From Nature to Maths: Improving Forecasting Performance in Subspace–based methods using Genetics Colonial Theory

Hossein Hassani[a], Zara Ghodsi[b], Emmanuel Sirimal Silva [c] and Saeed Heravi [d]

[a] *Institute for International Energy Studies, Tehran 1967743 711, Iran*
[b] *Translational Genetics Group, Bournemouth University, Fern Barrow, Poole, BH125BB, UK*
[c] *Fashion Business School, London College of Fashion, University of the Arts London, WC1V7EY, UK*
[d] *Cardiff Business School, Cardiff University, Cardiff, CF10 3EU, UK*

Abstract

Many scientific fields consider accurate and reliable forecasting methods as important decision-making tools in the modern age amidst increasing volatility and uncertainty. As such there exists an opportune demand for theoretical developments which can result in more accurate forecasts. Inspired by Colonial Theory, this paper seeks to bring about considerable improvements to the field of time series analysis and forecasting by identifying certain core characteristics of Colonial Theory which are subsequently exploited in introducing a novel approach for the grouping step of subspace based methods. The proposed algorithm shows promising results in terms of improved performances in noise filtering and forecasting of time series. The reliability and validity of the proposed algorithm is evaluated and compared with popular forecasting models with the results being thoroughly evaluated for statistical significance and thereby adding more confidence and value to the findings of this research.

Keywords: Colonial theory, forecasting, nature inspired algorithm, subspace methods.

1 Introduction

As a great source of inspiration, nature holds the key to many questions we face on a daily basis. Therefore, it is not entirely surprising that much of the novel problem solving techniques were initially inspired by nature (see for example [1–5]), even though credit is seldom given. In developing such intelligent solutions, nature provides us with effective background knowledge which follows from the profound observation and questioning of a natural phenomenon.

Quantum computing [1], genetic algorithms [2], neural networks [3], swarm algorithms [4] and ant colony optimization algorithms [5] are among the most

established nature inspired models which seek to imitate specific phenomenon from nature in order to provide simple solutions to complex problems. Although attempting to model natural phenomena has a long history, the recent application of nature inspired algorithms like firefly algorithm [6], neuro fuzzy technique [7] and genetic programming [8] in the area of soft computing and also recent improvements in forecasting approaches [9, 10], has lead to more accurate analysis and predictions, and thereby causing a noticeable growth of interest in this field [11–18]. However, attempts at improving signal extraction and forecasting using bio–inspired algorithms is a relatively new area of research. Moreover, it is noteworthy that in most of the nature inspired algorithms, the natural phenomenon of interest is the strategy taken by biological organisms after facing an environmental change which enables the organism to exploit an ingenious solution to meet the new specific conditions.

Even though such biological solutions have been many times by organisms over the evolution process, the most prominent one can be referred to as the multicellularity phenomenon which describes how multicellular organisms arose from a single cell and generated multi-celled organisms. Despite there being various theories that may be able to explain this mechanism, Colonial Theory (CT) has received most credit by developmental scientists [19]. Inspired by CT and by identifying certain characteristics of this theory, in this paper we seek to draw a line between nature and mathematics. The mathematical procedures which we specifically seek to link with nature consist of Singular Value Decomposition (SVD) based methods and signal subspace (SS) methods which form the basis of a general class of subspace-based noise reduction algorithms. The superior performance of this class of algorithms in noise reduction and forecasting has been proved by several studies [20–22].

In this context, the Singular Spectrum Analysis (SSA) technique, which is a SVD and SS based method has been considered as a powerful nonparametric tool [23]. In brief, the SSA technique begins by decomposing the original series into the sum of a small number of independent components. Thereafter, the selected components are used to reconstruct the less noisy series which can be used for forecasting future data points. However, due to the nature of least squares (LS) estimation method used in the current SSA procedure, the signal and noise separation is not optimum and the reconstructed series continues to hold some part of initial noise whilst the residual is not completely signal free. This paper seeks to consider an alternative approach which is based on CT in order to provide a more efficient outcome for the signal and noise separation issue in SSA.

The exploitation of CT towards improving the SSA process is made possible via our identification of a general similarity between CT and SSA. However, this similarity was not visible in the Grouping step of the basic SSA process, and therein lies our focus as we intend on defining a new approach to grouping in SSA by imitating one of the steps followed in CT. It is expected that this novel CT based approach to grouping will enable a more efficient separation of signal from noise which in turn will enhance the signal extraction and forecasting results.

The remainder of this paper is organized such that Section 2 presents an outline of CT along with the steps underlying the SSA process in order to

clearly illustrate similarities and discuss how the different steps of CT are fully consistent with the procedure underlying SSA. Section 3 describes the newly introduced approach for grouping in SSA. Section 4 provides the theoretical presentation of the algorithm which is followed by several applications in Section 5 and the paper concludes with a concise summary in Section 6.

2 Similarities between SSA and CT

This section focusses on providing a clear view on the similarities between SSA and CT as portrayed in Figure 1. In what follows, the information contained in Figure 1 is expanded upon as we present a detailed explanation of the linkages between SSA and CT.

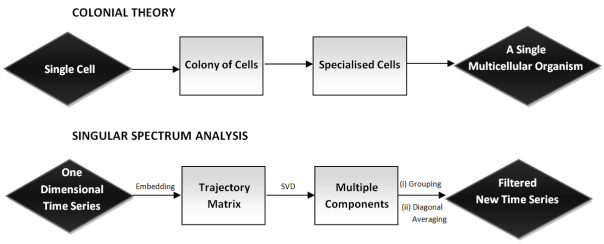

Figure 1: T
he linkage between CT and SSA.

The SSA method is made up of two complementary stages: Decomposition and Reconstruction; each stage consists of two compatible steps. At the first stage a group of small number of independent and interpretable components is achieved by decomposing the main series [23], which is followed by the reconstruction of a less noisy series at the second stage [24]. Thereafter, this noise free series is used for forecasting future data points.

2.1 Stage 1: Decomposition

We begin with a one dimensional time series, $Y_N = (y_1, \ldots, y_N)$ where N is the length of the series. The SSA technique consists of two choices, the window length L and the number of eigenvalues r [25]. In SSA the number of components are related to the selection of the proper window length L which should be defined such that it minimises the signal distortion and maximises the residual noise level. However, we cannot impose a general rule in selecting L for different time series with different structure. For example, in instances where there is a periodic component with an integer period like a seasonal component, to obtain a higher separability the tradition is to select L proportional to that period [23].

Likewise, the starting point of CT is a single cell which evolves over time and generates a multi-celled organism. Similar to SSA, there is also a limit on the number of cell types and different kinds of organisms necessitate different numbers of cell types. It is assumed that this number is determined by the

balance between selective pressure and functional requirements, whilst variety is favoured by selection, functional needs limit the number of cell types [26].

2.1.1 1st step: Embedding

Here we take the one dimensional time series Y_N, and map it in order to create a multi-dimensional variable of $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})' \in \mathbf{R}^L$. It is clear that this can be viewed as the creation of the colony from the initial single cell as we take Y_N and create multiple dimensions from the same series. This step provides us with a trajectory matrix, $\mathbf{X}$ which is a Hankel matrix that captures all information contained in Y_N.

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix}. \tag{1}$$

It should be noted that unlike the Symbiotic theory [27], which assumes that the symbiosis of various species caused a multicellular organism, in CT it is the symbiosis of many cells of the same species that forms a multicellular organism. This point is interesting as it can be referred to as the first and main difference between SSA and principal component analysis (PCA). In the latter, the obtained matrix is achieved by considering different time series (multiple cells) whilst in SSA we consider one time series (single cell).

Moreover, transferring a one dimensional time series into a trajectory matrix will enable us to significantly reduce the computation time required for running the algorithm, as it eliminates the need for running the algorithm over a wide range of values for the hidden state dimension. Furthermore, by analysing the eigenvalues with the aim of filtering the signal and noise, the signal to noise ratio (SNR) will be optimised in the newly reconstructed time series. Likewise, increasing in size is initially favoured by individual cells since multicellular organisms do not have the size limit which is mainly imposed by diffusion. As the surface-to-volume ratio decreases in a given single cell, with increased size they will experience difficulty in obtaining the required nutrients and transporting the cellular waste products out the cell [26, 28].

2.1.2 2nd step: Singular Value Decomposition

SVD is a procedure which is performed on $\mathbf{X}$ and provides us with several eigenvalues or components. The components obtained via this step are identified as trend, periodic, quasi-periodic component, or noise. In CT, $\mathbf{X}$ can represent the entire colony of cells generated from the original single cell. By increasing the interdependency level in a colony some of the cells specialize to do different tasks and by obtaining ever more complexity level, cells form tissues and then organs [26].

2.2 Stage 2: Reconstruction

In SSA, this is the stage where we seek to analyze the eigenvalues extracted via the SVD step to differentiate between noise and signal in a time series. In CT this would correspond to identifying which of the specialised cells are able to successfully carry out the reproductive task and which cells are responsible for viability.

2.2.1 1st step: Grouping

Grouping is a very important step in SSA as the quality of the filtering achieved via this technique depends on the successful analysis of eigenvalues and selection of appropriate groups of them to rebuild the less noisy time series. In brief, this step involves grouping together the eigenvalues with similar characteristics i.e. signal and harmonic components whilst leaving out those corresponding to noise. Likewise, the grouping step plays a significant role in CT as it determines which of the specialised cells are successful in carrying out the reproduction task and which of these fail along that way. Here it is imperative to note that in spite of the general similarity between SSA and CT in the grouping step, there is an important fundamental difference between the current version of SSA and CT which is discussed in detail in Section 4.

2.2.2 2nd step: Diagonal Averaging

The diagonal averaging step in SSA transforms the matrix of grouped eigenvalues back into a Hankel matrix which can later be converted into a time series. The resulting time series will be the less noisy, filtered time series corresponding to the original one-dimensional time series that was applied to the SSA process at the beginning. This step is important and similar to the final stage of CT where a single multiple organism is formed after defining the productive cells in order to compensate for the increased cost of reproduction imposed by increasing the size of the colony size.

According to the role of a small variant in CT, adding a single cell will only have a slight impact on the performance of a large organism [29] which means after achieving the major functional specializations, there would presumably be a decline in adding capabilities by increasing the number of cell lines [26]. Similarly, in a time series, after extracting different components related to the trend, oscillation and noise, increasing the number of observations will give us more components but all of these will be categorised in the previously defined groups of components.

3 A New Approach for Grouping

It is widely accepted that the first grouping in nature happened when multicellular organisms arose from a single cell and generated a multi-celled organism [30]. At the very beginning of life there were only single cells. Today, after millions of years, most animals, plants, fungi, and algae are made up of multiple cells that work together as a single being [31].

Presented in this section is a brief explanation on developing functional specialization and grouping the specialized cells of CT. Following this approach which is called changes in the level of complexity [32], we describe a novel approach for grouping in SSA.

In order to present a clearer view, as a model system, we consider Volvocalean green algae which are well suited for studying the transition as they provide different ranges from unicells to multicellular organisms with the explicit specification between germ and soma cells. In Volvocalean green algae, multicellular organisms are formed clonally from a single cell [33]. Here, we mainly focus on one member of this lineage named *V. carteri* which exhibits a range of development in different cell types [32].

In [34] a twelve-step program for the grouping step in *V. carteri* is considered. In this process, motility and mitosis activity compete for the same cellular machinery and cell destiny is determined finally by the location of microtubule organizing centre. The activity of the microtubule organizing centre mainly depends on its location in the cell and can serve either as a basal body which is related to the flagellar synthesis and consequently cell motion or a mitotic spindle which aids the segregation of the chromosomes during mitosis [35]. This germsoma dichotomy is generated during the early embryogenesis and results in specifying two cell types; the somatic cells which are non-reproductive and only vegetative and the germ cell which performs the exclusively reproductive task [34].

Here, as the germ cells execute the reproductive function and cell divisions which are required to produce a new daughter colony we consider them as those signal components which are later selected for series reconstruction. Differentiating the somatic and germ cells in *V. carteri* is largely dependent on the genetic differentiation which happens in somatic cells. During this process *regA* which is a regulatory gene and encodes a transcriptional repressor begins to express [36]. As a result, several nuclear genes which are responsible for coding the chloroplast proteins are suppressed [37]. Consequently, these somatic cells will not go under the cell growth or division. Lacking the division ability, they do not participate in the reproductive functions and offspring but accomplish the survival task by flagellar action [34].

When comparing the grouping step between CT and SSA, two important points must to be highlighted:

1. Following differentiation germ cells will not stay attached together. In a colony, those cells underpinning reproductive functions may divide on (a) the colony surface, (b) introgress or (c) in the interior of the colony as shown in Figure 2 [35].

2. In each round of reproduction, germ cells will produce both future germ cells and soma cells.

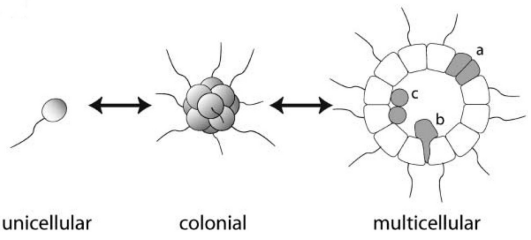

Figure 2: Genetic variants produced differentiated cells in colonial flagellates [35].

However, the current grouping stage of the SSA method is based on the least squares (LS) estimator i.e., choosing the leading eigentriple which describes the general tendency of the series [23]. Accordingly:

- Selecting the signal components follows a binary approach. In other words, by estimating the signal rank r, the first r eigenvalues are always selected as signal components and the rest is considered to be noise.
- The leading components of $I_1, \ldots, I_m$ are related only to signal, hence, it is assumed that the reconstructed signal is not perturbed by noise.

Taking into consideration these points of comparison and our assumption which states germs cell from CT are equal to signal components, we can conclude that following the LS estimator the first r selected eigenvalues will not produce a clear signal because germ cells will produce both future germ cells and soma cells. This statement makes sense when we take the noise perturbation into account. Even in the leading components of $I_1, \ldots, I_m$, we have to exclude some eigenvalues with less information about the series, seek to find the other related signal components which capture more information and use all of them for reconstructing the series.

4 SSA–CT Algorithm

Presented below is a concise explanation of the SSA–CT algorithm which has also been depicted in Figure 3. In this paper, we use root mean squared error (RMSE) criterion to determine the optimal choices of L and r (it is also possible to use any other criteria to determine the error as explained below in the algorithm). Accordingly, in each L we are looking for a combination of eigenvalues r which provides the lowest RMSE, and this in turn represents the optimal decomposition and reconstruction choices for the SSA model. The automated SSA-CT code is able to perform this task by evaluating all possible L and r choices for a given time series. [1]

[1]The SSA-CT code used in this study is available upon request.

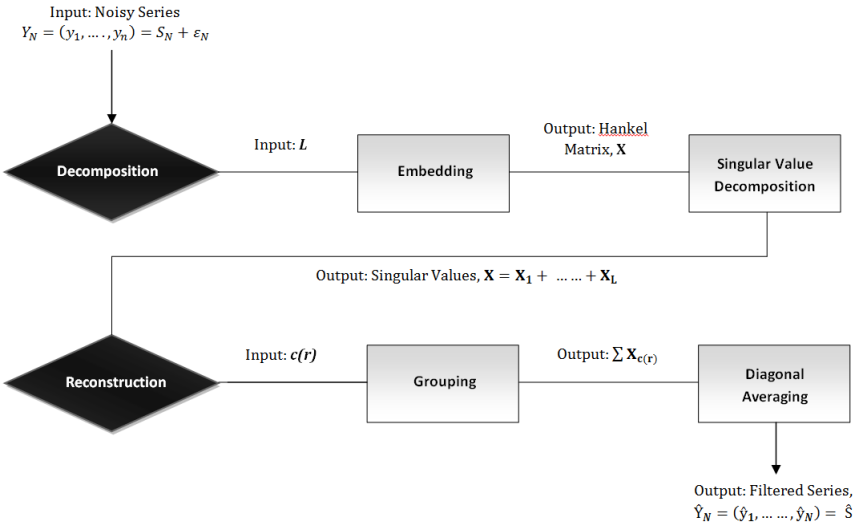

Figure 3: SSA-CT flowchart which depicts different computational steps of this algorithm.

1. Consider a real-valued nonzero time series $Y_N = (y_1, \ldots, y_N)$ of length N.

2. If the aim is signal extraction, consider the whole series and if the aim is forecasting, divide the time series into two parts; $\frac{2}{3}^{rd}$ of observations for training the SSA-CT model and $\frac{1}{3}^{rd}$ for testing the forecast accuracy.

3. Use the training data to construct the trajectory matrix $\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = [X_1, ..., X_K]$, where $X_j = (y_j, ..., y_{L+j-1})^T$ and $K = N - L + 1$. Initially, we begin with $L = 2$ ($2 \leq L \leq \frac{N}{2}$) and in the process, evaluate all possible values of L for Y_N.

4. Obtain the SVD of $\mathbf{X}$ by calculating $\mathbf{X}\mathbf{X}^T$ for which $\lambda_1, \ldots, \lambda_L$ denotes the eigenvalues in decreasing order ($\lambda_1 \geq \ldots \lambda_L \geq 0$) and by $U_1, \ldots, U_L$ the corresponding eigenvectors. The output of this stage is $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L$ where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$.

5. Adapt a diagonal matrix containing the weights $\mathbf{W}_{K\times K}$ as follows:

$$\hat{\mathbf{S}}_{CT} = \mathbf{U}_1(\mathbf{W}_{CT}\mathbf{\Sigma}_1)\mathbf{V}_1' \quad .$$

6. Choose the weight matrix $\mathbf{W}_{K\times K}$.

$$\mathbf{W}_{CT} = \text{diag}\left((1 \vee 0), \ldots, (1 \vee 0)\right) \quad .$$

7. Evaluate all possible combinations of $\mathbf{W}_{CT}$ (step by step) for the selected L and split the elementary matrices $\mathbf{X}_i$ ($i = 1, \ldots, L$) into several groups and sum the matrices within each group.

8. Perform diagonal averaging to transform the matrix into a Hankel matrix which can then be converted into a time series.

9. Find the RMSE value for each reconstructed series and report the L and the selected combination of r attributed to the minimum RMSE as optimal choices.

10. The output is a filtered series that can be used for forecasting.

The optimal choices of L and r can be used for forecasting via vector or recurrent SSA (SSA-V,SSA-R) techniques. More detailed information of these techniques can be found in [38].

5 Applications

Presented in this section is the application of the new algorithm in forecasting. In order to evaluate the results we rely on the Ratio of the Root Mean Squared Error (RRMSE) criterions which is a frequently used measure for the accuracy of forecasted values.

$$\text{RRMSE} = \frac{RMSE(SSA-CT)}{RMSE(SSA)} = \frac{\left(\sum_{i=1}^{N}(s_i-\widehat{s}_i)^2\right)^{1/2}}{\left(\sum_{i=1}^{N}(s_i-\widetilde{s}_i)^2\right)^{1/2}},$$

where, $\widehat{s}_i$ are the estimated values of s_i obtained via SSA-CT and $\widetilde{s}_i$ are the estimated values of s_i obtained through basic SSA, and N is the series length. If RRMSE < 1, then SSA-CT outperforms the SSA method. In contrast, when RRMSE > 1 it would indicate that SSA-CT fails to outperform the basic SSA model.

5.1 Forecasting

In this paper, several well known series are used to evaluate the performance of the newly introduced idea for the grouping step in SSA (SSA-CT). These include Death series, Petroleum series, Unemployment series and Industrial Production series. The results are then examined in a forecasting scenario.

5.1.1 Death series

The Death series reports the monthly accidental deaths in the USA between 1973 and 1978. This time series has been previously used by many authors (see for example, [39]). According to Table 1, the best SSA forecast for this series is achieved by SSA–CT considering $L = 14$ and the following combinations of $r : 1, 2, 3, 5, 6, 7, 8, 10, 12, 14$. It should also be noted that beside the least RMSE which is related to $L = 14$, the best forecast for this series in each different L (except $L = 5$ *and* 8) is achieved by considering a combination of eigenvalues instead of selecting the signal components using the binary approach. Figure 4 shows the Death series along with the forecast of six data points achieved using seven different forecasting methods. These include, two versions of the basic

SSA technique; Vector SSA (SSA-V) and Recurrent SSA (SSA-R) and also several well-known methods namely, an optimized version of the Autoregressive Integrated Moving Average (ARIMA) [40], Exponential Smoothing (ETS) [41], Holt-Winters (HW) [40] and the ARAR Algorithm [42].

Figure 4: Actual and forecasts for monthly accidental deaths in USA (June 1978 - June 1979).

Table 1: The RRMSE results for different L and r for Death series.

SSA-CT		**SSA**	**RRMSE**
L	**r**	**r**	
3	1	1:2	0.618
4	1, 2, 4	1:2	0.997
5	1:2	1:2	1
6	1, 2, 5	1:2	0.988
7	1, 2, 7	1:2	0.993
8	1:6	1:6	1
9	1, 2, 3, 5, 7, 9	1:3	0.498
10	1, 2, 3, 6, 7	1:3	0.777
11	1, 2, 3, 5, 7, 10, 11	1:3	0.77
12	1, 3, 5, 11, 12	1:3	0.492
13	1, 2, 3, 6, 7, 12	1:2	0.557
14	1, 2, 3, 5, 6, 7, 8, 10, 14	1:10	0.329

Table 2 shows the forecasts obtained by the various models. According to this table the forecasted values from the SSA–CT approach are very close to the original data and is the best among the methods considered, for example, the RMSE value from SSA–CT is 4 times less than the value of the RMSE for ARIMA and 3 times less than that of SSA-R. All outcomes are tested for

statistical significance using both the DM test and the KSPA test. This is important, especially to ensure that the newly proposed SSA–CT algorithm is able to provide significantly better results than the basic SSA-R and SSA-V models as otherwise this exercise bears no fruit if the results are mere chance occurrences. In the case of the Death series, except for in relation to the forecasts obtained via ETS, there is sound evidence to conclude that SSA-CT provides a significantly better forecast than ARIMA, HW, ARAR, SSA-V and SSA-R techniques based on the DM and KSPA tests.

Table 2: Death series forecasts and RMSE for six observations via several methods.

Actual	ARIMA	ETS	HW	ARAR	SSA-V	SSA-R	SSA-CT
7798.00	8337.15	8257.85	7975.63	8167.81	8211.60	8184.49	7651.20
7406.00	7534.21	7496.61	7012.07	7195.76	7336.44	7374.76	7478.49
8363.00	8317.62	8293.41	7763.98	7982.03	8012.39	8009.97	8383.60
8460.00	8589.01	8513.28	8082.34	8283.53	8305.86	8489.40	8518.86
9217.00	9490.16	9418.72	8945.36	9144.07	9200.47	9333.71	9233.50
9316.00	9860.70	9898.85	9324.54	9464.87	9579.78	9664.46	9412.11
RMSE	340.86	318.26	356.38	253.21	255.74	261.68	81.85
RRMSE	0.24^{*}	0.26	$0.23^{*,\dagger}$	$0.32^{\dagger}$	$0.32^{\dagger}$	$0.31^{*,\dagger}$	N/A

Note: * indicates a statistically significant difference between the forecasts based on the two-sided and one-sided KSPA tests in [43] at a p-value of 0.05. † indicates a statistically significant difference between the forecasts based on the Diebold Mariano test in [44] at a p-value of 0.10.

5.1.2 Petroleum series

Considered here is a monthly time series related to international petroleum consumption in OECD countries from August 2007 - May 2013 and was extracted via the U.S. Energy Information Administration. Data up until May 2012 were used to train the models and the last 12 observations from June 2012 - May 2013 were set aside for forecasting. Table 3 reports the best RMSE results achievable at each window length for this series when considering SSA–CT and basic SSA approaches. The lowest RMSE is achieved when $L = 12$ and r uses the combinations of $1, 2, 4, 6, 7, 9$ and interestingly it is only one instances that the basic SSA approach of binary grouping enables the lowest RMSE at a particular window length.

Table 3: The optimal RMSE results for different L and r for the Petroleum series.

L	r	RMSE
3	1,3	958.4
4	1,4	693.59
5	1,4,5	653.31
6	1,4,5,6	672.46
7	1,4,5,6	633.33
8	1,3,6,7	612.8
9	1,5,6,8	607.59
10	1:3	539.23
11	1,2,4,7,8,9,10	434.23
12	1,2,4,6,7,9	325.1

Table 4 presents the forecasts achieved by various methods for the last twelve observations in the petroleum consumption series. The SSA–CT forecast records a very low RMSE in relation to all other models whilst ARIMA and SSA-R are seen providing the worst forecasts for this series. Once again, all forecasts are evaluated for statistically significant differences using the DM and KSPA tests as indicated at the bottom of this table. Accordingly there is sufficient evidence to conclude that the SSA–CT forecasts are significantly better than ARIMA, ETS and HW forecasts. However, there is no such evidence in relation to ARAR, SSA-V and SSA-R in this case even though the reported RMSE values differ greatly in relation to that of SSA–CT. This could be a result of the low number of forecasts available for comparison and could indicate the sensitiveness of both tests to sample sizes.

Accordingly, we take into consideration Figure 5 which plots the forecasts from all models in relation to the actual data. This figure gives a clear indication of the accuracy associated with the SSA–CT forecast in relation to the rest of the models and provides more confidence in the reported RMSE values.

Table 4: Petroleum consumption series forecasts and RMSE for twelve observations via several methods.

	Actual	ARIMA	ETS	HW	ARAR	SSA-V	SSA-R	SSA-CT
1	46009.11	46280.52	46129.50	46570.59	45874.03	45482.35	45568.59	45984.45
2	45790.35	46597.05	46241.02	46359.95	45297.25	45839.38	46324.55	45568.38
3	46607.89	48000.75	46380.09	47113.64	46482.30	48536.40	49148.67	46533.51
4	45092.28	47409.37	46414.14	46371.06	45031.23	45889.26	46196.70	45418.91
5	46225.25	46533.94	46565.02	45401.51	45436.21	45991.01	45802.36	46041.64
6	46383.06	47144.63	46693.91	45989.31	45916.41	46950.66	46766.91	45777.45
7	45766.15	47499.15	47566.06	46606.12	45516.69	45363.41	45660.63	45656.44
8	45524.67	45559.74	46020.05	44753.90	44542.40	44807.37	45446.25	45104.66
9	46392.59	48240.58	47522.46	47214.36	46311.87	47186.14	48194.74	45833.27
10	44962.43	46343.67	46465.88	45759.52	44411.56	44947.74	45774.60	45199.25
11	45756.31	45273.49	45686.86	44766.40	44734.03	45676.37	45792.55	46116.82
12	45063.56	45969.44	44897.70	45095.67	45027.24	45493.44	45238.77	45199.11
	RMSE	1227.80	879.55	761.34	540.42	736.49	1018.07	325.10
	RRMSE	$0.26^{*,\dagger}$	$0.37^{\dagger}$	$0.43^{*,\dagger}$	0.60	0.44	0.32	N/A

Note: * indicates a statistically significant difference between the forecasts based on the two-sided and one-sided KSPA tests in [43] at a p-value of 0.05. † indicates a statistically significant difference between the forecasts based on the Diebold Mariano test in [44] at a p-value of 0.10.

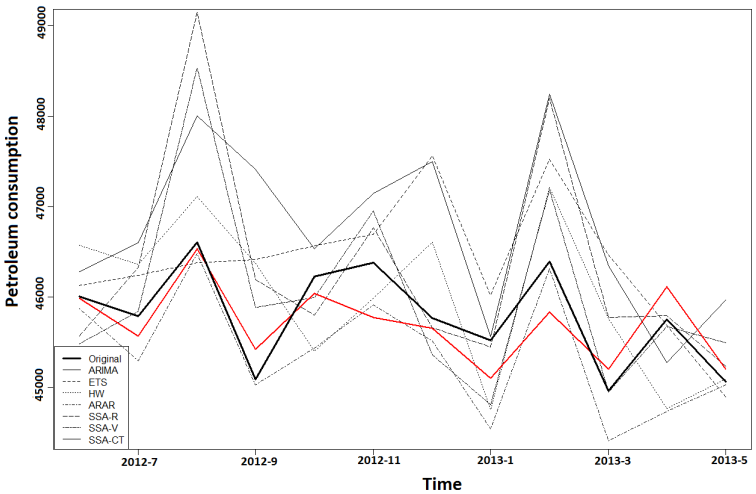

Figure 5: Petroleum consumption forecasts.

5.1.3 Unemployment series

The next application considers modelling the monthly U.S. Unemployment rate from November 2006 - April 2013 which was extracted via the U.S. Department of Labor. All data up until April 2012 is used for model training leaving aside the last 12 monthly observations to evaluate forecast accuracy. Table 5 below shows the lowest RMSE recorded by SSA–CT or basic SSA approaches for different L. The minimum RMSE is achieved when $L = 11$ and $r = 1, 2, 3, 4, 9$ which is effectively following the proposed SSA–CT grouping.

Table 5: The optimal RMSE results for different L and r for the U.S. Unemployment series.

L	r	RMSE
3	1:2	0.74931
4	1:3	0.36501
5	1:3	0.33927
6	1:4	0.43986
7	1,2,3,5	0.76149
8	1,2,3,4,5,7	0.30995
9	1,2,3,7,8	0.37552
10	1,2,3,7,8,10	0.2213
11	1,2,3,4,9	0.10905
12	1,2,3,4,7,10,12	0.15469

Next, we consider generating forecasts following the decomposition and reconstruction of the U.S. Unemployment series along with the SSA–CT choices and compare the forecasts with several other methods. The results are reported in Table 6. On this occasion there are a high number of statistically significant outcomes based on the DM and KSPA tests. As such, we are able to conclude that SSA–CT forecasts for the U.S. Unemployment rate are significantly better than those from ARIMA, ETS, ARAR, SSA-V and SSA-R models.

Figure 6 plots the forecasts from all models in relation to the actual values. Interestingly, even though the actual values appear to be almost linear with a downward sloping trend, except for HW and SSA–CT models all other techniques experience difficulties in accurately predicting this series. This application and the success of SSA–CT on this occasion adds more value to the results as it is evident that the series is difficult to predict regardless of its simple structure.

Table 6: U.S. Unemployment series forecasts and RMSE for twelve observations via several methods.

	Actual	ARIMA	ETS	HW	ARAR	SSA-V	SSA-R	SSA-CT
1	8.20	8.04	8.00	8.03	8.04	7.99	7.98	8.11
2	8.20	7.94	7.90	8.11	7.94	7.88	7.86	8.09
3	8.20	7.89	7.80	8.17	7.94	7.76	7.74	8.06
4	8.10	7.80	7.71	8.15	7.91	7.63	7.60	8.02
5	7.80	7.70	7.61	8.09	7.90	7.50	7.46	7.97
6	7.90	7.64	7.51	8.03	7.91	7.36	7.32	7.90
7	7.80	7.64	7.41	7.93	7.97	7.22	7.17	7.83
8	7.80	7.73	7.31	7.90	8.03	7.07	7.01	7.76
9	7.90	7.80	7.21	7.88	8.12	6.92	6.85	7.68
10	7.70	7.73	7.12	7.79	8.19	6.76	6.69	7.60
11	7.60	7.70	7.02	7.69	8.28	6.59	6.52	7.53
12	7.50	7.65	6.92	7.51	8.38	6.42	6.34	7.47
	RMSE	0.189	0.457	0.124	0.387	0.698	0.749	0.109
	RRMSE	$0.58^{\dagger}$	$0.24^{*,\dagger}$	0.88	$0.28^{*,\dagger}$	$0.16^{*,\dagger}$	$0.15^{*,\dagger}$	N/A

Note: * indicates a statistically significant difference between the forecasts based on the two-sided and one-sided KSPA tests in [43] at a p-value of 0.05. † indicates a statistically significant difference between the forecasts based on the Diebold Mariano test in [44] at a p-value of 0.10.

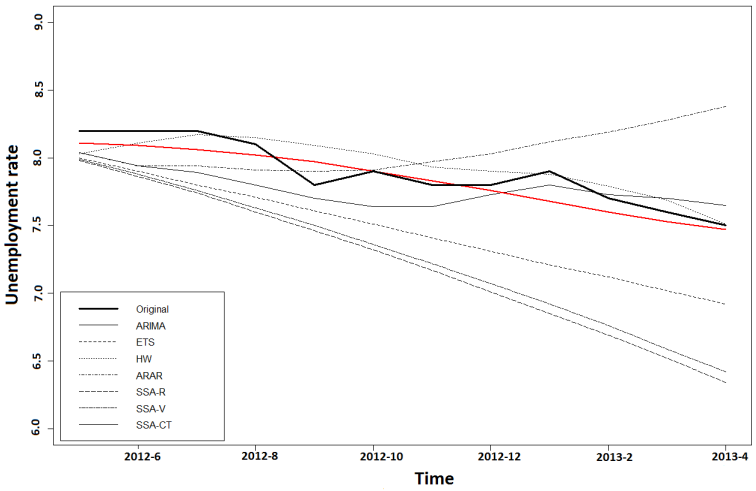

Figure 6: Forecasts for the U.S. Unemployment rate series.

5.1.4 Industrial Production series

The final application to real data considers the monthly growth rate of the U.S. Industrial Production index from March 2008 to December 2013. The data was compiled via the OECD Main Economic Indicators database and data up until December 2012 were used to train the models whilst observations from January 2013 to December 2013 formed the test set. Table 7 reports the minimum RMSE for each L when considering all possible combinations of r. In this instance the lowest RMSE is attainable when $L = 11$ and $r = 1, 3, 4, 5, 7, 9$ as achieved via the SSA–CT approach.

Table 7: The optimal RMSE results for different L and r for the U.S. Industrial Production series.

L	r	RMSE
3	1	0.30
4	1	0.29
5	1:3	0.3
6	1	0.31
7	1,5	0.31
8	1,3,7	0.34
9	1,3,6,7	0.32
10	1,3,4,5	0.34
11	1,3,4,5,7,9	0.23
12	1,3,4,5,6,7,9,10,11	0.30

Table 8 considers forecasting the last 12 observations of the U.S. Industrial production series using various methods including the SSA–CT grouping as

identified previously. The forecasts are once again tested for statistical significance using KSPA and DM tests. We find sound evidence to conclude that SSA–CT provides significantly better forecasts for U.S. Industrial production in comparison to ARIMA, HW, ARAR, SSA-V and SSA-R forecasts based on the DM and KSPA tests. Figure 7 presents a graphical representation of the forecasts from the various models.

Table 8: Forecast data and RMSE for twelve data points by several methods for the U.S. Industrial Production series.

	Actual	ARIMA	ETS	HW	ARAR	SSA-V	SSA-R	SSA-CT
1	0.05	0.20	0.27	0.64	0.07	0.10	0.02	0.21
2	0.64	0.55	0.27	-0.01	0.18	0.07	-0.02	0.57
3	0.46	0.62	0.27	-0.31	0.59	0.04	-0.03	0.24
4	-0.18	0.42	0.27	0.41	0.07	0.02	-0.02	-0.10
5	0.11	0.45	0.27	0.08	0.12	-0.01	-0.03	0.33
6	0.19	0.67	0.27	-0.14	0.04	-0.03	-0.05	0.07
7	-0.16	0.38	0.27	0.38	0.22	-0.05	-0.06	-0.04
8	0.56	0.94	0.27	-0.45	0.07	-0.06	-0.07	0.27
9	0.72	0.71	0.27	0.29	0.01	-0.07	-0.08	0.19
10	0.10	0.82	0.27	0.07	-0.07	-0.08	-0.07	-0.04
11	0.56	0.48	0.27	0.80	0.08	-0.09	-0.06	0.45
12	0.18	0.80	0.27	0.11	0.00	-0.09	-0.05	0.49
	RMSE	0.42	0.30	0.53	0.35	0.42	0.44	0.23
	RRMSE	$0.55^{*,\dagger}$	0.76	$0.43^{*,\dagger}$	$0.66^{\dagger}$	$0.55^{\dagger}$	$0.52^{\dagger}$	N/A

Note: * indicates a statistically significant difference between the forecasts based on the two-sided and one-sided KSPA tests in [43] at a p-value of 0.05. † indicates a statistically significant difference between the forecasts based on the Diebold Mariano test in [44] at a p-value of 0.10.

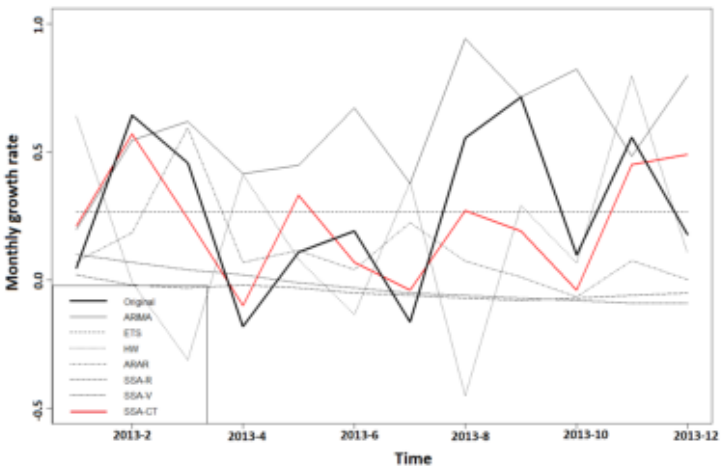

Figure 7: Forecasts for the U.S. Industrial production series.

6 Conclusion

This paper presents a novel approach for enhancing the accuracy of SSA-based forecasts based on the foundations of CT. Initially we draw upon the general similarity between CT and SSA, and then exploit these similarities, particularly certain characteristics of CT in the grouping step which is the most important step in the SSA procedure.

In brief, we suggest that relying on a binary approach of differentiating between signal and noise at the grouping step is not necessarily the best approach as this assumes there is no useful information contained in the selected noise components. Instead, based on CT we propose a different grouping approach which considers analysing all eigenvalues and selecting those which have useful information for grouping in SSA. The results achieved for several series have shown that the new idea of grouping has the potential to enable us to obtain a more efficient forecasts in comparison to the existing approach for grouping in SSA which is based on LS.

Moreover, the results, when compared with other basic versions of SSA and popular time series analysis and forecasting models further portray the superiority of the new approach and provides sound evidence of its significantly better performance in practice.

References

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 3rd ed. UK: CUP, 2000.

[2] A. E. Eiben et al., "Genetic algorithms with multi-parent recombination," in P*arallel Problem Solving from NaturePPSN III.*, Springer Berlin Heidelberg, 1994, pp. 78-87.

[3] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity,"*Bull Math Biophys*, vol. 5, no. 4, pp. 115-133, 1943.

[4] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, 1st ed. New York, USA: OUP, 1999.

[5] M. Dorigo, Optimization, "Learning and Natural Algorithms," Ph.D. dissertation, Politecnico di Milano, Italy, 1992.

[6] X. S. Yang. "Nature-inspired metaheuristic algorithms". Luniver press, 2010.

[7] JSR Jang, CT Sun, E. Mizutani . "Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence." (1997).

[8] J.R. Koza. "Genetic programming III: Darwinian invention and problem solving". Vol. 3. Morgan Kaufmann, 1999.

[9] Martens, M. and Zein, J. (2002). Predicting financial volatility: high-frequency time-series forecasts vis-a-vis implied volatility. Available at SSRN 301382.

[10] Benedetto, F., Giunta, G. and Mastroeni, L. (2015). A maximum entropy method to assess the predictability of financial and commodity prices. Digital Signal Processing, 46, 19-31.

[11] S. Motamedi, S. Shamshirband, D. Petkovi, & R. Hashim, "Application of adaptive neuro-fuzzy technique to predict the unconfined compressive strength of PFA-sand-cement mixture". *Powder Technology*, 278, pp. 278-285. 2015.

[12] M. Proti, S. Shamshirband, M. H. Anisi, D. Petkovi, D. Miti, M. Raos, ... & K. A. Alam, "Appraisal of soft computing methods for short term consumers' heat load prediction in district heating systems". *Energy*, 82, pp. 697-704. 2015.

[13] M. Goci, S. Motamedi, S. Shamshirband, D. Petkovi, S. Ch, R. Hashim, & M. Arif, "Soft computing approaches for forecasting reference evapotranspiration". *Computers and Electronics in Agriculture*, 113, pp. 164-173. 2015.

[14] K. Mohammadi, S. Shamshirband, C. W. Tong, K. A., Alam, & D. Petkovi, "Potential of adaptive neuro-fuzzy system for prediction of daily global solar radiation by day of the year". *Energy Conversion and Management*, 93, pp. 406-413. 2015.

[15] D. Petkovi, & S. Shamshirband, "Soft methodology selection of wind turbine parameters to large affect wind energy conversion". *International Journal of Electrical Power & Energy Systems*, 69, pp. 98-103. 2015.

[16] V.Nikoli, S. Shamshirband, D. Petkovi, K.Mohammadi, . ojbai, T. A. Altameem, & A. Gani, "Wind wake influence estimation on energy production of wind farm by adaptive neuro-fuzzy methodology". *Energy*, 80, pp. 361-372. 2015.

[17] K. Mohammadi, S. Shamshirband, C. W. Tong, M. Arif, D. Petkovi, & S. Ch, "A new hybrid support vector machinewavelet transform approach for estimation of horizontal global solar radiation". *Energy Conversion and Management*, 92, pp. 162-171. 2015.

[18] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petkovi, & C. Sudheer, "A support vector machinefirefly algorithm-based model for global solar radiation prediction". *Solar Energy*, 115, pp. 632-644. 2015.

[19] L. Wolpert, E. Szathmry, "Multicellularity: evolution and the egg," *Nature*, vol. 420, no. 9617, pp. 745-745, 2002.

[20] A. Soofi, L. Cao, Nonlinear forecasting of noisy financial data, in: Soofi, Cao (Eds.), Modeling and Forecasting Financial Data: Techniques of Nonlinear Dynamics, Kluwer Academic Publishers, Boston, 2002.

[21] H. Hassani, Z. Xu, A. Zhigljavsky. "Singular spectrum analysis based on the perturbation theory." Nonlinear Analysis: Real World Applications, 12, 2752-2766. 2011.

[22] Sanei, Saeid, and Hossein Hassani. Singular spectrum analysis of biomedical signals. CRC Press, 2015.

[23] H. Hassani, "Singular spectrum analysis: Methodology and comparison," *J. Data Sci.*, vol.5, pp.239-257, 2007.

[24] H. Hassani, Thomakos, D. "A review on singular spectrum analysis for economic and financial time series," *Statist. Interface* , vol. 3, no. 3, pp. 377-397, 2010.

[25] H. Hassani and R. Mahmoudvand, "Multivariate Singular Spectrum Analysis: A general view and new vector forecasting approach," *IJES*, vol. 1, no. 1, pp. 55-83, 2013.

[26] R. K. Grosberg and R. R. Strathmann, "The evolution of multicellularity: a minor major transition?" *Annual Review of Ecology, Evolution, and Systematics*, 621-654, 2007.

[27] L. S. Roberts, F. Miller and C. P. Hickman Jr, *Integrated Principles of Zoology*. 2001.

[28] D. L. Kirk, "A twelve step program for evolving multicellularity and a division of labor," *BioEssays*, vol. 27, no. 3, pp. 299-310, 2005.

[29] G, Bell and AO. Mooers, "Size and complexity among multicellular organisms," *Biol. J. Linn. Soc. London*, vol. 60, no. 3, pp. 34563, 1997.

[30] R. E. Michod, "Evolution of individuality during the transition from unicellular to multicellular life," in *NAS*, 2007, pp. 8613-8618.

[31] S. M. Adl, A. G. Simpson, M. A. Farmer et al., "The new higher level classification of eukaryotes with emphasis on the taxonomy of protists," *J. Eukaryot. Microbiol.*, vol. 52, no. 5, pp. 399-451, 2005.

[32] M. D. Herron and Michod, R. E. "Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye" *Evolution*, vol. 60, no. 3, pp. 436-451. 2008.

[33] R. E. Michod and A. M. Nedelcu, "On the reorganization of fitness during evolutionary transitions in individuality," *ICB*, vol. 43, no. 1, pp. 64-73. 2003.

[34] D. L. Kirk, "A twelve step program for evolving multicellularity and a division of labor," *BioEssays*, vol. 27, no. 3, pp. 299-310, 2005.

[35] N. King, "The unicellular ancestry of animal development," *Developmental cell*, vol. 7, no. 3, pp. 313-325, 2004.

[36] M. M. Kirk, K. Stark, S. M. Miller, Muller W, B. E. Taillon, H. Gruber et al., "RegA, a Volvox gene that plays a central role in germ-soma differentiation, encodes a novel regulatory protein," *Development*, vol. 126, no. 4, pp. 639647. 1999.

[37] M. Meissner, K. Stark, B. Cresnar, D. L. Kirk, R. Schmitt, "Volvox germline-specific genes that are putative targets of RegA repression encode chloroplast proteins," *Curr. Genet.*, vol. 36, no. 6, pp. 363370, 1999.

[38] S. Sanei and H. Hassani, *"Singular Spectrum Analysis of Biomedical Signals"*, CRC Press, 2015.

[39] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed, Springer, 2002.

[40] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, vol. 27, pp. 1-22, 2008.

[41] R. J. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice, OTexts: Australia," Available via: www.OTexts.com/fpp. 2013.

[42] H. J. Newton and E. Parzen, "Forecasting and time series model types of economic time series," In Major Time Series Methods and Their Relative Accuracy (Edited by S. Makridakis, et al.), pp. 267-287,1984.

[43] H. Hassani and E. S. Silva , "A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts." *Econometrics*, **3**(3), 590–609.(2015)

[44] Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics,* **13***(3), 253–263.*