.

In this sense, when a series is classified as deterministic, models from Dynamical Systems, specifically those from the Chaos Theory [1], are more adequate as they provide better results. When a series presents stochastic behavior, statistical models, such as the ones proposed by Box & Jenkins [5], are taken as the most appropriated ones. However, time series obtained from real-world systems usually present a mixture of both components, i.e., the value of a single observation is influenced by deterministic and stochastic components. In such situation, the discard of any component can jeopardize the quality of the resultant model [16], i.e., the application of Dynamical System techniques tends to generate malformed attractors, whereas statistical techniques tend to underestimate the deterministic part of the system.

Aiming at overcoming this issue, the study of system behavior must be performed according to three well-defined steps [41]: i) firstly, the system is decomposed in parts or components; ii) secondly, models are used to represent the behavior of each component; iii) at last, partial models are combined to describe the global behavior of the system. However, the decomposition of time series is not a simple task, once we do not know the influences of every component on observations.

In this sense, many researchers have been proposing new techniques to decompose time series, aiming at separating stochastic and deterministic components. This increasing number of publications motivated this work, whose main objective is to identify the most useful techniques to perform such decomposition. In that sense, we conducted a structured and documented study of published papers following the rules defined by the method of Systematic Literature Review (SLR) [22].

The SLR method is commonly employed in many areas, such as Medicine. Its goal is to define a rigorous and well-structured search for published papers in order to collect and evaluate evidences on a given subject. The main advantage of SLR is to perform a wide search, recovering not only well-known papers but also other related ones. Papers collected by SLR can be either directly read and analyzed or organized as a Systematic Mapping, as presented in [34, 10].

By employing SLR, this paper presents strong evidences of the importance of decomposing time series and also identifies the most common methods considered for this purpose, the frequency of published papers, the relation in between the number of papers and the type of publication, and the most important gaps of each method.

The remainder of this paper is organized as follows: Section 2 presents a brief overview about the concepts of Systematic Literature Review; Section 3 shows the parameters defined to conduct the search for documents in well-known databases; in Section 4, the found documents are analyzed according to their relevance to the subject of interest; Section 5 presents a summary of the most commonly considered techniques when decomposing time series; Section 6 presents concluding remarks and future work, followed by an appendix that summarizes the general results obtained by SLR and shows an analysis executed on the selected papers.

2 Systematic Review

In Medicine, the process of clinical decision is commonly performed according to the methodology called Evidence-Based Medicine (EBM) [13]. This methodology was developed after researchers concluded that the specialist's opinion based on a medical advice is as trustful as the opinion obtained from scientific experiments [22]. Considering this assumption, EBM was created to reduce the importance of non-systematic clinical experience on decision making. EBM also increased the importance of the evidence analysis obtained from clinical research. The application of this methodology requires new abilities of the physician as, for instance, the efficient research in scientific literature and the application of formal rules when evaluating evidences [11].

The advantages obtained with EBM have motivated the adoption of this methodology in other areas such as Criminology, Economy and Nursing [22]. In Computer Science, specifically in Software Engineering, Dyba *et al.* [11] defined, based on EBM, a methodology called Evidence-Based Software Engineering (EBSE). According to authors, the main objective is to provide ways by which the best evidences gathered from different researches can be integrated to human values and practice experiences, improving software development and maintenance.

According to Kitchenham *et al.* [22], an evidence can be defined as a synthesis of high-quality scientific studies about a given research topic. The main synthesis method defined by this methodology is the Systematic Literature Review (SLR), which provides guidelines to make a rigorous review of studies related to the topic of interest. It is important to highlight that SLR cannot substitute the traditional review of literature, since this review is necessary before executing a rigorous search and defining which papers are relevant for the review. In addition to that, traditional review provides some initial concepts about the studied subject, which makes it possible to choose keywords and define inclusion and exclusion criteria used in SLR, and, eventually, to add some specific and important studies, which were not retrieved by SLR.

However, SLR is limited in the sense that is restricted to some previously defined keywords, which can eventually lead to the lack of some important related work. An approach commonly considered to overcome such problem is to execute a recursive search; thus, references of a selected paper are then used to retrieve further studies.

In this sense, aiming at developing a systematic and rigorous research on time series decomposition, the guidelines provided by SLR, presented in [22, 12], were adapted and divided in three phases: in the first, called Search, some criteria were defined to seek related work; in the second, called Analysis, the quality of gathered papers was analyzed and quantified; finally, the last phase presents conclusions obtained from the systematic study. In the following sections, each phase is presented and discussed in details.

3 Phase I: Search

In the first phase, we defined the general scope of the research, i.e., we selected the research criteria to characterize whether a paper is related to this study or not. In this sense, we defined the objective of this research, the main and secondary research questions, the search repositories, the standard language, the list of keywords, the search query, the inclusion and exclusion criteria, and, finally, the general process of execution.

In general, the main objective of this research is to find techniques related to the decomposition of time series in terms of stochastic and deterministic components. Based on this objective, the Main Research Question (MRQ) to guide this research is:

> *What are the techniques used to decompose time series in stochastic and deterministic components?*

In addition to this main question, it is relevant to define a set of Secondary Questions (SQ), which are directly associated to the validity of the proposed research. This set of questions is used to discover practical applications of the research, evaluating techniques, and understanding publication trends. The secondary questions are:

SQ.1 - What types of practical applications can take advantage of this decomposition process?

SQ.2 - How are techniques evaluated?

SQ.3 - Why are time series decomposed?

SQ.4 - What are the main techniques used to decompose time series?

SQ.5 - What is the frequency of published papers per year?

SQ.6 - Who are the main researchers in this area?

SQ.7 - What are the limitations of techniques?

Due to the complexity and importance of the main objective of this research, the last secondary question was divided, as in the study presented in [22], into the following four questions:

SQ.7.1 - Was the search query limited?

SQ.7.2 - Is there any evidence that the studied subject was limited by the lack of primary studies[1]?

SQ.7.3 - Is the quality of the studied subject good enough?

SQ.7.4 - What are the main limitations of the found approaches/techniques?

After defining these questions, the next step of this first phase was to choose the search repositories, from which related studies were obtained. We selected repositories which provide Web search engines, accept queries using keywords, and are commonly used by the scientific community. Based on these restrictions, the following repositories were chosen:

- Scopus (http://www.scopus.com/home.url);
- ACM Digital Library (http://portal.acm.org/);
- IEEE Xplore Digital Library (http://ieeexplore.ieee.org/).

The standard language used in this systematic review was English, i.e., all papers written in other languages were discarded. As the next step, the following keywords were chosen considering the hypothesis and the main research question:

- Data organization: *time series*;
- Goals: *decomposition*, *filter*;
- Results: *deterministic*, and *stochastic*.

We have also decided to introduce the term "filter" as synonym to decomposition because, according to our previous studies, this term is frequently used to refer to techniques that decompose time series into stochastic and deterministic components. Of course, there are several synonyms that could be added as keywords, but they would considerably increase the number of returned papers, making it almost impossible to perform this study. This is also foreseen by SLR, which limits the number of keywords and make the systematic research feasible.

Based on keywords, the following search query was defined:

```
("time series") AND
(decomposition OR filter) AND
(deterministic AND stochastic)
```

We then use this query to obtain relevant papers from repositories. In spite of the reduced number of keywords, our query has returned a high number of papers, however many of them were not relevant to SLR. To reduce this number, we defined inclusion and exclusion criteria. Therefore, a work was added to SLR when it satisfied the following conditions:

- Does the work separate stochastic and deterministic components from a time series?
- Is it a primary study?
- Is the work focused on the modeling process of time series?

On the other hand, a paper must be discarded if any of the following exclusion criteria is applied:

- The work presents a technique which is very restricted to a specific problem;
- The work does not present a well-defined analytical model;
- The technique evaluation process is not satisfactory;
- The work does not have a comprehensive literature review.

Finally, after defining all initial conditions, the first phase of SLR, the search phase, was carried out according to the following steps:

1. The defined query was applied in the selected repositories;
2. The title and abstract of each paper were analyzed, considering the inclusion and exclusion criteria;
3. The redundant papers were removed from SLR;
4. The remaining papers were completely read.

[1] According to the guidelines of the SLR presented in [22], a primary study is a work that contributes to a new approach or technique of a given subject. On the other hand, reviews of existing approaches, such as SLRs and Surveys, are considered secondary studies.

It is important to highlight that a given paper, selected to compose SLR, can still be removed in step 4 if an exclusion criteria is applied to it, even after being completely read. In the same way, if references of selected papers are considered relevant to SLR, they are indeed included in the list of selected papers, aggregating studies to SLR that were not found when searching in repositories. The next section presents the analysis of papers selected in this first phase.

4 Phase II: Analysis

The application of the search query, defined in the first phase, on the selected repositories returned 852 papers, which were distributed in repositories as presented in Table 1. However, some of the obtained references were not valid, because they were proceedings, information on editorial boards or redundant papers. Therefore, after removing these invalid references, the total of remaining papers was 749. We then analyzed every paper, by reading their titles and abstracts, as well as using the inclusion criteria. After this analysis, 77 papers were selected to be completely read and evaluated. At last, part of these papers were still rejected after applying the exclusion criteria. Hence, we finally classified 26 papers as strongly related to the subject presented in this work.

Table 1: Number of papers returned by SLR. "Recovered papers" represents the number of papers returned by SLR, after removing invalid references. "Inclusion criteria" represents the number of papers selected after reading titles and abstracts. Finally, "Exclusion criteria" shows the number of remaining papers after being completely read.

Repository	Number of papers
ACM	773
IEEE	66
Scopus	13
Total	852
Recovered papers	749
After inclusion criteria	77
After exclusion criteria	26

First of all, we analyzed the 26 selected papers to understand the frequency distribution of authors per country, which is presented in Table 2. All the analysis performed on the 26 selected papers were also executed on the other papers: all the 749 papers and those ones selected by the inclusion criteria (77). That is made available in Appendix A.

Table 2: Frequency distribution of authors per country.

	Frequency distribution
US - United States	20
CN - China	6
DE - Germany	6
FR - France	6
UK - United Kingdom	5
ES - Spain	4
IT - Italy	4
BR - Brazil	3
GR - Greece	3
CA - Canada	2
CH - Switzerland	2
IN - India	2
PT - Portugal	2
SE - Sweden	2
AE - United Arab Emirates	1
AT - Austria	1
KR - Korea (South)	1
NL - Netherlands	1
TW - Taiwan	1

Table 3 presents another interesting result, ordered by researchers with more papers related to the time series decomposition in terms of stochastic and deterministic components. This analysis was useful to answer the secondary question SQ.6, which aims at presenting the main researchers in this area.

Table 3: List of authors who published more papers.

	Frequency distribution
Norden E. Huang	4
George Tzagkarakis	2
Maria Papadopouli	2
Panagiotis Tsakalides	2
Steven R. Long	2
Zhaohua Wu	2

Table 4 presents the frequency of papers published per year, what answers secondary question SQ.5. According to this table, 2006, followed by 2009, were the years with more publications.

Table 5 shows the frequency distribution of selected papers considering the type of publication. In this case, we observe that most of the papers were published in journals.

The next analysis (Table 6), executed after identifying the publication type, was performed to know how

Table 4: Number of papers published per year.

	Frequency distribution
1995	1
1998	1
2000	3
2001	2
2003	1
2004	1
2005	1
2006	6
2007	1
2009	5
2010	2
2011	2

Table 5: Relation between the number of papers and the publication type.

	Frequency distribution
Journal	21
Conference	3
Book Chapter	1
Symposium	1

many papers were primary studies.

Table 6: Type of published papers.

	Frequency distribution
Primary Study	22
Application	4

In order to answer secondary question SQ.3, in Table 7, we analyzed the relation between published papers and goals when applying decomposition techniques.

All these analysis compose an overview on papers obtained using SLR. However, to proceed with our study, we had to read and understand the details of each selected paper. This further analysis was conducted based on some questions, which were elaborated to answers secondary questions SQ.1, SQ.2, and SQ.4. Hence, we defined the following questions:

- What is the context of the paper?
- What are the techniques used by the new method?
- How is the new method described?

Table 7: Goals of publications.

	Frequency distribution
Denoising	8
Filtering	2
Decomposition	13
Modeling	3

- What are the goals?
- How is the new technique evaluated?
- What is the score attributed to the paper?

The score of every paper was limited in interval $[0, 6]$, in which 0 means very poor and 6 very good. This score was defined by the sum of other specific questions, whose objective was to evaluate the quality of the paper in relation to references, the reproducibility of results, the quality of obtained results, the description, formalism and analysis of the proposed method. So, aiming at quantifying the quality of each paper, every specific question was classified as fair (0), average (0.5), or good (1).

The general results of this analysis are summarized in Table 18, at the end of this paper. From these results, we observed decomposition techniques are very useful to improve a variety of applications, answering in this way secondary question SQ.1. For example, according to [9], the decomposition of time series in stochastic and deterministic components has been used to study tremor in the human bodies. In this situation, every component represents a type of tremor (voluntary or not), thus, decomposition supports such identification. In another scenario [29], decomposition is used to understand the dynamics of heartbeat interval fluctuations in awake unrestrained mice, following the intracerebroventricular application of the neuropeptide Corticotropin-Releasing Factor (CRF). Other two examples in Biology, discussed by Kopsinis and McLaughlin [23] and Gruber *et al.* [15], used stochastic and deterministic components to, respectively, understand the behavior of bats and analyze protein chains. Another very interesting study using the decomposition of time series is presented in [3], which models automotive engine sound. In Computer Science, time series decomposition has been used in several studies as, for instance, to understand behavior patterns of messages in Wifi networks [39, 7].

Besides providing some real and practical examples of time series decomposition, the selected papers

were still useful to show us the most common techniques used to evaluate experimental results. In this sense, after decomposing a time series, each component was evaluated using not only a visual inspection but also statistical techniques, Mean Squared Error (MSE) [9, 33, 3, 27], Signal-to-Noise Ratio (SNR) [33, 23, 32] and Recurrence Quantification Analysis (RQA) [26]. MSE is used, in general, to evaluate the results obtained when predicting new observations. In this situation, a time series is decomposed and, then, each component is individually modeled and used to predict future values. Thus, MSE is employed to compare expected values against predicted ones. In contrast, SNR is normally considered to assess the precision of removing the stochastic component from a time series. In such situation, a time series is decomposed into stochastic and deterministic components. The expected values of the deterministic component, which is known a priori, are compared against the estimated ones. At last, RQA is a set of measures, defined on the results of the Recurrence Plot (RP) technique [28], which quantifies the relation between stochastic and deterministic signals. One of the most used measure is the determinism rate (DET), that quantifies the influences of noise (or stochastic component) on time series, i.e. stochastic components present lower DET values than deterministic ones.

Finally, secondary question SQ.4 aims at determining what techniques are most used to decompose time series. As seen in Table 18, the most used techniques are: Fourier [27] and Wavelet [2, 16, 23] Transforms, Principal Component Analysis (PCA) [26, 39] and Empirical Mode Decomposition (EMD) [9, 23]. The next section provides a brief description of these techniques.

5 Phase III: Conclusion

This section presents the last phase performed using the SLR, which briefly describes some of the most commonly considered techniques to decompose time series. Although we have found many techniques for this porpose, as seen in Table 18, we decided to show only a subset of them, selecting the ones with better results. The discussion presented in this section is important to answer the main question of this work, whose objective was to identify techniques to decompose time series in terms of stochastic and deterministic components.

5.1 Principal Component Analysis

According to Andrew [4], the Principal Component Analysis (PCA) is often used as first stage in data analysis. The central idea of Principal Component Analysis is to reduce the data dimensionality, retaining as much information as possible [21].

The main reason for applying PCA on time series is to identify a group of underlying components to describe observations. In a general way, PCA attempts to linearly transform data into uncorrelated data (feature space) [35]. In PCA, a data vector is represented in an orthogonal basis system such that the projected information has maximal variance [35].

In summary, PCA is a non-parametric method for extracting relevant information from data composed of mixtures of different signals [35]. This technique is limited to linear time series.

5.2 Fourier Transform

Fourier Transform (FT) is one of the most used techniques to decompose time series. It has been widely used since its introduction and its main advantage is the possibility of examining a time series in terms of global energy-frequency distributions [19]. Moreover, FT has been applied to all sorts of data, mainly because of its well-defined mathematical formalism and simplicity [19].

In a general way, FT provides a function with series spectral information. In its basic algorithm, Fourier transform translates a series from the time domain to its frequency representation. This transformation is executed using sines and cosines as basis functions and produces spectral information on the analyzed signal, which is represented as waves. By analyzing the amplitude and phase of these waves, some researchers attempt to distinguish stochastic and deterministic components [14].

The application of FT in decomposition has some crucial restrictions, as discussed by Huang *et al.* [19], such as: the analyzed series must be linear and it must be strictly periodic or stationary.

5.3 Wavelet Transform

An alternative and more efficient method used to overcome the restrictions imposed by Fourier transform is provided by Wavelet Transform (WT). This method is defined by mathematical functions that decompose time series in different scales and resolutions [14]. This decomposition makes possible to analyze time series not only in the frequency domain, like Fourier Transform, but also in time domain, keeping, in this way, temporal relations and features among observations.

Another important peculiarity of WT, in contrast with FT that uses sines and cosines as basis functions, is the function used to analyze time series, which is based

on coefficients and wavelets. Coefficients are ordered using two dominant patterns: i) the first one works as result to a smoothing filter, and ii) the second one provides details on data [14]. Details represent information variations, which are used to detect noise in time series [17].

Although WT is considered a recent method, it has become extremely popular and useful to analyze data at gradual frequency changes [19]. As well as the Fourier Transform, WT has a well-defined mathematical formalism, which has called the attention of several researchers in different areas such as mathematicians, biologists, electrical engineers, physicists and statisticians.

However, this strong relation with Fourier Transform restricts its application to the same problems of Fourier spectral analysis, i.e., this technique is strongly limited to time series with linear behavior [19]. Another problem is related to the analysis of the wavelets generated by this method. They are used to solve the interwave frequency modulation provided by the gradual frequency variation, but it cannot solve the intrawave frequency modulation when the wavelet has a continuous length [19].

5.4 Empirical Mode Decomposition

Spectral representation has been a standard method for analyzing time series in many different areas, mainly, in Mathematics and Statistics [19]. In a general way, spectral analysis methods rely on a pre-defined basis, which is a collection of linearly independent vectors used to represent data [18, 19]. Example of spectral analysis are Fourier and Wavelet transforms, discussed in previous sections.

The Fourier transform, however, does not provide good results when the time series is not stationary nor generated from a linear process. To overcome these limitations, researchers have been applying Wavelet transform instead, which has providing good results to non-stationary processes [18]. However, WT is considered a Fourier-based method, this is, it performs data transformations considering a priori selected functions, working well for linear series [18, 19].

Due to such issues, Huang *et al.* [19] developed a new method called the Empirical Mode Decomposition (EMD), which, combined with the Hilbert Spectral Analysis (HSA), allows to analyze non-stationary and non-linear data. EMD decomposes any arbitrary time series into a set of components, called Intrinsic Mode Functions (IMFs), what makes possible to identify different data frequency bandwidths [19]. IMFs reveal important information embedded in the original series [19]. In the same sense, EMD can be used to extract stochastic and deterministic components. A disadvantage of this method is the complexity to combine IMFs related to the stochastic component and those related to the deterministic one to form only two components: one stochastic and one deterministic.

In the context of this work, EMD has been considered as the most useful tool to decompose time series in terms of stochastic and deterministic components, once it is not limited to linear nor stationary time series.

6 Concluding remarks

In this paper we present a wide and systematic research for papers related to time series decomposition. In this study, we used the guidelines provided by the Systematic Literature Review method [22]. This method has been commonly considered by several areas such as Medicine, Computing, and Nursing. This method takes advantage of well-defined phases to look for and analyze papers related to a specific subject. It is important to highlight that this method is extremely useful to start a new study on a specific topic, however it requires researchers to have previous knowledge on that, what is obtained through a first traditional review. This knowledge is important to define some particular criteria for the subject under study, as presented in Section 3.

By applying SLR in this work, we found a set of papers related to the problem of decomposing time series into stochastic and deterministic components. The most related papers were discussed and detailed in this work, but the complete list with all papers was omitted due to space restrictions.

After this analysis, we noticed there are few studies developed to decompose time series, underlining the influences of stochastic and deterministic components in separate. In general, there are lots of studies to smooth or remove noise from time series. Such studies deal with the stochastic component as a noise which can be discarded without causing any influence on modeling. Furthermore, we also found studies developed to decompose time series, however they were mostly directed to specific problems, thus, the established constraints avoid them to be employed on other time series.

Although the found techniques present limitations, answering secondary question SQ.7, there are important studies conducted in this research area. For instance, the method developed by Huang *et al.* [19] can be used to decompose non-stationary and non-linear time series. Moreover, it was also possible to find papers using other important approaches such as Fourier and Wavelet transforms [14], Principal Component Analysis [21], and so on. The fact techniques

present limitations makes evident the need for new researches in this area. Therefore, the decomposition problem is still considered an open problem.

As future work, we plan to extend this systematic review. For this, we intend to consider other keywords to find more important papers and, consequently, more techniques to decompose time series. We also plan to execute our query on other repositories such as Springer Link [2], Journal of Statistical Mechanics: Theory and Experiment [3] and Biometrika [4]. However, we are aware that it will return a large number of papers, which will demand plenty of time and people involvement.

Acknowledgments

This paper is based upon the work supported by CAPES (Coordination for the Improvement of Higher Level – or Education – Personnel) and FAPESP (São Paulo Research Foundation), under the grants 2011/02655-9 and 2009/18293-9. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of CAPES and FAPESP.

References

[1] Alligood, K. T., Sauer, T. D., and Yorke, J. A. *Chaos: An Introduction to Dynamical Systems*. Springer, 1997.

[2] Aminghafari, M., Cheze, N., and Poggi, J.-M. Multivariate denoising using wavelets and principal component analysis. *Comput. Stat. Data Anal.*, 50:2381–2398, May 2006.

[3] Amman, S. and Das, M. An efficient technique for modeling and synthesis of automotive engine sounds. *Industrial Electronics, IEEE Transactions on*, 48(1):225 –234, feb 2001.

[4] Andrew, A. M. Statistical pattern recognition. *Robotica*, 18:219–223, March 2000.

[5] Box, G., Jenkins, G. M., and Reinsel, G. *Time Series Analysis: Forecasting & Control*. Prentice Hall, 3^a edition, February 1994.

[6] Chatfield, C. *The Analysis of Time Series: An Introduction*. CRC Press LLC, 2004.

[7] Chung, P.-J., Viberg, M., and Mecklenbräuker, C. F. Broadband ml estimation under model order uncertainty. *Signal Process.*, 90:1350–1356, May 2010.

[8] D'Alessandro, C., Yegnanarayana, B., and Darsinos, V. Decomposition of speech signals into deterministic and stochastic components. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 760 –763 vol.1, may 1995.

[9] de Lima, E., Andrade, A., Pons, J., Kyberd, P., and Nasuto, S. Empirical mode decomposition: a novel technique for the study of tremor time series. *Medical and Biological Engineering and Computing*, 44:569–582, 2006. 10.1007/s11517-006-0065-x.

[10] Durelli, V. H. S., Felizardo, K. R., and Delamaro, M. E. Systematic mapping study on high-level language virtual machines. In *Virtual Machines and Intermediate Languages*, VMIL '10, pages 4:1–4:6, New York, NY, USA, 2010. ACM.

[11] Dyba, T., Kitchenham, B. A., and Jorgensen, M. Evidence-based software engineering for practitioners. *IEEE Software*, 22:58–65, 2005.

[12] Endo, A. T. and da Silva Simão, A. Formal testing approaches for service-oriented architectures and web services: a systematic review. Technical Report 348, Institute of Mathematic and Computer Sciences, University of São Paulo, March 2010.

[13] Evidence-based Medicine Working Group. Evidence-based medicine - a new approach to teaching the practice of medicine. *The Journal of the American Medical Association (JAMA)*, 268(17):2420–25, 1992.

[14] Graps, A. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2:50–61, 1995.

[15] Gruber, P., Stadlthanner, K., Böhm, M., Theis, F., Lang, E., Tomé, A., Teixeira, A., Puntonet, C., and Saéz, J. G. Denoising using local projective subspace methods. *Neurocomputing*, 69(13-15):1485 – 1501, 2006. Blind Source Separation and Independent Component Analysis - Selected papers from the ICA 2004 meeting, Granada, Spain, Blind Source Separation and Independent Component Analysis.

[2] Springer Link – http://www.springerlink.com/

[3] Journal of Statistical Mechanics – http://iopscience.iop.org/1742-5468

[4] Biometrika – http://biomet.oxfordjournals.org/

[16] Han, M. and Liu, Y. Noise reduction method for chaotic signals based on dual-wavelet and spatial correlation. *Expert Syst. Appl.*, 36:10060–10067, August 2009.

[17] Huang, H.-C. and Cressie, N. Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, 42:262–276, 1998.

[18] Huang, N. E., Chern, C. C., Huang, K., Salvino, L. W., Long, S. R., and Fan, K. L. A New Spectral Representation of Earthquake Data: Hilbert Spectral Analysis of Station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bulletin of the Seismological Society of America*, 91(5):1310–1338, 2001.

[19] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Royal Society of London Proceedings Series A*, 454:903–995, 1998.

[20] ISHII, R. P., RIOS, R. A., and MELLO, R. F. Classification of time series generation processes using experimental tools: a survey and proposal of an automatic and systematic approach. *International Journal of Computational Science and Engineering*, 1:1–21, 2011.

[21] Jolliffe, I. T. Introduction. In *Principal Component Analysis*, Springer Series in Statistics, pages 1–9. Springer New York, 2002.

[22] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. Systematic literature reviews in software engineering - a systematic literature review. *Information and Software Technology*, 51(1):7 – 15, 2009. Special Section - Most Cited Articles in 2002 and Regular Research Papers.

[23] Kopsinis, Y. and McLaughlin, S. Development of emd-based denoising methods inspired by wavelet thresholding. *Signal Processing, IEEE Transactions on*, 57(4):1351 –1362, april 2009.

[24] Lee, T. and Ouarda, T. B. M. J. An emd and pca hybrid approach for separating noise from signal, and signal in climate change detection. *International Journal of Climatology*, 1:n/a–n/a, 2011.

[25] Liszka, L., Pacholczyk, A. G., and Stoeger, W. R. Extraction of a deterministic component from rosat x-ray data using a wavelet transform and the principal component analysis. *Astronomy and Astrophysics*, 354:847–852, 2000.

[26] Liu, Y. and Liao, X. Adaptive chaotic noise reduction method based on dual-lifting wavelet. *Expert Syst. Appl.*, 38:1346–1355, March 2011.

[27] Macciotta, N. P., Cappio-Borlino, A., and Pulina, G. Time series autoregressive integrated moving average modeling of test-day milk yields of dairy ewes. *Journal of dairy science*, 83(5):1094–103, 2000.

[28] Marwan, N., Romano, M., Thiel, M., and Kurths, J. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, January 2007.

[29] Meyer, M. and Stiedl, O. Fractal rigidity by enhanced sympatho-vagal antagonism in heartbeat interval dynamics elicited by central application of corticotropin-releasing factor in mice. *Journal of mathematical biology*, 52(6):830–74, 2006.

[30] Minerva, T. Wavelet filtering for prediction in time series analysis. In *Non-Linear Systems & Wavelet Analysis*, pages 89–94, 2010.

[31] Mitra, A., Kundu, D., and Agrawal, G. Frequency estimation of undamped exponential signals using genetic algorithms. *Computational Statistics & Data Analysis*, 51(3):1965 – 1985, 2006.

[32] Moon, T. and Weissman, T. Universal fir mmse filtering. *Signal Processing, IEEE Transactions on*, 57(3):1068 –1083, march 2009.

[33] Nounou, M. N. Multiscale finite impulse response modeling. *Engineering Applications of Artificial Intelligence*, 19(3):289 – 304, 2006.

[34] Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. Systematic Mapping Studies in Software engineering. In *EASE '08: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, 2008.

[35] Shlens, J. A Tutorial on Principal Component Analysis, 2005.

[36] Shumway, R. H. and Stoffer, D. S. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2^a edition, May 2006.

[37] Small, M., Tse, C. K., and Member, S. Detecting determinism in time series : The method of surrogate data. *IEEE Trans. on Circuits and Systems-I Fundamental Theory and Applications*, 50:663–672, 2003.

[38] Soriano, D. C., Suyama, R., and Attux, R. Blind extraction of chaotic sources from white gaussian noise based on a measure of determinism. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, ICA '09, pages 122–129, Berlin, Heidelberg, 2009. Springer-Verlag.

[39] Tzagkarakis, G., Papadopouli, M., and Tsakalides, P. Singular spectrum analysis of traffic workload in a large-scale wireless lan. In *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, MSWiM '07, pages 99–108, New York, NY, USA, 2007. ACM.

[40] Tzagkarakis, G., Papadopouli, M., and Tsakalides, P. Trend forecasting based on singular spectrum analysis of traffic workload in a large-scale wireless lan. *Performance Evaluation*, 66(3-5):173 – 190, 2009. Modeling and Analysis of Wireless Networks: Selected Papers from MSWiM 2007.

[41] Wangler, B. and Backlund, A. Information systems engineering: What is it? In *CAiSE Workshops (2)*, pages 427–437, 2005.

[42] Wu, Z. H. and Huang, N. E. A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. Lond. A*, 460:1597–1611, 2004.

[43] Zampa, P. and Arnost, R. A new approach to system structure definition. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 7, page 5, October 2002.

A Appendix - Systematic Review Results

This appendix presents some complementary tables considering all papers selected in the first phase of SLR. After reading in detail each selected paper, we organized the most relevant ones in 6 topics: Context, Techniques, Method, Goal, Evaluation, and Score. All the resultant tables are presented next.

Table 8: Frequency distribution of authors per country, considering all selected papers. This table only presents the first 20 countries.

	Frequency distribution
US - United States	559
UK - United Kingdom	134
IT - Italy	112
CN - China	105
DE - Germany	103
FR - France	89
CA - Canada	66
ES - Spain	61
AU - Australia	59
BE - Belgium	38
IN - India	36
BR - Brazil	32
NL - Netherlands	31
JP - Japan	26
SE - Sweden	26
TW - Taiwan	26
KR - Korea (South)	24
GR - Greece	23
IR - Iran	22
TR - Turkey	20

Table 9: Frequency distribution of authors per country, considering papers selected according to the inclusion criteria. This table only presents the first 20 countries.

	Frequency distribution
US - United States	45
CN - China	22
UK - United Kingdom	17
DE - Germany	9
BR - Brazil	8
IN - India	8
IT - Italy	8
PL - Poland	7
FR - France	6
SE - Sweden	6
ES - Spain	5
PK - Pakistan	4
PT - Portugal	4
CH - Switzerland	3
FI - Finland	3
GR - Greece	3
IR - Iran	3
KR - Korea (South)	3
MY - Malaysia	3
RO - Romania	3

Table 10: Number of published papers per year, considering the ones selected according to the inclusion criteria.

	Frequency distribution
1992	1
1995	2
1996	1
1997	2
1998	1
1999	2
2000	3
2001	2
2002	2
2003	1
2004	4
2005	8
2006	8
2007	6
2008	5
2009	14
2010	13
2011	2

Table 11: Number of published papers per year, considering all selected papers.

	Frequency distribution
1974	1
1979	1
1981	1
1987	4
1988	1
1989	2
1991	2
1992	2
1994	3
1995	5
1996	3
1997	10
1998	9
1999	9
2000	14
2001	17
2002	18
2003	22
2004	26
2005	60
2006	66
2007	107
2008	97
2009	146
2010	94
2011	29

Table 12: Relation between the number of papers and the publication type, considering papers selected according to the inclusion criteria.

	Frequency distribution
Journal	65
Conference	10
Book Chapter	1
Symposium	1

Table 13: Relation between the number of papers and the publication type, considering all selected papers.

	Frequency distribution
Journal	624
Conference	105
Workshop	4
Book Chapter	4
Symposium	5
Technical Report	4
Newsletter	2
Thesis/Dissertation	1

Table 14: Type of published papers, considering papers selected according to the inclusion criteria.

	Frequency distribution
Primary Study	70
Application	5
Survey/SLR	2

Table 15: Type of published papers, considering all selected papers.

	Frequency distribution
Primary Study	419
Application	287
Survey/SLR	43

Table 16: Goals of publications, considering papers selected according to the inclusion criteria.

	Frequency distribution
Smoothing	4
Denoising	19
Filtering	15
Decomposition	33
Modeling	5
Feature extraction	1

Table 17: Goals of publications, considering all selected papers.

	Frequency distribution
Smoothing	49
Denoising	31
Filtering	42
Segmentation	2
Decomposition	45
Modeling	430
Feature extraction	2
Other	148

Table 18: A summary of all papers collected by the systematic literature review.

Author	Context	Techniques	Method	Goal	Evaluation	Score
[9]	To analyze the influences of the stochastic and deterministic components.	Empirical Mode Decomposition (EMD) and Hilbert Spectrum (HS).	By applying the EMD method, a time series is decomposed into different IMFs. According to authors, the first IMF represents involuntary tremor and the sum of other IMFs represents voluntary movements, that is, the first IMF is the stochastic behavior and the other ones are the deterministic components. Besides this analysis, the HS tool is used to understand the time series behavior over time and frequency domains.	To distinguish tremors from voluntary behavior in people with neurological problems.	Squared error between the expected and predict values.	5.00
[29]	The dynamics of heartbeat interval fluctuations were studied in awake unrestrained mice following the intracerebroventricular application of the neuropeptide corticotropin-releasing factor (CRF).	Delay-vector variance (DVV) analysis, higher-order variability (HOV) analysis, empirical mode decomposition (EMD), multiscale embedding-space decomposition (MESD) and multiexponent multifractal (MEMF) analysis.	The application allowed to understand intrinsic characteristics of the analyzed time series. For instance, the influences of the stochastic and deterministic components are determined by the EMD method, whereas the linearity of data is analyzed using DVV. The application of this technique provides important information on data, improving model accuracy.	The study using these techniques allowed to understand the impact of dynamical neurocardiopathy in living beings, explaining, for example, whether a man has a precipitating factor for the propensity of cardiac arrhythmias or sudden cardiac death.	Visual inspection and statistical tools to calculate error.	4.50
[37]	To remove noise from time series.	Surrogate data and Minimum Description Length (MDL).	Authors use surrogate data and hypothesis test to determine whether the time series was generated by a stochastic or deterministic process. After that, they use MDL to estimate the parameters and models, which are employed to reconstruct data, suppressing noise.	Synthetic time series.	The evaluation is performed by visual inspection and statistical tools.	5.50
[31]	To decompose a time series, underlining the stochastic and deterministic components.	Genetic algorithm (GA).	The proposed methods use GA, with elitism, to obtain least squares and approximate least squares estimates. The proposed method uses the elitist GA stochastic search procedure to locate optima, with respect to M unknown frequencies, of the concentrated likelihood function. The algorithm discovers the unknown frequencies through a sequential M one-dimensional GA for finding components.	Synthetic time series.	Cramér-Rao lower bound.	4.50
[23]	To decompose a time series, underlining the stochastic and deterministic components.	EMD and Wavelets thresholding.	Authors proposed a new decomposition algorithm using EMD and Wavelet thresholding.	The method is used to understand the behavior of bats.	Signal-to-noise ratio (SNR).	5.50
[33]	To model time series according to the influences of the stochastic and deterministic components.	FIR filter.	Authors present a technique to, empirically, construct time series models at multiple scales.	Synthetic time series.	SNR, MSE, and visual Inspection.	5.50
[3]	To decompose a time series, underlining the stochastic and deterministic components.	Fourier transform.	In the first step, a time series is analyzed through SDFT (Synchronous Discrete Fourier Transform) technique to decompose a time series in different frequencies. After that, a set of stochastic pulses are removed using the multipulse excitation technique. Finally, these removed pulses are modeled using the ARX model.	To model automotive engine sounds.	Visual inspection, specific techniques to analysis sound signals, compression rate, MSE and statistical tools.	5.50
[27]	To estimate the model parameters with high accuracy.	Fourier transform and Box&Jenkins models.	In spite of that paper do not propose a technique to decompose a time series, it was considered useful because the estimation of the white noise, which can be used to understand the influences of the stochastic component in a further estimation step, once the Fourier Transform was applied on experiments.	To model and understand the milk production evolution.	MSE.	3.00

[32]	To decompose a time series, underlining the stochastic and deterministic components.	MSE and FIR filter.	A universal filter is proposed, which per-symbol squared error, for every bounded underlying signal, is as small as the best finite-duration impulse response (FIR) filter of a given order. The application of the technique provides good results for linear and non-linear time series.	Synthetic time series.	SNR.	5.50
[39]	To extract components (or features) of time series.	PSA (SSA).	First, authors apply the SSA technique on the analyzed time series. The components extracted in the previous step are characterized as either stochastic or deterministic, and then, they are modeled according to their influences.	To study and model send-and-receive messages in wireless networks.	Statistical tools, visual inspection and amplitude probability density (APD).	6.00
[40]	The paper is an extension of [39].	PSA (SSA).	Authors use the same techniques previously presented in [39], but in this new paper, the obtained model is used as a predictor.	To study and model send-and-receive messages in wireless networks.	Statistical tools, visual inspection, and amplitude probability density (APD).	6.00
[26]	To decompose chaotic and noisy time series.	SSA, gradient decent, and dual-lifitting wavelet.	It is proposed an adaptive chaotic noise reduction method that uses Dual-Lifting Wavelets. In order to reduce noise, the singular spectrum analysis (SSA) and gradient decent algorithm are respectively employed in the analysis of coefficients and details obtained by Dual-Lifting Wavelet Transform.	Time series which represent the changing phenomenon of solar activity.	SNR, RQA, and visual inspection.	5.50
[15]	To reduce the noise presents in time series.	ICA, PCA, and MDL.	Authors propose three different techniques: the first one, Local ICA is used to detect the statistically most interesting signal+noise subspace. The parameters used by the technique are selected using Minimum Description Length (MDL) criterion. In the second technique, authors combine ideas of solving Blind Source Separation (BSS) problems algebraically using a Generalized Eigenvector Decomposition (GEVD) with local projective denoising techniques. In the last one, the decomposition is performed using KPCA-based denoising techniques which have been proven to outperform linear PCA.	To analyze protein chains.	SNR, MSE, Kurtosis, and visual inspection.	4.50
[2]	To reduce the noise presents in time series.	Wavelet and PCA.	It is presented a new soft-threshold and a new filter, which uses PCA to analyze details generated by the application of Wavelets.	fRMI time series.	SNR.	5.50
[7]	To separate unknown signal from a time series.	Broadband ML estimation, and hypothesis test.	It is developed a broadband ML estimation procedure for an unknown numbers of signals. The proposed algorithm computes ML estimates only for the maximally hypothesized number of signals. The test threshold is approximated by the Cornish-Fisher expansion.	To identify different communication channels in wireless networks.	SNR.	4.50
[16]	To decompose noisy time series.	Dual-Wavelet and SSA.	The proposed method determines the optimal decomposition scales. For this, the approximate coefficients are handled by the Singular Spectrum Analysis (SSA). For extracting useful signals, the spatial correlation is used for the analysis of the detail parts in adjacent scales, while an adaptively discriminant factor is introduced for selecting Wavelet coefficients. After noise reduction, the signal is reconstructed by using the average values of coefficients and details, respectively.	Time series about the changing phenomenon of solar activity.	SNR, MSE, and visual inspection.	6.00

[38]	To decompose a time series, underlining the stochastic and deterministic components.	Recurrence plot and ICA.	Authors apply a recurrence analysis in time series and, the resultant textures are analyzed using Independent Component Analysis (ICA) to remove noise from the deterministic signal.	Synthetic time series.	RQA and MSE.	4.00
[42]	To determine whether a time series and its components contain useful information.	EMD.	Authors carried out numerical experiments by decomposing uniformly distributed white noise into IMFs, which were obtain from the application of EMD method. Based on the empirical finding, they have deduced additional analytical expressions for the various statistical properties of white noise. From the properties of the Fourier spectra of IMFs, they derived an equation to state that the product of the energy density of IMFs by the corresponding average period of IMFs is constant.	Southern Oscillation Index (SOI) data.	Visual inspection and statistical tools.	5.50
[18]	To decompose a time series, underlining the stochastic and deterministic components.	EMD and HS.	The method was detailed in Section 5.4.	Earthquake record from station TCU129, at Chi-Chi, Taiwan.	Visual inspection and statistical tools.	6.00
[24]	To decompose a time series, underlining the stochastic and deterministic components.	EMD and PCA.	Initially, the high-frequency components of climate datasets induced from the random noise process are extracted by EMD analysis, and then, statistical detection techniques are applied to remove high-frequency random components.	To detect climate change signal (global surface temperature anomaly data - GSTA).	Visual inspection and tools.	5.50
[17]	To decompose a time series, underlining the stochastic and deterministic components.	Wavelet Transform.	Authors propose a new Wavelet shrinkage method for recovering unknown piecewise-smooth signals from noisy data. For this, authors assume that the observed process can be decomposed into a large-scale deterministic trend plus a zero-mean Gaussian stochastic process plus noise.	Synthetic time series.	MSE and statistical tools.	5.00
[25]	To extract the deterministic components from a noisy time series.	Wavelet Transform and PCA.	The method uses Wavelet Transform to analyze time series. After that, authors use PCA to define a good threshold to be applied on Wavelet coefficients and, consequently, to extract the deterministic components from a noisy time series.	To extract the deterministic component in the X-ray photon series from Seyfert 1 galaxies and quasars.	Visual inspection.	5.00
[8]	To extract the deterministic components from a noisy time series.	Fourier Transform and Linear Predictive Analysis.	Initially, the methods execute a separation of speech into an approximate excitation and filter components using Linear Predictive (LP) analysis; After that, it executes an identification of frequency regions of noise and deterministic components of excitation using Cepstrum; Then, the two excitation components of the residual are reconstructed using an iterative algorithm; Finally, the deterministic and stochastic components of the excitation are then obtained by combining the reconstructed frames of data using an overlap-add procedure.	To decompose several of natural and synthetic speech signals.	Visual inspection.	4.50
[30]	To separate a time series (also referred to as signal) into its main components.	Wavelet Transform.	The proposed decomposition approach is based on a thresholding rule related to the detail level. This thresholding method allows to obtain a signal decomposition based on frequency separation, to identify stochastic components and to recognize several sources of deterministic components.	To model and understand the behavior of financial time series in order to execute long-term predictions.	Visual inspection.	5.00