"The Portuguese pharmaceutical market in the near future – a time series exploration approach"

by

Maria Helena Miranda Flores Baptista

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Mestre em Estatística e Gestão de Informação

(Master in Statistics and Information Management)

by

Instituto Superior de Estatística e Gestão de Informação

da

Universidade Nova de Lisboa

“The Portuguese pharmaceutical market in the near future – a time series exploration approach”

Dissertation Supervised by

Professor Doutor Jorge M. Mendes

July 2008

Acknowledgments

I would like to acknowledge all of those who helped and support me throughout the last few months while I developed the present work.

I would like to start by expressing my gratitude to my supervisor, Prof. Jorge Mendes, for his continuous help, encouragement and valuable inputs.

My sincere thanks to my family for their support and understanding, not being there for them during this period, was more frequent than I could predict and would like to have happened.

Last but not least I would like to thank IMS health, and its General Manager, Dr. Carlos Mocho for allowing me to use the data for the present work.

Abstract

Using a novel exploratory technique for time series analysis, Single Spectrum Analysis (SSA), it was possible to develop a comprehensive study of the Portuguese pharmaceutical market.

In the introductory chapter this technique is described in detail, for the decomposition step, homogeneity structure testing and forecasting. A bibliography review was conducted on the technique. To the best of our knowledge this was the first time that SSA was applied to any pharmaceutical market, so it was not possible to compare results with other published work.

A detailed explanation on the Portuguese pharmaceutical market is provided in order to allow comprehensiveness of the work. The Portuguese pharmaceutical market is divided in 15 classes, which aggregates all drugs sold in the country. The technique was applied to those 15 time series plus the “Total Market” time series.

Applying SSA, time series were decomposed in the respective components, which can be described as trend, cyclical movements and seasonality. The structure of all time series was tested for homogeneity. With those steps concluded, a monthly forecast, for the years 2008 and 2009 (with the respective confidence bounds) were produced for all the 16 time series.

As a complex methodology, decisions need to be taken in several steps of the study. Even if not all possible choices are presented in the work, lengthy analyses were done to reach the best possible results. In fact, choosing between possible window lengths, Singular Value Decomposition (SVD) approaches, and eigentriples to be grouped together is sometimes more an “art” than a science; experience and previous knowledge of the actual phenomena can and should help.

For confidentiality reasons the raw data is not provided in this work, but both forecast values and confidence bounds are presented.

Key Words: SSA; Time series; Forecasting; Pharmaceutical Market;

Sumário

Utilizando uma nova técnica exploratória para análise de séries temporais, Single Spectrum Analysis (SSA), foi possível desenvolver um estudo aprofundado do mercado farmacêutico Português.

No capítulo introdutório é descrita, em detalhe, a técnica, para a fase de decomposição, o teste de homogeneidade da estrutura e a previsão. A revisão bibliográfica foi efectuada para a metodologia. Desconhecemos uma aplicação anterior desta técnica a qualquer mercado farmacêutico, pelo que, tendo em princípio sido esta a primeira vez que tal sucedeu, não foi possível comparar os resultados obtidos com outros trabalhos publicados.

Para permitir uma melhor compreensão deste trabalho é apresentada uma explicação detalhada do mercado farmacêutico Português. Este mercado está dividido em 15 classes que agrupam as vendas realizadas pela totalidade das especialidades farmacêuticas existentes. A técnica SSA, foi aplicada a todas as 15 classes, bem como à série temporal “Vendas Totais”.

Aplicando a técnica SSA, as séries temporais foram decompostas nos seus respectivos componentes, que podem ser descritos como tendência, movimentos cíclicos e sazonalidade. Foi testada a homogeneidade da estrutura de cada série temporal. Após concluída esta fase, foram produzidas previsões de vendas por mês, para os anos de 2008 e 2009 (com os respectivos intervalos de confiança) para todas as 16 séries temporais.

Pelo facto de se tratar de uma metodologia complexa, é sempre necessário optar entre múltiplas alternativas nas diversas fases do estudo. Mesmo que todas as diferentes opções não estejam mencionadas no trabalho, uma análise aprofundada foi sempre realizada, para que os melhores resultados fossem atingidos. Na realidade, a escolha entre diversos “tamanhos de janela”, várias abordagens de Decomposição do Valor Singular (DVS), criação de diferentes agrupamentos com diversos “trio-próprio”, é por vezes mais uma “arte” do que uma ciência; a experiência e o conhecimento prévio do fenómeno podem e devem ajudar.

Por razões de confidencialidade os valores das séries temporais não são disponibilizados no trabalho, no entanto, quer os valores de previsão quer os intervalos de confiança estão incluídos.

Palavras-Chave: SSA; Séries temporais; Previsão; Mercado Farmacêutico;

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Time series

A *time series* is defined as a group of observations seen in points or periods in time for a definitive interval, beginning at a specific starting point. In some cases, the observations are equally spaced in time, for example, the data in study, where the time interval is the month. In other cases, the observations occur almost continuously and can be seen as evolving continuously over time, for example, ECG's.

Basically, a time series results from the observation during a determine period of time of a real situation with a stable structure. Consequently the observations are not independent and the temporal order a fundamental aspect to be taken in account.

Usually all records which fall within the field of time series analysis are influenced, at least in part, by sources of random variation, which do not disappear as soon as they happen, but are incorporated in the future development of the phenomenon. Therefore we call the sequence of random variables in time $\{X_t\}_t \in Z$, a stochastic process and its realization $\{x_t\}_t \in Z$ is then a time series.

The main motivators to study a time series are:

- Description, the basic task to better understand the time series;
- Explanation, to create the best fitting model;
- Prediction, to predict the future behaviour of the time series;
- Control, to constantly evaluate the stability of the time series.

In all cases the purpose is to create a model that fits the time series adequately. One of the many fitting models is based on the decomposition of the stochastic process $\{X_t\}_t \in Z$, into 4 distinct parts: $X_t = M_t + C_t + S_t + N_t$. The 4 components of the model can be group in two parts, the "*dynamical"* part and the "*random*" part. The first 3 components, which represent the so called *"dynamical"* part, are: trend M_t, cyclical movements C_t and seasonality S_t. The last part is the random variation term, also known as *random noise term or error term (noise)* and describes random fluctuations of the series.

The trend component of the decomposition has an intuitive meaning and can be described as the inertia of the series, the main pathway or the "average" variation throughout time. It comprehends the mild and consistent movements for long periods of time, and can be modelled by a low-order polynomial function.

The cyclical component consists of quasi-periodic functions of varying amplitude and duration, so it is not modelled by simple periodic functions.

The seasonal component explains the periodical behaviour terms and effects which occur regularly over a period with pronounced short-term fluctuations in time series. It can be modelled by a simple periodic function with know period, for example, annual, quarterly, weekly…

Statisticians usually try to explain time series processes from the point of view that there exists some relation (correlation) between successive observations.

1.2. Singular Spectrum Analysis (SSA) and Multichannel Singular Spectrum Analysis (MSSA)

1.2.1. Singular Spectrum Analysis (SSA)

The singular-spectrum analysis (SSA) methodology is a novel exploratory technique of time series analysis incorporating the elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems, and signal processing. It is a nonparametric method.

The main idea is to expand a single univariate or multivariate time series into orthogonal vectors and interpret them by the PCA point of view, using lag-correlation structures. The final purpose is to decompose, by data-adaptive filters, a time series into several components, which usually can be identified as been the trend, seasonality, cycling movements or noise. It generates statistical significance information on these components, and provides a reconstruction of those.

This methodology had is "official" beginning with the publication of papers from Broomhead and King (1986) and further developed by Vautard and Ghil (1989).

The basic version of SSA consists of 4 steps, which are performed as follows:

- **The construction of the trajectory matrix – the embedding step**

In this step the idea is to create a multidimensional series from a one-dimensional series. The dimension of the series is called the *window length*. This multidimensional time series forms the *trajectory matrix.*

Let $F = \left(f_0, f_1, \ldots f_{j,} f_{N-1}\right)$ be a time series of length *N*, and *L* be an integer, which is the "window length", with$1 \prec L \prec n$. The choice of *L* is not obvious and further discussion around it will arise further ahead.

After setting $K = N - L + 1$ and after defining the *K* *L*-lagged vectors $X_j = \left(f_{j-1}, \ldots j_{j+L-2)}\right)^T$, $j = 1,2,\ldots K$, the trajectory matrix is:

$$\mathrm{X} = \left(\mathrm{f}_{\mathrm{i+j-2}}\right)_{i,j=1}^{L,K} = \left[X_1 : \ldots : X_K\right] = \begin{bmatrix} x_1 & x_2 & \ldots & x_k \\ x_2 & x_3 & \ldots & x_{k+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_l & x_{l+1} & \ldots & x_n \end{bmatrix}$$

This trajectory matrix **X** is an Hankel matrix, meaning that all the elements along the diagonal *i*+*j* = constants and are equal.

- **SVD (singular value decomposition)**

Using the PCA theory the singular value decomposition of a matrix **X** is done by calculating the *eigenvalues* and *eigenvectors* of the matrix **S = XX$^{\mathbf{T}}$** of size *L x L*.

The basic SSA might also use *the lag-covariance matrix* $C = S/K$ (the only difference is the magnitude of the corresponding eigenvalues which in **S** are K times larger).

There are several versions to calculate the lag-covariance matrix, with both advantages and disadvantages for each. We will return later to this matter.

The representation of X is then the sum of rank-one biorthogonal matrices X_i *(i = 1, ..., d)*, where *d* (*d*≤ *L*) is the number of nonzero singular values of **X**.

After doing that, a collection of *L* singular values will be found, represented by $\sqrt{\lambda_i}$, which are the square roots of the eigenvalues of the matrix **S**, and the corresponding left and right vectors, represented respectively by U_i and V_i.

The left singular vectors of **X,** U_i, are the orthonormal eigenvectors of **S**, commonly called the "empirical orthogonal functions".

The right singular vectors, V_i, can be seen as the eigenvectors of the matrix X$^{\mathrm{T}}$X.

By considering that $V_i = \dfrac{X^T U_i}{\sqrt{\lambda_i}}, i = 1,...,d$, the SVD of **X** can be written like $X = X_1 + ... + X_d$ where $X_i = \sqrt{\lambda_i U_i V_i^T}$. The **eigentriple** of the SVD is then the collection of $\sqrt{\lambda_i U_i V_i}$.

These two steps form the ***reconstruction stage***. The ***grouping stage*** corresponds to the following two steps.

- **Grouping of matrices**

This step corresponds to splitting the matrices, computed in the previous step $X = X_1 + \ldots + X_d$, into *d* groups from $\{1,\ldots,d\}$ and summing the matrices within *m* disjoints subsets $I_1,\ldots,I_m$. These matrices are computed for $I = I_1,\ldots,I_m$ and the previous decomposition leads to the following decomposition $X = X_{I_1} + \ldots + X_{I_m}$.

This process of choosing the group $I_1,\ldots,I_m$ is called eigentriple grouping. The purpose of this step is to separate the additive components of the time series. The concept of *separability* will be further discussed later.

- **Diagonal averaging**

This step transfers each resultant matrix, which is an additive component of the initial series, into a new time series with dimension *n*. It is a linear operation and maps the trajectory matrix of the initial series into the initial series itself.

This is done by averaging (we will return later to the methodology to do this) over the diagonals *i* + *j* = const of the matrices X_{I_k} obtaining the series $\widetilde{F}^{(k)} = \left(\widetilde{f}_0^{(k)} + \ldots + \widetilde{f}_{n-1}^{(k)}\right)$ and the initial series is decomposed into a sum of *m* series:

$$f_n = \sum_{k=1}^{m} \widetilde{f}_n^{(k)} \;, n = 0, \ldots, N-1$$

This equality only occurs when *m*=*L*. Where for each *k* the series $f_n^{(k)}$ is the result of diagonal averaging of the matrix X_{I_k}.

These *m* time series represent the *m* first principal components.

The general purpose of the SSA analysis is to reach the 4th step with additive components $f_n^{(k)}$ which are "independent" and "identifiable" time series.

This new time series serves the only purpose of analyzing the structure of the time series. As a result we can then have a $f_n^{(k)}$ component that can be identified as the trend of the original series, an oscillatory series or noise. Figure 1 shows the trend identification in the time series B.

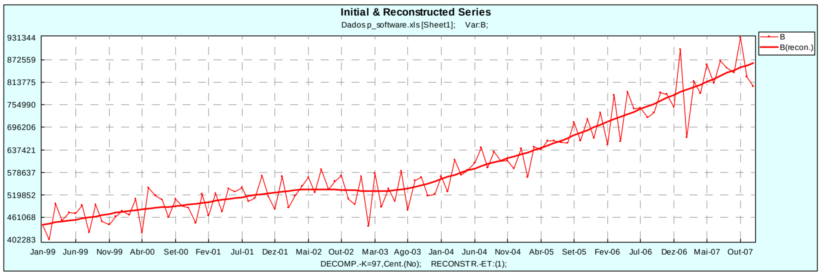


Figure 1 - Time series B – Trend

These components are produced by the series itself (no parametric model is fixed), so it can not be expected to get, in real life series, the components as exact harmonics or linear trend, even if these harmonic or linear trends are present in the series. This is both because of the presence of noise and the non-parametric nature of the method.

The two most important moments in the SSA "world" are:

- **The choice of the "*window length*";**
- **The "*separability*" of the components.**

The "*window length*" is the main parameter of basic SSA, in the sense that its wrong choice would imply that no grouping activities could be performed to obtain a good SSA decomposition.
Have an incorrect "*window length*" can mean that the separability of the components might not occur. This is a critical point since achieving "*independence*" of the components is of fundamental importance to the process.

There are several notions of separability, but the most important is weak separability, defined as:

Provided that the original time series f_n is a sum of *m* series $f_n^{(k)}$ $\left(k=1,\ldots,m\right)$, for a fixed window length *L*, weak *L*-separability means that any subseries of length *L* of the *k*th series $f_n^{(k)}$ is orthogonal to any subseries of length *L* of the *l*th series $f_n^{(l)}$ with $l \neq k$, and the same holds for their subseries of length $K = N - L + 1$.

The only problem is that exact separability rarely happens in practice. Therefore an approximate separability is more important and achievable. Several different characteristics are used to measure the degree of separability.

In fact, if two or more of the singular values of the trajectory matrices $X^{(k)}$ and $X^{(l)}$ corresponding to two different components of $f_n^{(k)}$ and $f_n^{(l)}$ of the original series are equal or close, then the SVD is not uniquely defined and those two series $f_n^{(k)}$ and $f_n^{(l)}$ are mixed up, and an additional analysis is required to separate them. Several options exist for the additional analysis.

1.2.2. SSA forecasting of time series

A forecast can only be build if the model found fits appropriately the data, meaning that the structure of the data was found and is defined by a model. The model can derive from the data or at least can be checked against the data. In SSA forecasting, these models can be described with the help of the **linear recurrent formulae (LRF)**.

The series governed by LRF's admits natural *recurrent continuation* since each term of such a series is equal to a linear combination of several preceding terms.

So, if the original series f_n satisfies a linear recurrent formula $f_n = a_1 f_{n-1} + \ldots + a_d f_{n-d}$ of some dimension *d* with some coefficients $a_1,\ldots,a_d$, then for any *N* and *L* there are at most *d* nonzero singular values in the SVD of the trajectory matrix **X**; therefore, even if the window length *L* and $K = N - L + 1$ are larger than *d*, we only need at most *d* matrices $\mathbf{X}_i$ to reconstruct the series.

If we have a series satisfying a LRF then we can continue it for an arbitrary number of steps using the same LRF.

But there is another way of forecasting with SSA. It is the vector forecasting algorithm. While the recurrent forecasting algorithm explained above performs a recurrent continuation of a one-dimensional series, the vector forecasting algorithm does that by the continuation of the vectors in an *r*-dimensional space and only then returns to the time-series representation. Apparently this option is better for long-term forecasting.

Creating confidence intervals for this forecast is not only needed but is desirable to assess quality. Two methodologies can be used, one by using the recurrent forecast process to forecast the periods already known, and after that comparing the values achieved with the real ones. Assuming that the residual series is stationary and ergodic, quantiles of the related marginal distribution can be estimated and therefore confidence bounds can be created. The second technique is the bootstrap, which uses Monte Carlo simulations to create the confidence intervals.

1.2.3. Multichannel Singular Spectrum Analysis (MSSA)

If instead of having one time series, we have *p* time series, it can be used a modified version of the SSA technique called MSSA (Multichannel SSA).

This methodology allows to correlate not only observation but also to correlate variables (time series).

MSSA is used in the same way that SSA is used, it analyses each time series with *n* observations (assuming that all time series has the same number of observations) until a specific lag *l*, which implies that the covariance matrix has information on interrelations between lagged versions of the original variables as well as between different variables.

The technique can be described as follows.

Consider an *l*-variate time series $f_n = \left(f_n^1, \ldots, f_n^l\right)$, where *n = 0,1,...,N-1.* Then for a fixed window length *L* define the trajectory matrices $X^{(i)}(i = 1, \ldots, l)$ of the one-dimensional time series $f_n^{(i)}$. The trajectory matrix X can be defined as:

$$X = \begin{pmatrix} X^{(1)} \\ \ldots \\ X^{(l)} \end{pmatrix}$$

The lagged matrix $X_{Kxp'}$, where $K = n - L + 1$ and $p' = Lp$ can be seen as:

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,L} & x_{2,1} & \cdots & x_{2,L} & \cdots & x_{p,1} & \cdots & x_{p,L} \\ x_{1,2} & \cdots & x_{1,l+1} & x_{2,2} & \cdots & x_{2,l+1} & \cdots & x_{p,2} & \cdots & x_{p,L+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1,k} & \cdots & x_{1,n} & x_{2,k} & \cdots & x_{2,n} & \cdots & x_{p,k} & \cdots & x_{p,n} \end{bmatrix}$$

The generalization of SSA to a multivariate time series requires the construction of an augmented block-matrix S$_X$, with the dimension *pL x pL*:

$$S_{X=} \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Each **S**$_{kl}$ is the matrix that contains estimates of the lag covariance between *k* and *l*.

All the following steps follow the same theory as in SSA.

1.2.4. Caterpillar SSA

The caterpillar SSA, version 3.30, Professional M Edition, is the software used to develop the present work. All graphs and results presented are coming directly from the software.

This software was developed by the Gista T Group, which is a group of 3 Russian scientists (Nina Golyandina, Vladimir Nekrutkin and Kirill Braulov) working in the University of St. Petersburg since the middle of 1990's on the development of SSA and MSSA and the corresponding software.
The program performs extended analysis, forecasting for both one-dimensional and multi-dimensional time series. The program also performs change-point detection for one-dimensional time series.

There is another very active group of scientist working in UCLA (University of California) the Theoretical Climate Dynamics (TCD) group which developed also software for SSA. This software follows a slightly different approach especially on the forecasting part.

1.3. Bibliography review

This work will mainly follow the work of Golyandina *et al.* (2001) in the book "Analysis of Time Series Structure – SSA and related techniques". The aim of this book is to **explain the methodology and theory of SSA**. The main topics are SSA analysis, SSA forecasting and SSA detection of structural changes.

Broomhead and King (1986) and Broomhead *et al.* (1987) publications were the first ones related with this subject. In fact, Broomhead and King (1986) started by developing a singular system analysis based on the method of delays. The method of delays was introduced initially by Takens (1981). The singular system analysis was developed by Bertero, Pike and co-workers (1982). The method presented is therefore based on the Takens (1981) proof and on the ideas from Bertero, Pike and co-workers. After introducing: a) some of the relevant language of dynamical systems theory, b) the definition of qualitative dynamics, c) the concept of equivalence relations, d) the discussion of Whitney's embedding theorem and e) the review of the method of delay, they developed **a full theoretical approach and have applied it to a time series**, obtained from the Lorenz model.

The first step of the SSA method is called the embedding step. Embedding can be regarded as a mapping that transfers a one-dimensional time series to a multidimensional series. The theoretical justification of data embedding techniques used by experimentalists to reconstruct dynamical information from time series is provided in a paper of Sauer *et al* (1991), expanding on the work of Whitney (1936) and Takens (1981).

Vautard and Ghil (1989) developed further the Singular-spectrum analysis (SSA). They refined certain aspects of its application, such as the influence of the window size, sampling interval, and length of the sample on the results of SSA. One of the objectives of this paper was to **explore fully the potential of SSA in studying the dynamics recorded in the data**. All the series considered here were zero-mean, continuous, infinite and ergodic. In this article authors investigated the properties of SSA first for simple phenomena, such as pure oscillation, a red-noise process, or the Lorenz system. Next, they applied it to four paleoclimatic records from the Quaternary. They concluded that SSA is a powerful descriptive tool for nonlinear dynamics in general and climate dynamics in particular.

A good example of the capabilities of the SSA method is given by Ghil and Vautard (1991) when they used it to analyze the time series of global surface air temperatures for the past 135 years, allowing a secular warming trend and a small number of oscillatory modes to be separated from the noise.

Vautard *et al.* (1992) continued the work in this area and proved that **SSA works well for short, noisy time series**. In this article SSA is combined with advanced spectral-analysis methods – the maximum entropy method (MEM) and the multi-taper method (MTM) – to refine the interpretation of oscillatory behavior. A combined SSA-MEM method is also used for the prediction of selected subsets of RC's (Reconstructed components). They have proven that SSA extracts as much reliable information as possible from short and noisy time series without using prior knowledge about the underlying physics or biology of the system; based on this information, it also provides prediction models. The superiority of this method over classical spectral methods lies in the data-adaptive character of the eigenelements it is based on. They have also proved that SSA can provide useful physical insight and modest, but unprecedented, medium-term predictive skills starting with the few hundred data points typically available for geophysical and other natural systems. All the work is based on the assumption that the process *x* under study is stationary in the weak sense, i.e., that the second order moments are invariant under translation. One of the results of this paper is the presentation of a **particular method of estimating of the Toeplitz matrix** shown to have little bias compared to other estimates.

Plaut and Vautard (1994) used successfully not the SSA method but the **Multichannel SSA** - MSSA method to identify dynamically relevant space-time patterns and to provide an adaptive filtering technique. One of the aims of this paper is to provide a manual of MSSA; on Section 2 emphasis is put on the mathematical formulation of the method, with all the technical details being provided. In the multichannel case, the separation property acts both in time and space- MSSA is capable of distinguishing two oscillations with the same spatial patterns but with different periods, as well as oscillations with the same period and spatially orthogonal patterns. This method is mathematically equivalent to the extended EOF analysis of Weare and Nasstrom (1982). The spirit of extended EOF's, however, is different and aims at including temporal information in the EOF's, by adding a few lags in the state vectors. MSSA essentially differs by the use of large number of lags from which spectral properties can be drawn. For more information on EOF's see also the work of Lau and Chan (1985) and Chen and Harr (1993).

Allen and Smith (1996) showed how the basic formalism of SSA provides a natural test for modulated oscillations against an arbitrary "**colored noise**" null hypothesis. This test is called **Monte Carlo SSA** and the authors illustrate their use in 3 situations. A method of distinguishing signals from arbitrary noise processes via SSA, based on the notion of "surrogate data" (surrogate data is random data generated to have the same mean, variance, and autocorrelation function as the original data) is introduced. A Monte Carlo ensemble of surrogate data is generated using the null hypothesis as a model, and a test is applied to establish whether it is possible to distinguish the data series from a member of the ensemble. The approach proposed is a method of fitting AR (1) parameters to the data such that the process tested is on some measure, that which is most likely to cause a failure to reject the null hypothesis. In this way, if the process is rejected, there is a reason to believe that all other AR (1) processes would also be rejected at the same or higher confidence level. The algorithm proposed makes unnecessary to preprocess data to remove a trend or annual cycle before the analysis. The basic principle of surrogate data testing is that both data and surrogates must be treated in exactly the same way. To achieve that, a variant of SSA is needed, because SSA selects the EOF basis that compresses the maximum possible variance in the data series into the highest-ranked EOF's, implicitly assuming that none of the data is noise. Therefore a variant of SSA was introduced in order to assume that all of the data is noise *except* that which is established as signal. This method also provides a way to build *confidence bounds* for the forecast. Another way to construct confidence bounds is the bootstrap variant. This method is explained by Efron and Tibshirani (1986).

Yiou *et al*. (1996) published a paper where **several modern time series analysis methods were compared with each other**. The methods compared are: Fourier techniques (Blackman-Tukey and Multi-Taper), Maximum Entropy technique, Singular-spectrum techniques and Wavelet analysis. Their final recommendation is that all of those methods should be used in conjunction with each other for better results, because by confronting those methods, more information can be extracted from the system generating the analyzed signal, and the possibility of spurious results due to biases of one particular method is reduced. Nevertheless, they mentioned two major problems that can arise; a) when the time series are relatively short and b) the stationarity hypothesis which is implicitly made when classical methods are applied. For both problems SSA is mentioned as a robust method to be used.

In Lisi (1996) a criterion to choose the number of components which leads to the best filtering is purposed. The selection is made by minimizing the prediction error.

Elsner and Tsonis (1996) published a book called: "Singular Spectrum Analysis. A new tool in Time Series Analysis", each provides elementary introduction to the subject.

Varadi *et al*. (1999) proposed to generalize SSA from short and noisy time series to long and noisy time series. They called it **Random-Lag SSA**. SSA is based on a fixed sequence of lags, 1, 2, …, up to some maximum M. One then computes the eigenvalues-eigenvector decomposition of a Toeplitz matrix of size MxM, consisting of the autocorrelations up to the lag $M-1$. Random-lag SSA employs multiple random sequences of lags in which the average difference between consecutives lags is typically larger than the unit. The maximum lag can be large, while the number of lags can be kept small. The matrix to be decomposed in not Toeplitz, and it can incorporate a large number of autocorrelations at different lags. The randomness in the selection of lags is actually an advantage, since one can average the results of signal-noise decomposition over many sets of lags. This is, of course, important when the time series requires M larger than 2000-3000.

Yiou *et al*. (2000) continued their work on this subject and published a paper with some developments. The idea is to **extend the singular-spectrum analysis to the study of nonstationary time series**, including the case where intermittency gives rise to the divergence of their variance. In SSA the largest scale at which the signal X is analyzed is approximately N (the length of the time series), and the largest period is M. As a consequence, the EOF's p_k contain information from the whole time series. The proposal is to extending global SSA analysis to a local one. In fact, they proposed to extend the SSA methodology by using a time-frequency analysis within a running time window whose size *W* is proportional to the order *M* of the correlation matrix. Varying *M*, and thus *W* in proportion, they obtain a multi-scale representation of the data. They perform ***local*** **SSA** on a time series by sliding windows of length $W \leq N$, centered on times $b = \frac{1}{2}W, ..., N - \frac{1}{2}W$. When using this method, they assume that considerable information content resides in the local variance structure and the time series is locally the sum of a trend, statistically significant variability, and noise. The crucial difference between this local version and the global SSA is that the Reconstructed Components are obtained here from local lag-correlation matrices. As b varies from $\frac{1}{2}W$ to $N - \frac{1}{2}W$, this implies that the RC's will be truncated near the edges of the time series. Therefore with this new method, authors were able to reconstruct in an exact way the initial signal of the time series. The new method was also helpful in revealing key properties of a few irregular time series which conventional single-scale spectrum-analysis techniques would not reveal. **Multi-scale SSA** solves objectively the delicate problem of optimizing the analyzing wavelet in the time-frequency domain by a suitable localization of the signal's correlation matrix.

Ghil *et al*. (2002) published a review where they describe the connections between time series analysis and nonlinear dynamics, discuss signal-to-noise enhancement, and present some of the novel methods for spectral analysis. The various steps, as well as the advantages of these methods, are illustrated by their application to an important climatic time series, the Southern Oscillation Index. For enhancing the Signal-to-Noise Ratio they used SSA, Monte Carlo SSA and Multiscale SSA and wavelet analysis. As Spectral analysis methods they used the Classical spectral estimates, Maximum entropy method (MEM) and Multitaper method (MTM). As Multivariate methods they used Principal Oscillation patterns (POP's) and Multichannel SSA. This is a good review because they not only provide the theory of the most recent developments in the spectral analysis but they also provide up-to-date information on the most refined and robust statistical significance tests available for each one of the **three methods discussed in depth (SSA, MEM, and MTM)**. They also confirmed as a reliable way of forecasting ("*relative high accuracy*") the combination of SSA-MEM.

SSA has been widely used for several different purposes in the past few years. Here are only a few examples:

- In providing a qualitative decomposition of the signal into significant and noise components of ultrasound biomedical echoes, by Maciel and Pereira (2000).
- To reduce the effects of the possible discontinuity of the signal and to implement an efficient ensemble method to forecast individual rain-fall intensities series distributed in the Tiber basin, by Baratta *et al* (2003).
- To denoising chaotic data, by Liu and Zhao (2005).
- To smooth raw kinematic signals, by Alonso *et al* (2005).
- To forecast chaotic time series that contains short time surges with high amplitudes, by Ivanov *et al* (2005).
- To extrapolate time series, by Istomin *et al* (2005).

- To fill the gaps in several types of data sets, by Schoelhamer (2001) and by Kondrashov and Ghil (2006).
- To forecast the number of monthly accidental deaths in the USA, and to compare the results with those obtained using Box-Jenkins SARIMA models, the ARAR algorithm and the Holt-Winter algorithm, by Hassani (2007).

An important basis for this work was also the thesis presented for the degree of Doctor of Philosophy in Statistics at the University of Aberdeen by Oliveira (2003), each main theme is how to deal with PCA for non-independent observations. In this work and after the presentation of PCA and its relationship with time series datasets, the most important existing techniques in the field were presented: Singular Spectrum Analysis (SSA), Hilbert EOF, Extended EOF and Multichannel Singular Spectrum Analysis (MSSA), Principal Oscillation Pattern Analysis (POP Analysis).

On the PCA field the main sources of information used were the book "Applied Multivariate Techniques" by Sharma (1996) and the manuscript by Gomes, "Análise em Componentes Principais" (in Portuguese) (2006).

For basic Time Series analysis the main source of information was the book in Portuguese "Análise de Sucessões Cronológicas" published by Murteira *et al* (2000).

1.4. Single-spectrum analysis – the methodology

After the presentation of the model, done in section 1.2.1., some more details needs to be given in order to understand and implement this methodology.

In fact, the method is complex, therefore a full in depth explanation can be found both in the book "Analysis of Time Series Structure – SSA and Related Techniques" by Golyandina *et al.* (2001), and in the Annex 1 of the present work. These explanations are needed in order to understand the coming chapters, where real time series are analyzed using SSA.

In the Annex 2 can be found both the theoretical explanation of the method and the implications of those in the real world analysis, in what concerns the:

a) Window length – having an improper window length can mean that the separability of the components will not be achieved and the grouping of the eigentriples will not be successful. The success of the method relies on a correct window length size. As basic rule it can be said that the window length should never be greater than $N/2$. The dimension of the window length is determined by the problem in hand. A large L will provide separation results more stable (with respect to small perturbations), the information extracted will be larger and the components will be less mixed up. On another hand a small L will help on the proper definition of the noise floor. If the time series has a seasonal component the window length needs to be proportional to that period. For more details see section 6.1. - The window length.

b) SVD – several different matrices can be used to calculate the singular value decomposition, depending on the type of time series in study. Different methodologies will, of course, create different results. To choose the most adequate matrix some theoretical questions needs to be evaluated. The Basic SSA will be the most used but others techniques including Single centering SSA (which should be used in cases of time series with a constant component and a component that oscillates around zero), Double centering SSA (which should be used to extract the linear component of the times series) and Toeplitz SSA (which can only be used in stationary time series), could be alternatives. Description of those can be found on section 6.2. – SVD.

c) Separability – To decompose the time series in its additive components it is absolutely needed that those components be separable. There exist several types of separability and several ways to identify them. W-correlations are used as the best way to measure separability. In section 6.3. Separability, all definitions can be found.

d) Grouping – After having the eigentriples identified, and proved separable, one need to group them as the second part of the process. There are several ways to identify the eigentriples that should be grouped together, namely the scatter plots for the eigenfunctions and the EVAL percents, and those are explained in section 6.4. – Grouping.

e) The final step of the method is the diagonal averaging. There is need for a formal procedure to transform an arbitrary matrix into a Hankel matrix and therefore into a series. This formal procedure is provided in section 6.5. – Diagonal Averaging.

The desired output of the above methodology is a reconstructed homogeneous time series governed by a *linear recurrent formula*, with a small dimension relative to *N*. To get to the point when the above can justifiably be said one need to evaluate several aspects.

Structural changes can happen when transforming an homogeneous time series into an heterogeneous one, therefore a way to detect those changes is needed. The heterogeneity matrix is the way to solve this problem. More details about that matrix can be found in section 6.6.1 – Heterogeneity matrix and section 6.6.2. – Heterogeneity functions.

Because the point where the change happens is important, especially for forecasting, the detection functions play a great role here. The detection functions determine the specific point where an homogeneous time series become an heterogeneous one. The theory of detection functions is provided in section 6.6.3. – Detection functions.

The type of violations on the homogeneity of a time series and the linkage between the homogeneity of the time series and the separability of its components are described in the last two sections of the Annex 2. The general form of the H-matrix is presented and explained, being the “heterogeneity cross” the most helpful visual aspect on the detection of violation.

One of the most important aspects when confirming homogeneity of a time series is the choice of the parameters, which will help to determine the number of change-points, their location and if violation is permanent or temporary. The renormalization of the heterogeneity matrix is also important in order to evaluate correctly the possible heterogeneities.

1.5. Single-spectrum analysis - Forecasting

An acceptable forecast can only be performed if the conditions that follow are met:

The series has a structure;
A method or algorithm identifying this structure is found;
A method of the time series continuation, based on the identified structure is available;
The structure is preserved for the time period over which the forecast will be done.

The structure mentioned is usually hard to find and definitively is not unique, since most of the series has a noise component. That creates the opportunity for existence of different and even contradictory forecasts. One of the most important tasks related with the structure of the series is not only found it but also to check its stability.
The method that identifies the structure can derive from the series data or at least be checked against that data. In SSA forecasting these models are described with the help of the *Linear Recurrent Formulae*.

The series governed by LRF's admits natural *recurrent continuation* because each term of the time series is equal to a linear combination of several preceding terms.

The idea behind the searching of the LRF's is as follows:

If *d* is the minimal dimension (or order) of all LRF's governing *F*, it can be proved that if the window length *L* is larger than *d*, and the length of the series is sufficiently large, than the trajectory space for the series *F* is *d*-dimensional. The trajectory space determines a LRF of dimension *L-1* that governs the series. When this LRF is applied to the last terms of the initial series *F*, a continuation of *F* is obtained.

Usually what is obtained from the basic steps of SSA are additive components of the series F, for example $F = F^{(1)} + F^{(2)}$ where $F^{(2)}$ is residual series. If the component $F^{(1)}$ is governed by a LRF and is strongly separable from $F^{(2)}$ for the selected value of the window length *L*, then each of them must satisfy some LRF.

In practice, and for a certain window length *L*, and assuming that the series components $F^{(1)}$ and $F^{(2)}$ are approximately strong separable, the series $F^{(1)}$ is reconstructed with the help of a selected set of eigentriples and an approximation to the series $F^{(1)}$ and his trajectory space is obtained. This basically means that a LRF, approximately governing $F^{(1)}$, and the initial data for this formula are found, providing the possibility to have a forecast.

A theoretical description of the SSA recurrent forecasting algorithm is available both in the book "Analysis of Time Series Structure – SSA and Related Techniques" by Golyandina *et al.* (2001), and in the Annex 2 of the present work, in section 7.1 - SSA recurrent forecasting algorithm. Section 7.2 - Approximate continuation, introduces the concept of approximate continuation because the exact continuation is mainly methodological and theoretical.

There exists another way to forecast with SSA, is the method V-Forecasting, in opposition to the above mentioned R-Forecasting. For R-Forecasting, diagonal averaging is used to obtain the reconstructed series, and continuation is performed by applying the LRF. In the V-Forecasting, these two stages are used in the reverse order. More details are provided in the section 7.3 – Modifications to basic SSA R-algorithm. V-forecasting tends to be more "conservative" in cases of rapid increase or decrease of the R-forecasting values. V-forecasting tends to be better to forecast on the long-term than the R-forecasting.

Forecast needs to be presented together with its confidence bounds. There are two variants to the construction of those. The empirical and the Bootstrap, both are explained it in the section 7.4 – Forecast confidence bounds and in its sub-sections. It needs to be said that the empirical variant can only be used for short-term forecasting.

To assess the forecast stability and its reliability it can be said that:
Different algorithms: If the results of V and R-forecasting coincide then forecasting is stable;
Different initial data: Using different points of the reconstructed series as the base of the forecasting. Comparing results can give insights to the stability of the forecast;
Different window lengths: If the separability characteristics are stable under a small variation in the window length *L*, than forecasts for different *L* can be compared;
Forecast of truncated series: If the results of the forecast from the series truncated by removing the last few terms of it can be compared with the results of the forecast from the non truncated series than the forecast can be regarded as adequate and stable.

2. Outline

The present work can be divided in two different sections. Section 1, which could be called the “Theory part” is composed with chapter 1, and builds all the basic theoretical background for the remaining of the work. More detailed information is provided in Annex 1 and 2.

Part two, which could be called the “Practical part” is the remaining of the work and condenses chapters 3 and 4.

Chapter 1 is, in a short summary, the basis to understand what is a time series and what the methodology SSA stands for. It also contains explanations of the software used in the present work. The bibliography review intends to provide an overview of everything that has been published regarding SSA, since the beginning to nowadays.

Section 1.4 together with Annex 1 are an in depth review of the theoretical fundaments of the methodology. Explained in detail, has the aim of providing enough information on how the methodology is developed in order to be adequately used in the real cases presented.

In Section 1.5 and Annex 2 the same is done but now for the second part of the methodology – the forecasting. An in depth review of the several ways of forecasting is provided, including the different methodologies of calculating confidence bounds.

The second part, in chapter 3 starts by providing all necessary information to understand the time series that will be studied. Background information on the pharmaceutical market in Portugal is provided because this information is needed to fully understand the evolution of the market, both past and future.

At this time, three times series plus the sum of the available 15 were selected to be analyzed in detail and all steps of the process are conducted and explained. The final part of this chapter is dedicated to the forecast of the selected time series. By the end of the chapter all steps of the SSA method have been fully developed and presented.

Chapter 4 is dedicated to the Discussion and Conclusions, providing the final comments and thoughts of the present work.

All the 15+1 time series went through the same in depth analysis, each time series was decomposed, grouped, reconstructed and forecasted with the same level of attention and care. In order to do not transform this work in an endless list of justifications for each parameters choice, the results are presented in the Annex 3. There, all parameters and results are presented but no graphs or explanations are provided.

3. Data analysis

3.1. The Portuguese Pharmaceutical Market

The WHO (World Health Organization) Collaborating Centre for Drug Statistics Methodology develops and maintains the ATC/DDD (Anatomical Therapeutic Chemical classification/Defined Daily Dose) classification system. By doing this it classifies all and every drug existing in the market in their respective ATC. This system is widely use and known by all players in the pharmaceutical market. They are 5 levels in the ATC system. 1st level represents the anatomical main group, 2nd level represents the therapeutic subgroup, 3rd level represents the pharmacological subgroup, 4th level represents the chemical subgroup and finally 5th level represents the chemical substance.

The Portuguese pharmaceutical market works through two different channels, retail and hospital. The Retail market represents the sales of all drugs sold in a Retail Pharmacy with or without medical prescription. The Hospital channel represents the sales of all drugs sold directly by the Pharmaceutical Industry to the Hospital Pharmacies in order to be administered to inpatients.

According to IMS Health and regarding size, the total Portuguese pharmaceutical market value was more than 3.5 billion Euro in the year 2007. From those three quarters are sold in the Retail Market and the remaining in the Hospital segment.

From this point onwards, everything mentioned relates only to the **retail** Pharmaceutical Market.

According to INFARMED (Autoridade Nacional do Medicamento e Produtos da Saúde I.P., The Portuguese Drugs Authority) from that market, 12% of the total number of packs sold in 2007 were of generic products (generics are products with the same active ingredient of those that have seen their patent protection expired). This is important to be mentioned once the market as been largely influenced during the last 3 years by several institutional campaigns run to increase the utilization of those products, increasing the percentage of packs sold of generics products from 5% in 2004 to the already mentioned 12% in 2007.

IMS Health publishes monthly the sales in Portugal (continental and the islands), of all pharmaceutical products grouped in the above mentioned classes.

There are 15 ATC1, and they represent the Portuguese retail pharmaceutical market, namely:

A - Alimentary track and metabolism
B - Blood and blood forming organs
C - Cardiovascular system
D - Dermatologicals
G - Genito urinary system and sex hormones
H - Systemic hormonal preparations, excl. sex hormones and insulins
J - Antiinfectives for systemic use
L - Antineoplastic and immunomodulating agents
M - Musculo-skeletal system
N - Nervous system

P - Antiparasitic products, insecticides and repellents
R - Respiratory system
S - Sensory organs
T - Products to perform diagnosis
V - Various

All data that will be used in this work represents the total monthly packs sold per ATC1, in Portugal, from January 1999 until December 2007.

As have already been said, total amount of drugs sold in Portugal for the year of 2007 in the retail segment was more than 2,5 billion Euro, and they represent less than 1% of the total drugs sold in the world. The market grew from 1999 to 2007 at a 7% CAGR (Compound Annual Growth Rate). Biggest ATC's in 2007 are (in descending order): Cardiovascular system (28%), Nervous System (17%) and Alimentary track and metabolism (14%). First two grew above the market for the same period and the 3rd one grew slightly less (5%).

For the Retail pharmaceutical market products the distribution channel is: Pharmaceutical Industry => Wholesalers => Pharmacies => Patients. All products follow this path. IMS Health provides the sales at the Wholesalers to Pharmacies point with coverage just over 96% of total market. The remaining is projected, in order to achieve the total market. Once the projection method is not the subject of this work we will not go further into its explanation.

According to IMS Health, in 2007, there were 110 pharmaceutical companies selling above 1 billion Euro. According with INFARMED there were 334 Wholesalers and 2,666 Pharmacies by the end of 2006, representing coverage per Pharmacy of 3,782 Inhabitants.

Before being sold in the country all products are approved by INFARMED. In this context "a product" represents all pack sizes, of all formulations, of all strengths. This means that, for example, the sales (considered in this work) for the well known product Aspirin will be the sum of the total packs sold for all presentations in the market, which will be for 2007:

Aspirin = 674731 units

Aspirin 500 mg (500mg of active ingredient, acetylsalicylic acid) x 20 pills = 661323 units

Aspirin 500 mg (500mg of active ingredient, acetylsalicylic acid) x 10 pills = 13408 units

In Portugal and in 2006 were 11,984 products with an authorization to be marketed, with 38,481 different packs sizes.
The data used represents, therefore, the total number of packs sold, as defined above, in a monthly basis nth in Portugal.

The price of pharmaceutical products is defined in two steps. Firstly, the Minister of Economy defines the maximum public price for the pack. For the products that do not have a co-payment from the SNS (Serviço Nacional de Saúde, National Health Service) the process stops here. From this point onwards all products can be sold in a retail pharmacy. For the products that are a co-paid by the SNS another steps is needed.

INFARMED determines both the co-payment level and the final public price. This co-payment only applies when the product is dispensed in a pharmacy with a medical prescription. In Portugal, in 2006, existed 4,176 products (8,117 different packs sizes) with a reimbursement granted.

Together with the public prices also the margins for the wholesalers and pharmacies are defined by law. Nevertheless, those margins differ from product groups. Due to all this specificities and once the Portuguese government has full control on prices and margins increases and decreases, the unit used in this work is **packs** sold and not Euro sold. The purpose is to determine the market movements despite the price changes. At the end and after finding the total number of packs that are expected to be sold in 2009, some assumptions needs to be made regarding prices evolution, still this is not the aim of this work.

Is well known that once a drug, touches one of the most sacred area for humans, their health and well being, those should be used with extreme caution. Drugs have, as chemical entities, side effects and only after exhaustive study they can be widely used. It is then expected that sales of a specific drug will increase over time, after medical doctors have learnt how to use it in an efficacious and safe way. So, it is fair to say that the utilization of a drug today is the result of the accumulated experience over the past years. It is only of common sense to agree that there exists a correlation between successive observations.

It is therefore easy to accept that we are in presence of time series and that those should be studied taking into account the temporal correlation between successive observations. Due to the above mentioned we can say that we have to study 16 time series, namely:

- 1 time series that represents sales of the total market, meaning the sum of the 15 ATC1;
- 15 times series, each one representing the sum of the products grouped in each ATC1.

The purpose of this work is to find the best fitting model for those series and to predict the future behaviour of each of them.

We can not say that the behaviour of one of the 15 times series is completely not correlated with another one. In fact, these time series might even have high correlations among each other. There are several reasons that can lead to drug co-prescriptions. Concomitant diseases and population aging (leading to several diseases in the same person) are only two examples of co-prescription causes.
PCA is a technique used as an exploratory multivariate technique to reduce the dimension of a large set of variables into a small set of principal components that synthesise the information of the original data set.

Is true that in the present case we do have 15 variables, the ATC1 groups, and if our aim is to analyse the interrelationship among those variables, this technique would be perfect. The technique would project the data onto a lower dimension space in which the variability of the original data set would be as large as possible, and the new uncorrelated variables would be arranged in order of decreasing variance.

The problem with the use of the classical PCA in this case is that the study of the covariance ignores the fact that the observations may be correlated (which we believe they are, as already mentioned). If more than a weak dependence is present between observations than the standard inference procedures in PCA are invalid. There are existing techniques of PCA that are used to study data sets where time series are treated as variables as already described in the previous chapters.

3.2. Preliminary Data Analysis

The first step of a time series analysis should be a graph showing its development. Figure 2 shows all the 15 time series. Figure 3 show the time series resulting from the sum of the above 15 times series which is the total pharmaceutical market in Portugal.

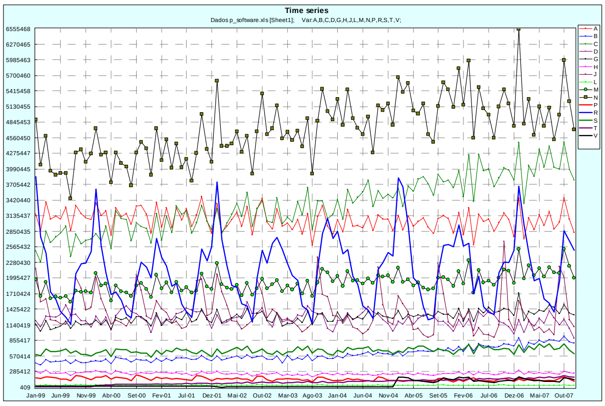


Figure 2 – Time series All 15 – Initial

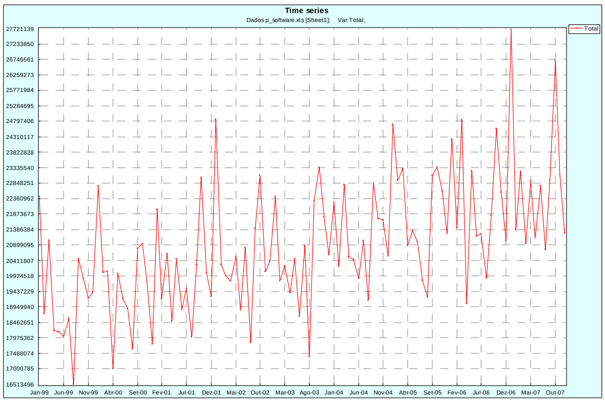


Figure 3 –Time series Total – Initial

One of the most important analysis that needs to be done before the first step of SSA is the seasonality analysis. This is important due to the fact that the window length should be in line with this seasonality. This means that the window length should be 12 or multiples of 12, due to the monthly presentation of the data. Using a simple calculation of $S_j = I_j \big/ \sqrt[12]{I_1 I_2 \ldots I_{12}}$, with $I_j = \dfrac{1}{N}\sum_{j=1}^{12} n_j$, being N the number of total years observed, one can analyze series seasonality. The values of S_j are presented in Table 1.

	1	2	3	4	5	6	7	8	9	10	11	12
A	1,07	0,94	1,04	0,95	1,03	0,98	1,04	0,93	1,03	1,08	1,00	0,93
B	1,01	0,88	1,02	0,94	1,04	1,01	1,05	0,98	1,04	1,06	1,01	0,98
C	1,02	0,89	1,02	0,95	1,06	1,02	1,08	0,94	1,03	1,05	1,00	0,97
D	1,02	0,92	1,01	0,94	1,04	1,06	1,17	1,08	1,04	1,03	0,93	0,82
G	1,08	0,91	1,00	0,94	1,02	0,98	1,07	0,97	1,03	1,06	0,99	0,95
H	1,13	0,98	1,06	0,96	1,04	0,96	1,00	0,87	0,95	1,06	1,01	0,99
J	1,25	1,07	1,03	0,84	0,85	0,78	0,80	0,86	1,51	1,40	0,98	0,93
L	1,04	0,93	1,03	0,92	1,00	0,96	1,01	0,87	1,17	1,18	1,01	0,93
M	1,12	0,97	1,05	0,94	1,01	0,94	0,99	0,93	1,00	1,08	1,03	0,97
N	1,14	0,98	1,06	0,94	1,00	0,95	0,97	0,86	1,01	1,10	1,03	0,98
P	1,04	0,97	1,12	1,03	1,04	0,92	0,92	0,73	1,30	1,28	1,01	0,78
R	1,61	1,32	1,17	0,91	0,90	0,73	0,67	0,60	0,98	1,20	1,19	1,20
S	1,01	0,93	1,06	0,98	1,09	1,05	1,09	0,96	1,00	1,02	0,96	0,88
T	0,93	0,82	0,96	0,91	1,03	1,00	1,09	0,98	1,05	1,14	1,09	1,02
V	1,37	1,25	1,23	0,89	0,94	0,85	0,78	0,73	0,84	1,18	1,12	1,04
Total	1,14	0,98	1,05	0,93	1,00	0,94	0,98	0,89	1,04	1,10	1,02	0,97

Table 1 - Time series All - Table of seasonality

Is easy to see that in the summer, comprehended between June (6) and August (8) is the period where more values below 1 are concentrated, therefore were less drugs are sold. The period between September (9) and January (1), with the exception of December (12) is when more drugs are sold.

This seasonality is more relevant if the products are antibiotics (time series J) or products for the respiratory system (time series R).

3.3. The time series selected

In order to show the most significant aspects of the method three individual (B, R and V) plus the total time series were selected to be analyzed in this work. The remaining of the series was also analyzed and the results of that analysis are shown in the Annex 4.

3.4. Change-point detection

3.4.1. Time series B – Blood and Blood Forming Organs

This time series represents the sales in packs of all products indicated mainly to treat and to prevent atherothrombotic events. In 1998 it was approved by EMEA (European medicines agency) a new product to this class, called Plavix, a trade mark of Bristol Myers Squibb, with the active ingredient clopidogrel. This product was only introduced reimbursed in Portugal in 2004. This product was considered revolutionary in the treatment and prevention of the above mentioned pathology. Therefore, the authorities, both the Ministry of Economy and the Ministry of Health approved a significant higher price than other products already in the class. Before clopidogrel the standard treatment was the well known Aspirin (acetylsalicylic acid) which costs per day around 22 cents of €. The clopidogrel cost is about 1.82€ a day, more than 8 times the Aspirin cost.

Having all of this in mind and before starting to forecast this time series it was necessary to understand if there existed structural changes.

The time series development is shown in the Figure 4:

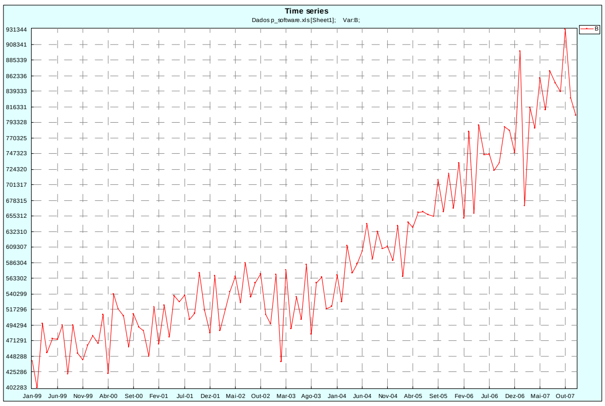


Figure 4 - Times series B - Initial

It looks easy to identify the moment of clopidogrel entrance in the market.

A *window length* of 12 was used, for two main reasons, a) the tendency seems quite "simple" which requires a smaller *L*, in order to easily identify the "noise floor"; b) the series has seasonality, so *L* should be proportional to that.

The Decomposition stage, using the Basic SSA SVD (long time series nonstationary) produced very well defined eigentriples where the first one represents the leading tendency of the series and the followings ones the seasonality. In the Figure 5 the main tendency of the time series is shown.

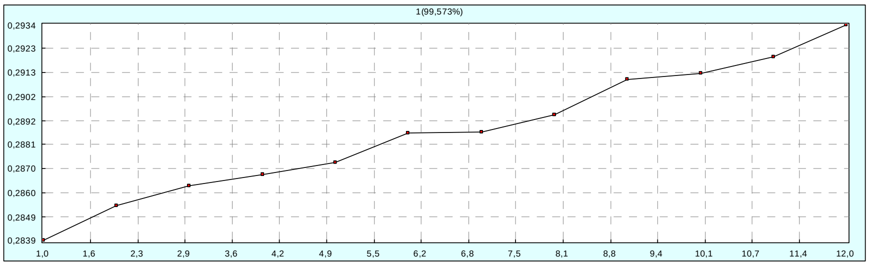


Figure 5 - Time series B- Tendency

In order to evaluate the structural changes after the entrance in the market of clopidogrel it was needed to define the base part of the series (the one to which the second part of the series will be compared with) to be from January 1999 to December 2003, which represents 60 months.

Therefore B is defined as 61, in order to get the base space from 1 to 60.

The base set of eigentriples (*I*) needs to be less than the minimum between *L* and *K*, when *K=B-L+1*, in this specific case *I* needs to be lower than the minimum between 12 and 50. It was chosen 6.

The *T*, meaning the test subseries of the time series, which needs to be at least equal to *L*, was chosen 12.

By definition the number of vector for averaging is equal to K, which is, as already mentioned, 50.

The row detection function is the most reliable one to identify the structural changes in the time series.

The diagonal detection function is useful in detection abrupt structural changes against the background of slow structural changes and the symmetric functions is good to measure the quality of approximation of the base series by the chosen eigentriples.

Therefore in Figure 6 all 3 functions are shown.

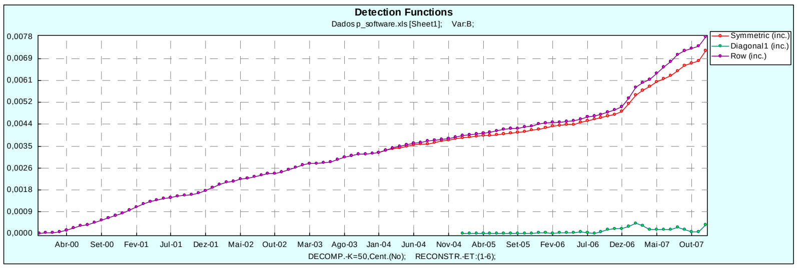


Figure 6 – Time Series B - Row, Symmetric and Diagonal Detection functions

The heterogeneity matrix is shown in the Figure 7.

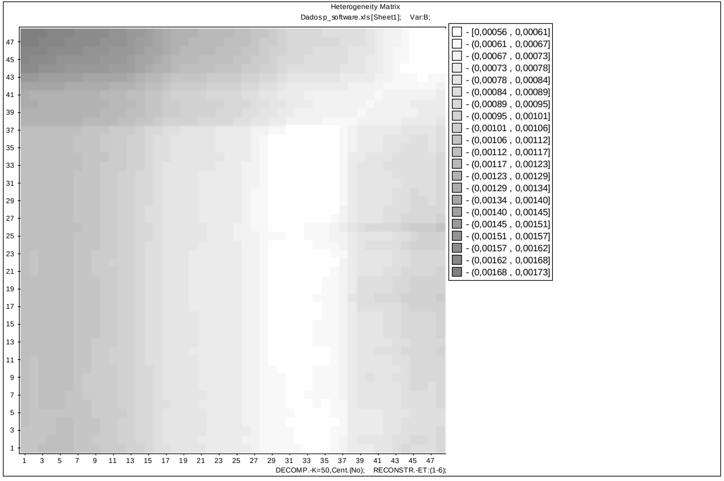


Figure 7 – Time series B - H-Matrix

The values of the matrix are very, very small, showing a very homogeneous structure. The values of this matrix have been renormalized, because the time series is positive and monotone increasing, in order to avoid that the denominator of the row detection function increases just because of that.

Based on the above results the study of the series can go on in order to achieve a stable forecast. The entrance of a “breakthrough” product did not seem to have an effect on the trajectory of the time series. Despite the fact that the value of this class increased immensely due to the launch of the new product, the number of patients treated increased accordingly with the previous tendency.

3.4.2. Time series R – Respiratory System

This time series represents the sales of products indicated in the treatment of respiratory system diseases, and includes well known products like Cegripe and Ilvico, and caught products like Bisolvon. Most of those products are sold without the need of a medical prescription and are mainly used during flow seasons but they are not antibiotics. The large majority of those products have more than 20 years in the market, so this is a very mature class of products without any new entrances.

So, is easy to understand the seasonality of the class, presented at Table 1. The time series development is shown in Figure 8.

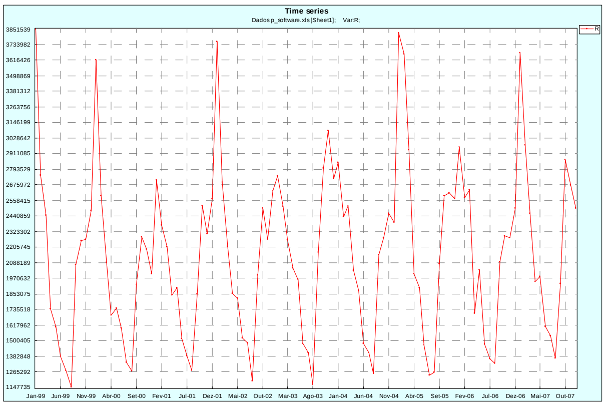


Figure 8 - Time series R – Initial

From the time series periodogram is possible to identify the following seasonalties: 6,12 18 month periods. The Figure 9 shows the periodogram for the times series.

It is easy to see that this series is most probably stationary and the main components will be related with the strong seasonality.

The decomposition method used was the Basic SSA with a *window length* of 18. The fact that the time series is stationary could lead us to use the Toeplitz decomposition method, but due to the fact that the time series is long, the Basic SSA returns better results.

Because there were no major events happening during the total period, like new product launches, the decision to evaluate the homogeneity of the structure of the series were to have the following sizes of the subseries:

Base subseries: 1 to 54, the first half of the series; B: 55; T: 18; I: 11.

The detection functions are shown in Figure 10 and the H-Matrix in Figure 11.

There are no abrupt changes in any of the detection functions indicating that there are no structural changes points. Especially the diagonal detection function that is very helpful in detecting abrupt changes on a slow changing structure is not showing any change point.

All values of the heterogeneity matrix are very small showing that the time series do not have significant structural changes which corroborates the very stable class of products. These values have been renormalized.

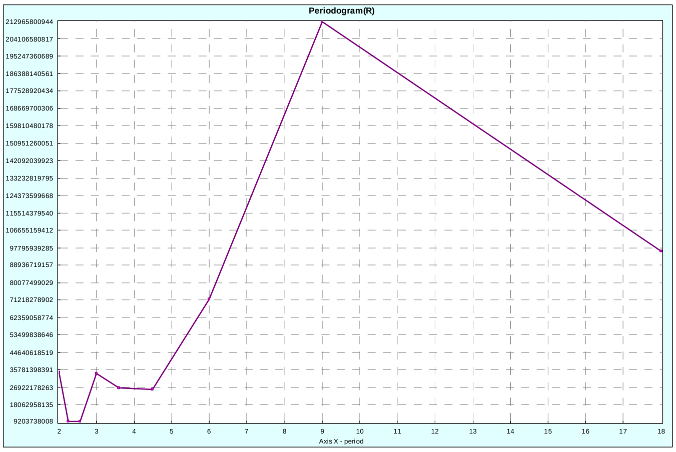


Figure 9 – Time series R – Periodogram

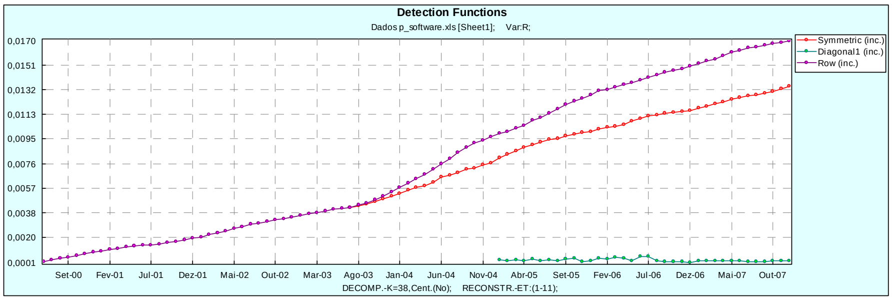


Figure 10 –Time series R - Detection functions

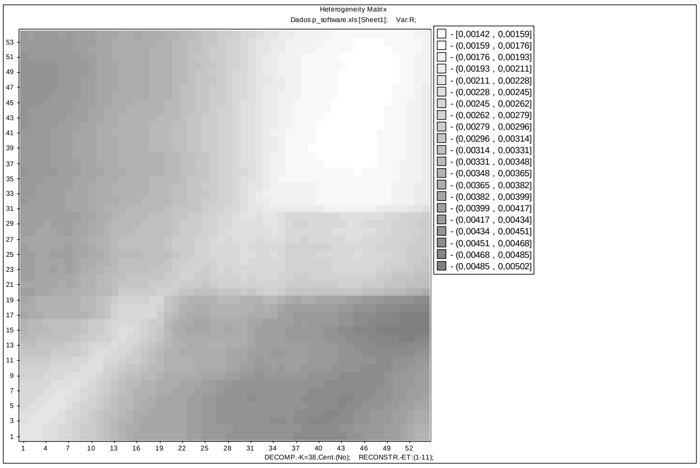


Figure 11 –Time series R - H- Matrix

3.4.3. Time series V – Various

For the V time series the approach is slightly different because this time series is not a real class of products. In fact this class, called Various, includes products that are not related to each other in any way. Therefore, this time series study will only happen to illustrate the homogeneity/heterogeneity of the time series structure.

The time series evolution is depicted in Figure 12 below, where is easy to see that there are three different periods of time in this series of 3 years each.

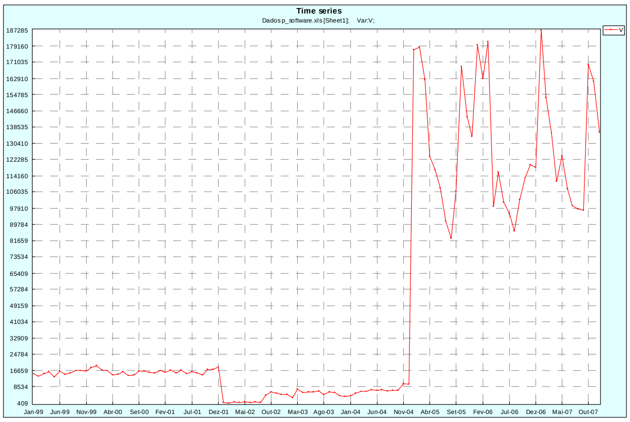


Figure 12 - Time series V- Initial

The time series has, as previously shown seasonality, therefore the *window length L* used is 12.

The base subseries will be the first two years of the time series, in order to identify the two major changes on the time series evolution. So, the *B* is 25.

It will be used a *T* of 12, and the only eigentriple used will be the first one, which represents the tendency of the series. The reason for that choice is that what is to be proven is the dramatic change on the tendency of the time series. So, *I* will be equal to 1.

The row, symmetric and diagonal detection functions are shown in the Figure 13. The values are significantly high; in fact the highest value is 0.96. The values of the heterogeneity matrix vary from 0, when the homogeneity of the time series structure exists and 1 when the structure of the time series is heterogeneous.

Therefore, it can be said that this time series, reconstructed using only the first eigenvalue is very heterogeneous.

Figure 14 shows the first row and the first column of the Heterogeneity matrix where is very easy to identify the two change points in the structure of the series by the two abrupt jumps on both lines.

It is not a surprise that the H-matrix shows two different “heterogeneity crosses”. This matrix is shown in Figure 15. These two crosses identify clearly the two changing points in the time series structure.

Based on these results if the objective was forecasting, one need to identify the part of the time series that have an homogeneous structure. If by the knowledge of the market is expected that the last 3 years of the time series will continue that only that part should be used for the forecast.

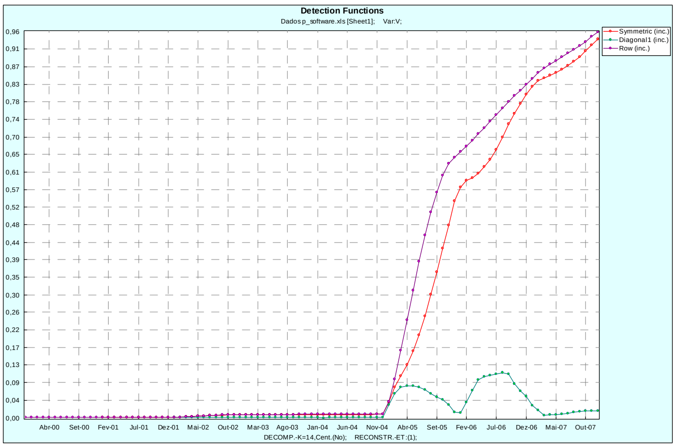


Figure 13 – Time series V - Row, symmetric and diagonal detection functions

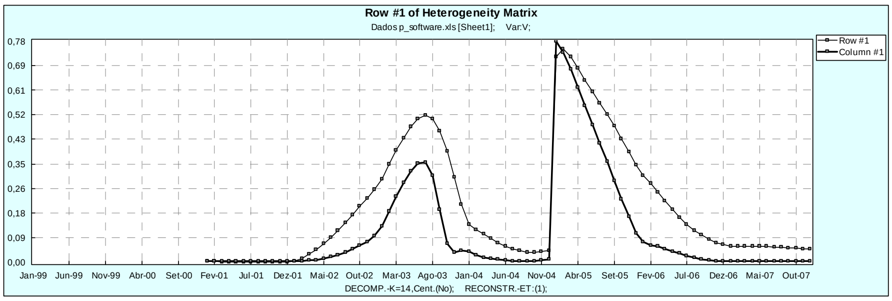


Figure 14 –Time series V - 1st Row and 1st Column of the H-Matrix

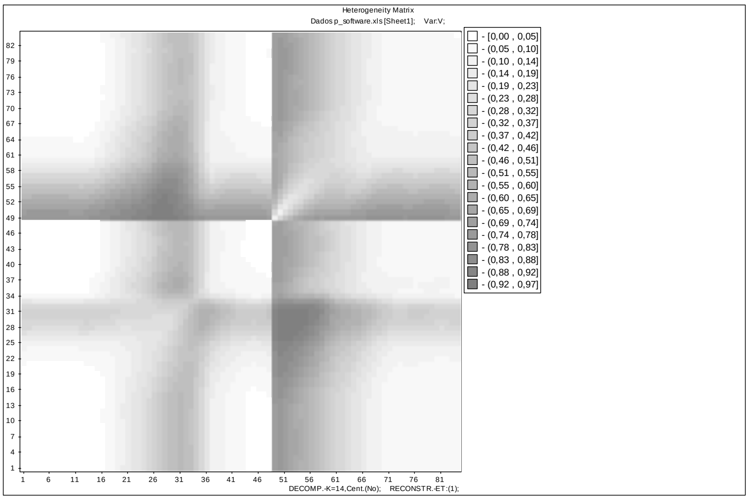


Figure 15 -Time series V - H – Matrix

3.4.4. Time series Total – Sum of the 15 ATC's

This time series is the result of the sum of the 15 ATC's, this means that it represents the evolution of the total retail pharmaceutical market in Portugal. So, it is expected that the time series has a quite homogeneous structure. In fact due to the size of the market is not expected that this time series suffer big changes in the structure. Nevertheless, due to the fact that the Health Authorities have initiated a campaign to increase the percentage of generics sold is a good idea to try to understand if that created changes in the structure of the series.

The time series development is shown in the Figure 16.

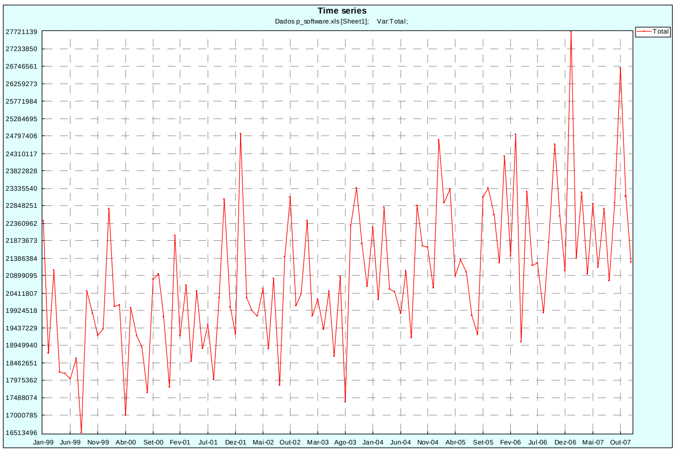


Figure 16 - Time series Total – Initial

For the decomposition stage, it was used a *window length* of 24, due to the seasonality existing in the time series. The series is not stationary, so it was used the Basic SSA SVD.

In order to evaluate the structural changes after the generics campaigns implementation the base part of the series was defined (the one to which the second part of the series will be compared with) to be from January 1999 to December 2003, which represents 60 months.

Therefore the B is defined as 61, in order to get the base space from 1 to 60.

The base set of eigentriples needs to be less than the minimum between *L* and *K*, when *K*=*B*-*L*+*1*, in this specific case *I* needs to be lower than the minimum between 24 and 48. It was chosen 8.

The *T*, meaning the test subseries of the time series, which needs to be at least equal to *L*, was chosen to be 24.

The row detection function is the most reliable one to identify the structural changes in the time series.

The diagonal detection function is useful in detection abrupt structural changes against the background of slow structural changes and the symmetric functions is good to measure the quality of approximation of the base series by the chosen eigentriples. Therefore, in Figure 17 all 3 functions are shown.

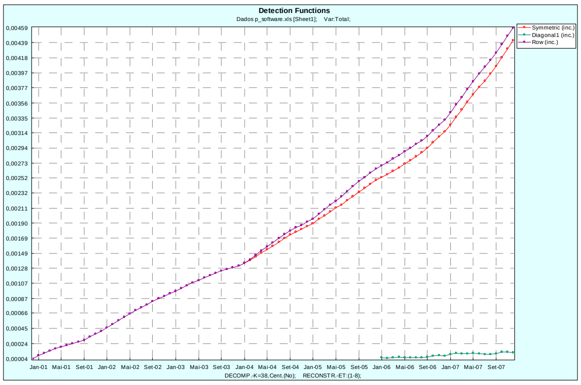


Figure 17 - – Time series Total - Row, symmetric and diagonal detection functions

The heterogeneity matrix is shown in the Figure 18.

The values of the matrix are very, very small, showing a very homogeneous structure. The values of this matrix have been renormalized, because the time series is positive and monotone increasing in order to avoid that the denominator of the row detection function increases just because of that.

Based on the above results the study of the series can go on in order to achieve a stable forecast. The campaign has not created an extra demand for pharmaceutical products. In fact, the market did not increase more than the expected tendency of the time series.

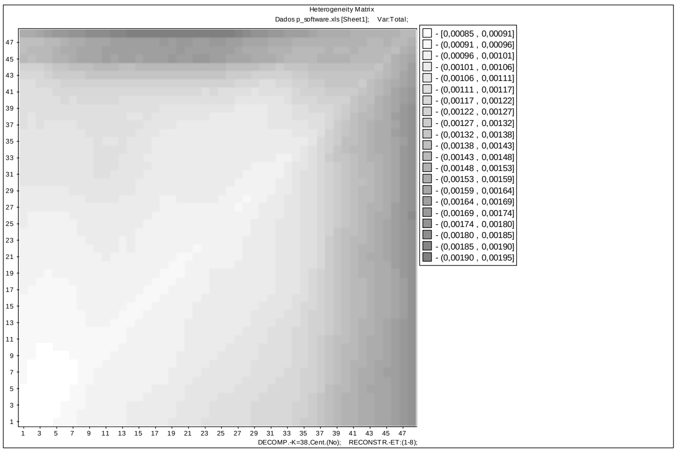


Figure 18 – Time series Total - H – Matrix

3.5. Decomposition and Reconstruction of the Time series

3.5.1. Time series B – Blood and Blood Forming Organs

The four steps of the Basic SSA, applied to the Time series B, produced the following results.

Embedding
Window length – 12, as already explained in section 3.4.
SVD
The SVD used to decompose this time series is the basic SSA, as also already explained in section 3.4.
As a result of this procedure the time series is now decomposed in several eigenvectors that identify the major components of the time series.
In fact, it is clear that eigentriple 1 (Figure 19, left graph) corresponds to the trend of the time series, eigentriple 2 (Figure 19, right graph) corresponds to the fact that two months always alternate in a cyclical movement of increasing/decreasing, meaning that the average value sold in February is lower than the average value sold in January, and lower than the value sold in March, and so one, as shown in the Table 2. This movement does not happen in November and December what is also reflected in the eigentriple 2.

	January	February	March	April	May	June	July	August	September	October	November	December
Average	601.896	519.831	608.347	559.686	616.184	596.884	625.051	580.235	615.698	626.186	601.274	580.214
Difference		-82.066	88.517	-48.661	56.498	-19.300	28.168	-44.817	35.464	10.488	-24.912	-21.060

Table 2 - Time series B - Monthly movements

The eigentriple 3-6 (Figure 20) corresponds to the seasonality that exists in the time series, as already shown in Table 1.

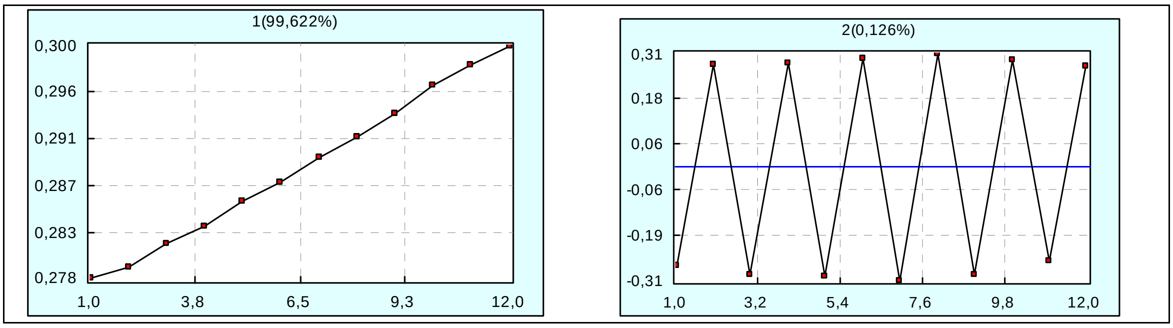


Figure 19 - Time series B – Trend and Cyclical movement

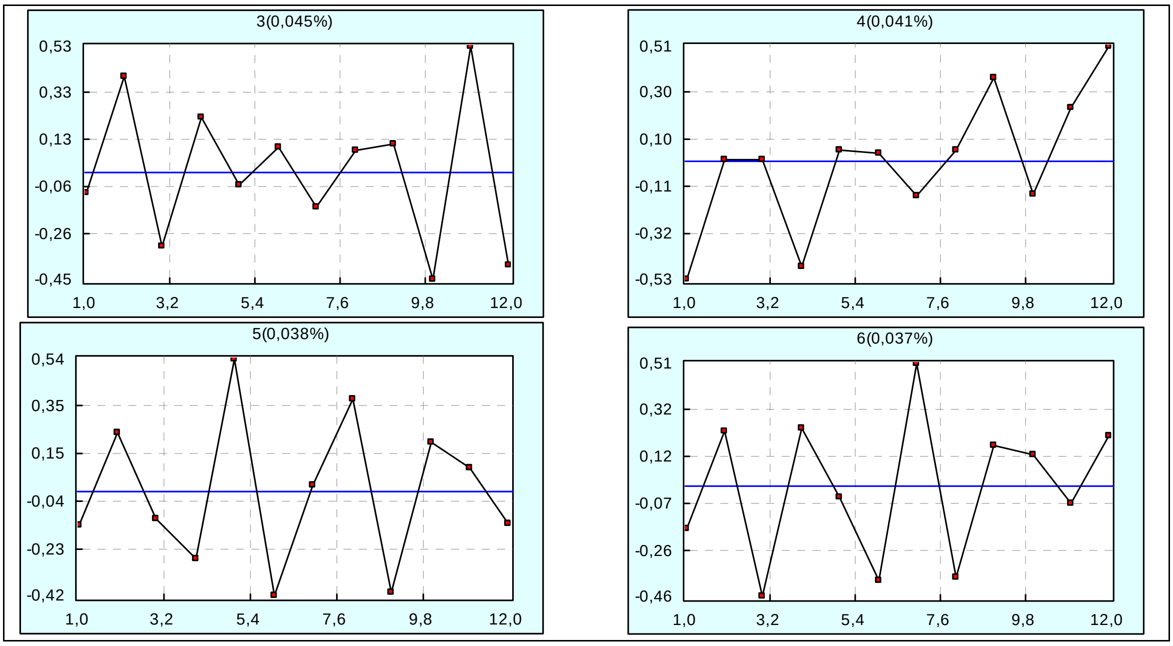


Figure 20 - Time series B - Seasonality

Grouping

The decomposition of the time series is only successful if the resulting additive components of the series are approximately separable from each other.

The first part of this step is to try to identify the components that should be aggregated to each other.

The grouping was done by having eigentriple 1 and eigentriple 2 alone, and by summing eigentriples 3 to 6.

To assess the quality of this grouping it was produced the w-correlations matrix. As already mentioned, the w-correlations between the groups should be close to zero, meaning that the correlations between rows and columns of the trajectory matrices are close to zero.

In fact, as shown in Figure 21 the w-correlations of the above defined groups are close to zero.

Group 1 represents the eigentriple 1, group 2 represents the eigentriple 2, group 3 the eigentriples 3-6.

Therefore, it can be said that the groups are separable from each other.

The remaining 6 were considered to be noise and were left out of the reconstruction part.

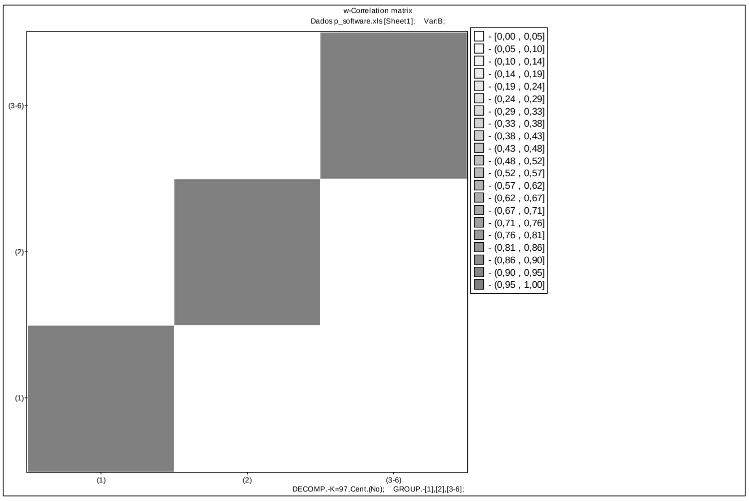


Figure 21 - Time series B - W-Correlations

Reconstruction

Using the above identified components it is possible to reconstruct the initial time series. In Figure 1 is possible to see the reconstruction of the trend of the time series. In Figure 22 is possible to see the good reconstruction of the time series using the above mentioned eigentriples. Figure 22 represents the initial and the reconstructed series. Figure 23 represents (top plot) the relative errors of reconstructing averaging, and in the same figure (bottom plot) the absolute errors are shown. It can be seen that the errors are relatively small, never going above 3.3%.

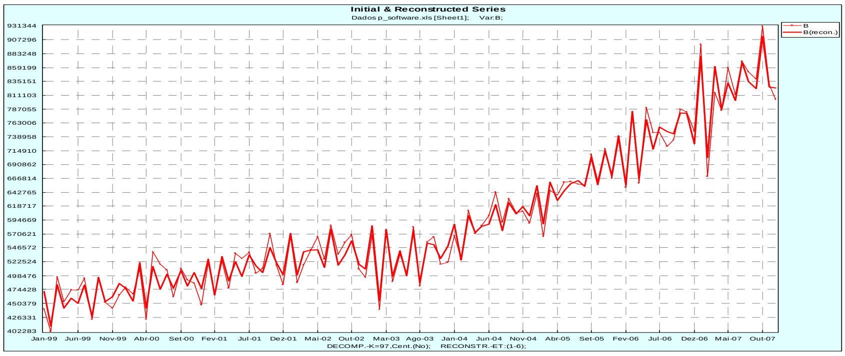


Figure 22 - Time series B - Initial and reconstructed Time Series

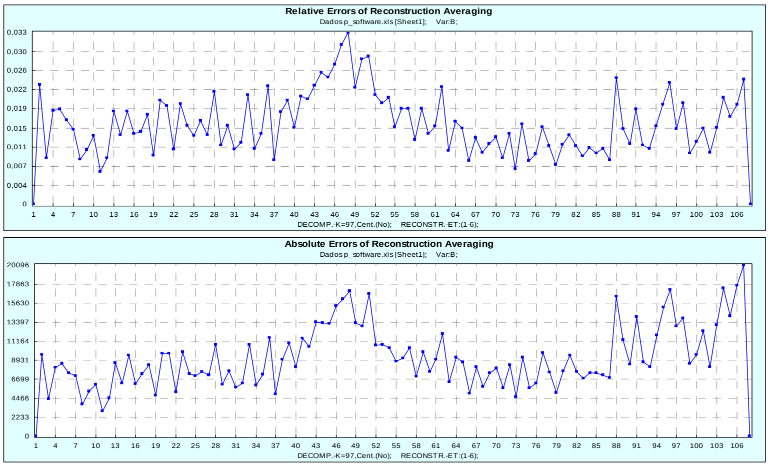


Figure 23 - Time series B - Relative and Absolute errors of reconstruction averaging

3.5.2. Time series R – Respiratory System

The four steps of the Basic SSA, applied to the Time series R, produced the following results.

Embedding

Window length – 18, as already explained in section 3.4.

SVD

The SVD used to decompose this time series is the Basic SSA, as also already explained in section 3.4.

As a result of this procedure the time series is now decomposed in several eigenvectors that identify the major components of the time series.

Despite the fact that the time series is almost stationary the first eigentriple corresponds to the small increasing pattern that can be found in the last part of the time series. The following 2 and 3 eigentriples correspond clearly to the seasonality of the series.

Figure 24 shows the eigentriples 2 and 3 where this seasonality can be found.

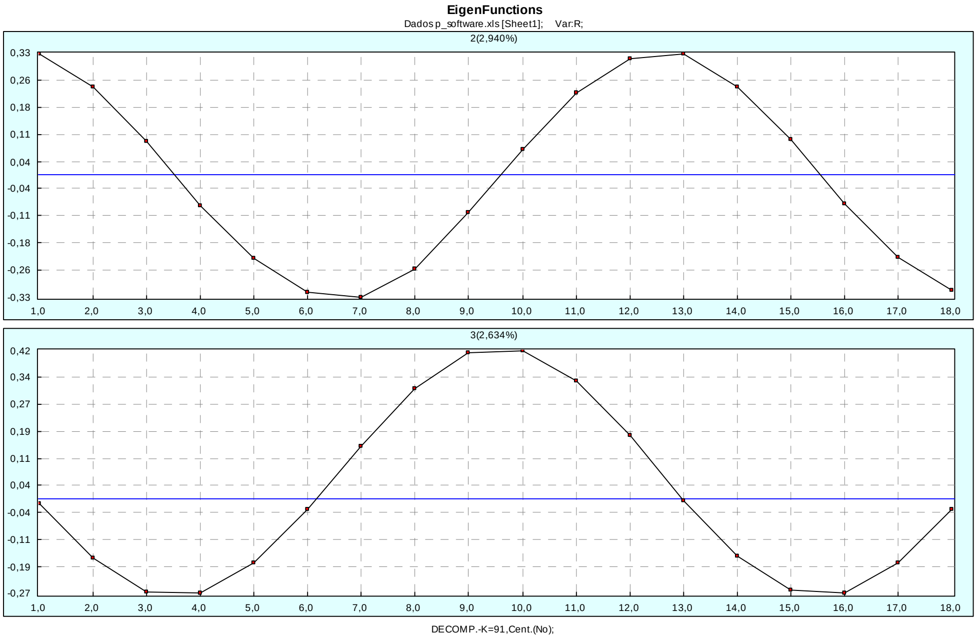


Figure 24 - Time series R - Seasonality

Grouping

This complex time series will need a high number of eigentriples groupings in order to get to a “good” reconstruction.

Looking to the scatter plots of the eigentriples is easy to identify the ones that should be grouped together. Figure 25 represents the scatter plot of eigentriples 2 and 3, where a perfect match between them can be seen.

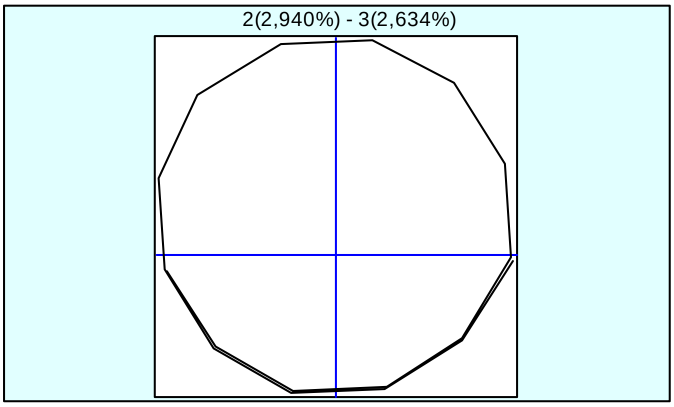


Figure 25 - Time series R - Scatter plot of eigentriples 2 and 3

By analyzing scatter plots, singular values closeness and w-correlations it is possible to get to the best grouping possible that in this case is the one shown in Figure 26, with the w-correlations matrix.

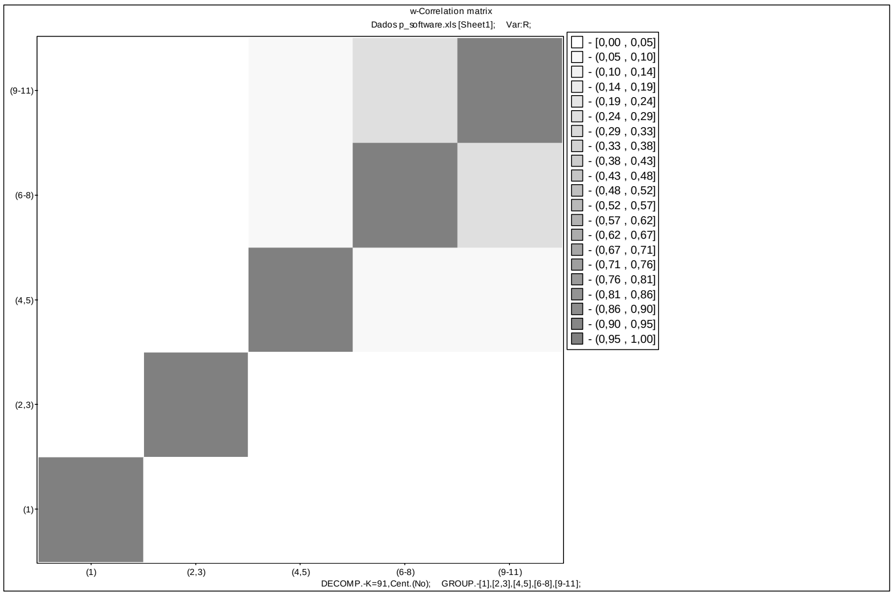


Figure 26 - Time series R - W-Correlations Matrix

Reconstruction

The reconstruction of this time series will be done with the above mentioned groups which include the first 11 eigentriples of the SVD. The remaining 7 were considered noise and were not included. Figure 27 shows the initial and the reconstructed series. The reconstruction seems quite close to the initial series, nevertheless and due to the fact that there exists a large variability in this times series, the errors of the averaging reconstruction, which can be seen in Figure 28 can reach levels above the 5.7%.

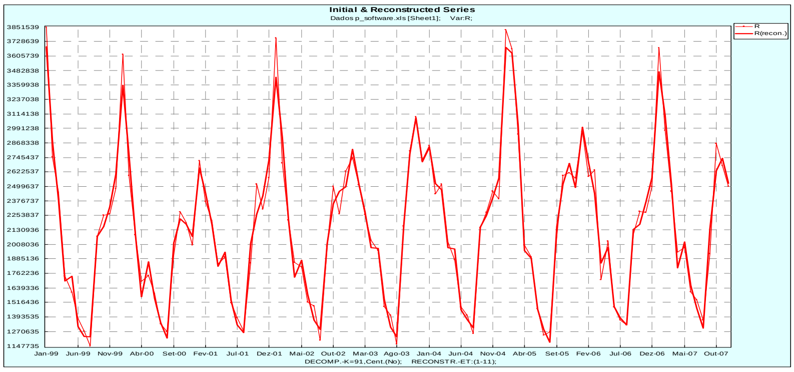


Figure 27 - Time series R - Initial and reconstructed time series

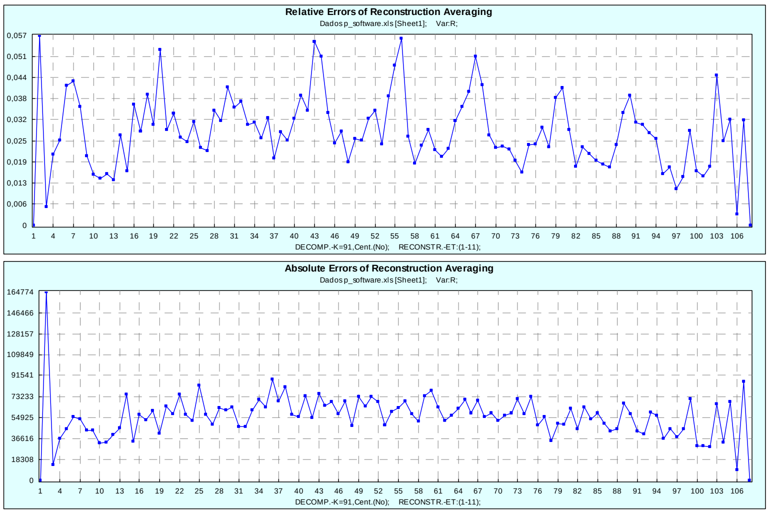


Figure 28 - Time series R - Relative and Absolute errors

3.5.3. Time series V – Various

As mentioned in section 3.4 the time series V does not have an homogenous structure when the first eigentriple is used to reconstruct the time series. Nevertheless, if the decomposition and reconstruction steps are repeated again, without ignoring all eigentriple except the first one, the results could be different. So, the study was conducted changing some of the parameters.

The four steps of the Basic SSA, applied to the Time series V, produced therefore the following results.

Embedding

Window length – 36, different from the one mentioned in section 3.4. The window length should be as large as possible in order to be stable in presence of small changes. This time series is very irregular, so the window length should be as large as possible, maintaining the seasonality proportion.

SVD

The SVD methodology used to decompose this times series was the basic SSA.
The leading singular values shown in Figure 29 led us to conclude that all eigentriples from 1 to 8 should be kept as important components of the time series.

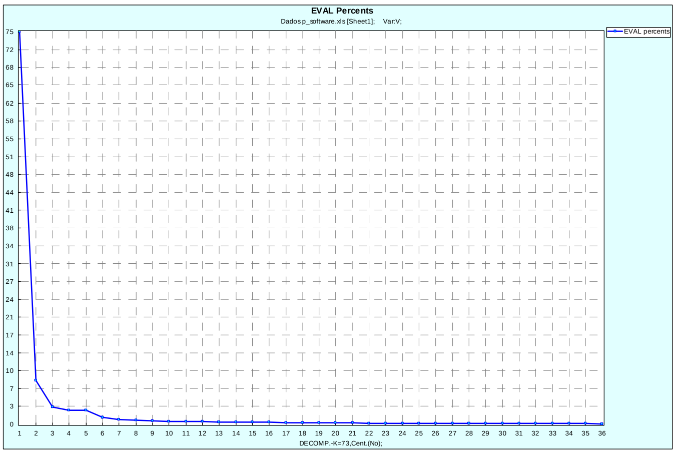


Figure 29 - Time series V - Singular values in percentage

Grouping

The w-correlations matrix of these 8 eigentriples provides significant help for the grouping stage of this time series study.
The Figure 30 shows the w-correlations between the 8 selected eigentriples. After evaluating the values obtained it was decided to group the eigentriples in the following way: eigentriple 1, eigentriple 2, eigentriple 3, eigentriple 4-5, eigentriple 6 and eigentriple 7-8.

Reconstructing

The reconstruction results are (as expected) quite weak and the errors are significant. Reconstructed times series and the residuals are shown in Figure 31. With such high residuals a forecast would never be accurate enough to be reliable.

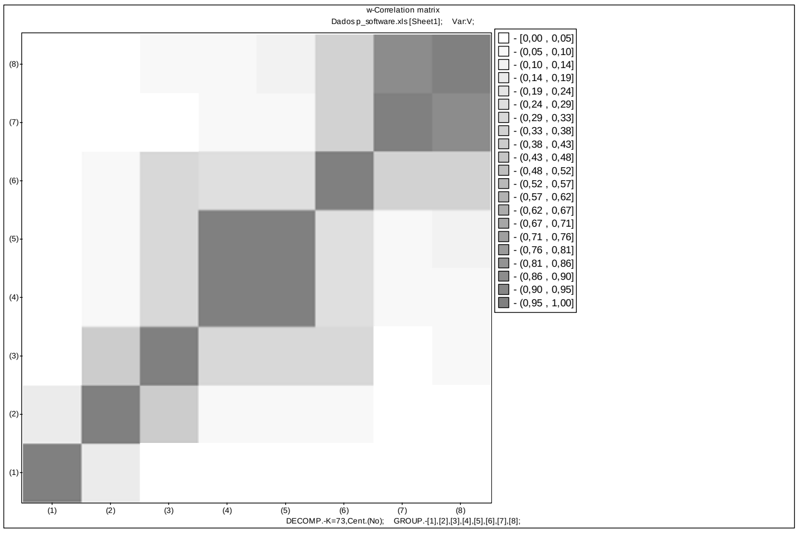


Figure 30 - Time series V - H-matrix

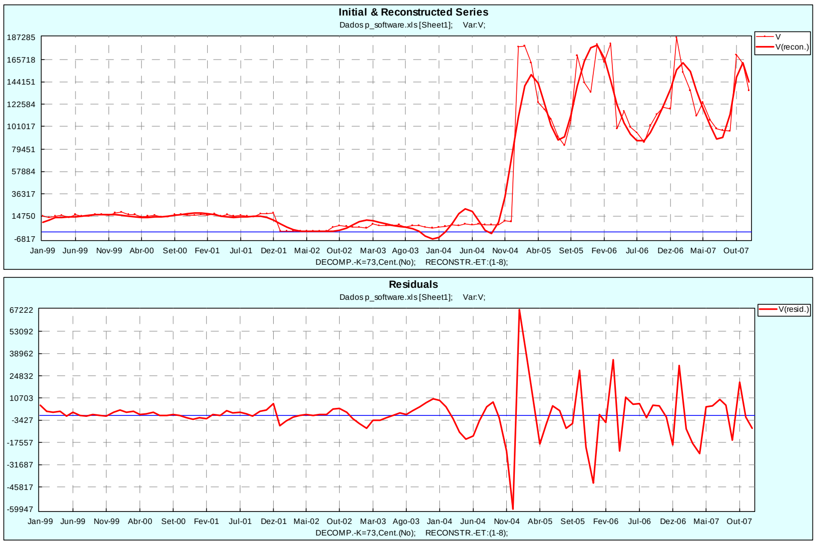


Figure 31 - Time series V - Reconstructed series and residuals

3.5.4. Time series Total – Sum of the 15 ATC’s

As already said the time series Total is the result of the summing of the 15 time series that represents sales of all pharmaceutical products in Portugal.

The 4 steps of the SSA methodology were also applied to this time series and the results are the following.

Embedding
Window length – 24, as already explained in section 3.4.

SVD
The series is not stationary, so it was used the Basic SSA SVD, as already mentioned in section 3.4
The SVD produced a first distinct eigentriple which corresponds to the trend of the time series, as shown in Figure 32.

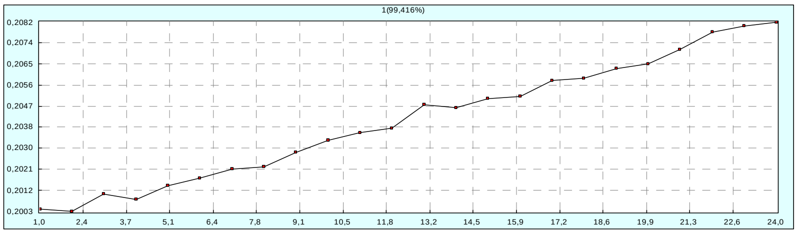


Figure 32 - Time series Total - eigentriple 1 - Trend

Eigentriple 2 corresponds to the monthly movements of up and down also seen in time series B which can be again seen in Table 3.

	January	February	March	April	May	June	July	August	September	October	November	December
Average	23.709.429	20.442.009	21.790.560	19.359.943	20.840.981	19.634.333	20.390.931	18.482.954	21.770.370	22.956.094	21.200.456	20.166.460
Difference		-3.267.419	1.348.550	-2.430.617	1.481.038	-1.206.648	756.598	-1.907.977	3.287.417	1.185.724	-1.755.638	-1.033.996

Table 3 - Time series Total - Monthly movements

The fact that between October and December the movements are not the same is well seen in the representation of the second eigentriple where the last points representing those periods show different amplitudes (Figure 33).

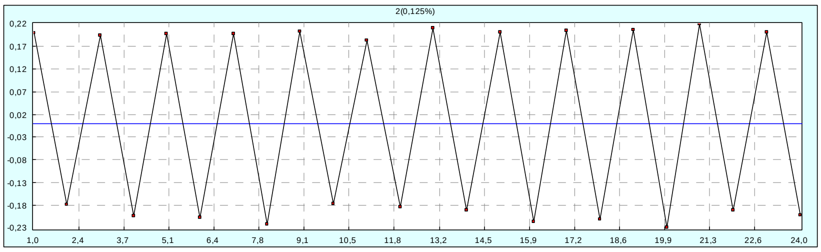


Figure 33 - Time series Total - Eigentriple 2

The following eigentriples corresponds to the seasonality that exists in the time series. In Figure 34 is well seen the "lower period" of summer. The following eigentriples corresponds to the refined seasonality of the time series.

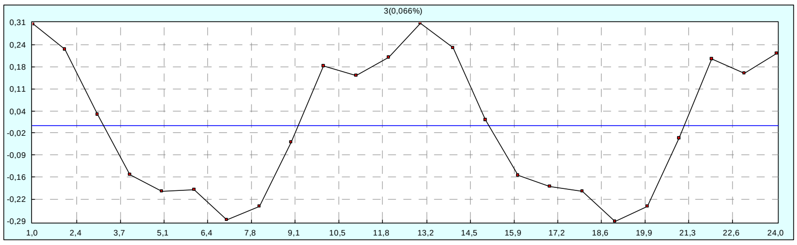


Figure 34 - Time series Total - Eigentriple 3

Grouping

In order to get the right grouping it was created the w-correlations matrix of all 24 eigentriples, which is represented in Figure 35.

It was also analyzed the scatter plots of the eigentriples, and the following groupings seems evident: eigentriple 1, eigentriple 2, eigentriple 3-4, eigentriple 5-6, eigentriple 7-8. The remaining were considered noise and left out of the reconstruction. Figure 36 shows eigentriple 3 and 4, and eigentriple 7 and 8 scatter plots.

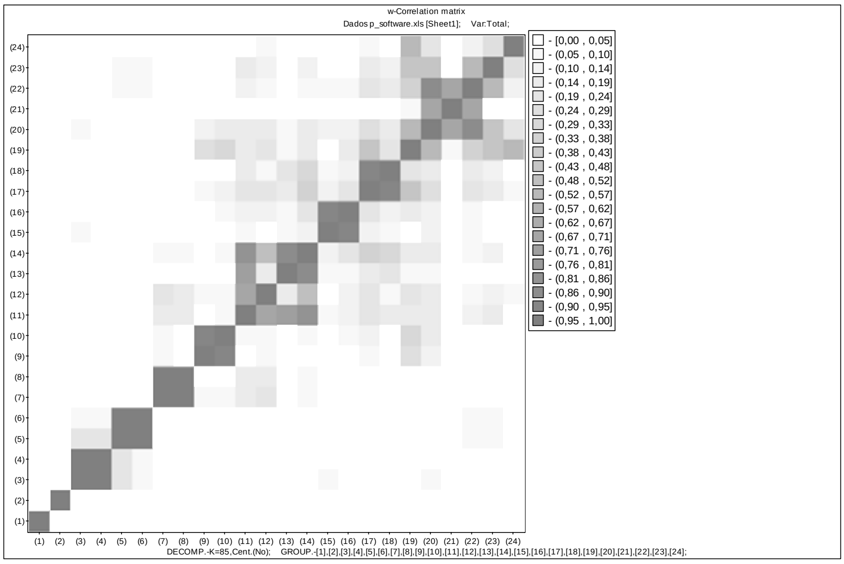


Figure 35 - Time series Total - W-Correlations Matrix

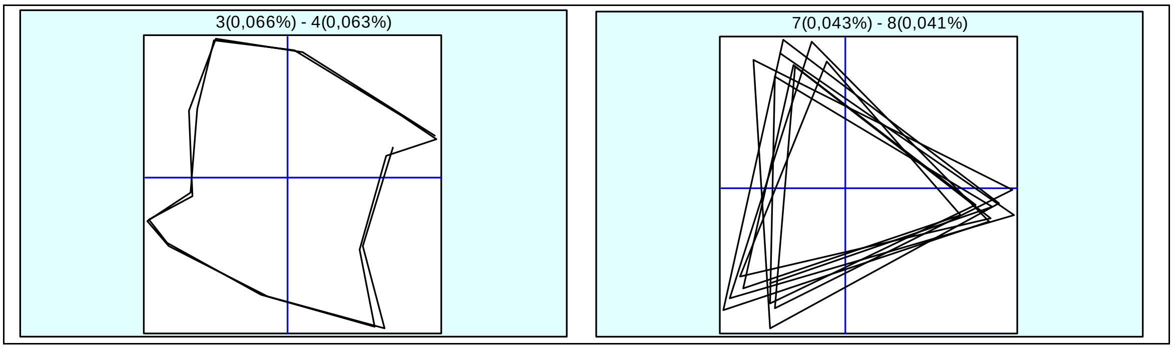


Figure 36 - Time series Total - Scatter plots

Reconstruction

Reconstruction of the time series was done with the 8 major eigentriples (the ones that in section 3.4 proved to have an homogeneous structure) and the result of that reconstruction is shown in Figure 37.

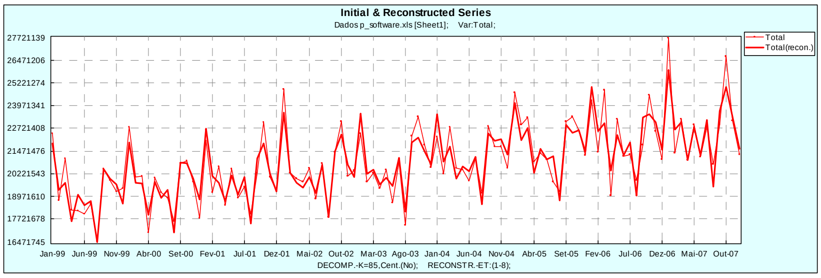


Figure 37 - Time series Total - Initial and Reconstructed time series

The relative and absolute errors of the averaging reconstruction are quite small, with the highest being below 2.9%, as shown Figure 38(top).

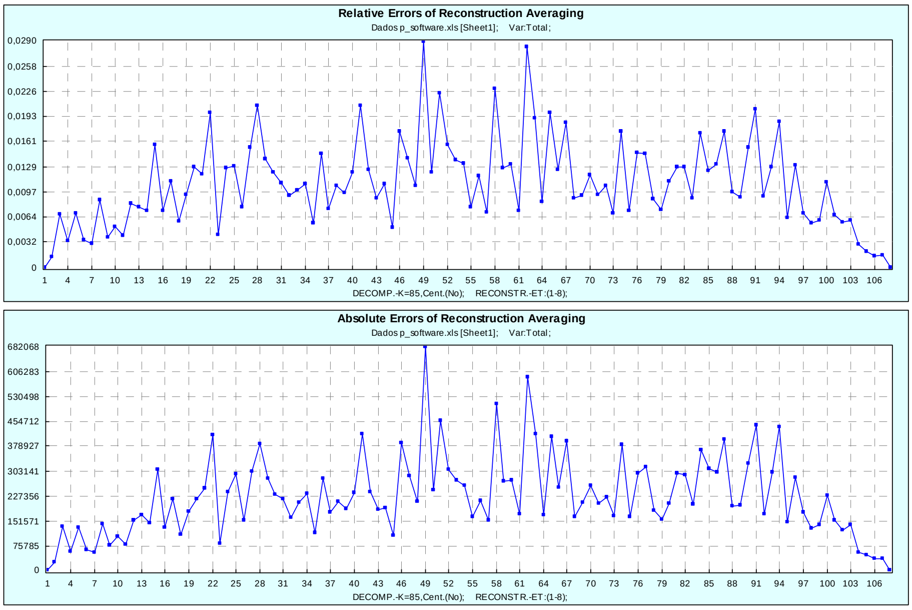


Figure 38 - Time series Total - Relative and Absolute errors

3.6. Forecast for the Time series

As explained in Chapter 1, there are two different methodologies to calculate the forecast of a time series, and two different types of forecast confidence bounds calculation.

In the coming part, the process of choosing the “best forecast” will be developed; by presenting the several steps run to achieve the most apparently stable forecast.

3.6.1. Time series B – Blood and Blood Forming Organs

As presented before the following decisions were taken to decompose and reconstruct the Time series B:

L = 12

SVD = Basic SSA

Grouping and reconstructing: Eigentriple 1; Eigentriple 2; Eigentriple 3-6.

The structure homogeneity of this decomposition and reconstruction was test and proved real.

The verticality coefficient, v^2, which represents the squared cosine of the angle between the space L_r and the vector e_L, is presented in Figure 39. The condition that $v^2 \prec 1$is necessary to forecasting. If the expected behavior of the forecast does not suggest a rapid increase or decrease, then a large value of the verticality coefficient indicates a possible difficulty with the forecast. In this case the time series development suggests that an increase is expected which corroborates the coefficient value near 0.5. The verticality coefficient is at eigentriple 4 quite high.

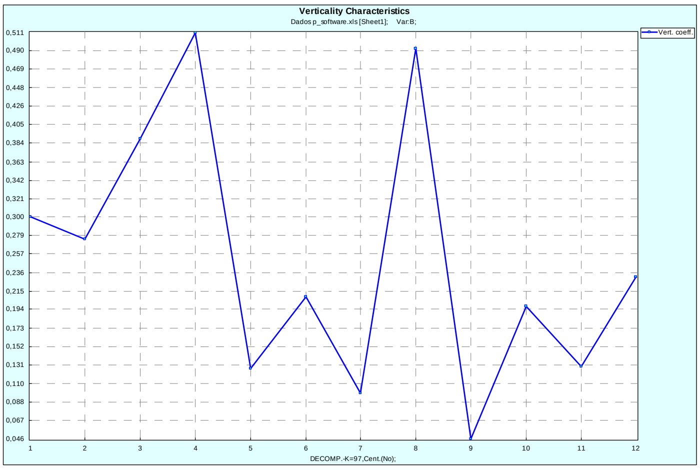


Figure 39 - Time Series B - Verticality coefficient

As already mentioned the SSA V-forecasting tends to be better for long-term forecasting and the empirical confidence intervals should not be used for long-term forecasting. These two rules proved true in the forecasting for the time series B. Figure 40 shows the V-forecast with the empirical confidence intervals and Figure 41 shows the same forecasting but using the bootstrap methodology to create the confidence intervals, with 1000 interactions. Graphs have been truncated for the clearness of the work.

As it can be easily seen in the graph of Figure 40, the confidence intervals are growing through out the forecast points. In fact, the last points of the forecast are varying between +11% of forecast and -14% of forecast. This implies that the value in December 2009 can vary between 1.122.469 packs sold and 870.889 packs sold. When using the bootstrap methodology, those intervals are significantly reduced and the forecast of December 2009 only varies between +4% and -6%. Therefore, December forecast range between 1.056.005 and 952.700.

In fact, bootstrap methodology seems to be more stable in this case even on the short-term forecast.

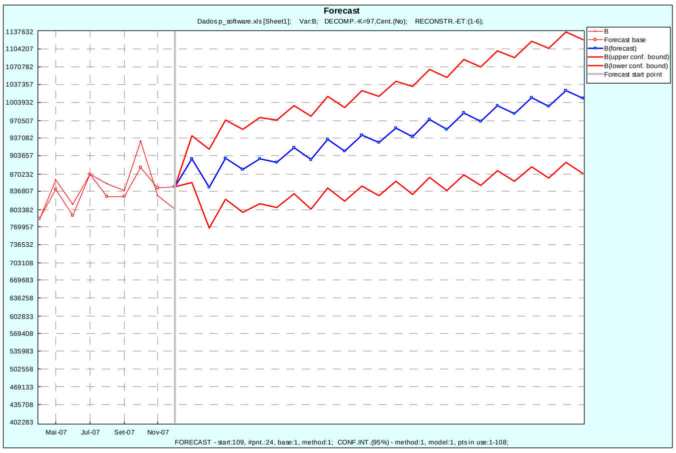


Figure 40 - Time Series B - V forecast with empirical confidence intervals

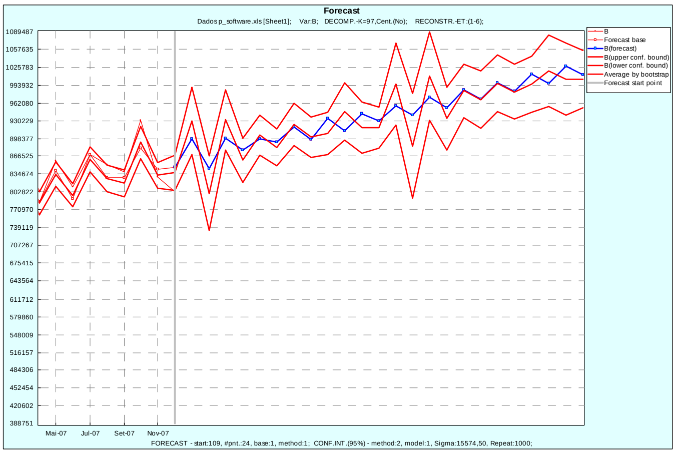


Figure 41 - Time Series B - V forecast with Bootstrap confidence intervals

Nevertheless, the SSA R-forecasting was also tested but results were quite weak with very large confidence intervals on the last forecasting points reaching values of more than 50% difference. Figure 42 shows R-forecasting with empirical confidence intervals for illustration purposes only.

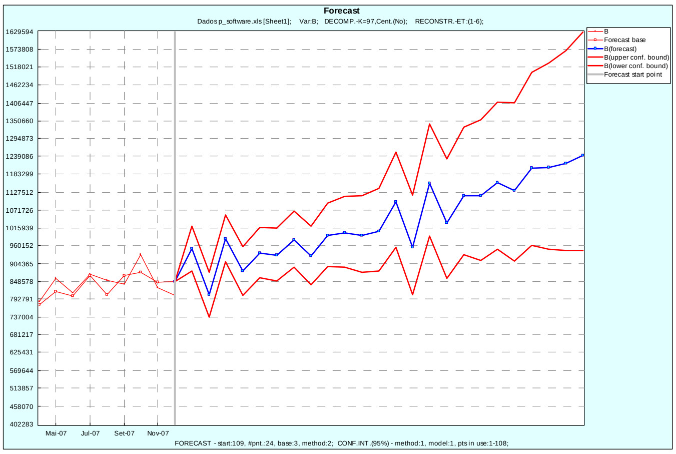


Figure 42 - Time series B – R forecast with empirical confidence intervals

3.6.2. Time series R – Respiratory System

As presented before the following decisions were taken to decompose and reconstruct the Time series R:
L = 18
SVD = Basic SSA
Grouping and reconstructing: Eigentriple 1; Eigentriple 2-3; Eigentriple 4-5; Eigentriple 6-8; Eigentriple 9-11.

The structure homogeneity of this decomposition and reconstruction was test and proved real.

The verticality coefficient is quite low, which is good because the time series do not have a strong increasing or decreasing tendency. Figure 43 shows the verticality coefficient.

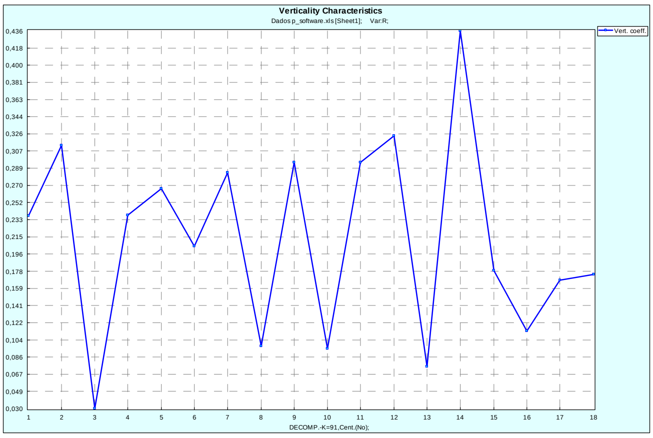


Figure 43 - Time series R - Verticality coefficient

The highest value of the verticality coefficient is reached at eigentriple 14, which is not included in the reconstruction. The highest v^2 included in the reconstruction is the one of eigentriple 2 with 0,301.

The reconstruction tested was the one used for forecasting, and due to the same reasons presented at Time Series B section, the forecast which shows more stability was the obtained with the V-methodology with confidence intervals calculated with the bootstrap method, with 1000 interactions.

Figure 44 shows the results of this forecast. Nevertheless, this is a time series with large variability which makes the forecasting exercise very, very difficult. The confidence intervals are quite large and become larger for the long-term forecast. It is needed to be extra careful with this numbers. The graph has been truncated to start at point 100 in other to make the forecast points more visible.

This forecast is also less stable in the long-term. Forecast values for the end of the period, December 2009, ranges from +15% to -12%. In fact, there are periods like August 2008 that can range between +27% and -22%. In order to obtain accuracy it was needed to increase the confidence bounds.

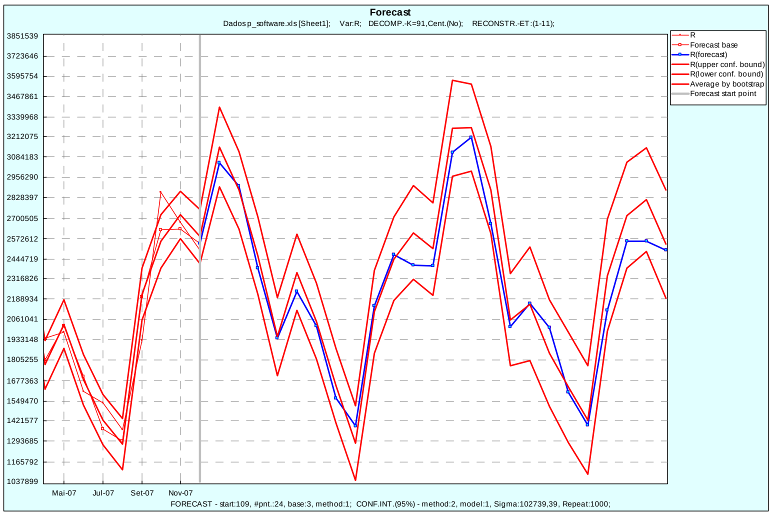


Figure 44 - Time series R - V forecast with Bootstrap confidence intervals

3.6.3. Time series V – Various

As already mentioned this time series do not have an homogeneous structure and therefore forecast can not be done using the totality of the time series. Therefore, there is no forecast prepared for this time series, due to the fact that the continuity of the time series is expected to be the same as the followed in the last 3 years, but this would only provide 36 points, which is quite small for forecasting. Indeed, forecast was tested and confidence intervals range from 36% and -30%.

3.6.4. Time series Total – Sum of the 15 ATC's

As presented before the following decisions were taken to decompose and reconstruct the Time series Total:
L = 24
SVD = Basic SSA
Grouping and reconstructing: Eigentriple 1; Eigentriple 2; Eigentriple 3-4; Eigentriple 5-6; Eigentriple 7-8;

The structure homogeneity of this decomposition and reconstruction was test and proved real.

The highest verticality coefficient is seen in eigentriple 5, and is around 0.34. As already mentioned the time series shows an increasing trend, therefore this level of verticality coefficient is expected. Figure 45 shows the verticality coefficient.

Both the V-forecast and the R-forecast were tested and both with empirical and bootstrap confidence intervals were tested.
In fact, both forecast values are very close to each other, which is a very good indicator of the stability of the forecast.
Also, empirical confidence bounds provided stable but higher intervals both on short and long-term.

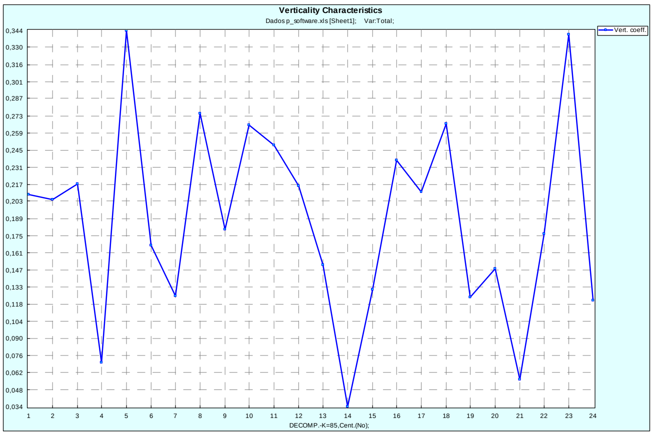


Figure 45 - Time series Total - Verticality coefficient

Therefore, the chosen methodology was the V-forecast with bootstrap confidence intervals, shown in Figure 46, again the graph was truncated to show only the last part of the time series to increase visibility and bootstrap interactions were of 1000.

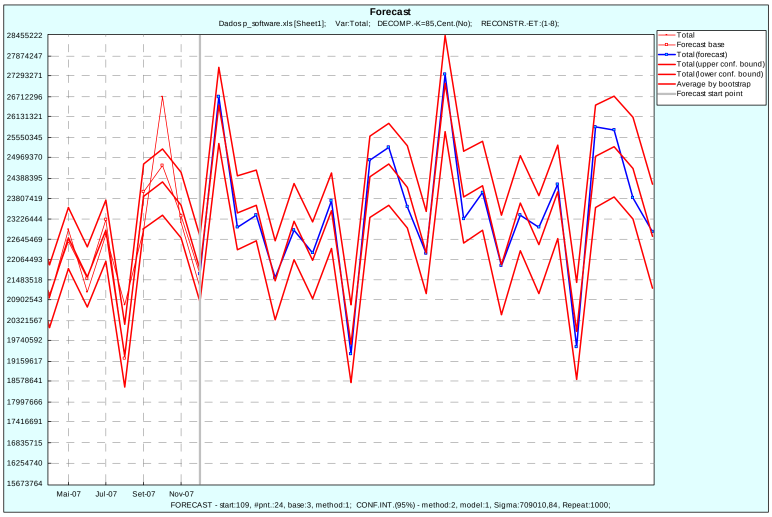


Figure 46 - Time series Total - V forecast with bootstrap intervals

4. Discussion and Conclusions

As proved in many other published papers SSA is a reliable nonparametric method to forecast Time Series.

Along the past years several time series have been forecasted using SSA method with strong and accurate outcomes.

As far as we know, the present work is the first one to deal with time series which evolution depends significantly on direct actions of economic players.

As shown in the bibliography review, SSA was been mainly used in time series like: paleoclimate records of deap-sea cores; global surface air temperature – IPCC; geopotential heights at 700hPa covering the Northern Hemisphere extratropics; tropical Pacific Ocean surface temperatures; and other climate data. Some other time series like: ultrasound biomedical echoes, monthly accidental deaths in USA; Births; etc, were also explained and forecasted with the help of SSA.

The common point on those series relies on the fact that there is very little intervention of men on the series evolution. Most of them are naturally occurring phenomena.

The challenge here is to try to prove that despite the fact that there is an enormous number of factors that can change the evolution of the, here studied, time series, SSA method still applies and the forecast is reliable.

In fact, things like new product launches, governmental cuts on health expenditures; price and reimbursement laws changes; can transform the shape of the time series evolution. Still, has not all of the previous already happened in the past? Can not the method incorporate that information in the results? Unless the paradigm changes in the future, there is no reason to expect that the method will fail on forecasting the coming two years of the retail pharmaceutical market in Portugal.

Time will confirm or not the reliability of the forecast.

Nevertheless, there is no need to wait till 2010 to find out that as an analysis tool, SSA has proven reliable to decompose and reconstruct the present time series. The present work has provided the composition of time series structure, and now is possible to identify trends, oscillations, seasonalities on all of them.

Understand the past is, also here, needed to predict the future.

By the obtained results it can be said that the Portuguese retail pharmaceutical market will growth at a rate of 1% in 2008 $\left[-3\%,+6\%\right]$and a rate of 2% in 2009 $\left[-4\%,+8\%\right]$.

This means that the Portuguese pharmaceutical market is not a fast growing market as seen in other parts of the globe. It is a mature and stable market.

To grow, pharmaceutical companies will have to develop their marketing strategies based on cannibalization objectives, because the natural growth of the market is going to be small. The fastest growing products classes will be the ATC B and the ATC C, which will grow respectively 9% $\left[+3\%,+9\%\right]$and 10%$\left[0\%,+10\%\right]$ in 2008 and 2% $\left[+2\%,+9\%\right]$and 11% $\left[-1\%,+11\%\right]$in 2009.

In what relates to the composition of the market, classes C, N, M and R will increase theirs percentage and mainly class J will decrease. This is also expected because all drugs related with the cardiovascular (C) and muscular (M) pathologies are expected to grow due to the aging of the population and drugs related with the central nervous system (N) are expected to grow due to the increased stress of life style.

The decrease of the class J is not only expected but already seen in more developed countries, antibiotics are less and less prescribed because they are less and less effective due to increase on bacterial resistance.

By the end of the decade the Portuguese retail pharmaceutical market will be built almost in the same way that it was in the beginning of the decade, in percentage of total units sold and by descending order, ATC N will maintain leadership with 23% in 2009 vs. 21% in 2000; the ATC C will be second larger with 19% in 2009 vs. 15% in 2000; the ATC A will have be third with 13% in 2009 vs. 16% (second position) in 2000. In fact, by the end of the decade it will be sold more than 280 million packs of pharmaceutical products in a year timeframe. This will represent a CAGR (2009/2000) of only $2.2\%\left[1.5\%,2.8\%\right]$. These results imply that the average number of packs sold per inhabitant in Portugal will increase from 23 packs/year, in 2000, to 27 $\left[25,28\right]$packs/year in 2009 (the source for the both total inhabitants in Portugal, in 2000, and the estimation of the total inhabitants in Portugal in 2010 is of INE, the Portuguese Statistical Institute).

The methodology, as already mentioned, proved to be quite powerful on the decomposition of the time series. All time series were successfully decomposed as initially proposed. The forecast part of the method, in the other hand, proved to be stable on the short-term but less stable on the long-term, especially in some of the time series. In fact, all series show an increase on the confidence bounds in year two (2009) vs. year one (2008) of the forecast. Due to this phenomenon it seems reasonable to purpose that if more complete years needed to be forecasted a different set of data would need to be obtained. A good approach would be to collect data from previous years before 1999 and forecast with that data, instead of using monthly data.

The decision to take monthly data and not yearly data was taken based on two aspects, first because it is the lowest level of granulation available, second because it is stable enough to be analyzed and to build a forecast. The main idea was to provide both monthly and yearly (by the sum of the monthly results) forecast to the years of 2008 and 2009. The results proved that the method can be applied to the available monthly data with a good level of confidence. In fact the relative error on reconstruction on all time series is quite low. The method also proved to be able to maintain the trend, cyclical movements and seasonality of the several time series. Comparing the S_j values found in the period of 1999 to 2007 with the S_j values found in the period 1999 to 2009 leads us to the conclusion that the seasonalities were kept by the model, the months with higher/lower values are the same in both periods. This helps to prove that the monthly forecast is valid and reliable.

As a conclusion if the pricing strategies of the government are well thought out, it will not be because of the unit growth that the expenditures of the National Health Service will increase significantly.

5. References

Alonso, F.J, Del Castillo, J.M., Pintado, P., (2005). “Application of singular spectrum analysis to the smoothing of raw kinematic signals”. Journal of Biomechanics; 38:1085-1092.

Allen, M.R., and Smith, L.A., (1996). “Monte Carlo SSA: Detecting Irregular Oscillations in the Presence of Colored Noise”. Journal of Climate; vol. 9: 3373-3404.

Baratta, D., Cicioni, G., Masulli, F., and Studer, L., (2003). “Application of an ensemble technique based on singular spectrum analysis to daily rainfall forecasting”. Neural Networks; 16:375-387.

Bertero, M., and Pike, E.R., (1982). “Resolution in diffraction-limited imaging, a singular value analysis. I: The case of coherent illumination”, Opt. Acta. 29 (1982) p. 727.

Broomhead, D.S. and King, G.P., (1986). “Extracting Qualitative Dynamics From Experimental Data”. Pshysica D; 20:217-236.

Broomhead, D.S. and King, G.P., (1986). “On the qualitative analysis of experimental dynamical systems”. In S.Sakar (Ed.), Nonlinear Phenomena and Chaos, pp. 113-144. Adam Hilger, Bristol.

Broomhead, D.S., Jones, R., King, G.P., and Pike, E.R., (1987). “Singular system analysis with application to dynamical systems”. In E.R. Pike and L.A. Lugaito (Eds.), Chaos, Noise and Fractals, pp. 15-27. IOP Publishing, Bristol.

Chen, Jeng-Ming and Harr, P.A., (1993). “Interpretation of Extended Empirical Ortoghonal Function (EEOF) Analysis”. Notes and Correspondence; September; 2631-2636.

Efron, B., and Tibshirani, R., (1986). “Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy”. Statistical Science; Vol. 1, No. 1, 54-77.

Elsner, J.B., and Tsonis, A.A., (1996). “Singular Spectrum Analysis. A New Tool in Time Series Analysis”. Plenum Press, New York.

Ghil, M., and Vautard, R., (1991). “Interdecadal oscillations and the warming trend in global temperature time series”. Letters to Nature; Vol. 350:324-327.

Ghil, M., Allen, M.R., Dettinger, M.D., Ide, K., Kondrashov, D., Mann, M.E., Robertson, A.W., Saunders, A., Tian, Y., Varadi F., and Yiou, P., (2002). “Advanced Spectral methods for Climatic Time Series”. Reviews of Geophysics; 40:1/Month pages 1-1-1-41.

Golyandina, N., Nekrutkin, V., Zhigljavsky, A., (2001). “Analysis of Time Series Structures”, Chapman & Hall/CRC.

Gomes, P., (2006). “Análise em Componentes Principais”. Manuscript.

Hassani, H., (2007). “Singular Spectrum Analysis: Methodology and Comparison”. MPRA Paper No. 4991, November.

Istomin, I.A., Kotlyarov, O.L., and Loskutov, Y., (2005). “The problem of processing time series: extending possibilities of the local approximation method using singular spectrum analysis”. Theoretical and Mathematical Physics; 142:128-137.

Ivanov, V.V., Kryanev, A.V., and Lukin, G.V., (2005). “Robust Non-stationary Singular Spectrum Analysis of Chaotic Time Series”.

Kondrashov, D., and Ghil, M., (2006). “Spatio-temporal filling of missing points in geophysical data sets”. Nonlinear Processes in Geophysics; 13:151-159.

Lau, Ka-Ming, and Chan, P.H., (1985). “Aspects of the 40-50 Day Oscillation during the Northern Winter as Inferred from Outgoing Longwave Radiation”. American Meteorological Society; November; 1889-1909.

Liu, Yuan-Feng, and Mei, Z., (2005). “ Denoising method based on Singular Spectrum Analysis and its Applications in calculation of Maximal Liapunov Exponent”. Applied Mathematics and Mechanics; Sanghai University; Vol. 26, No. 2; Feb.

Lisi, F., (1996). “Statistical Dimension Estimation in Singular Spectrum Analysis”. J. Ital. Statist. Soc.; 2:203-209.

Maciel, C.D., and Pereira, W.C.A., (2000). “RF Ultrasound Echo Decomposition Using Singular-Spectrum analysis”. Acoustical Imaging, Vol. 24, Edited by Hua Lee, Kluwer Academic/Plenum Publishers.

Murteira, B.J.F., Muller, D.A., and Turkman, K.F., (2000). “Análise de Sucessões Cronológicas”. McGraw-Hill (Eds.).

Oliveira, I., (2003). “Correlated data in Multivariate Analysis”. Thesis presented at the University of Aberdeen in June.

Plaut, G., and Vautard, R., (1994). “Spells of Low-Frequency Oscillations and Weather Regimes in the Northern Hemisphere”. Journal of the atmospheric sciences; Vol. 51, No.2:210-236.

Sauer, T., Yorke, J.A., and Casdagli, M., (1991). “Embedology”. Journal of Statistical Physics; Vol. 65, Nos. 3/4: 579-616.

Schoelhamer, D.H., (2001). “Singular spectrum analysis for time series with missing data“. Geophysical Res. Letter; 28-3187-3190.

Sharma, S., (1996). “Applied Multivariate Techniques”. John Wiley & Sons, Inc (Eds.). Chapter 4, pp 58-89.

Takens, F., (1981). “Detecting strange attractors in turbulence”. Lecture Notes in Mathematics; D.A. Rand and L.-S. Young, eds. (Springer, Berlin, 1981) p. 366.

Varadi, F., Pap, J.M., Ulrich, R.K., Bertelo, L., and Henney, J.C., (1999). “Seraching for Signal in Noise by random-lag Singular Spectrum Analysis”. The Astrophysical Journal; 526: 1052-1061, December 1.

Vautard, R., and Ghil, M., (1989). “Singular Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time Series”. Physica D; 35:395-424.

Vautard, R., Yiou, P., and Ghil, M., (1992). “Singular-spectrum analysis: A toolkit for short, noisy chaotic signals”. Physica D; 58:95-126.

Weare, B.C., and Nasstrom, J.S., (1982). “Examples of Extended Empirical Orthogonal Function Analyses”. American Meteorological Society; June; 481-485.

Whitney, H., (1936). “Differentiable manifolds”. Ann. Math.; 37:645-680.

Yiou, P., Baert, E., and Loutre, M.F., (1996). “Spectral Analysis of Climate Data”. Surveys in Geophysics; 17:619-663.

Yiou, P., Sornette, D., and Ghil, M., (2000). “Data-adaptive wavelets and multi-scale singular-spectrum analysis”. Physica D; 142:254-290.

6. Annex 1 – Single-Spectrum analysis – the methodology theory

6.1. The window length

SSA consists of two complementary stages: decomposition and reconstruction.

There are two different steps on the decomposition stage: embedding and reconstruction. Considering now the **Embedding step**:

Considering a real-valued time series $F = (f_0, f_1, \dots, f_{N-1})$of length N with $N \succ 2$. Assuming that F is a nonzero series, meaning that there exists at least one i such that $f_i \neq 0$.

The embedding step maps the original time series to a sequence of multidimensional lagged vectors.

Embedding techniques are used to reconstruct dynamical information from time series. The dimension of the *embedding space* is called the *embedding dimension* or the *window length*, which make visible L elements of the time series. At any stage the elements visible in the *L- window* constitute the components of a vector in the embedding space. As the time series is advanced step-wise through the window, a sequence of vectors in the embedding space is generated. These form a discrete trajectory.
The above sequence can be used to construct a *trajectory matrix,* **X**, which contains the complete record of patterns which have occurred within the window.

Let L be an integer (*window length*), $1 \prec L \prec N$. The embedding procedure forms $K = N - L + 1$ lagged vectors,

$$X_i = \left(f_{(i-1)}, \dots, f_{(i+L-2)}\right)^T, 1 \le i \le K$$

which have dimension *L.*

The *L-trajectory matrix* of the series F:

$$X = [X_1 : \dots : X_K]$$

has lagged vectors as its columns. The trajectory matrix can also be:

$$X = \left(x_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix}$$

The trajectory matrix is an *Hankel matrix*, because has equal elements on the 'diagonals' $i + j = const.$. As N and L are fixed, then there is a one-to-one correspondence between the trajectory matrix and the time series.

The most important parameter of this step is obviously the *window length*. The choice of *L*, corresponds to a compromise between the amount of significant information that needs to be retained - the larger the *L* the better, and the statistical confidence that needs to be achieved – the smallest the *L* the better.

When choosing the *window length* several aspects need to be taken in consideration. One of the most important is the problem in hand, meaning that it depends on the purpose of the exercise and the nature of the time series. Nevertheless, some '*rules'* needs to be considered always:

The window length *L* should be sufficiently large so that each *L*-lagged vector incorporates an essential part of the behavior of the initial series $F = (f_0, \ldots, f_{N-1})$.
If what is needed is to analyzed the time series structure so *it is meaningless to take the window length larger than half of the time series length*. This is due to the fact that the SVD's of the trajectory matrices, corresponding to the window lengths L and $K = N - L + 1$, are equivalent (up to the symmetry: left singular vectors $\leftrightarrow$ right singular vectors).
If what is needed is a very must detailed decomposition of the time series than the larger the L the better, close to $L \approx N / 2$ (there are some exceptions which will be mentioned later), as long as statistical errors do not dominate the last values of the autocovariance function. Therefore to prevent this, is advisable to not exceed $L = \frac{1}{3} N$.
As it will be seen in the further ahead, the second step of SSA is to find the components of the time series. These components need to be *separable*. The larger the window length (having in mind the previous comments) the better because:
A small window length could mix up interpretable components, meaning that some components would not be separated from each other, providing the needed information to understand the structure of the time series; on the other hand a small window length will make that the separation results will not be stable to small perturbations in *L*.
If what is needed is to properly define the noise floor, than a large window is not advisable. A large window will create an entire spectrum much flatter, and may exhibit a smooth transition to the noise floor, making it much more difficult to identify the noise floor.

If what is needed is to extract trend let $F = F^{(1)} + F^{(2)}$ where $F^{(1)}$ is a trend and $F^{(2)}$ is the residual. If the series $F^{(1)}$ is 'simple', meaning that (a) $F^{(1)}$ is well approximated by a series with finite and small rank *d* (for example it looks like an exponential, $d = 1$, a linear function, $d = 2$, etc); (b) the general tendency is the only one of interest; (c) in terms of SSA decomposition, the first few eigentriples of the decomposition of the trajectory matrix are enough for a reasonable good approximation of it; and the series $F^{(1)}$ is much 'larger' than the series $F^{(2)}$, than the window length *L* should not be very large.

However if we need to extract refined trend $F^{(1)}$, when the residual $F^{(2)}$ has a complex structure, then a large L can cause not only mixing of the ordinal numbers of the eigentriples corresponding to $F^{(1)}$ and $F^{(2)}$, but also closeness of the corresponding singular values, and therefore a lack of strong separability, even though a larger *L* is needed due to the complexity of the trend. This is the most difficult case and needs to be treated with caution.

If the interest lies in a periodic component $F^{(1)}$ out of the sum $F = F^{(1)} + F^{(2)}$ than some care is needed between the window length and the period. If the time series has a periodic component with an integer period *T,* than it is better to take the window length *L* proportional to that period.

In cases where there are more than one periodicity, for example time series with weekly and annual periodicities, the window length should be multiple of the larger periodicity, in this case the annual one.

In summary, the window length should be as small as possible keeping the separability of the components. A control of the correct choice of the window length is made at the grouping stage; the possibility of a successful grouping of the eigentriples means that the window length has been properly selected.

As a summary:

L > N/2	Meaningless because SVD of matrices *L* and *K=N-L+1* are equivalent (up to the symmetry).
L = N/2	Gives the most detailed information.
L > N/3	SSA does not resolve periods longer than the window length. So, the larger the L the better as long as statistical errors do not dominate the last values of the autocovariance functions.

Large L	The separation results are stable with respect to small perturbations.
	More quantity of information extracted.
	Will not mix up interpretable components.

Small L	Helps on proper definition of the noise floor (with large window the entire spectrum is much flatter and may exhibit a smooth transition to the noise floor).

Comments:	If the time series has a seasonal component it is advisable to take the window length proportional to that period.
	The window length has to be chosen between the period of the oscillation and the average time of its spells; SSA is typically successful at analyzing periods in the range (L/5, L).

6.2. SVD

The second step of the first stage is the SVD, Singular Value Decomposition. For the Basic SSA the matrix used to calculate the SVD of the trajectory matrix **X** is the matrix **S**, defined as follows:

$$S = XX^T$$

It can also be used the *lag-covariance matrix* **C,** where **C=S/*K*,** with exactly the same results. The only difference is the magnitude of the corresponding eigenvalues (for **S** they are *K* times larger), the singular vectors of both matrices are the same.

Therefore the SVD of an arbitrary nonzero *LxK* matrix $X = [X_1 : \ldots X_K]$is a decomposition of the matrix X in the form:

$$X = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T$$

Where $\lambda_i (i = 1, \ldots, L)$ are eigenvalues of the matrix **S,** arranged in decreasing order of magnitude.

From the above expression *d* is the rank of the matrix **X** and is the maximum value of *i,* such that $\lambda_i \succ 0$.

At the same time, $\{U_1, \ldots, U_d\}$ is the corresponding orthonormal system of the eigenvectors of the matrix **S**, and $V_i = X^T U_i / \sqrt{\lambda_i}$ **.**

From the standard terminology of SVD, the $\sqrt{\lambda_i}$ are the *singular values*; the U_i the *left singular vectors*; the U_i the *right singular vectors* of the matrix **X**. And therefore the $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ is called the *ith eigentriple* of the matrix **X**.

Sometimes is needed to use transformations of the above matrices to work with specific classes of time series and with time series of a complex structure.
Several techniques can be used to overcome these problems, namely the single and double centring SSA, and the Toeplitz SSA.

Centring means to introduce a new matrix A, of dimension *LxK* and pass from the trajectory matrix X of the time series F to the matrix X* = X – A. The decomposition obtain is therefore:

$$X = A + \sum_{i=1}^{d} X_i^* \text{ where } X_i^* = \sqrt{\lambda_i} U_i V_i^T$$

- **Single centring SSA**

Single centring is shifting the center of gravity of the lagged vectors and then uses the SVD of the obtained matrix.
It means that **A** (above) is equal to $A = \mathrm{A}(X) = [\varepsilon_1(X) : \ldots \varepsilon_1(X)]$.
It is the row centring of the trajectory matrix by having $X_i^{(c)} = X_i - \varepsilon_1(X)$ with the vector $\varepsilon_1(X)$ $(i = 1, \ldots L)$equal to the average of the *i*th components of the lagged vector $X_1, \ldots, X_K$.

The advantage of using the *Single centring* is easy to understand if the series F can be expressed in the form of $F = F^{(1)} + F^{(2)}$, where $F^{(1)}$is a constant series and $F^{(2)}$oscillates around zero.
If the series length *N* is large enough, its additive constant component will definitively be extracted by Basic SSA (which do not use any centring), but, for short series, single centring SSA can work better.

- **Double centring SSA**

In double centring the trajectory matrix X suffers the following change: to each element is subtracted the corresponding row and column averages and is added the total matrix average.
This means that **A** (above) is equal to $A = \mathrm{A}(X) + \mathrm{B}(X)$where A(X) is defined above e B(X) = $[\varepsilon_{12}(X):\ldots:\varepsilon_{12}(X)]^T$ where the *j*th component of the vector $\varepsilon_{12}(X)$ $(j = 1,\ldots,K)$ is equal to the average of all components of the vector $X_j^{(c)}$.
Double centring leads to an asymptotic extraction of the linear component of the series, if the initial series is a linear one. In fact, taking X as the trajectory matrix and A defined as before, then for $F = F^{(1)} + F^{(2)}$with linear $F^{(1)}$, the matrix *A* contains the entire linear part of the series F.
This extraction of the linear component can not be compared with the linear regression method. While the linear regression is a linear approximation by the least-squares method and gives a linear function of time for any series, even if the series does not have a linear tendency at all, the double centring SSA estimates the values of a linear function at each point, and only if strong linear tendency is really present.
Again, these two methods produce quite similar results on long series. When the time series is short is best to use double centring SSA.

Centring is more appropriated to short time series. Single centring is more appropriate to series F that can be expressed in the form $F = F^{(1)} + F^{(2)}$, where $F^{(1)}$is a constant series and $F^{(2)}$oscillates around zero. Double centring is more appropriate for the linear component extraction, meaning that for linear-like tendency, this approach is better than Basic SSA.

- **Stationary series and Toeplitz SSA**

If the time series is not sufficiently large and the series is assumed to be stationary, then the Basic SSA should be replaced by the Toeplitz version of the lag co-variance matrix **C**=**S**/K.

The Basic SSA matrix *C* where the elements are:

$$c_{ij} = \frac{1}{K}\sum_{m=0}^{K-1} f_{m+i-q} f_{m+j-q}\ ,\ 1 \le i, j \le L.$$

To get a Toeplitz lag-covariance matrix there are several ways, but the most common is the one that use the standard estimate of the covariance function of the series and to transform it into an *L x L* matrix. So, for the time series $F = (f_0,\ldots,f_{N-1})$and a fixed window length *L*, the matrix is then $\tilde{C}$ with the elements

$$\tilde{c}_{ij} = \frac{1}{N-|i.j|} \sum_{m=0}^{N-|i-j|-1} f_m f_{m+|i-j|}, 1 \le i, j \le L.$$

The main idea is to put equal values $\tilde{c}_{ij}$ in each matrix diagonal $|i-j| = k$.

Having obtained the Toeplitz lag-covariance matrix $\tilde{C}$ the orthonormal eigenvectors are calculated and they are $H_1, \ldots, H_L$. The decomposition of the trajectory matrix is then:

$$X = \sum_{i=1}^{L} H_i Z_i^T \text{ being } Z_i = X^T H_i.$$

If the initial series is a sum of a constant series with the general term c_o and a stationary series, then centring seems to be a convenient procedure. One way is to centre the entire series before calculating the matrix $\tilde{C}$ mentioned above.

The other possible method is to apply the single centring. This means that for the matrix described above $\tilde{C}$ the following product is extracted:

$$M_{ij} = \left(\frac{1}{n(i,j)} \sum_{m=0}^{n(i,j)-1} f_m \right) \left(\frac{1}{n(i,j)} \sum_{m=0}^{n(i,j)-1} f_{m+|i-j|} \right) \text{ being } n(i,j) = N - |i-j| \text{ from } \tilde{c}_{ij}.$$

The basic problem with this approach is that is not designed for non stationary series, so if the series has a strong nonstationary component Basis SSA should be used.

Toeplitz produces a non optimal orthogonal decomposition of the trajectory matrix. Nevertheless for stationary, short and noisy series Toeplitz SSA can be advantageous.

As a summary:

	Time series				
	Short	Long	Stationary	Linear Trend	Noisy
Single centring	X				
Double centring	X			X	
Toeplitz	X		X		X
Basic		X			

6.3. Separability

The main purpose of the SSA is a decomposition of the original series into a sum of components, so that each component in this sum can be identified as either a trend, or a periodic or quasi-periodic component or noise.

Each additive component of the series F needs to be separable from each other in order to have a successful SSA decomposition.

There are two different types of separability, the weak and the strong separability.

For a fixed window length L, let's consider a certain SVD of the L-trajectory matrix **X** of the initial series F of length N, and assume that the series F is a sum of two series $F^{(1)}$and $F^{(2)}$, that is, $F = F^{(1)} + F^{(2)}$.

Separability of the series $F^{(1)}$ and $F^{(2)}$ means that the matrix terms of the SVD of the trajectory matrix X can be split into two different groups, so that the terms within the groups give the trajectory matrices $X^{(1)}$and $X^{(2)}$of the series $F^{(1)}$ and $F^{(2)}$, respectively.

The separability implies that both for the rows and the columns of the trajectory matrix $X^{(1)}$of the first series are orthogonal to each row and column of the trajectory matrix $X^{(2)}$of the second series, if this orthogonality holds, then the series $F^{(1)}$ and $F^{(2)}$ are *weakly separable*.

Another condition for separability (necessary but not sufficient condition) is the **w**-orthogonality:

Let L* = min (L,K) and K* = max (L,K). Let:

$$w_i = \begin{cases} i+1 & \text{for } 0 \le i \le L^*+1, \\ L^* & \text{for } L^* \le i \le K^*, \\ N-i & \text{for } K^* \le i \le N-1. \end{cases}$$

be a set of weights.

Define the inner product of series $F^{(1)}$ and $F^{(2)}$ of length N as

$$\left(F^{(1)}, F^{(2)}\right)_w \underline{\underline{def}} \sum_{i=0}^{N-1} w_i f_i^{(1)} f_i^{(2)}$$

and call the series $F^{(1)}$ and $F^{(2)}$ *w-orthogonal* if

$$\left(F^{(1)}, F^{(2)}\right)_w = 0$$

The exact separability does not happen for real-life series and in practice only approximate separability is possible.

In case of exact separability, the orthogonality of rows and columns of the trajectory matrices $X^{(1)}$and $X^{(2)}$means that all pairwise inner products of their rows and columns are zero. This implies that a characteristic of separability of two series $F^{(1)}$ and $F^{(2)}$ is the maximum correlation coefficient $\rho^{(L.K)}$. So, $F^{(1)}$ and $F^{(2)}$ are *approximately separable* if all correlations between the rows and the columns of the trajectory matrices $X^{(1)}$and $X^{(2)}$are close to zero.

Weighted correlation or w-correlation, is a natural measure of deviation of two series $F^{(1)}$ and $F^{(2)}$ from w-orthogonality and is:

$$\rho_{12}^{(w)} = \frac{\left(F^{(1)}, F^{(2)}\right)_w}{\left\|F^{(1)}\right\|_w \left\|F^{(2)}\right\|_w}, \text{ where } \left\|F^{(i)}\right\|_w = \sqrt{\left(F^{(i)}, F^{(i)}\right)_w}, i = 1{,}2$$

If the absolute value of the w-correlation is small, then the two series are almost w-orthogonal, and therefore separable. Figure 47 shows a W-correlations matrix for the time series B, as an example. The darkness of the squares indicates the values of the W-correlations.

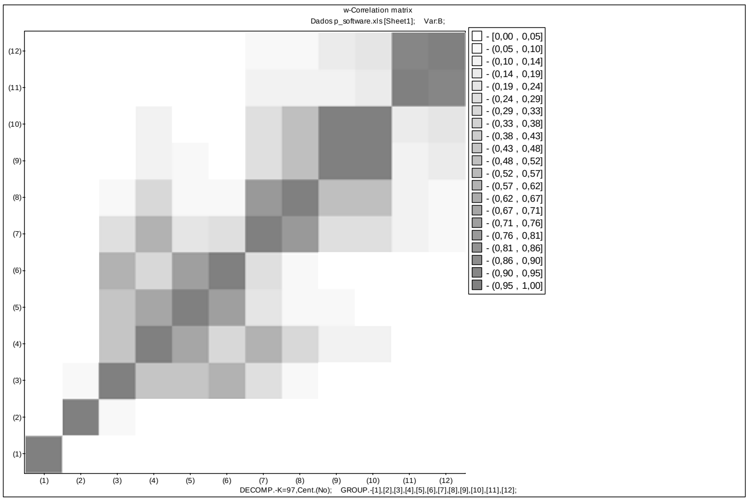


Figure 47 - Time series B - W-Correlations

Asymptotically separable series are the series that the maximum $\rho^{(L.K)}$ of the absolute values of the correlation coefficients between the rows/columns of the trajectory matrices of the series $F^{(1)}$ and $F^{(2)}$ tends to zero, as $N \to \infty$.

If $F^{(1)}$ and $F^{(2)}$ are weakly separable and all the singular values of the trajectory matrix **X** are different, then *strong separability* exists as well.

If two series $F^{(1)}$ and $F^{(2)}$ fulfill the following two conditions than strong separability exists: (a) the series $F^{(1)}$ and $F^{(2)}$ are weakly separable and (b) the collection of the singular values of the trajectory matrices $X^{(1)}$ and $X^{(2)}$ are disjoint.

Let $X^{(1)} = \sum_k X_k^{(1)}, X^{(2)} = \sum_m X_m^{(2)}$ are the SVD's of the trajectory matrices $X^{(1)}$ and $X^{(2)}$ of the series $F^{(1)}$ and $F^{(2)}$, respectively. If the series are weakly separable, then $X = \sum_k X_k^{(1)} + \sum_m X_m^{(2)}$ is the SVD of the trajectory matrix X of the series $F = F^{(1)} + F^{(2)}$.

If the singular values corresponding to the elementary matrices $X^{(1)}$ and $X^{(2)}$ coincide, this means that the terms $X_1^{(1)}$ and $X_1^{(2)}$ in the sum $X_1^{(1)} + X_1^{(2)}$ are not uniquely identified, since these two matrices correspond to the same eigenvalues of the matrix XX^T. If the series $F^{(1)}$ and $F^{(2)}$ are weakly separable, then a constant $c \neq o$ can always be found such that the series $F^{(1)}$ and $cF^{(2)}$ are strongly separable.

6.4. Grouping

One of the two only parameters of SSA is the way of grouping in the second stage of the process.

For a general series F it can be typically assumed that its trend component $F^{(1)}$ is approximately strongly separable from all other components. Therefore for extracting a trend of a series, all the elementary matrices related to slowly varying singular vectors needs to be collected. This is because a trend is a slowly varying component of a time series which do not contain oscillatory components.

If the time series F has a strong tendency $F^{(1)}$ and a relatively small oscillatory-and-noise component $F^{(2)}$ then most of the trend eigentriples will have the leading positions in the SVD of the whole series F. This does not mean that the singular values are large, in case that the trend is refined, the singular values can be small.

On the other hand if the time series has high oscillations on the background of a small and slow general tendency, the leading elementary matrices describe oscillations, while the trend eigentriples can have small singular values and can be far from the top in the ordered list of eigentriples.

To identify the harmonic components of the series, an analysis of the scatter plots of the singular values allows identification of those eigentriples that correspond to these components, provided these are separable from the residual component. In practice the singular values of the two eigentriples of an harmonic series are often close to each other, and the corresponding eigentriples are, as a rule, consecutive in the SVD order. This occurs when both L and K are several times greater than$1/\omega$, being $\omega$the frequency of the harmonic component. If the harmonic period is comparable to N, the above will not happen and therefore the two eigentriples may not be consecutive and the two singular values are small and comparable to the singular values of the component noise. Figure 48 shows a scatter of two eigenvalues for a time series.

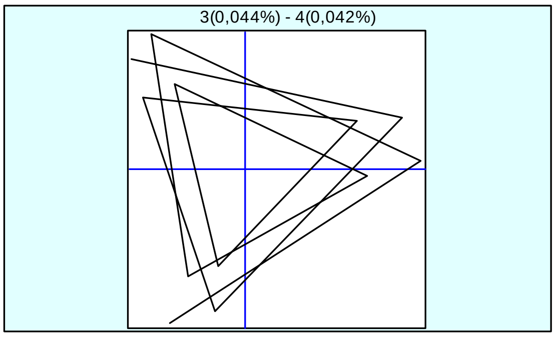


Figure 48 - Time series A - scatter plot for eigenfunctions 3 and 4

If N, L and K are sufficiently large than each harmonic different from the saw-tooth one, produces two eigentriples with close singular values. Also, a pure noise series produces a slowly decreasing sequence of singular values. If such a noise is added to a signal, described by a few eigentriples with large singular values, then a break in the eigenvalues spectrum can distinguish signal eigentriples from the noise ones. Figure 49 shows this effect.

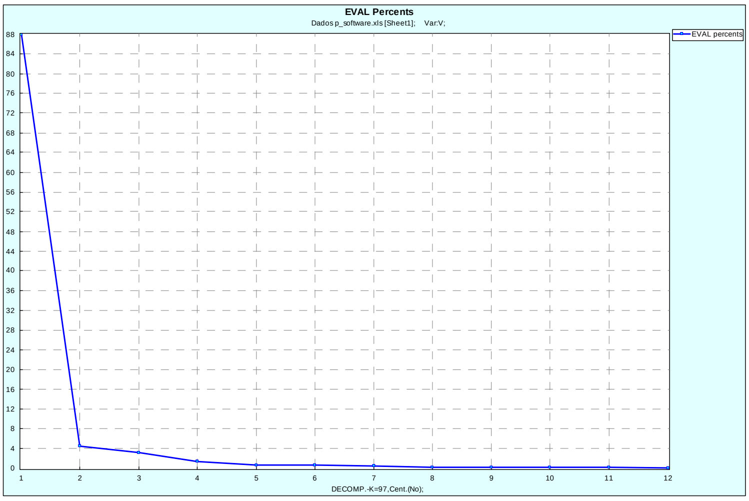


Figure 49 - Time series V - EVAL Percents

A word of caution, for complex signals and large noise: the signal and noise eigentriples can be mixed up with respect to the order of their singular values.

As already mentioned in the section dedicated to separability, a necessary condition for the approximate separability of two series is the approximate zero **w**-correlation of the reconstructed components. On the other hand, the eigentriples entering the same group correspond to highly correlated components of the series.

Therefore a natural help for grouping is the matrix of the absolute values of the w-correlations, corresponding to the full decomposition, where each group corresponds to only one matrix component of the SVD.

6.5. Diagonal averaging

The final step of the final stage of SSA is the diagonal averaging.

If the components of the series are separable and the indices are being split up accordingly, then all the matrices in the expansion $X = X_{I_1} + \ldots + X_{I_m}$ are Hankel matrices and therefore the initial series $f_o, \ldots, f_{N-1}$ is decomposed into the sum of m series: $f_n = \sum_{k=1}^{m} \tilde{f}_n^{(k)}$, and for every k and n, $\tilde{f}_n^{(k)}$ is equal to all entries $x_{ij}^{(k)}$ along the secondary diagonal $\{(i, j)$ such that $i + j = n + 2\}$ of the matrix X_{I_k} .

In practice, however, this situation never happens. In general no secondary diagonal exists of equal elements. Therefore a formal procedure of transforming an arbitrary matrix into a Hankel matrix and therefore into a series is needed. Is exactly here that the diagonal averaging enters, defining the values of the time series $\tilde{F}^{(K)}$as averages of the corresponding diagonals of the matrices X_{I_k} .

Assuming that the *Hankelization* operator, *H;* the (LxK) matrix $Y=(y_{ij})$, $L\le K$; $i+j=s$ and $N=L+K-1$; the elements $\tilde{y}_{ij}$ of the matrix *H***Y** be:

$$\tilde{y}_{ij}=\begin{cases}\dfrac{1}{s-1}\sum_{l=1}^{s-1}y_{l,s-l} & \text{for} \quad 2\le s\le L-1,\\ \dfrac{1}{L}\sum_{l=1}^{L}y_{l,s-l} & \text{for} \quad L\le s\le K+1,\\ \dfrac{1}{K+L-s+1}\sum_{l=s-K}^{L}y_{l,s-l} & \text{for} \quad K+2\le s\le K+L\end{cases}$$

For $L\succ K$ the expression for the elements of the matrix *H***Y** is analogous, the changes are the substitution $L\leftrightarrow K$ and the use of the transposition of the original matrix **Y.**

Applying this procedure to all matrix components the following expansion is obtained:

$X=\tilde{X}_{I_1}+\ldots+\tilde{X}_{I_m}$, where $\tilde{X}_{I_l}=HX_{I_l}$.

The procedure of computing the time series $\tilde{F}^{(k)}$ is called the *reconstruction of a series component* $\tilde{F}^{(k)}$ *by the eigentriples* with indices in I_k.

6.6. Detection of Structural changes

The result of the embedding procedure in a real-valued sequence series $F_N=(f_0,\ldots,f_{N-1})$, with $N\ge 3$ and fix window length $L(1\prec L\prec N)$ is a sequence of L-lagged vectors of the series F_N:

$$X_i^{(L)}=X_i=(f_{i-1},\ldots,f_{i+L-2})^T, \qquad i=1,\ldots,K.$$

If we denote $L^{(L)}(F_N)\overset{def}{=}span(X_1,\ldots,X_K)$ the trajectory space of the series F_N and if $\dim L^{(L)}=d$, with $0\le d\le L$, then it can be said that the series F_N has *L-rank d* and write this as $rank_L(F_N)=d$, assuming that $d\ne 0$ which means that not all the f_n are zero.

The equality $rank_L(F_N)=d$ is true only if $d\le\min(L,K)$. If this is true to all the appropriate L, then *the series* F_N *has rank d.*

Is also true that $rank_L(F_N)$ is the order of the SVD decomposition of the trajectory matrix X.

It can also be said that F_N has difference dimension not larger than d (fdim (F_N) $\le$ d) if $1\le d\prec N-1$ and there are numbers $a_1,\ldots,a_d$ such that

$$f_{i+d} = \sum_{k=1}^{d} a_k f_{i+d-k}\ , \quad 0 \le i \le N-d-1\ , \quad a_d \ne 0$$

The above formula is called the *linear recurrent formula (LRF)*. The LRF with d = fdim(F_N) is the *minimal* LRF.

If the above formula is valid that it can be said that the series F_N is *governed* by that LRF.

A time series F_N is *homogeneous* if it is governed by some linear recurrent formula whose dimension is small relative to *N*.

A violation in the homogeneity of the series will force the lagged vectors to leave the space $L^{(L)}$. This will make the homogeneous series to transform in an heterogeneous series. The detection of the structural changes is crucial.

There are two possibilities:

- In a determined point in time the series stops following the original LRF, and after a certain time period it again becomes governed by an LRF equal to the previous one.
- In a determined point in time the series stops following the original LRF, and after a certain time period it again becomes governed by an LRF, which is not equal to the previous one.

It really does not matter each one happens, because in both cases the series as a whole stops being homogeneous and the problem of studying this heterogeneity arises.

To detect this changes a heterogeneity matrix is build and heterogeneity functions are studied.

6.6.1. Heterogeneity matrix

This matrix characterizes the discrepancy between the series $F^{(2)}$ and the structure of the series $F^{(1)}$. Let's define the mentioned series:
Let's consider two time series $F^{(1)} = F^{(1)}_{N_1}$ and $F^{(2)} = F^{(2)}_{N_2}$ and take an integer L with $2 \le L \le \min(N_1 - 1, N_2)$. The linear space spanned by the L-lagged vectors of the series $F^{(1)}$ is $L^{(L,1)}$.
The eigenvectors of the SVD of the trajectory matrix of the series $F^{(1)}$ are $U_l^{(1)}(l = 1,\dots,L)$, for $l \succ d \overset{def}{=} \dim L^{(L,1)}$, we take vectors from any orthonormal basis of the space orthogonal to $L^{(L,1)}$ as the eigenvectors $U_l^{(1)}$.

With $I = \{i_1,\dots,i_r\}$as a subset of $\{1,\dots,L\}$and$L_r^{(1)} \overset{def}{=} span\left(U_l^{(1)}, l \in I\right)$. The lagged vectors of the time series $F^{(2)}$ are$X_1^{(2)},\dots,X_{K_2}^{(2)}$ $\left(K_2 = N_2 - L + 1\right)$.

The matrix is therefore:

$$g(F^{(1)};F^{(2)})=\frac{\sum_{l=1}^{k_2} dist^2\left(X_l^{(2)},L_r^{(1)}\right)}{\sum_{l=1}^{K_2}\left\|X_l^{(2)}\right\|^2}$$

where $dist(X,L)$is the Euclidean distance between the vector $X\in\Re^L$ and the linear space $L\subset\Re^L$.

The index *g* is the relative error of the optimal approximation of the L-lagged vectors of the time series $F^{(2)}$ by vectors from the space $L_r^{(1)}$.

The values of *g* belong to the interval $[0,1]$.

To define the *heterogeneity matrix* (*H matrix*) of a time series F$_N$, where the elements are values of the heterogeneity index *g* for different pairs of subseries of the series F$_N$, is needed to introduce the following objects:

- The initial series $F_N : F_N=(f_0,\ldots,f_{N-1}), N\succ 2$;
- The subseries (intervals) $F_{i,j}$ of the time series $F_N : F_{i,j}=(f_{i-1},\ldots f_{j-1})$ for $1\le i\le j\le N$;
- The window length $L:1\prec L\prec N$;
- The length B of the base subseries of the series F_N: $B\succ L$;
- The length T of the test subseries of the series F_N: $T\ge L$;
- The collection I of different positive integers $I=\{j_1,\ldots,j_r\}$; assuming that I is such that $j\prec\min(L,B-L+1)$for each $j\in I$;
- The base spaces $(i=1,\ldots,N-B+1)$are spanned by the eigenvectors with the indices *I*, obtained by the SVD of the trajectory matrices $X^{(i,B)}$ of the series $F_{i,i+B-1}$ with the window length *L*. The corresponding set of eigentriples is called the *base set of eigentriples*.

In these terms the elements $g_{i,j}$ of the heterogeneity matrix G = G$_{B,T}$ are $g_{i,j}=g(F_{i,i+B-1};F_{j,j+T-1})$, with $i=1,\ldots,N-B+1$and $j=1,\ldots,N-T+1$. The series $F_{i,i+B-1}$ is the *base subseries* of the series F_N and $F_{j,j+T-1}$is the *test subseries*.

6.6.2. Heterogeneity functions

Based on the H-matrix there are several heterogeneity functions:

Row heterogeneity functions

It is a series $H_{N-T+1}^{(r,i)}$for fixed $i\in[1,N-B+1]$, which corresponds to the *i*th row of the matrix **G**, with the general term $h_{n-1}^{(r,i)}\overset{def}{=}g_{in}=g(F_{i,i+B-1};F_{n,n+T-1}), n=1,\ldots,N-T+1$, and reflects the homogeneity of the series F_N (of its test subseries $F_{n,n+T-1}$) relative to a fixed base subseries $F_{i,i+B-1}$ (relative to the base space $L_{I,B}^{(L,i)}$).

Column heterogeneity functions

Corresponds to the *j*th column of the matrix **G** which for fixed $j \in [1, N-T+1]$ the column heterogeneity function $H_{N-B+1}^{(r,j)}$ with the general term $h_{n-1}^{(c,j)} \overset{def}{=} g_{nj} = g\left(F_{n,n+B-1}; F_{j,j+T-1}\right)$, , and reflects the homogeneity of the series F_N (the base space $L_{I,B}^{(L,n)}$) relative to its fixed test subseries $F_{n,n+T-1}$.

Diagonal heterogeneity functions

Is a time series $H_{N-T-\delta+1}^{(d,\delta)}$ with the parameter $0 \le \delta \le N-T$, for $n = 1,\dots,N-T+\delta+1$, where $j = i+\delta$ corresponds to the diagonal of the matrix **G**. The general term is $h_{n-1}^{(d,\delta)} \overset{def}{=} g_{n,n+\delta} = g\left(F_{n,n+B-1}; F_{n+\delta,n+\delta+T-1}\right)$.

This series reflects the local heterogeneity of the series, since both the base and the test subseries of the series F_N vary at the same time.

Symmetric heterogeneity functions

When $\delta = 0$ and $T = B$ the base subseries of the series coincide with the test subseries. The matrix **G** becomes a square matrix and the series $H_{N-B+1}^{(s)} \overset{def}{=} H_{N-B+1}^{(d,0)}$ corresponds to its principal diagonal. The general term $h_{n-1}^{(s)} \overset{def}{=} g_{n-1}^{(d,0)} = g(F_{n,n+B-1}; F_{n,n+B-1})$ of the series $H_{N-B+1}^{(s)}$ is equal to the eigenvalues share $h_{n-1}^{(s)} = 1 - \frac{\sum_{l \in I} \lambda_l^{(n)}}{\sum_{l} \lambda_l^{(n)}}$ where $\lambda_l^{(n)}$ are the eigenvalues of the SVD of the trajectory matrix of the series $F_{n,n+B-1}$ with window length *L*. Therefore the series $H_{N-B+1}^{(s)}$ is the symmetric heterogeneity function.

6.6.3. Detection functions

There are two types of change detection, the ‘forward’ change detection and the ‘backward’ change detection.

When the first one is applied what is being tested is the homogeneity of the series relatively to the initial part of the series. By definition the ‘backward’ change is the test of the homogeneity of the series relatively to the terminal part of the series.

This last option is important specially on forecasting, when finding the original part of the series that can be used for forecast.

Nevertheless, the ‘forward’ change detection problem can easily be transformed into the ‘backward’ problem by inverting the time; this is, by considering the series $f_i' \overset{def}{=} f_{N-i-1}$.

There are also 4 types of detection functions which differ from the heterogeneity functions in several aspects, which are: a) assuming only ‘forward’ changes so only the series $F_{1,B}$ should be used as the base part of the series for both row and column heterogeneity functions; b) for the diagonal (but not symmetric) heterogeneity functions should be assumed that $\delta = B$, meaning that there is no gap between the base and the test intervals, or that what is being compared are neighboring parts of the time series; c) the domain is different, the interest is almost only in the first ‘forward’ change in the series.

Row detection function

Is the series $D_{T,N}^{(r)}$ with the terms $d_{n-1}^{(r)} \overset{def}{=} h_{n-T}^{(r,1)} = g\left(F_{1,B}; F_{n-T+1,n}\right)$ with $T \leq n \leq N$. This corresponds to the detection of changes with respect to the initial part of the series, to its first B terms.

Column detection function

Is the series $D_{B,N}^{(c)}$ with the terms $d_{n-1}^{(c)} \overset{def}{=} h_{n-B}^{(1,c)} = g\left(F_{n-B+1}; F_{1,T}\right)$ with $B \leq n \leq N$.

Diagonal detection function

Is the series $D_{T+B,N}^{(d)}$ with the terms $d_{n-1}^{(d)} \overset{def}{=} h_{n-T-B}^{(d,B)} = g\left(F_{n-T-B+1,n-T+1}; F_{n-T+1,n}\right)$ with $T + B \leq n \leq N$. As mentioned in the beginning of this part there is no gap between the base and test intervals, therefore this function can be used to detect abrupt structural changes against the background of slow structural changes.

Symmetric detection function

If T=B, then the terms of the series $D_{B,N}^{(s)}$ are defined by $d_{n-1}^{(s)} \overset{def}{=} h_{n-B}^{(s)} = g\left(F_{n-B+1,n}; F_{n-B+1,n}\right)$ with $B \leq n \leq N$. This function measures the quality of approximation of the bases series by the chosen eigentriples.

6.6.4. Homogeneity and Heterogeneity

For a time series F_N homogenous, and governed by a LRF of dimension *d*, with L and *r*, so that $L \geq d$ and $d \leq r \leq \min(L, N-L+1)$; and $I = \{1,2,\ldots,r\}$,then the heterogeneity matrix is the zero matrix, since $B \geq L$, then for any i, $\Im^{(L)}\left(F_{i,j+B-1}\right) = \Im^{(L)}\left(F_N\right)$, and therefore all the L-lagged vectors of the series $F_{j,j+T-1}$ belong to the space $L^{(L)}\left(F_{i,i+B-1}\right)$for all i, j.

Conclusion: Any homogeneous series F_N gives rise to a zero heterogeneity matrix, and the presence of nonzero elements $g_{i,j}$ in this matrix is an indication of a violation of homogeneity.

The types of violations are two:

- If the same LRF is restored, after the perturbation as taken place, then the violations are *temporary*.
- If a different LRF from the original one appears after the perturbation then the violations are *permanent*.

The moment of perturbation is called the *change-point* Q and it is the maximal moment of time such the series $F_{1,Q-1}$ is homogenous, d = fdim$\left(F_{1,Q-1}\right)$. If after some time S $\left(S \geq 0\right)$, the time series becomes homogeneous again, meaning that the series $F_{Q+S,N}$ is homogeneous, and let d$_1$ = fdim$\left(F_{Q+s,N}\right)$then the time interval $\left[Q, Q+S\right]$is the *transition interval*.

Assuming that $L \geq \max\left(d, d_1\right)$ and that $L \leq Q-1$and $L \leq N-Q-S+1$, then if the L-lagged vectors of the series F_N span the original subspace $L^{(L)}\left(F_{1,Q-1}\right)$after they have left the transition interval $\left[Q, Q+S\right]$, then both homogeneous parts of the time series are governed by the same minimal LRF, and this is a case of temporary homogeneity. Examples of those are changes in the phase of one of the harmonic components, and a change in the slope of a linear additive component of the series.

Examples of permanent homogeneity are a change in the period of the harmonic components of the series and a change in the number of harmonic components.

The Figure 50 represents the general form of the heterogeneity matrix of a locally perturbed homogeneous series, assuming that the lengths of both the base and the test intervals satisfy the condition $\max\left(B, T\right) \prec Q$.

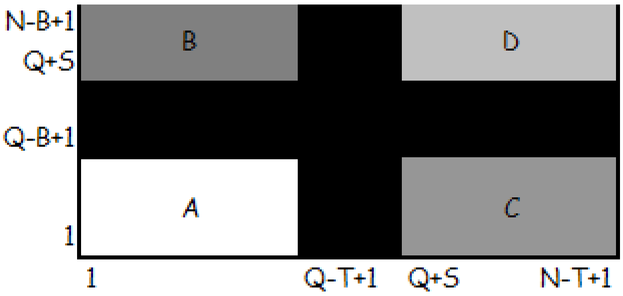


Figure 50 - General form of the H-matrix

First note: in case of temporary heterogeneity all four regions A, B, C, and D are zero regions.

Region A corresponds to the elements $g_{i,j}$ of the H-matrix where the series $F_{i,i+B-1}$ and $F_{j,j+T-1}$ are subseries of the homogeneous series $F_{1,Q-1}$. Therefore all $g_{i,j}$ are equal to zero.

In region D both series $F_{i,i+B-1}$ and $F_{j,j+T-1}$ are intervals of the series $F_{Q+S,N}$, if the dimension d_1 of the series $F_{Q+S,N}$ is not larger than the dimension d of the series $F_{1,Q-1}$, then this region is also zero.

‘The heterogeneity cross’ is the region of the elements $g_{i,j}$ of the H-matrix with indices $\left(i, j\right)$such that $Q-B+1 \leq i \leq Q+S-1$, $Q-T+1 \leq j \leq Q+s-1$. It corresponds to those $\left(i, j\right)$ where either the base or the test interval has a nonempty intersection with the transition interval.

The width of the vertical strip of the cross is equal to $T+S-1$, and the height of its horizontal strip is $B+S-1$.

In the case of permanent violations, and because the dimension of the LRF reflects the complexity of the related series, the main classification of this cases will be done in terms of the correspondence between the dimension d = $\mathrm{fdim}\left(F_{1,Q-1}\right)$ and the dimension d$_1$ = $\mathrm{fdim}\left(F_{Q+s,N}\right)$. The various cases can be:

- Preservation of dimension, meaning that the $\mathrm{fdim}\left(F_{1,Q-1}\right) = \mathrm{fdim}\left(F_{Q+s,N}\right)$. In this case the blocks A and D are zero blocks and the blocks B and C are generally not.
- Reduction of dimension, meaning that the $\mathrm{fdim}\left(F_{1,Q-1}\right) > \mathrm{fdim}\left(F_{Q+s,N}\right)$, but $L^{(L)}\left(F_{Q,N}\right) \not\subset L^{(L)}\left(F_{1,Q-1}\right)$. Also in this case the blocks A and D are zero blocks and the blocks B and C are generally not.
- Reduction of dimension (in the specific case when this reduction is caused by the disappearance of one of the series components), in this case the blocks A, C, and D are zero blocks.
- Increase of dimension, meaning that the $\mathrm{fdim}\left(F_{1,Q-1}\right) < \mathrm{fdim}\left(F_{Q+s,N}\right)$ and $L^{(L)}\left(F_{Q,N}\right) \not\subset L^{(L)}\left(F_{1,Q-1}\right)$. In this case only the block A is a zero block.
- Increase of dimension (in the specific case when this increase is caused by the adding of one of the series components), in this case the blocks A and B are zero blocks.

When the violation is temporary then all four blocks of the H-matrix are zero blocks, hence the pictorial representation of this matrix has the form of a cross. The horizontal strip reflects the transition interval, and the vertical strip shows what kind of influence the heterogeneity has on the lagged vectors of the series.

Up to now only single heterogeneity has been considered, but multiple heterogeneities can happen. That means that there are several local regions of heterogeneity in the time series. The heterogeneity matrix contains submatrices corresponding to matrices represented in Figure 5. When this happens, H-matrix has more than one cross.

6.6.5. Heterogeneity and separability

Up to now a series F_N as been considered, but realistically what happens is $F_N = F_N^{(1)} + F_N^{(2)}$, where the additive component $F_N^{(1)}$ is subject to a perturbation and the series $F_N^{(2)}$ has a sense of nuisance (for example, $F_N^{(2)}$ is noise). To describe the various forms of the 'background' H-matrices for the problem of detection of structural changes in the series components is needed that firstly the case of an homogeneous series $F_N^{(1)}$ whose subseries are (approximately) separable from the corresponding subseries of the series $F_N^{(2)}$ is studied.

The case of *stably separable* series will be specified now. Assuming that:

$F_N = F_N^{(1)} + F_N^{(2)}$ and $F_N^{(1)}$ are homogeneous;
d = $\mathrm{fdim}\left(F_N^{(1)}\right)$.

For all $i = 1, \ldots, N - B + 1$the subseries $F^{(1)}_{i,i+B-1}$and $F^{(2)}_{i,i+B-1}$are strongly separable for some window length L such that $d \prec L \prec B$.
As B>d, the subseries $F^{(1)}_{i,i+B-1}$ are governed by the same LRF that governs the series $F^{(1)}_{N}$, and therefore fdim($F^{(1)}_{i,i+B-1}$) = *d.*
r = *d.*
For any *i*, the subseries $F^{(1)}_{i,i+B-1}$is described in the SVD of the *L*-trajectory matrix of the series $F_{i,i+B-1}$ by the eigentriples indexed by the numbers in $I = \{i_1, \ldots i_r\}$, which are the same as for the series $F^{(1)}_{1,B}$.

Then the series $F^{(1)}_{N}$ and $F^{(2)}_{N}$ can be called *stably separable*, and for all $1 \le i \le N - B + 1$, the $g_{i,j}$ elements are $g_{i,j} = \dfrac{\sum_{l=1}^{T-L+1} \left\| X^{(2)}_{l,j} \right\|^2}{\sum_{l=1}^{T-L+1} \left\| X_{l,j} \right\|^2}$ where $X_{l,j}$ and $X^{(2)}_{l,j}$are the *l*-lagged vectors of the time series $F_{j,j+T-1}$and $F^{(2)}_{j,j+T-1}$, respectively.

There are two possible cases of stable separable series, the case of nonperiodic series and the case of periodic series:
nonperiodic series: stable separability of the components leads only to the equality of all row heterogeneity functions.
periodic series: will guarantee the constancy of the $g_{i,j}$.

Still assuming homogeneity, there are different kinds of deviation to the stable separability possible. Deviations from weak separability, not related to the ordering of the eigenvalues, and the effects of coincidence and rearrangements of the eigenvalues, which have influence on both strong separability and the constancy of the sets of I_i.

Examples:

Approximate weak separability - nonperiodic series: The H-matrix will be the same as above, but the row heterogeneity functions are no longer equal.
Approximate weak separability - periodic series: In these case there exists a dependence of the H-matrix on T. The smaller T (and therefore T – L + 1), the larger fluctuations the elements of the matrix H-matrix may have.

Asymptotic separability: when the time series is large asymptotic weak separability is more natural than approximate weak separability. As a rule, asymptotic separability implies that there are small fluctuations around the limiting H-matrices, which are either constant or have the form of the stable separation with nonperiodic series. Natural cases of asymptotic separability are the noisy series, meaning the series is corrupted by noise. Asymptotically, the values of the H-matrix have a constant limit:

$$\lim g_{ij} = \frac{2R_\varepsilon(0)}{c^2 + 2R_\varepsilon(0)}$$

In practice, the closeness of the elements of the H-matrix to the constant value mentioned above is achieved due to the large value of the series length *N* and the small (relative to C^2) value of the variance $R_\varepsilon(0)$.

Rearrangement of the eigentriples: it can occur either by the increase of dimension of the I , for example a series that is described by only one eigentriple and starts in a defined point to be defined by the two leading eigentriples, or by maintaining the dimension but the eigentriples that describes the series change.

Nevertheless, it is important to identify, when they exists, the intervals of heterogeneity. The series $F_N^{(1)}$ and $F_N^{(2)}$ themselves can be either stably separable on the homogeneity intervals or have discrepancies from this ideal situation. As always the heterogeneity matrix G gives the best description of the entire situation.

The most important is the choice of the detection parameters. The idea is to determine if a violation of homogeneity did occur and if it did when is needed to know: the number of change-points; their location; and if the violation is permanent or temporary.
Let's divide this subject in two possible situations, a single heterogeneity and a multiple one.

- **Single heterogeneity**

a) The 'ideal' detection - It happens when the series $F_N^{(2)}$ is the zero series.
The series $F_N^{(1)} = F_N$ and assuming that exists $Q' \prec N$ such that the series $F_{1,Q'-1}$ is a homogeneous series, and the dimension *d* of its minimal LRF is less than $Q'/2$. By definition, the maximal Q' coincides with Q.
B and *L* should be chosen accordingly with the knowledge that any $B \succ 2d$ and *L* such that $d \prec \min(L, B-L+1)$.
Any subseries $F_{i,i+B-1}$, such that $i \le Q - B$, and considering the SVD of its trajectory *L*-trajectory matrix, then for some *r* all the eigenvalues $\lambda_s^{(i)}$ with $s \succ r$ are equal to zero. Therefore, $I = [1,2,\ldots r]$ and $r = d$. All the nonzero elements of the matrix indicate the existence of some heterogeneity in the series.

b) Nonzero $F_N^{(2)}$: identification – To solve this problem, is better to assume that $F_N = F_N^{(1)} + F_N^{(2)}$ holds but the addends are unknown. In this way what will be studied is the entire series F_N. Some assumptions needs to be taken: i) $F_{1,Q'-1}^{(1)}$ is homogeneous; ii) the chosen set *I* of eigentriples correspond to the subseries $F_{i,i+B-1}^{(1)}$ for all $i = 1,\ldots,Q'-B$ (as a rule it is a set of several eigentriples); iii) the heterogeneity under detection is happening at the series $F_N^{(1)}$.
By the above assumptions, for the selected B and *L* it exist certain eigentriples of the trajectory matrices of the subseries $f_{i,i+B-1}(i \le Q'-B)$ stably interpretable as (approximately) describing subseries of the same homogeneous series. Until the moment Q' the series $F_N^{(1)}$ is (approximately) identified by the obtained set *I* of the eigentriples, and thus have the detection parameters *r* and *I*. In some case the identification procedure is easy to perform in others can not be done.

c) Small noisy-like $F_N^{(2)}$ - This case is similar to the 'ideal' detection case.

To obtain (approximate) separability of the series $F_{i,i+B-1}^{(1)}$ and $F_{i,i+B-1}^{(2)}$ up to the change-point, a relatively large *B* needs to be taken (it can not be larger than the expected value of *Q*). *L* should be chosen to be approximately equal to *B/2*.

Since the series $F_N^{(2)}$ is small enough and in view of the (approximate) separability obtained, several *r* leading singular values produced by the series $F_{i,i+B-1}$, must be large enough and describe the series $F_{i,i+B-1}^{(1)}$, while the other singular values are expected to be small. Therefore and abrupt decrease of the singular values, placed in decreasing order of their magnitudes, may help finding the number *r* and the set $I = [1,2,\ldots,r]$.

If all the parameters were chosen correctly, then the corresponding heterogeneity matrix will have small elements in the block *A*.

If $F_{1,Q-1}^{(1)}$ is harmonic with amplitude C and $F_N^{(2)}$ is a white noise with variance σ^2, then asymptotically in N and other parameters, the elements of the block *A* are close to $\sigma^2 / \left(0.5C^2 + \sigma^2\right)$. If $\sigma^2 \prec\prec C^2$, the block *A* is zero-like.

d) General $F_N^{(2)}$ - The goal of obtaining Q' as large as possible it may contradict in the case of general $F_N^{(2)}$.

The detection problem is complicated in view of the possibility that the detection background (the block *A*) may contain large elements in certain columns. If separability is approximate, then the equal-row background is perturbed, and it is difficult to recognize a possible heterogeneity on the non-constant background.

The same applies when the entire series is generally increasing and decreasing, when the detection background varies in a monotone way, and the heterogeneity recognition is even more complicated.

However, there are no general rules for the choice of B and L in all 'simple' situations. As a result, if the detection is performed in a situation close to 'ideal', then large values of the heterogeneity index indicate heterogeneity, and the general form of the H-matrix can help to identify both the change-point and the type of the heterogeneity.

- **Multiple heterogeneity**

It is useful to search sequentially for the change-points and heterogeneities. This is done by producing sequential H-matrices until the end of the series is reached. The collection of H-matrices obtained in this way would give the entire description of the situation.

Detection functions

The row detection function is the best way to detect the first change-point of the series. The change-point coincides with the first point of sharp increase of this function;

The diagonal detection function indicates more clearly change-points if the series has a slowly varying structure. It needs to be taken with care because a single heterogeneity may give rise to several sharp peaks on the plot of this detection function.

The symmetric detection function can only be used to characterize the local description of the series F_N by a fixed set of eigentriples.

The column detection function despite the fact that is weak in detecting heterogeneities is often informative when the idea is to distinguish the heterogeneity from the eigentriple rearrangement.

The difference between the row and the column detection function is a good indicator of the true heterogeneity. If the column function has a sharp increase and the row function is slowly varying, then is must certain that there is an eigentriple rearrangement.

- **Heterogeneity in trends**

The trend of the series is associated with its low frequency component. To separate it from the other components of all series $F_{i,i+B-1}$ with the help of a stable set of eigentriples, a small *B* must be taken. A sufficiently large trend is described by the single leading eigentriple of the SVD of the trajectory matrix of the series $F_{i,i+B-1}$. Therefore the series $\widetilde{F}_{i,i+B-1}$, reconstructed from the leading eigentriple must be similar to the exponential series with some rate α_i, that is for the series of the form $g_n = c_i e^{\alpha_i n}$ with some c_i and $n = 1, \ldots i + B - 2$. On time intervals where the trend changes its behavior, the rates also change, and is the case of (approximate) permanent violations in the piecewise exponential series. Therefore, sharp changes in the trend behavior will be detected via the increase of the detection functions.

The H-matrix will have small values of the heterogeneity index in block *A* and *D*, and in all the rectangles along the main diagonal. Other blocks of homogeneity can also have small elements. Despite the fact that the heterogeneity under consideration is of a permanent type, the heterogeneity matrix is going to be ‘cross-structured’.

- **Heterogeneity in periodicities**

In view of the periodic feature of the signal, the detection parameters *B*, *L* and *T* should be proportional to the period of the series.

At any rate, at least the block A of the heterogeneity matrix will consist of approximately equal elements.

- **The role of the parameter *T***

Small values of *T* imply both a large contrast in the detection and a high sensitivity to small perturbations of the series. By enlarging *T*, the contrast between small and large values of the detection function is reduced and these functions are smoother.

The minimal *T* values is *T* = *L*.

When dealing with periodic series, T must be proportional to the period of the series.

- **Detection Characteristics**

There are various additional detection characteristics which can help to identify and interpret heterogeneities in time series. They can be divided in three major groups:

Renormalized heterogeneity matrices. By definition the heterogeneity index is normalized, because when pure homogeneity exists all values are zero, and in pure heterogeneity all values are one. When the series is positive and monotone increasing the denominator of the row detection function increases as well. Therefore, the heterogeneity index of the last part of the series is generally smaller than the analogous index of the initial interval of the series only because of the increase of the entire series. This makes that the background is non-constant and two ‘equivalent’ heterogeneities occur at the beginning and at the end of the series producing different increases of the heterogeneity characteristics. To avoid all these the heterogeneity index should be denormalized, by omitting the denominator of formula, which creates the following formula:

$$\tilde{g}_{ij}\, \underline{\underline{def}}\, \frac{\frac{1}{(T-L+1)} \sum_{l=1}^{T-L+1} dist^2\left(X_{l,j}, L_{I,B}^{(L,i)}\right)}{\frac{1}{N}\sum_{k=0}^{N-1} f_k^2}$$

where *B*, *L*, *T*, *r* and *I* are fixed detection parameters and $X_{l,j}$ are the L-lagged vectors of the series $F_{j,j+T-1}$. In this definition is used the squared sum of all the elements of the series F_N as the denominator and take averaging coefficients in agreement with the total number of the terms of the series in all the sums.

This new heterogeneity index may also be very helpful in the detection of change-points in the variance of the noise.

The Roots of characteristics polynomials is related with the variations in linear spaces $L_{I,B}^{(L,i)}$.

The root functions of the characteristics polynomial seem to be preferable for the purpose of monitoring the homogeneity of the series.

Characteristics related with moving periodograms which describes the changes in the spectral structure of the initial series in time.

7. Annex 2 – Single-spectrum analysis – Forecasting theory

7.1. SSA recurrent forecasting algorithm

Some of the algorithm inputs, notations, comments and properties:

Time series $F_N = \left(f_{0,\dots,f_{N-1}}\right), N \succ 2$.

Window length L, 1 < L < N.

Linear space $L_r \subset \Re^L$ of dimension *r* < *L*. It is assumed that $e_L \notin L_r$, where $e_L = (0,0,\dots,0,1)^T \in \Re^L$. In other terms, *L_r* is not a 'vertical' space but is defined by its certain orthonormal basis (the forecast result do not depend on this concrete basis).

Number M of points to forecast for.

$X = [X_1 : \dots : X_K]$ (where *K* = *N* – *L* + *1*) is the trajectory matrix of the time series F_N.

$P_1, \dots, P_r$ is an orthonormal basis in L_r.

$\hat{X}\, \underline{\underline{def}} \left[\hat{X}_1 : \dots : \hat{X}_K\right] = \sum P_i P_i^T X$. The vector $\hat{X}_i$ is the orthogonal projection of X_i onto the space L_r.

$\tilde{X} = \mathrm{H}\hat{X} = \left[\tilde{X}_1 : \ldots : \tilde{X}_K\right]$is the result of the Hankelization of the matrix $\hat{X}$. The matrix $\tilde{X}$ is the trajectory matrix of some time series $\tilde{F}_N = \left(\tilde{f}_0, \ldots, \tilde{f}_{N-1}\right)$.

For any vector $Y \in \Re^L$, $Y_\Delta \in \Re^{L-1}$ is the vector consisting of the last *L – 1* components of the vector *Y*, while $Y_\nabla \in \Re^{L-1}$ is the vector consisting of the first *L – 1* components of *Y*.

$v^2 = \pi_1^2 + \ldots + \pi_r^2$, where π_i is the last component of the vector $P_i(i = 1, \ldots, L)$. Since v^2 is the squared cosine of the angle between the vector e_L and the linear space *L_r*, it should be called the *verticality coefficient of L_r*.

As have already been said $e_L \notin L_r$ so $v^2 \prec 1$. In this case the last component y_L of any vector $Y = \left(y_1, \ldots, y_L\right)^T$ is a linear combination of the first components $y_1, \ldots, y_{L-1}$:

$y_L = a_1 y_{L-1} + a_2 y_{L-2} + \ldots + a_{L-1} Y_1$

Vector $R = \left(a_{L-1}, \ldots, a_1\right)^T$ can be expressed as $R = \frac{1}{1 - v^2} \sum_{i=1}^{r} \pi_i P_i^\nabla$ and does not depend on the choice of a basis $P_1, \ldots, P_r$ in the linear space L_r.

The series $G_{N+M} = \left(g_0, \ldots g_{N+M-1}\right)$ is the result of:

$$g_i = \begin{cases} \tilde{f}_i & \text{for i = 0, … , N – 1} \\ \sum_{j=1}^{L-1} a_j g_{i-j} & \text{for i = N, … , N + M - 1} \end{cases}$$

where the numbers $g_N, \ldots, g_{N+M-1}$ form the M terms of the SSA recurrent forecast (or only SSA R-forecasting algorithm).

If $P^{(r)} : L_r \to \Re^L$ is defined as a linear operator by the formula $P^{(r)}Y = \begin{pmatrix} Y_\Delta \\ R^T Y_\Delta \end{pmatrix}, Y \in L_r$, and setting

$$Z_i = \begin{cases} \tilde{X}_i & \text{for i = 0, … , K} \\ P^{(r)} Z_{i-1} & \text{for } i = K + 1, \ldots, K + M \end{cases}$$

The matrix $Z = \left[Z_1 : \ldots : Z_{K+M}\right]$is the trajectory matrix of the series G_{N+M}.

It is evident that the initial points $g_{N-L+1}, \ldots, g_{N-1}$ of the forecasting recurrent formula coincide with the last *L-1* terms of the series *F*.

The series F_N admits a continuation in L_r if there is an uniquely defined $\tilde{f}_N$ such that all *L*-lagged vectors of the series $\tilde{F}_{N+1} = \left(f_o, \ldots, f_{N-1}, \tilde{f}_N\right)$ belong to $L^{(L)}$. In this case the series $\tilde{F}_{N+1}$ will be called the *one-step L-continuation* of the series F_N.

If $e_L \in L_r$, then F_N does not admit *L*-continuation. Consequently if $d = L$, then the series cannot be *L*-continued since the uniqueness condition does not apply.

If $d \prec L \le K$ and $e_L \notin L_r$, then the series F_N admits *L*-continuation.

The one step continuation formula is: $\tilde{f}_N = \sum_{k=1}^{L-1} a_k f_{N-k}$ where the vector $R = \left(a_{L-1}, \ldots, a_1\right)^T$ which was identified above.

The series F_N is governed by the LRF: $f_{i+L} = \sum_{k=1}^{L-1} a_k f_{i+L-k}$, $0 \le i \le N - L - 1$.

If the series F_N admits a one-step *L*-continuation, then it can be *L*-continued for an arbitrary number of steps.

If a series F_N satisfy a LRF: $f_{i+d_o} = \sum_{k=1}^{d_0} b_k f_{i+d-k}$, $0 \le i \le N - d_o - 1$ and $d_0 \le \min(L-1, K)$, then $d \le d_o, e_L \notin L^{(L)}$ and the series will admit L-continuation, produced by the above formula.

The concepts of recurrent continuation and *L*-continuation are equivalent.

Nevertheless, it is not realistic to believe that the series are governed by some LRF of relatively small dimension; the exact continuation is mainly methodological and theoretical. Therefore, the concept of approximate continuation is more realistic and helpful.

7.2. Approximate continuation

Assuming that the following conditions hold:

The series of length *N* and window length *L* provide approximate strong separability of the series $F_N^{(1)}$ and $F_N^{(2)}$;

$X = \sum_i \sqrt{\lambda_i} U_i V_i^T$ is the SVD of the trajectory matrix X of the series F_N. The choice of the eigentriples $\left\{\sqrt{\lambda_i} U_i V_i^T\right\}_{i \in I}$, $I = (i_1, \ldots, i_r)$ associated with $F_N^{(1)}$ allows achieving approximate separability;

$d\,\underline{\underline{def}}\,f \dim\left(F_N^{(1)}\right) \le r \le L \le K$;

$e_L \notin span(U_i, i \in I)$, meaning that $\sum_{i \in I} u_{iL}^2 \prec 1$, where u_{iL} is the last component of the eigenvector U_i.

Than the Basic SSA R-forecasting algorithm can be applied, and the result $g_N, \ldots g_{N+M-1}$ is called the *approximate recurrent continuation* of the series F_N.

Usually and due to the fact that forecasting errors occur the forecast series g_N do not coincide with recurrent continuation of the series $F_N^{(1)}$. The errors can be of two types, first it happens because the LRF is produced by the vector *R* which is strongly related to the space L_r, and the discrepancy with this space and the space $L^{(L,1)}$ produces the error, in particular because the finite-difference dimension of the forecast series $g_N(n \ge N)$ is generally greater than *d*. Secondly, the error can be produced by the initial data for the forecast. For recurrent continuation, the initial data is $f_{N-L+1}^{(1)}, \ldots, f_{N-1}^{(1)}$, where $f_n^{(1)}$ is the *n*th term of the series $F_N^{(1)}$. In the Basic R-forecasting algorithm the initial data are the last L-1 terms $g_{N-L+1}, \ldots, g_{N-1}$ of the reconstructed series. Since $f_n^{(1)} \ne g_n$, the initial series produces its own error.

When the quality of the approximate separability is "good" is expected that the SSA R-forecasting produces a reasonable approximation to recurrent continuation of $F_N^{(1)}$.

7.3. Modifications to Basic SSA R-algorithm

There are some specific situations when some modifications to the Basic SSA R-algorithm can be helpful in forecasting more accurately.

SSA V-forecasting: Both V and R forecasting works with two general stages, diagonal averaging and continuation. For R-forecasting, diagonal averaging is used to obtain the reconstructed series, and continuation is performed by applying the LRF. In the V-forecasting, these two stages are used in the reverse order, first vector continuation in L_r is performed and then diagonal averaging gives the forecast values. When a series admit recurrent continuation, both forecast methods provide the same results. In case of only approximate continuation than the results differ. As poor as the approximations is as large the difference between forecast values will be. The forecast stability can be "proved" if the two forecasting values are close. In cases of rapid increase or decrease of the R-forecasting values, V-forecasting tends to be more "conservative".
Toeplitz SSA forecasting: both V and R-forecasting can be applied to a series which have been decomposed using Toeplitz SSA. Therefore, for stationary time series Toeplitz SSA forecasting may give more stable results.
Centring in SSA forecasting: When reconstructing a component of a time series with the help of the single centring variant of the Basic or Toeplitz SSA, the average triple can be either included into the list of the eigentriples selected for reconstruction or not. In the case when the average triple is not taken for reconstruction everything holds the same for both V and R-forecasting, except the matrix $\hat{X}$, which is modified for: $\hat{X} = \left[\hat{X}_1 : \ldots \hat{X}_K\right] = \sum_{i=1}^{r} P_i P_i^T (X - A)$, where $A = [\varepsilon : \ldots \varepsilon]$and the vector ε has the form $\varepsilon = (X_1 + \ldots + X_K)/K$. In the case that the average triple is included in the reconstructing than the matrix $\hat{X}$ takes the following definition: $\hat{X} = \left[\hat{X}_1 : \ldots \hat{X}_K\right] = \sum_{i=1}^{r} P_i P_i^T (X - A) + A$, with the same notation as above. Some of the formulas for both V and R-forecasting are also changed to include the centring.

Also, very important is to mention that double centring can not be used for forecasting. The main reason for that is that the double centring is applied to both the rows and the columns of the trajectory matrix, while the SSA forecasting algorithm and all its modifications and variants are based on the linear space L_r, which is associated only with the columns of the trajectory matrix.

7.4. Forecast Confidence Bounds

There are two different problems when constructing confidence bounds for the forecast. The first is to construct confidence interval for the entire series $F = F^{(1)} + F^{(2)}$ at some future point in time *N+M*. The second is to construct confidence intervals for the signal $F^{(1)}$ at the same future point in time.

The first problem will be solved by using the information about the forecast errors obtained by processing the series. This can be called the empirical variant.
The second problem needs some additional information about the model governing the series $\tilde F_N^{(2)}$ to apply a bootstrap simulation of the series F_N.

7.4.1. Empirical variant

The *multistart M-step recurrent continuation* procedure stats that: taking a relatively small integer M and apply M steps of recurrent continuation produced by the forecasting LRF modifying the initial data from $\left(\tilde f_o^{(1)},\ldots,\tilde f_{L-2}^{(1)}\right)$to $\left(\tilde f_{K-M}^{(1)},\ldots,\tilde f_{N-M-1}^{(1)}\right)$, $K = N - L + 1$. The last points $g_{j+M+L-1}$ of these continuations can be compared with the values $f_{j+M+L-1}$ of the initial series F_N. A *multistrat M-step residual series* H_{K-M+1} with $h_j^{(M)} = f_{j+M+L-2} - g_{j+M+L-2}, j = 0,\ldots,K-M$.

If the reconstructed series $\tilde F_N^{(1)}$ coincides with $F_N^{(1)}$ and the forecasting LRF governs it, than $g_k = f_k^{(1)}$ and the multistrat M-step residual series coincides with the last K – M +1 terms of the stationary noise series $F_N^{(2)}$. If this is not true, but assuming that the multistrat series is stationary and ergodic in the sense that its empirical cumulative function tends to the theoretical empirical cumulative function of the series as $N \to \infty$ .Then, having the series H_{K-M+1} means that certain of its quantiles can be estimated.

Because the terms $g_{j+M+L-2}$ are obtained through the same number of steps with the same LRF as the forecast values $\tilde f_{N+M-1}^{(1)}$, and their initial data is taken from the same reconstructed series, and because the forecasting requires the assumption that the series structure is kept in the future, the obtained empirical cumulative distribution function of the multistrat *M*-step residual series can be used to construct the empirical confidence interval for f_{N+M-1}.

The empirical confidence interval is: $\left(\tilde f_{N+M-1}^{(1)} + c_{\alpha/2}^{-}, \tilde f_{N+M-1}^{(1)} + c_{\alpha/2}^{+}\right)$, with the confidence level $\gamma(0 \prec \gamma \prec 1)$, and $\alpha = 1 - \gamma$, $c_{\alpha/2}^{-}$ and $c_{\alpha/2}^{+}$ the lower and upper $\alpha/2$ - quantiles.

If the multistrat *M*-step residual series can be regarded as white noise, then the other variant of empirical confidence intervals is meaningful.

This type of confidence intervals can only be used for short-term forecasting.

These confidence intervals are constructed for the entire series F_N.

7.4.2. Bootstrap confidence bounds for the forecast of a signal

If it could be assumed (unrealistically) that the rules of the eigentriples selection are fixed, S independent copies $F_{N,i}^{(2)}$ of the process $F_N^{(2)}$ could be simulated. The forecasting procedure would then be applied to the S independent time series $F_{N,i} \underline{\underline{def}} F_N^{(1)} + F_N^{(2)}$. The forecasting results would form a sample $\tilde f_{N+M-1,i}^{(1)} (1 \le i \le S)$, which should be compared against $f_{N+M-1}^{(1)}$. In this way the *Monte Carlo confidence bounds* for the forecast could be built up.

Since in practice the signal $F_N^{(1)}$ is not known, this procedure can not be applied.

But, under a suitable choice of the window length L and the corresponding eigentriples, the representation $F_N = \tilde{F}_N^{(1)} + \tilde{F}_N^{(2)}$ is known, where $\tilde{F}_N^{(1)}$ approximates $F_N^{(1)}$, and $\tilde{F}_N^{(2)}$ is the residual series. If a stochastic model of the residuals $\tilde{F}_N^{(2)}$ exists, for instance, a model can be postulated, and since $\tilde{F}_N^{(1)} \approx F_N^{(1)}$, the same model can be applied to $\tilde{F}_N^{(2)}$ with the estimated parameters.

After these steps simulating S independent copies $\tilde{F}_{N,i}^{(2)}$ of the series $F_N^{(2)}$, it will be obtained S series $F_{N,i}\,\underline{\underline{def}}\,\tilde{F}_N^{(1)} + \tilde{F}_N^{(2)}$, and S forecasting results $\tilde{f}_{N+M-1}^{(1)}$ are produced as in the Monte Carlo simulation variant.

As soon as the sample $\tilde{f}_{N+M-1,i}^{(1)}\left(1 \le i \le S\right)$ of the forecasting results is obtained, the lower and upper quantiles for a fixed level γ can be calculated and confidence intervals for the forecast can be obtained. The interval, called bootstrap confidence interval, can be compared with the forecast value $\tilde{f}_{N+M-1}^{(1)}$ obtained from the initial forecasting procedure, being the discrepancies between this value and the obtained confidence interval caused by the inaccuracy of the stochastic model for $\tilde{F}_N^{(2)}$.

The average of the bootstrap forecast sample estimates the mean value of the forecast and the mean square deviation of the sample shows the accuracy of the estimate.

The Monte Carlo forecast of the signal $F_N^{(1)}$ is useful in at least two respects: its average (Monte Carlo average forecast) shows the bias produced by the corresponding forecasting procedure, while the upper and lower quantiles indicate the role of the random component in the forecasting error.

The Bootstrap confidence intervals are built for the continuation of the signal $F_N^{(1)}$.

8. Annex 3 – Data analysis – All series

8.1. Time series A

N = 108; L = 12; B = 55; T = 12; K = 44; I = $\{1,2,3,4,5,6\}$

Decomposition method: Basic SSA

Maximum ν^2 = 0,46

Eigentriples for reconstruction = 1, 2, 3-4, 5-6

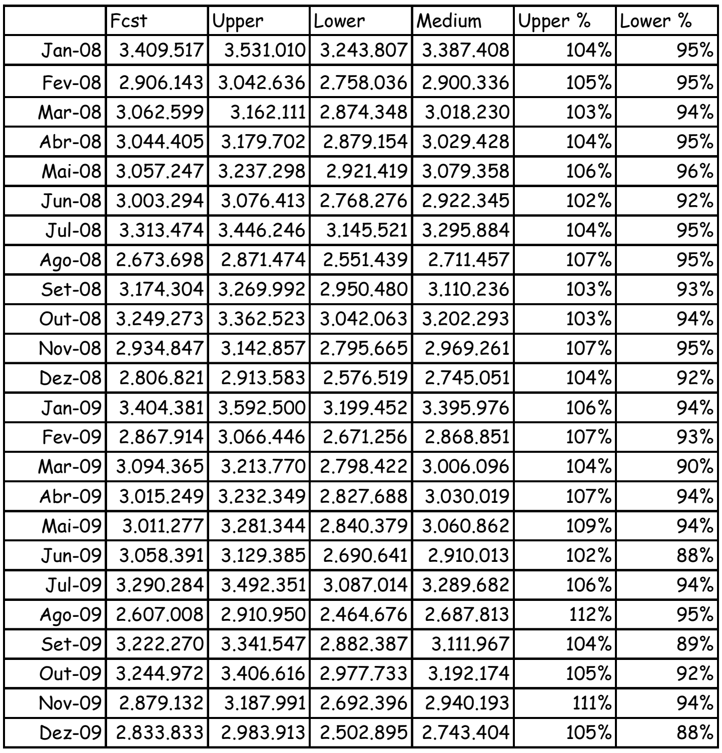
Maximum Relative error of reconstruction = 2,76%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 4, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.



	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	3.409.517	3.531.010	3.243.807	3.387.408	104%	95%
Fev-08	2.906.143	3.042.636	2.758.036	2.900.336	105%	95%
Mar-08	3.062.599	3.162.111	2.874.348	3.018.230	103%	94%
Abr-08	3.044.405	3.179.702	2.879.154	3.029.428	104%	95%
Mai-08	3.057.247	3.237.298	2.921.419	3.079.358	106%	96%
Jun-08	3.003.294	3.076.413	2.768.276	2.922.345	102%	92%
Jul-08	3.313.474	3.446.246	3.145.521	3.295.884	104%	95%
Ago-08	2.673.698	2.871.474	2.551.439	2.711.457	107%	95%
Set-08	3.174.304	3.269.992	2.950.480	3.110.236	103%	93%
Out-08	3.249.273	3.362.523	3.042.063	3.202.293	103%	94%
Nov-08	2.934.847	3.142.857	2.795.665	2.969.261	107%	95%
Dez-08	2.806.821	2.913.583	2.576.519	2.745.051	104%	92%
Jan-09	3.404.381	3.592.500	3.199.452	3.395.976	106%	94%
Fev-09	2.867.914	3.066.446	2.671.256	2.868.851	107%	93%
Mar-09	3.094.365	3.213.770	2.798.422	3.006.096	104%	90%
Abr-09	3.015.249	3.232.349	2.827.688	3.030.019	107%	94%
Mai-09	3.011.277	3.281.344	2.840.379	3.060.862	109%	94%
Jun-09	3.058.391	3.129.385	2.690.641	2.910.013	102%	88%
Jul-09	3.290.284	3.492.351	3.087.014	3.289.682	106%	94%
Ago-09	2.607.008	2.910.950	2.464.676	2.687.813	112%	95%
Set-09	3.222.270	3.341.547	2.882.387	3.111.967	104%	89%
Out-09	3.244.972	3.406.616	2.977.733	3.192.174	105%	92%
Nov-09	2.879.132	3.187.991	2.692.396	2.940.193	111%	94%
Dez-09	2.833.833	2.983.913	2.502.895	2.743.404	105%	88%

Table 4 - Times series A-Forecast

8.2. Time series B

N = 108; L = 12; B = 61; T = 12; K = 50; I = $\{1,2,3,4,5,6\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,51

Eigentriples for reconstruction = 1, 2, 3-6

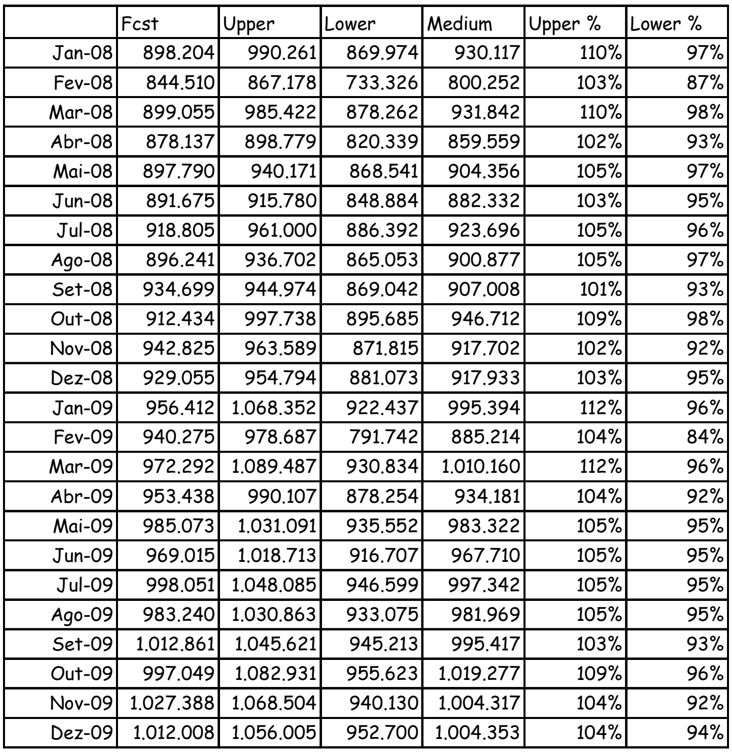
Maximum Relative error of reconstruction = 3,30%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 5, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.



	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	898.204	990.261	869.974	930.117	110%	97%
Fev-08	844.510	867.178	733.326	800.252	103%	87%
Mar-08	899.055	985.422	878.262	931.842	110%	98%
Abr-08	878.137	898.779	820.339	859.559	102%	93%
Mai-08	897.790	940.171	868.541	904.356	105%	97%
Jun-08	891.675	915.780	848.884	882.332	103%	95%
Jul-08	918.805	961.000	886.392	923.696	105%	96%
Ago-08	896.241	936.702	865.053	900.877	105%	97%
Set-08	934.699	944.974	869.042	907.008	101%	93%
Out-08	912.434	997.738	895.685	946.712	109%	98%
Nov-08	942.825	963.589	871.815	917.702	102%	92%
Dez-08	929.055	954.794	881.073	917.933	103%	95%
Jan-09	956.412	1.068.352	922.437	995.394	112%	96%
Fev-09	940.275	978.687	791.742	885.214	104%	84%
Mar-09	972.292	1.089.487	930.834	1.010.160	112%	96%
Abr-09	953.438	990.107	878.254	934.181	104%	92%
Mai-09	985.073	1.031.091	935.552	983.322	105%	95%
Jun-09	969.015	1.018.713	916.707	967.710	105%	95%
Jul-09	998.051	1.048.085	946.599	997.342	105%	95%
Ago-09	983.240	1.030.863	933.075	981.969	105%	95%
Set-09	1.012.861	1.045.621	945.213	995.417	103%	93%
Out-09	997.049	1.082.931	955.623	1.019.277	109%	96%
Nov-09	1.027.388	1.068.504	940.130	1.004.317	104%	92%
Dez-09	1.012.008	1.056.005	952.700	1.004.353	104%	94%

Table 5 - Times series B-Forecast

8.3. Time series C

N = 108; L = 36; B = 55; T = 36; K = 20; I = $\{1,2,3,4,5,6,7\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,48

Eigentriples for reconstruction = 1, 2, 3-4, 5-7

Maximum Relative error of reconstruction = 3,05%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 6, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.4. Time series D

N = 108; L = 12; B = 73; T = 12; K = 62; I = $\{1,2,3,4,5,6\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,42

Eigentriples for reconstruction = 1, 2-3, 4, 5-6

Maximum Relative error of reconstruction = 2,45%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 7, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	4.376.902	4.519.128	4.056.711	4.287.920	103%	93%
Fev-08	3.985.470	4.165.960	3.762.529	3.964.244	105%	94%
Mar-08	4.382.484	4.588.268	4.138.361	4.363.314	105%	94%
Abr-08	4.160.313	4.348.975	3.944.139	4.146.557	105%	95%
Mai-08	4.320.681	4.480.931	4.111.700	4.296.315	104%	95%
Jun-08	4.160.235	4.429.058	4.048.085	4.238.571	106%	97%
Jul-08	4.494.295	4.632.367	4.230.895	4.431.631	103%	94%
Ago-08	4.088.656	4.374.097	3.935.711	4.154.904	107%	96%
Set-08	4.504.906	4.669.631	4.261.792	4.465.711	104%	95%
Out-08	4.258.408	4.462.133	4.060.351	4.261.242	105%	95%
Nov-08	4.432.441	4.563.233	4.151.479	4.357.356	103%	94%
Dez-08	4.274.501	4.494.843	4.084.004	4.289.423	105%	96%
Jan-09	4.602.865	4.785.558	4.205.271	4.495.415	104%	91%
Fev-09	4.195.910	4.461.817	3.928.549	4.195.183	106%	94%
Mar-09	4.628.591	4.862.637	4.304.744	4.583.690	105%	93%
Abr-09	4.361.980	4.609.821	4.102.701	4.356.261	106%	94%
Mai-09	4.548.221	4.744.066	4.273.131	4.508.599	104%	94%
Jun-09	4.393.285	4.712.124	4.220.369	4.466.247	107%	96%
Jul-09	4.713.633	4.899.039	4.360.847	4.629.943	104%	93%
Ago-09	4.307.548	4.667.296	4.094.810	4.381.053	108%	95%
Set-09	4.754.460	4.961.851	4.424.123	4.692.987	104%	93%
Out-09	4.468.178	4.740.646	4.204.082	4.472.364	106%	94%
Nov-09	4.667.472	4.873.340	4.317.818	4.595.579	104%	93%
Dez-09	4.514.764	4.799.292	4.245.697	4.522.495	106%	94%

Table 6 - Times series C-Forecast

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	1.122.424	1.181.647	991.205	1.086.426	105%	88%
Fev-08	1.066.946	1.127.509	948.171	1.037.840	106%	89%
Mar-08	1.176.240	1.282.821	1.106.648	1.194.734	109%	94%
Abr-08	1.087.623	1.211.464	1.028.062	1.119.763	111%	95%
Mai-08	1.223.808	1.334.142	1.158.655	1.246.398	109%	95%
Jun-08	1.197.741	1.258.804	1.078.602	1.168.703	105%	90%
Jul-08	1.348.199	1.398.964	1.217.257	1.308.111	104%	90%
Ago-08	1.238.546	1.306.870	1.121.951	1.214.410	106%	91%
Set-08	1.252.679	1.374.211	1.198.458	1.286.334	110%	96%
Out-08	1.054.788	1.191.876	1.004.994	1.098.435	113%	95%
Nov-08	1.089.992	1.202.543	1.022.568	1.112.556	110%	94%
Dez-08	989.257	1.038.031	856.078	947.055	105%	87%
Jan-09	1.110.793	1.166.057	946.602	1.056.330	105%	85%
Fev-09	1.032.777	1.112.210	886.352	999.281	108%	86%
Mar-09	1.139.109	1.282.525	1.062.929	1.172.727	113%	93%
Abr-09	1.069.412	1.227.585	1.002.512	1.115.049	115%	94%
Mai-09	1.227.957	1.365.275	1.141.034	1.253.155	111%	93%
Jun-09	1.205.597	1.276.992	1.049.859	1.163.426	106%	87%
Jul-09	1.340.559	1.400.252	1.177.986	1.289.119	104%	88%
Ago-09	1.213.851	1.300.961	1.067.243	1.184.102	107%	88%
Set-09	1.228.288	1.378.959	1.157.037	1.267.998	112%	94%
Out-09	1.045.809	1.210.032	978.960	1.094.496	116%	94%
Nov-09	1.094.020	1.240.521	1.005.535	1.123.028	113%	92%
Dez-09	987.808	1.059.459	835.479	947.469	107%	85%

Table 7 - Times series D-Forecast

8.5. Time series G

N = 108; L = 36; B = 55; T = 36; K = 20; I = $\{1,2,3,4\}$
Decomposition method: Basic SSA

Maximum v^2 = 0,23
Eigentriples for reconstruction = 1, 2, 3-4
Maximum Relative error of reconstruction = 2,30%
Forecast Type = V forecast
Confidence bounds type = Bootstrap
Interactions = 1000
Forecast – Table 8, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	1.511.720	1.545.081	1.438.758	1.491.920	102%	95%
Fev-08	1.333.471	1.371.264	1.261.807	1.316.536	103%	95%
Mar-08	1.443.303	1.505.825	1.397.928	1.451.877	104%	97%
Abr-08	1.425.315	1.445.282	1.333.525	1.389.404	101%	94%
Mai-08	1.435.460	1.489.937	1.376.204	1.433.071	104%	96%
Jun-08	1.356.345	1.404.571	1.293.661	1.349.116	104%	95%
Jul-08	1.528.542	1.564.361	1.448.511	1.506.436	102%	95%
Ago-08	1.347.457	1.388.776	1.269.126	1.328.951	103%	94%
Set-08	1.459.908	1.527.349	1.409.184	1.468.266	105%	97%
Out-08	1.440.891	1.461.341	1.339.531	1.400.436	101%	93%
Nov-08	1.450.860	1.511.077	1.387.172	1.449.124	104%	96%
Dez-08	1.371.705	1.422.823	1.301.153	1.361.988	104%	95%
Jan-09	1.545.544	1.584.447	1.457.889	1.521.168	103%	94%
Fev-09	1.361.592	1.406.581	1.276.155	1.341.368	103%	94%
Mar-09	1.476.709	1.549.807	1.420.191	1.484.999	105%	96%
Abr-09	1.456.629	1.477.890	1.344.868	1.411.379	101%	92%
Mai-09	1.466.427	1.533.127	1.397.882	1.465.505	105%	95%
Jun-09	1.387.245	1.441.681	1.308.086	1.374.884	104%	94%
Jul-09	1.562.728	1.605.259	1.466.923	1.536.091	103%	94%
Ago-09	1.375.877	1.425.109	1.282.677	1.353.893	104%	93%
Set-09	1.493.709	1.572.909	1.430.967	1.501.938	105%	96%
Out-09	1.472.531	1.494.991	1.349.638	1.422.314	102%	92%
Nov-09	1.482.165	1.556.186	1.408.319	1.482.253	105%	95%
Dez-09	1.402.964	1.460.781	1.314.565	1.387.673	104%	94%

Table 8 - Times series G-Forecast

8.6. Time series H

N = 108; L = 12; B = 73; T = 12; K = 62; I = $\{1,2,3,4\}$
Decomposition method: Basic SSA
Maximum v^2 = 0,42
Eigentriples for reconstruction = 1, 2, 3-4
Maximum Relative error of reconstruction = 2,02%
Forecast Type = V forecast
Confidence bounds type = Bootstrap
Interactions = 1000
Forecast – Table 9, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	280.335	298.508	261.977	280.243	106%	93%
Fev-08	264.607	279.143	243.431	261.287	105%	92%
Mar-08	275.383	294.445	258.617	276.531	107%	94%
Abr-08	255.552	271.081	234.774	252.927	106%	92%
Mai-08	263.846	283.457	247.848	265.653	107%	94%
Jun-08	244.812	261.205	222.955	242.080	107%	91%
Jul-08	255.731	275.167	239.700	257.434	108%	94%
Ago-08	241.603	256.849	219.616	238.232	106%	91%
Set-08	257.382	275.904	241.725	258.814	107%	94%
Out-08	248.058	262.462	226.001	244.231	106%	91%
Nov-08	266.400	286.693	249.198	267.946	108%	94%
Dez-08	257.974	273.803	234.377	254.090	106%	91%
Jan-09	274.343	297.016	255.568	276.292	108%	93%
Fev-09	262.695	279.521	238.152	258.837	106%	91%
Mar-09	274.334	298.122	254.980	276.551	109%	93%
Abr-09	258.623	276.783	232.261	254.522	107%	90%
Mai-09	266.750	291.941	246.043	268.992	109%	92%
Jun-09	249.871	269.284	221.554	245.419	108%	89%
Jul-09	258.348	283.278	238.131	260.704	110%	92%
Ago-09	244.188	262.470	216.851	239.660	107%	89%
Set-09	256.059	279.794	238.040	258.917	109%	93%
Out-09	246.130	264.396	219.326	241.861	107%	89%
Nov-09	261.111	287.098	242.120	264.609	110%	93%
Dez-09	253.450	273.733	224.859	249.296	108%	89%

Table 9 - Times series H-Forecast

8.7. Time series J

N = 108; L = 48; B = 55; T = 48; K = 8; I = $\{1,2,3,4,5,6,7\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,27

Eigentriples for reconstruction = 1, 2-3, 4-5, 6-7

Maximum Relative error of reconstruction = 4,90%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 10, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.8. Time series L

N = 108; L = 12; B = 73; T = 12; K = 62; I = $\{1,2,3,4,5,6,7,8\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,40

Eigentriples for reconstruction = 1, 2, 3-4, 5-6, 7-8

Maximum Relative error of reconstruction = 5,20%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 11, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	1.198.595	1.406.746	929.491	1.168.118	117%	78%
Fev-08	1.346.430	1.583.802	1.081.127	1.332.465	118%	80%
Mar-08	1.092.729	1.359.783	867.424	1.113.603	124%	79%
Abr-08	968.042	1.227.976	735.354	981.665	127%	76%
Mai-08	1.096.528	1.343.113	835.140	1.089.127	122%	76%
Jun-08	981.435	1.228.419	732.221	980.320	125%	75%
Jul-08	738.692	983.340	488.659	735.999	133%	66%
Ago-08	1.078.766	1.298.001	790.687	1.044.344	120%	73%
Set-08	1.832.131	2.039.546	1.554.976	1.797.261	111%	85%
Out-08	1.966.164	2.242.645	1.733.159	1.987.902	114%	88%
Nov-08	1.326.370	1.613.571	1.142.282	1.377.927	122%	86%
Dez-08	904.490	1.166.763	654.750	910.757	129%	72%
Jan-09	1.117.499	1.375.088	794.382	1.084.735	123%	71%
Fev-09	1.280.428	1.586.004	968.209	1.277.107	124%	76%
Mar-09	1.054.657	1.385.469	781.991	1.083.730	131%	74%
Abr-09	948.913	1.260.585	653.287	956.936	133%	69%
Mai-09	1.093.696	1.395.002	768.500	1.081.751	128%	70%
Jun-09	980.904	1.295.495	675.573	985.534	132%	69%
Jul-09	711.394	1.011.889	402.742	707.316	142%	57%
Ago-09	1.028.248	1.292.964	659.533	976.248	126%	64%
Set-09	1.804.710	2.056.028	1.460.611	1.758.319	114%	81%
Out-09	1.980.274	2.326.278	1.698.417	2.012.348	117%	86%
Nov-09	1.334.074	1.689.761	1.107.868	1.398.814	127%	83%
Dez-09	858.610	1.176.031	549.361	862.696	137%	64%

Table 10 - Times series J-Forecast

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	30.078	33.834	29.324	31.579	112%	97%
Fev-08	29.820	30.090	25.660	27.875	101%	86%
Mar-08	28.888	30.613	26.238	28.425	106%	91%
Abr-08	25.963	29.292	24.904	27.098	113%	96%
Mai-08	29.615	31.454	27.032	29.243	106%	91%
Jun-08	30.693	31.221	26.807	29.014	102%	87%
Jul-08	30.816	33.839	29.531	31.685	110%	96%
Ago-08	27.680	29.796	25.230	27.513	108%	91%
Set-08	30.795	33.802	29.175	31.488	110%	95%
Out-08	31.482	34.053	29.568	31.811	108%	94%
Nov-08	30.818	32.555	27.971	30.263	106%	91%
Dez-08	26.570	28.964	24.481	26.722	109%	92%
Jan-09	28.907	33.577	28.477	31.027	116%	99%
Fev-09	29.589	30.405	25.398	27.902	103%	86%
Mar-09	29.410	30.507	25.682	28.094	104%	87%
Abr-09	25.695	28.916	24.250	26.583	113%	94%
Mai-09	28.644	31.520	26.582	29.051	110%	93%
Jun-09	30.282	30.914	26.080	28.497	102%	86%
Jul-09	30.918	33.467	28.614	31.041	108%	93%
Ago-09	27.242	29.826	24.834	27.330	109%	91%
Set-09	29.649	33.456	28.363	30.909	113%	96%
Out-09	30.807	33.575	28.728	31.151	109%	93%
Nov-09	30.924	32.765	27.614	30.190	106%	89%
Dez-09	26.389	28.876	24.008	26.442	109%	91%

Table 11 - Times series L-Forecast

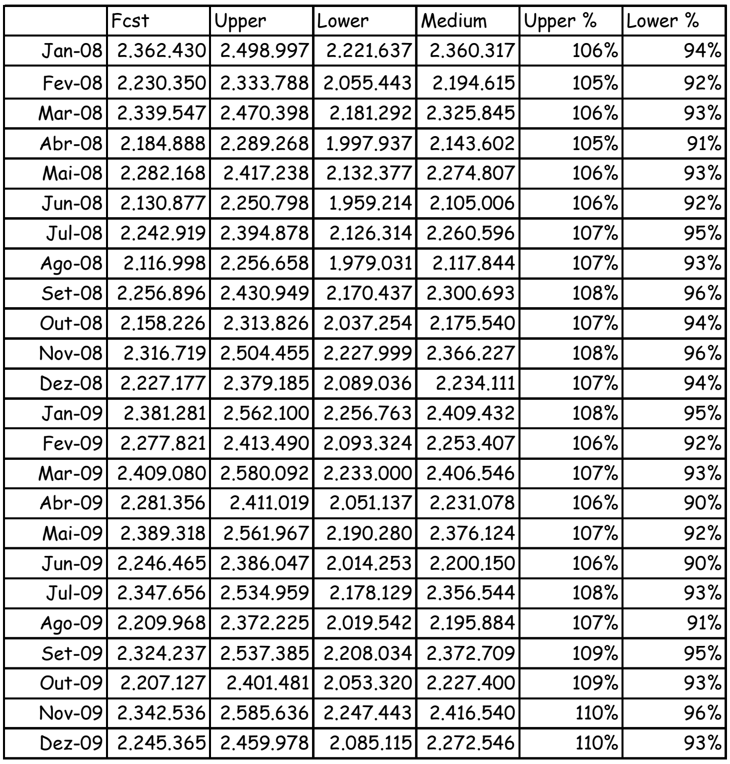
8.9. Time series M

$N = 108; L = 12; B = 73; T = 12; K = 62; I = \{1,2,3,4\}$

Decomposition method: Basic SSA
Maximum ν^2 = 0,42
Eigentriples for reconstruction = 1, 2, 3-4
Maximum Relative error of reconstruction = 2,30%
Forecast Type = V forecast
Confidence bounds type = Bootstrap
Interactions = 1000
Forecast – Table 12, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.10. Time series N

N = 108; L = 12; B = 37; T = 12; K = 72; I = $\{1,2,3,4,5,6,7,8\}$
Decomposition method: Basic SSA
Maximum ν^2 = 0,49
Eigentriples for reconstruction = 1, 2, 3-4, 5-6, 7-8
Maximum Relative error of reconstruction = 3,03%
Forecast Type = V forecast
Confidence bounds type = Bootstrap
Interactions = 1000
Forecast – Table 13, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.



	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	2.362.430	2.498.997	2.221.637	2.360.317	106%	94%
Fev-08	2.230.350	2.333.788	2.055.443	2.194.615	105%	92%
Mar-08	2.339.547	2.470.398	2.181.292	2.325.845	106%	93%
Abr-08	2.184.888	2.289.268	1.997.937	2.143.602	105%	91%
Mai-08	2.282.168	2.417.238	2.132.377	2.274.807	106%	93%
Jun-08	2.130.877	2.250.798	1.959.214	2.105.006	106%	92%
Jul-08	2.242.919	2.394.878	2.126.314	2.260.596	107%	95%
Ago-08	2.116.998	2.256.658	1.979.031	2.117.844	107%	93%
Set-08	2.256.896	2.430.949	2.170.437	2.300.693	108%	96%
Out-08	2.158.226	2.313.826	2.037.254	2.175.540	107%	94%
Nov-08	2.316.719	2.504.455	2.227.999	2.366.227	108%	96%
Dez-08	2.227.177	2.379.185	2.089.036	2.234.111	107%	94%
Jan-09	2.381.281	2.562.100	2.256.763	2.409.432	108%	95%
Fev-09	2.277.821	2.413.490	2.093.324	2.253.407	106%	92%
Mar-09	2.409.080	2.580.092	2.233.000	2.406.546	107%	93%
Abr-09	2.281.356	2.411.019	2.051.137	2.231.078	106%	90%
Mai-09	2.389.318	2.561.967	2.190.280	2.376.124	107%	92%
Jun-09	2.246.465	2.386.047	2.014.253	2.200.150	106%	90%
Jul-09	2.347.656	2.534.959	2.178.129	2.356.544	108%	93%
Ago-09	2.209.968	2.372.225	2.019.542	2.195.884	107%	91%
Set-09	2.324.237	2.537.385	2.208.034	2.372.709	109%	95%
Out-09	2.207.127	2.401.481	2.053.320	2.227.400	109%	93%
Nov-09	2.342.536	2.585.636	2.247.443	2.416.540	110%	96%
Dez-09	2.245.365	2.459.978	2.085.115	2.272.546	110%	93%

Table 12 - Times series M-Forecast

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	6.387.041	6.827.045	6.210.575	6.518.810	107%	97%
Fev-08	5.210.657	5.574.149	5.019.435	5.296.792	107%	96%
Mar-08	5.162.685	5.389.445	4.813.360	5.101.403	104%	93%
Abr-08	4.942.242	5.386.408	4.765.593	5.076.001	109%	96%
Mai-08	5.117.353	5.614.021	4.994.096	5.304.058	110%	98%
Jun-08	4.921.537	5.036.315	4.353.960	4.695.137	102%	88%
Jul-08	5.444.423	5.777.264	5.166.965	5.472.115	106%	95%
Ago-08	4.164.002	4.587.500	3.919.051	4.253.276	110%	94%
Set-08	5.621.669	5.813.226	5.139.282	5.476.254	103%	91%
Out-08	5.977.483	6.310.196	5.613.871	5.962.034	106%	94%
Nov-08	5.277.353	5.735.606	5.023.817	5.379.711	109%	95%
Dez-08	4.916.328	5.143.070	4.440.203	4.791.636	105%	90%
Jan-09	6.473.061	7.292.746	6.329.687	6.811.216	113%	98%
Fev-09	5.303.598	5.854.921	4.999.756	5.427.338	110%	94%
Mar-09	5.363.115	5.630.384	4.715.008	5.172.696	105%	88%
Abr-09	4.954.894	5.736.144	4.787.553	5.261.848	116%	97%
Mai-09	5.225.382	6.044.630	5.047.957	5.546.293	116%	97%
Jun-09	5.225.061	5.305.169	4.204.955	4.755.062	102%	80%
Jul-09	5.566.268	6.117.015	5.200.655	5.658.835	110%	93%
Ago-09	4.171.878	4.863.300	3.801.346	4.332.323	117%	91%
Set-09	5.869.534	6.133.438	5.089.534	5.611.486	104%	87%
Out-09	6.123.121	6.651.683	5.597.002	6.124.342	109%	91%
Nov-09	5.337.416	6.047.941	4.910.308	5.479.124	113%	92%
Dez-09	5.011.631	5.351.185	4.248.751	4.799.968	107%	85%

Table 13 - Times series N-Forecast

8.11. Time series P

N = 108; L = 36; B = 55; T = 36; K = 20; I = $\{1,2,3,4,5,6,7,8\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,24

Eigentriples for reconstruction = 1, 2-3, 4-5, 6, 7-8

Maximum Relative error of reconstruction = 4,10%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 14, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.12. Time series R

N = 108; L = 18; B = 55; T = 18; K = 38; I = $\{1,2,3,4,5,6,7,8,9,10,11\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,31

Eigentriples for reconstruction = 1, 2-3, 4-5, 6-8, 9-11

Maximum Relative error of reconstruction = 5,70%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 15, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	120.136	128.336	103.865	116.101	107%	86%
Fev-08	104.938	111.744	87.547	99.645	106%	83%
Mar-08	124.752	135.865	112.243	124.054	109%	90%
Abr-08	139.915	149.064	124.992	137.028	107%	89%
Mai-08	127.641	135.249	110.382	122.816	106%	86%
Jun-08	102.598	112.727	89.565	101.146	110%	87%
Jul-08	108.725	115.184	89.933	102.559	106%	83%
Ago-08	91.012	99.039	74.756	86.897	109%	82%
Set-08	129.619	143.395	118.636	131.016	111%	92%
Out-08	149.479	156.863	132.654	144.758	105%	89%
Nov-08	116.977	122.400	97.374	109.887	105%	83%
Dez-08	88.338	101.195	76.488	88.841	115%	87%
Jan-09	115.525	124.282	97.240	110.761	108%	84%
Fev-09	100.863	107.755	80.731	94.243	107%	80%
Mar-09	116.426	129.802	103.421	116.611	111%	89%
Abr-09	135.817	146.545	119.618	133.081	108%	88%
Mai-09	124.520	132.164	104.379	118.272	106%	84%
Jun-09	96.820	108.403	82.404	95.404	112%	85%
Jul-09	104.264	112.027	84.014	98.021	107%	81%
Ago-09	88.619	97.005	69.843	83.424	109%	79%
Set-09	121.324	136.831	109.284	123.057	113%	90%
Out-09	143.822	152.518	125.643	139.080	106%	87%
Nov-09	114.248	120.195	92.240	106.218	105%	81%
Dez-09	84.112	98.692	71.104	84.898	117%	85%

Table 14 - Times series P-Forecast

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	3.049.752	3.398.982	2.895.370	3.147.176	111%	95%
Fev-08	2.903.029	3.120.067	2.631.606	2.875.836	107%	91%
Mar-08	2.387.810	2.706.920	2.218.770	2.462.845	113%	93%
Abr-08	1.944.610	2.198.776	1.709.384	1.954.080	113%	88%
Mai-08	2.238.562	2.597.253	2.119.516	2.358.385	116%	95%
Jun-08	2.022.104	2.288.622	1.814.022	2.051.322	113%	90%
Jul-08	1.566.867	1.880.578	1.411.853	1.646.216	120%	90%
Ago-08	1.389.706	1.516.631	1.048.240	1.282.435	109%	75%
Set-08	2.147.086	2.370.881	1.848.972	2.109.926	110%	86%
Out-08	2.470.662	2.707.185	2.179.495	2.443.340	110%	88%
Nov-08	2.406.419	2.904.410	2.313.008	2.608.709	121%	96%
Dez-08	2.397.462	2.798.961	2.216.947	2.507.954	117%	92%
Jan-09	3.113.598	3.567.997	2.965.548	3.266.773	115%	95%
Fev-09	3.209.053	3.546.269	2.998.512	3.272.390	111%	93%
Mar-09	2.667.014	3.152.819	2.605.080	2.878.949	118%	98%
Abr-09	2.018.128	2.350.712	1.768.125	2.059.419	116%	88%
Mai-09	2.160.988	2.519.754	1.800.539	2.160.146	117%	83%
Jun-09	2.012.835	2.186.555	1.514.540	1.850.548	109%	75%
Jul-09	1.606.076	1.982.158	1.287.808	1.634.983	123%	80%
Ago-09	1.396.658	1.770.602	1.085.594	1.428.098	127%	78%
Set-09	2.121.926	2.692.580	1.988.188	2.340.384	127%	94%
Out-09	2.555.161	3.052.928	2.383.196	2.718.062	119%	93%
Nov-09	2.555.226	3.145.074	2.488.402	2.816.738	123%	97%
Dez-09	2.499.112	2.876.841	2.195.172	2.536.007	115%	88%

Table 15 - Times series R-Forecast

8.13. Time series S

N = 108; L = 12; B = 73; T = 12; K = 62; I = $\{1,2,3,4\}$

Decomposition method: Basic SSA

Maximum ν^2 = 0,42

Eigentriples for reconstruction = 1, 2, 3-4

Maximum Relative error of reconstruction = 1,77%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 16, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.14. Time series T

N = 96; L = 24; B = 49; T = 24; K = 26; I = $\{1,2,3,4,5\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,35

Eigentriples for reconstruction = 1, 2, 3, 4-5

Maximum Relative error of reconstruction = 5,80%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

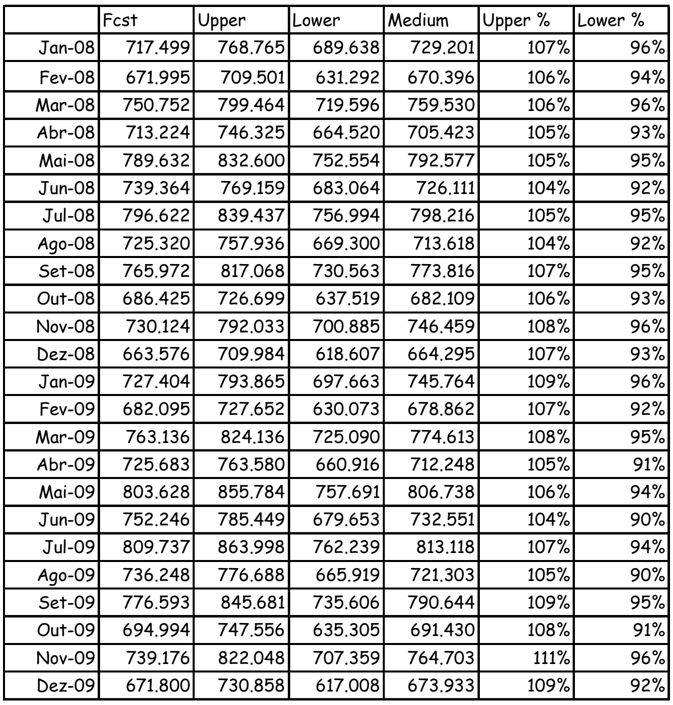
Forecast – Table 17, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

8.15. Time series V

N = 108; L = 12; B = 25; T = 12; K = 14; I = $\{1\}$

Decomposition method: Basic SSA

No reconstruction and Forecast was performed for this Time series because it could not be proven that there was a homogenous structure.



	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	717.499	768.765	689.638	729.201	107%	96%
Fev-08	671.995	709.501	631.292	670.396	106%	94%
Mar-08	750.752	799.464	719.596	759.530	106%	96%
Abr-08	713.224	746.325	664.520	705.423	105%	93%
Mai-08	789.632	832.600	752.554	792.577	105%	95%
Jun-08	739.364	769.159	683.064	726.111	104%	92%
Jul-08	796.622	839.437	756.994	798.216	105%	95%
Ago-08	725.320	757.936	669.300	713.618	104%	92%
Set-08	765.972	817.068	730.563	773.816	107%	95%
Out-08	686.425	726.699	637.519	682.109	106%	93%
Nov-08	730.124	792.033	700.885	746.459	108%	96%
Dez-08	663.576	709.984	618.607	664.295	107%	93%
Jan-09	727.404	793.865	697.663	745.764	109%	96%
Fev-09	682.095	727.652	630.073	678.862	107%	92%
Mar-09	763.136	824.136	725.090	774.613	108%	95%
Abr-09	725.683	763.580	660.916	712.248	105%	91%
Mai-09	803.628	855.784	757.691	806.738	106%	94%
Jun-09	752.246	785.449	679.653	732.551	104%	90%
Jul-09	809.737	863.998	762.239	813.118	107%	94%
Ago-09	736.248	776.688	665.919	721.303	105%	90%
Set-09	776.593	845.681	735.606	790.644	109%	95%
Out-09	694.994	747.556	635.305	691.430	108%	91%
Nov-09	739.176	822.048	707.359	764.703	111%	96%
Dez-09	671.800	730.858	617.008	673.933	109%	92%

Table 16 - Times series S-Forecast

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	210.440	212.660	196.916	204.788	101%	94%
Fev-08	186.323	195.819	181.816	188.817	105%	98%
Mar-08	214.551	215.319	200.798	208.059	100%	94%
Abr-08	201.858	206.485	190.472	198.479	102%	94%
Mai-08	212.344	221.100	205.391	213.246	104%	97%
Jun-08	208.489	209.391	194.172	201.782	100%	93%
Jul-08	227.183	233.584	214.022	223.803	103%	94%
Ago-08	204.848	215.270	197.244	206.257	105%	96%
Set-08	237.518	238.008	218.830	228.419	100%	92%
Out-08	217.130	226.385	206.181	216.283	104%	95%
Nov-08	234.575	243.693	223.406	233.549	104%	95%
Dez-08	229.501	230.918	210.830	220.874	101%	92%
Jan-09	245.308	256.741	232.180	244.461	105%	95%
Fev-09	226.648	236.574	213.982	225.278	104%	94%
Mar-09	260.659	263.019	238.092	250.555	101%	91%
Abr-09	234.199	248.247	222.808	235.528	106%	95%
Mai-09	260.123	268.786	242.989	255.888	103%	93%
Jun-09	250.388	254.288	228.647	241.468	102%	91%
Jul-09	266.083	282.255	251.825	267.040	106%	95%
Ago-09	251.219	260.275	231.917	246.096	104%	92%
Set-09	284.006	290.460	258.838	274.649	102%	91%
Out-09	254.202	272.313	240.448	256.380	107%	95%
Nov-09	288.337	296.852	264.102	280.477	103%	92%
Dez-09	271.430	279.899	247.462	263.680	103%	91%

Table 17 - Times series T-Forecast

8.16. Time series Total

N = 108; L = 24; B = 61; T = 24; K = 38; I = $\{1,2,3,4,5,6,7,8\}$

Decomposition method: Basic SSA

Maximum v^2 = 0,34

Eigentriples for reconstruction = 1, 2, 3-4, 5-6, 7-8

Maximum Relative error of reconstruction = 2,90%

Forecast Type = V forecast

Confidence bounds type = Bootstrap

Interactions = 1000

Forecast – Table 18, with Absolute Forecast, Absolute Upper confidence bound, Absolute Lower confidence bound, Relative Upper confidence bound vs. Forecast, and Relative Lower confidence bound vs. Forecast.

	Fcst	Upper	Lower	Medium	Upper %	Lower %
Jan-08	26.708.564	27.549.146	25.363.986	26.456.566	103%	95%
Fev-08	22.985.122	24.437.447	22.323.443	23.380.445	106%	97%
Mar-08	23.328.536	24.607.161	22.589.688	23.598.425	105%	97%
Abr-08	21.533.099	22.594.734	20.325.469	21.460.102	105%	94%
Mai-08	22.891.845	24.232.987	22.059.548	23.146.267	106%	96%
Jun-08	22.245.942	23.134.334	20.927.687	22.031.010	104%	94%
Jul-08	23.729.058	24.525.799	22.367.226	23.446.512	103%	94%
Ago-08	19.364.793	20.758.185	18.539.000	19.648.593	107%	96%
Set-08	24.895.115	25.584.899	23.266.405	24.425.652	103%	93%
Out-08	25.266.268	25.948.427	23.608.850	24.778.639	103%	93%
Nov-08	23.555.642	25.295.658	22.955.769	24.125.714	107%	97%
Dez-08	22.222.550	23.421.225	21.075.042	22.248.133	105%	95%
Jan-09	27.344.911	28.455.222	25.710.302	27.082.762	104%	94%
Fev-09	23.225.494	25.146.499	22.532.430	23.839.464	108%	97%
Mar-09	23.974.835	25.439.560	22.885.057	24.162.308	106%	95%
Abr-09	21.872.195	23.316.996	20.492.391	21.904.693	107%	94%
Mai-09	23.323.930	25.017.006	22.305.905	23.661.456	107%	96%
Jun-09	22.973.985	23.892.703	21.086.988	22.489.846	104%	92%
Jul-09	24.194.781	25.322.784	22.652.753	23.987.769	105%	94%
Ago-09	19.551.998	21.400.278	18.635.707	20.017.993	109%	95%
Set-09	25.835.551	26.472.262	23.543.360	25.007.811	102%	91%
Out-09	25.755.590	26.732.310	23.835.522	25.283.916	104%	93%
Nov-09	23.815.184	26.121.567	23.210.514	24.666.040	110%	97%
Dez-09	22.839.725	24.194.983	21.246.572	22.720.778	106%	93%

Table 18 - Times series Total-Forecast