European Commission

JRC TECHNICAL REPORTS

Identification of outliers in simultaneous time series using Wavelets and Forward Search

Christophe Damerval

2012

Report EUR 25493 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Christophe Damerval
Address: Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
E-mail: christophe.damerval@jrc.ec.europa.eu or spyros.arsenis@jrc.ec.europa.eu
Fax: +39 0332 78 5154

http://theseus.jrc.ec.europa.eu
http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission
is responsible for the use which might be made of this publication.

JRC72903

EUR 25493 EN

ISBN 978-92-79-26269-2

ISSN 1831-9424

doi:10.2788/47188

Luxembourg: Publications Office of the European Union, 2012

Printed in Italy

Identification of outliers in simultaneous time series using Wavelets and Forward Search

C. Damerval

Abstract

In this paper we present an original methodology for the processing of simultaneous time series. It allows to identify times at which two time series present different patterns – these being outliers in a certain sense. We first apply wavelet techniques to transform data into representations focusing on singular locations. Several approaches are proposed to obtain such representations of the data. Then we explore these with a modern tool in robust statistics, the Forward Search, which allows in particular to evidence outliers.

1 Introduction

The processing of time series is a wide topic of interest for a large scientific community (researchers, engineers, practionners). Many issues are raised by the variety of applications (sensored data, manually-entered) and the different settings (sampling regularity, stationarity, observed variables). Here we are interested in the detection of anomalies in simultaneous time series, for which several variables can be observed at each time

$$x_t^j \in \mathbb{R} \quad t \geq 0, j = 1..J \tag{1}$$

This problem is motivated by a particular application: the monitoring of trade data, with an emphasis on the identification of abnormal behavior. First let us precise that trade data report exchanges of products between two countries, in terms of quantity and value at different times (for instance: tons and euros, monthly). Such data are available from several sources, let us cite in particular the databases Eurostat COMEXT and UN COMTRADE. Now, let us point out that the identification of abnormal behavior in such data is related to an important EU policy issue to which the Joint Research Centre contributes: the fight against fraud and money-laundering. The challenges associated to this anomaly detection lie in the quality of the reported trade data, its heterogeneity,the proper definition of the anomalies one is looking for, and also the possible interpretation of the detected anomalies in terms of fraud. In this regard, let us mention concrete examples of such fraudulent behavior. A first example is an artificially low-priced product, so as to avoid customs tax. This results in discrepancies in a scatterplot representing V versus Q. A second example is the systematic underpricing of a product, which is a generalization of low-price outliers, with the difference it bears on a greater area (one country for example). This results in A third example is the deflection of trade: the trade of one product on a give destination changes origin, so as to avoid country-specific rules. This corresponds to a downward level-shift of quantity for one country and an upward level-shift of quantity for one country (the latter occurring slightly after the former).

Let us now present different points of view for detecting anomalies in trade data. A first approach for this task consists in considering data as records of quantity and value (thus ignoring the temporal aspect), and then apply robust regression techniques. The validity of this approach comes from the proportionality between quantity and value: a value divided by a quantity corresponds to a price, which should be stable under certain assumptions.

A second approach consists in considering data as time series (associated to quantity for instance), and then apply specific algorithms to detect singularities (spikes, level-shifts). This approach turns out as interesting for trade subject to a certain stability (seasonality, steady trend). In this case the occurrence of unexpected patterns can be efficiently detected. As a novel approach, we propose to use both aspects. We first extract features from time series (several approaches being possible), and then explore the space of features to identify anomalies.

2 Tools and algorithms

We present here the tools we use in our approach: Wavelets (WL) and Forward Search (FS). The data we consider are uniformly sampled time series

$$(x_i)_{i=1..n} \ , \ x_i \text{ value associated to instant } i \tag{2}$$

2.1 Wavelet analysis (WL)

Wavelets are a powerful tool for time-frequency analysis. The wavelet framework includes mathematical properties and efficient algorithms. Here we focus on wavelet analysis to extract features from time series. In particular we use a measure of regularity, which can characterize sharp changes such as spikes or clear level-shifts. We will use these tools to obtain other representations of simultaneous time series. Their relevance come from the fact they quantify how regular or irregular are time series. Later we first use wavelet analysis to extract features from time series, and then apply the Forward Search to evidence outliers in the space of features.

Wavelet transform

The Continuous Wavelet Transform (CWT) of a real function $f : \mathbb{R} \to \mathbb{R}$ using a wavelet function $\psi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\forall u \in \mathbb{R}, \forall s > 0 \quad Wf(u, s) = \frac{1}{\sqrt{s}} \int_{\mathbb{R}} f(t) \psi \left(\frac{t-u}{s} \right) dt \tag{3}$$

While f is the analyzed function (the data), the wavelet ψ is an analyzing function whose choice mainly depends on the application. The CWT defined on $\mathbb{R} \times \mathbb{R}_+^*$ consists in a scale-space representation, which gives multi-scale information on the analyzed function – see [7, 8] for details. When dealing with numerical data $(x_i)_{i=1..n}$, algorithms based on Fast Fourier Transforms (FFT) provide efficient computations of the CWT at any chosen scale $s > 0$, see [10]. Let us denote $(c_i)_{i=1..n}$ the computed values of the CWT (wavelet coefficients) of the data $(x_i)_{t=i..n}$ at locations $i = 1..n$ and scale $s = 1$. We denote $w_i = |c_i|$ their absolute value.

The CWT can be used to detect singularities, thanks to its modulus maxima (MM). These MM are defined as locations at which the response $u \mapsto |W(u, s)|$ attains a local maximum (in time, at fixed scale $s > 0$):

$$\mathcal{MM}(s) = \{t \in \mathbb{R}, |Wf(., s)| \text{ locally maximum at } t\} \tag{4}$$

To precisely identify the singularities, we focus on the finest scale $s = 1$. Concerning numerical data, let us define the set of MM as

$$MM = \{i \in 1..n, (w_{i-1} \leq w_i > w_{i+1}) \text{ or } (w_{i-1} < w_i \geq w_{i+1})\} \tag{5}$$

(using the convention $w_0 = 0$ and $w_{n+1} = 0$). Wavelet theory shows these locations correspond to singularities such as a spike or a clear level-shift. We note that in practice (real or simulated data), this set contains roughly $n/3$ elements (for a time series having n observations). This set is rarely empty: when using an appropriate wavelet (such as the Sombrero wavelet we use) such a case occurs only for constant time series – see annex for details. In conclusion this set contains relevant information on time series.

Regularity estimation

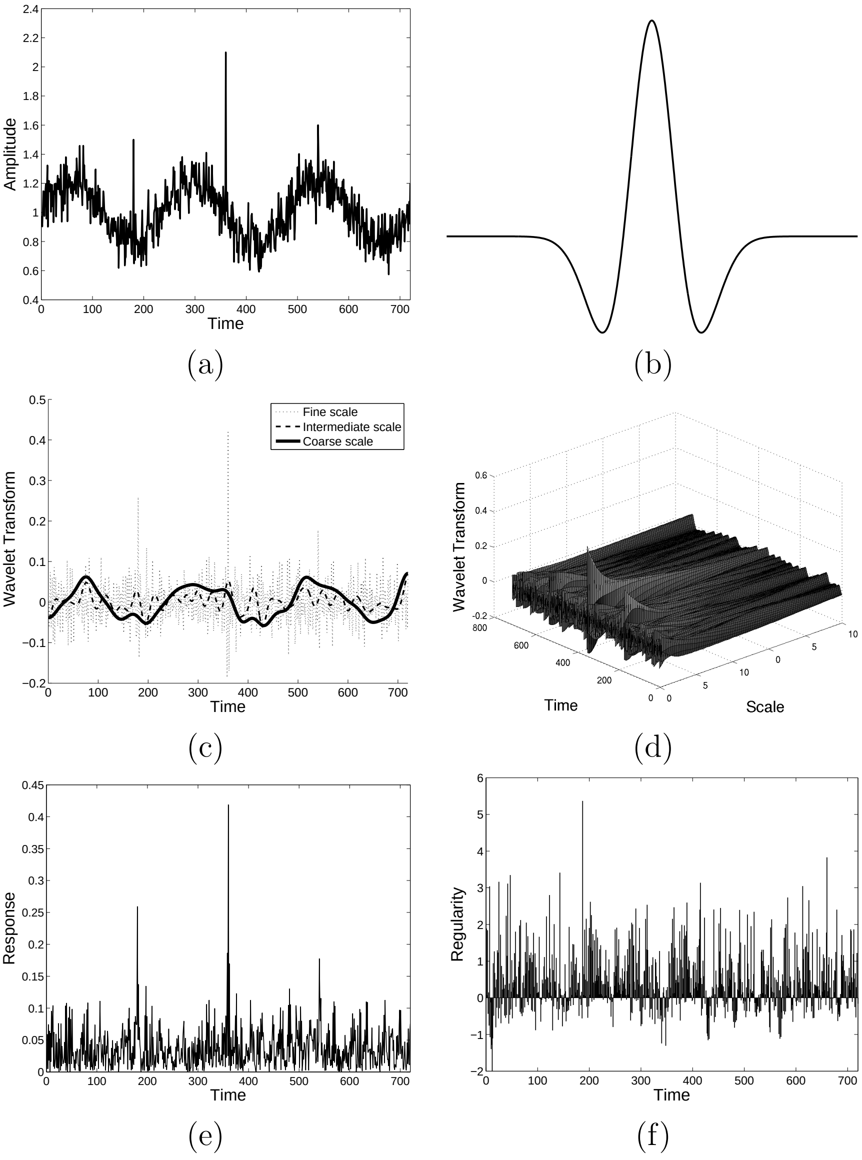
An appealing aspect of wavelet analysis lies in its ability to measure how regular or irregular is a a time series. This is performed through numerical estimation of the Lipschitz regularity exponent at certain instants [6, 4]. First let us recall this exponent of regularity is the value $\alpha \in \mathbb{R}$ appearing in the expression

$$|f(t) - f(t_0)| \leq C|t - t_0|^{\alpha} \qquad (6)$$

where $f : \mathbb{R} \to \mathbb{R}$, $t_0 \in \mathbb{R}$ and $C > 0$, for all t in a neighborhood of t_0. Wavelet methods allow to compute numerically this regularity, performing a linear regression at fine scales using the formula

$$\log W(u, s) = \alpha \log s + D \quad (D \in \mathbb{R}) \qquad (7)$$

In practice the estimated value of regularity allows to quantify how regular or singular a pattern is: this value indicates the sharpness of a spike. This comes from the fact the regularity α is a robust characteristic value [1, 5]. A positive value denotes a regular pattern (like a smooth evolution), a value close to zero a level-shift (like a Heaviside step) and a value close to -1 a spike (like a Dirac impulse). In practice this value of regularity is generally comprised between -2 and 2 (large positive or negative values are uncommon). The estimation through eq.(7) is possible at any instant. Furthermore it turns as very precise and also robust to noise when computed at modulus maxima (MM) seen in the preceding section. Since MM correspond to singularities, the computed values of regularity are often negative. In the following we focus on regularity estimated at MM.
Illustration. We represent on Figure 1 one time series, the Sombrero wavelet $\psi(t) = (1 - t^2) \exp{(t^2/2)}$ (used in all our experiments), the CWT associated to this time series, and the corresponding singularities (MM) with their estimated regularity α.



(a) (b)

(c) (d)

(e) (f)

Figure 1: Wavelet analysis of time series: (a) Time series; (b) Sombrero wavelet; (c) CWT $u \mapsto Wf(u,s)$ at different fixed scales (d) CWT $(u,s) \mapsto W(u,s)$ in 3D representation; (e) Wavelet response at fine scale $u \mapsto |W(u,s=1)|$; (f) Estimated regularity α.

2.2 Forward Search (FS)

The Forward Search (FS) is an adaptive approach allowing to perform important tasks in robust statistics [2, 3, 9]. The FS allows to perform robust regression and to evidence outliers. Besides, a measure of outlyingness can be computed for every outlier, thus indicating its strength. Here we focus on the univariate case, considering data

$$(x_i, y_i)_{i=1..N} \ , \ \ x_i, y_i \in \mathbb{R} \tag{8}$$

and a linear model

$$y_i = ax_i + \varepsilon_i \qquad \varepsilon_i : \mathcal{N}(0, \sigma^2) \tag{9}$$

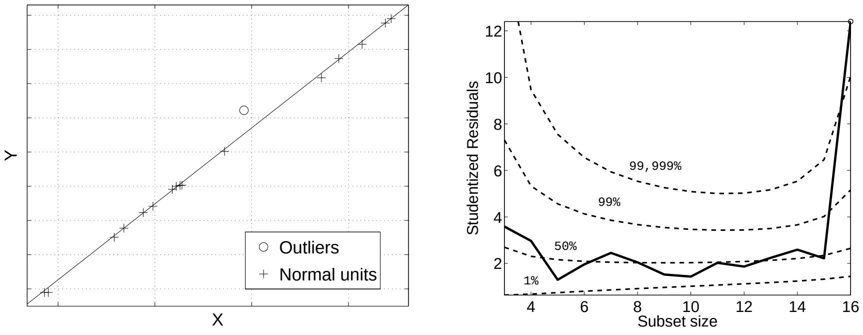
Let us recall in a general context the main steps of the Forward Search [2].

1. **Initialization**: choose an initial subset, ideally free of outliers. This can be done using methods such as least median of squares (LMS) or least trimmed squares (LTS).

2. **Iteration**: starting from the initial subset, increase its size by adding progressively observations (and sometimes removing some). This is performed by selecting those corresponding to the smallest squared residuals.

3. **Monitoring**: during the search, compute the evolution curve of standardized residuals with respect to subset size. This curve should be inside two known envelopes when the subset contains normal units, and outside when outliers enter the subset (the envelopes depending on the nominal test size of FS, typically 0.01).

Identifying outliers. The FS allows to divide data into two subsets: normal units on which a classical linear regression can be fitted, and outliers. A strong point of the FS lies in its ability to order the data, from the most normal units until the most salient outliers. The degree of sensitivity of FS concerning outlier detection depends on a parameter chosen by the user: the nominal test size $ts \in [0, 1]$ (typical value $ts = 0.01$). Let us recall this notion: considering a large number of datasets free of outliers, the FS will identify outliers in a fraction of them, on average equal to ts. For instance if we choose $ts = 0.01$, on average 1% of these datasets will be identified



(a) (b)

Figure 2: Forward Search applied to univariate data: (a) Data made of 17 observations, containing one clear outlier. The FS evidences 16 normal units and one outlier. A linear regression (b) Monitoring of the Forward Search: evolution of the studentized residuals with respect to subset size. The dotted curves are envelopes associated to different test sizes. When a residual surpasses the 99% curve, the corresponding observation is detected as an outlier.

as containing at least one outlier. Let us denote the set of outliers (in one dataset $(x_i, y_i)_{i=1..N}$) as

$$\mathcal{O} = \{(x_k, y_k) \text{ identified as outliers by FS using a test size } ts \in [0,1]\} \quad (10)$$

Computing a measure of outlyingness. Note the parameter ts influences the number of detected outliers, but does not quantify their strength. Given one outlier identified by FS, let us carry out the following operations: first fit a linear regression using normal units; second compute the deletion residual on this outlier (see [2] for details); third perform a statistical test (Student's t-test), leading to a p-value in $[0, 1]$. This value measures the outlier strength, a value close to zero indicating a strong outlier.

Illustration. We represent on Figure 2 results of the Forward Search on an univariate case. We mention the Forward Search algorithm is efficiently implemented in the Matlab toolbox Forward Search and Data Analysis (FSDA). This toolbox provides many function in robust statistics, along with dynamic visualization tools. It is available at the url `http://fsda.jrc.ec.europa.eu`.

3 Extracting features from simultaneous time series

Here we present several approaches that represent the data in a space made of features extracted from time series – these features will be further analyzed with the Forward Search. These approaches can be applied to simultaneous time series: when data consists of entities associated to at least two time series. Here we focus on the case of two simultaneous time series. This is especially relevant for the analysis of EU trade data, for which we observe the evolution over time of quantity $Q_t^k \geq 0$ and value $V_t^k \geq 0$ (k=1..N, N: number of POD, and t=1..n, n: number of observations per POD) given for a great number of entities called POD (see section 1)

$$(Q_t^k, V_t^k)_{k=1..N}, \quad Q_t^k \in \mathbb{R}, V_t^k \in \mathbb{R} \tag{11}$$

These approaches rely on wavelet analysis and regularity estimation, focusing on the significant singularities of the time series. Let us define the set of singularities as

$$S = \left\{ i \in \text{MM and } w_i > 3 \ \frac{\text{MAD}}{0.6745} \right\} \tag{12}$$

where MM are the modulus maxima as seen in eq.(5) and MAD is the median absolute value of $(w_i)_{i=1..N}$. This means we focus on the singularities that surpass the noise level. The main idea lies in identifying all time instants at which there is a significant singularity in either $Q_.^k$ or $V_.^k$, and then compute the same quantity on both time series. Let us explore now different possibilities for extracting features of one entity (one POD in the case of trade data). In this regard, let us consider the wavelet coefficients and values of regularity computed at each time instant

$$w_i(Q), w_i(V) \in \mathbb{R} \ \ (i = 1..n) \tag{13}$$

$$\alpha_i(Q), \alpha_i(V) \in \mathbb{R} \ \ (i = 1..n) \tag{14}$$

and the set of singularities

$$S_Q, S_V \subset \{1..n\} \tag{15}$$

associated to time series representing Q and V respectively.

3.1 Approach using wavelet coefficients

A first approach consists in using wavelet coefficients, so that for each entity (POD) we extract

$$(w_i(Q), w_i(V)) \quad i \in S_Q \cup S_V \tag{16}$$

In this representation, singularities of equal (or similar) strengths and same direction (upward or downward) will lead to equal (or similar) wavelet coefficients. On the contrary, an upward spike in Q and a downward spike in V will result in positive and negative wavelet coefficients. Note also that for one spike in Q and one level shift in V (in the same direction) the features extracted $w_i(Q)$ and $w_i(V)$ will be similar. In the case of a significant singularity in V and a regular pattern in Q (at the same time instant), this leads to a high value of $|w_i(Q)|$ and a low value of $|w_i(V)|$. In summary this approach focuses on the strength of singularities present in Q and V.

3.2 Approach using normalized wavelet coefficients

A variant of the preceding approach consists in normalizing wavelet coefficients, defining features of one entity (POD) as

$$\left(\frac{w_i(Q)}{\sum_i |w_i(Q)|}, \frac{w_i(V)}{\sum_i |w_i(V)|}\right) \quad i \in S_Q \cup S_V \tag{17}$$

Naturally these features bear similar properties compared to the preceding ones: it focuses on singularities present in the time series representing Q and V. A consequence of this normalization

Another consequence of this is that if Q and V present the same patterns but at different scales, these features are equal. On the contrary, if Q presents one strong singularity while V present only weak ones, then we will obtain a high value for $|\frac{w_i(Q)}{\sum_i |w_i(Q)|}$ and a low value for $\frac{w_i(V)}{\sum_i |w_i(V)|}$. As opposed to the previously defined features, all entities (POD) are treated equally: should the time series associated to Q and V be multiplied by a constant factor, these features will remain the same. In summary this approach focuses on the relative strength of singular patterns compared to the other singularities present in the same time series (either Q and V).

3.3 Approach using regularity at singular locations

Here we use the regularity α seen in section 2.1. While wavelet coefficients quantify the strength of singularities, the regularity α quantifies their sharpness. We define the following features

$$(\alpha_i(Q), \alpha_i(V)) \ \ i \in S_Q \cup S_V \tag{18}$$

This representation focuses on the regularity computed only at singular locations. We mention that the regularity estimation is more precise at these singular locations (compared to regular locations).

3.4 Approach using one feature per time series

Here we define a novel feature based on wavelets, allowing to summarize one time series into a real number. Having identified the set of wavelet modulus maxima MM, we compute values of regularity $(\alpha_i)_{i \in MM}$ and wavelet coefficients $(w_i)_{i \in MM}$. Let us define the feature $M_\alpha \in \mathbb{R}$ as

$$M_\alpha = \frac{1}{\sum\limits_{i \in S} |w_i|} \sum_{i \in S} |w_i|(1 - \alpha_i) \tag{19}$$

when the set MM is non-empty and $M_\alpha = 0$ otherwise. The relevance of this feature lies in the fact that it measures the irregularities within a time series, weighed by their relative strength: typically $M_\alpha > 1$ for a time series with clear singularities, $M_\alpha < 1$ for a time series with smooth variations. Considering data as in eq.(11), let us denote $M_\alpha(X, k)$ and $M_\alpha(Y, k)$ the features corresponding to the time series (Q_t^k) and (V_t^k) respectively ($k \in 1..N$) – computed using eq.(19). This allows to transform the data in eq.(11) into new data made of wavelet features

$$(M_\alpha(Q), M_\alpha(V)) \tag{20}$$

Important note. In the context of EU trade data, if quantity and value over time vary in a similar way, then the associated features will be similar. However, a discrepancy between the two time series can be considered as a signal of abnormal behavior. This motivates the use of regression techniques and outlier detection methods such as the Forward Search.

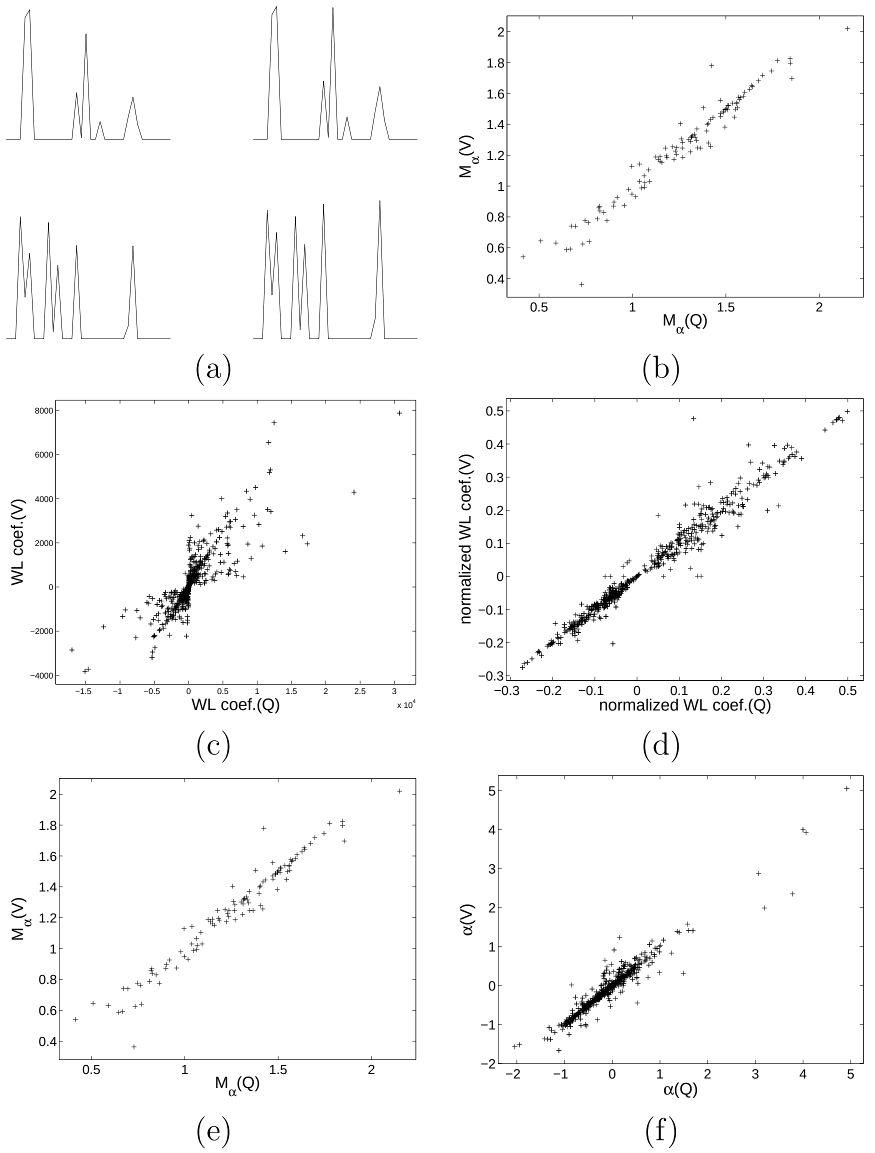
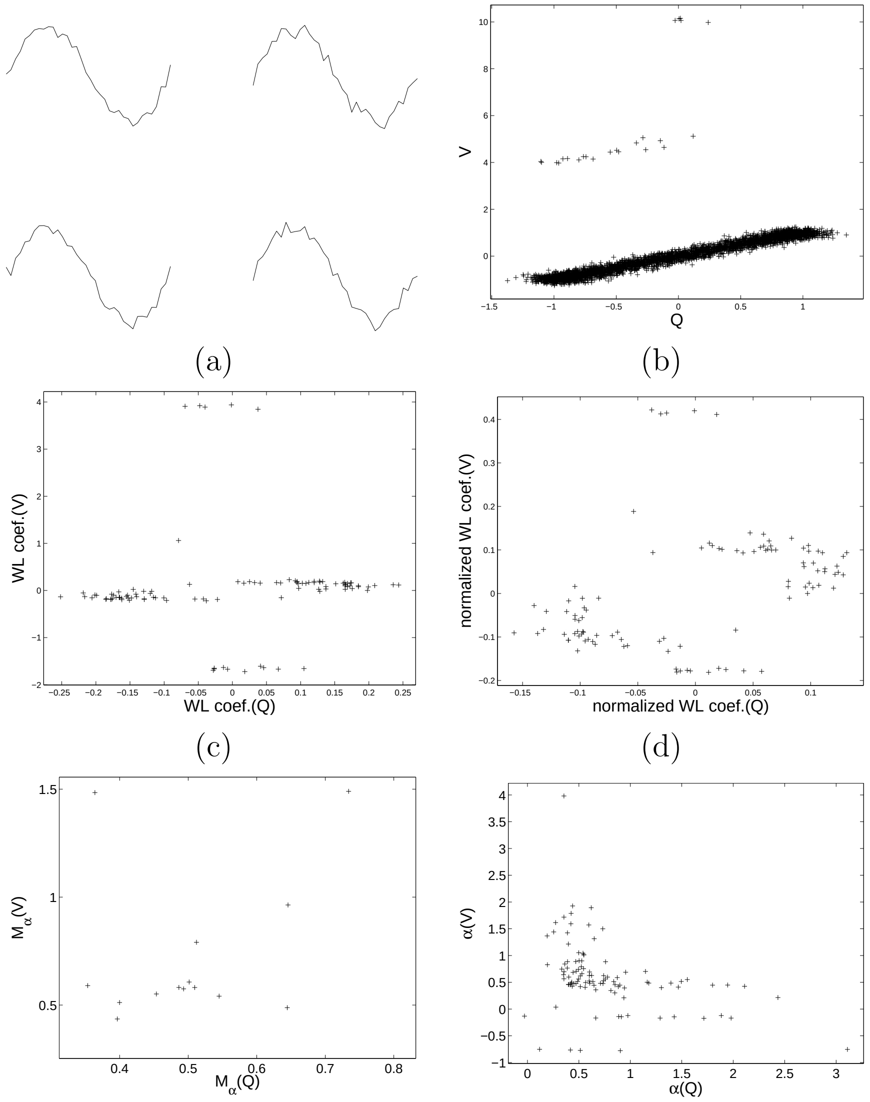


Figure 3: Illustration on real data. (a) Time series representing quantity Q (left) and value V (right) for two real POD (top/bottom); scatterplots representing: (b) V vs Q; (c) wavelet coefficients of Q vs those of V; (d) normalized wavelet coefficients of Q vs those of V; (e) composite measure of Q vs the one of V; (f) values of regularity α corresponding to $t \mapsto Q_t$ vs those corresponding to $t \mapsto V_t$, computed at locations associated to a singularity either of $t \mapsto Q_t$ or of $t \mapsto V_t$.



(a) (b)

(c) (d)

(e) (f)

Figure 4: Illustration on simulated data. (a) Time series representing quantity Q (left) and value V (right) for two simulated POD (top/bottom); scatterplots representing: (b) V vs Q; (c) wavelet coefficients of Q vs those of V; (d) normalized wavelet coefficients of Q vs those of V; (e) composite measure of Q vs the one of V; (f) values of regularity α corresponding to $t \mapsto Q_t$ vs those corresponding to $t \mapsto V_t$, computed at locations associated to a singularity either of $t \mapsto Q_t$ or of $t \mapsto V_t$.

4 Identifying outliers with Forward Search

The Forward Search (FS) proved its efficiency to identify outliers in complex data, especially when there are different groups of outliers. Let us present how we apply the FS to the features extracted with wavelets.

1. Extract features $(a_i, b_i)_{i=1..N}$ with wavelets from two time series $(x_i, y_i)_{i=1..n}$
2. Perform FS on the set $\mathcal{F} = (a_i, b_i)_{i=1..N}$ using a nominal test size ts ($ts \in [0, 1]$, typically $ts = 0.01$). This evidences two distinct subsets: one of normal units, denoted $\mathcal{N}$; another made of outliers, denoted $\mathcal{O}$.
$$\mathcal{F} = \mathcal{N} \cup \mathcal{O} \qquad\qquad \mathcal{N} \cap \mathcal{O} = \varnothing \tag{21}$$
3. Concerning the set of normal units $\mathcal{N}$, perform a linear regression (ordinary least squares). This allows to assess the homogeneity of the whole dataset.
4. Concerning the set of outliers $\mathcal{O}$, compute for each detected outlier the normalized residual and corresponding p-value of a t-test. This allows to quantify the outlyingness of each outlier.
$$\begin{cases} \text{Outlier} & : \ (x_k, y_k) \ \ (k \in 1..N) \\ \text{Fitted value} & : \ (x_k^{fit}, y_k^{fit}) \\ \text{Residual} & : \ r_k = x_k - x_k^{fit} \\ \text{p-value} & : \ p_k \in [0, 1] \end{cases} \tag{22}$$

Note 1: an alternative approach consists in applying the FS directly on the data seen in eq.(11), considering them as univariate data observations of size $n \times N$ (n: number of observations per time series, N: number of entities associated with two time series). Although this allows to correctly detect outliers in scatterplot representations, it does not take into account the temporal aspect – whereas our approach does. We mention another advantage of our approach concerning the processing of massive datasets: while the FS can be computationally expensive, the WL computations are extremely fast in practice. As a consequence, the computational time of our approach is clearly reduced compared to the alternative one.

Note 2: this approach presented for two simultaneous time series can be generalized to the case of several simultaneous time series (using FS in the multivariate case).

5 Conclusion

In this paper we presented an overview concerning the extraction of features from time series. We presented several approaches based on wavelet analysis. These provide an alternative representation of data consisting in simultaneous time series. Considering the case of two simultaneous time series, we presented the obtained features from simulated and real data. In this case, these are univariate data, which can be further processed by regression techniques. Finally we described a methodology based on the Forward Search, allowing to evidence outliers in simultaneaous time series. Future works will allow to better understand how robust regression techniques can highlight specific outliers in simultanenous time series.

References

[1] Patrik Andersson. Characterization of pointwise hlder regularity. *Applied and Computational Harmonic Analysis*, 4(4):429–443, 1997.

[2] AC Atkinson, A Cerioli, and M Riani. *Exploring Multivariate Data With the Forward Search Springer Series in Statistics.* Springer, 2004.

[3] AC Atkinson and M Riani. *Robust Diagnostic Regression Analysis.* Computational Statistics, 2000.

[4] A. Benassi, S. Cohen, J. Istas, and S. Jaffard. Identification of filtered white noises. *Stochastic Process. Appl.*, 75(1):31–49, 1998.

[5] C. Damerval and S. Meignen. Study of a robust feature: The pointwise lipschitz regularity. *International Journal of Computer Vision*, 88(3):363–381, 2009.

[6] S. Jaffard and Y. Meyer. Wavelet methods for pointwise regularity and local oscillations of functions. *American Mathematical Society*, 1996.

[7] S Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[8] S Mallat and WL Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, 1992.

[9] P Rousseeuw and A Leroy. *Robust regression and outlier detection.* Wiley, New York, 1987.

[10] G Strang and T Nguyen. *Wavelets and filter banks.* Wellesley-Cambridge Press, USA, 1996.

European Commission

EUR 25493 EN Joint Research Centre – Institute for the Protection and Security of the Citizen

Title: **Identification of outliers in simultaneous time series using Wavelets and Forward Search**

Authors: Christophe Damerval

Luxembourg: Publications Office of the European Union

2012 – 20 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online), ISSN 1018-5593 (print)

ISBN 978-92-79-26269-2

doi:10.2788/47188

Abstract

In this paper we present an original methodology for the processing of simultaneous time series. It allows to identify times at which two time series present different patterns - these being outliers in a certain sense. We first apply wavelet techniques to transform data into representations focusing on singular locations. Several approaches are proposed to obtain such representations of the data. Then we explore these with a modern tool in robust statistics, the Forward Search, which allows in particular to evidence outliers.

LB-NA-25493-EN-N

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.

ISBN 978-92-79-26269-2

9 789279 262692

Publications Office